Using Random Forests to Multiply Impute Missing Data in an Online Patient-Centered Support Platform

A thesis presented

by

Nathan E. Hall

to

the Department of Biostatistics

in partial fulfillment of the requirements

for the degree of

Master of Science

in the subject of

Biostatistics

Harvard T.H. Chan School of Public Health

Boston, Massachusetts

December 2018

Acknowledgments

It is with the deepest gratitude that I would like to thank all of those who supported me throughout my studies:

First, I would like to thank Dr. David Schoenfeld, for being my dedicated mentor throughout my thesis project. Without his knowledge, guidance, and support, this thesis would not have been possible. I would also like to give a special thank you to two of my thesis advisors, Dr. Andy Nierenberg and Dr. Louisa Sylvia. I greatly appreciate everything they both have done for me and have taught me.

Next, I want to thank both of my final thesis advisors, Dr. Erin Lake and Dr. David Wypij. They both have both been very supportive and great mentors throughout my time at Harvard. I would also like to thank Dr. Heather Mattie, for being extremely helpful with both my classes and with this thesis.

I want to also thank the entire staff at the MGH Biostatistics Center and Dauten Family Center, all of my professors and colleagues at Harvard, and all of my previous instructors and teachers who have all guided me throughout the entirety of my academic career.

Lastly, I would like to thank my family and friends for supporting me throughout the years. I would like to give a special thank you to my wonderful girlfriend, Ashley, for always being there for me and believing in me.

Thesis Advisor: Dr. David Schoenfeld Nathan E. Hall

Using Random Forests to Multiply Impute Missing Data in an Online Patient-Centered Support Platform

Abstract

Background: Approximately 360 million people across the world are affected by mood disorders such as depression and bipolar affective disorder. There is a growing need for effective interventions in order to reduce this burden and help guide future research efforts. Online patient-centered studies offer a new and increasingly popular way of collecting these types of data, but psychiatric studies such as these often suffer from large amounts of missing data. Establishing a sound statistical framework to properly handle the missing data in these settings is important for achieving valid inference. Using multiple imputation in combination with random forests to impute these missing values offers a statistically sound and flexible solution to this problem.

Methods: Multiple imputation by chained equations (MICE) [59] using random forests (RFs) [6] were used to multiply impute the missing values associated with the mixed-type (both qualitative and quantitative) data in an online patient-centered support platform called MoodNetwork (MN) [54]. Data was collected from 4,344 participants and consisted of both demographic information and scores on mood assessment instruments. The pre and post-imputation densities of the data were compared using kernel density plots. Additionally, a series of twelve univariate linear

regression models as well as a single, more complex multiple linear regression model were fit to both the raw data and each of the $m = 30$ imputed data sets. Parameter estimates and their associated standard errors and p-values were pooled using Rubin's rules [48], and were compared between the raw data results and the imputed data results.

Results: The post-imputation densities of each imputed variable closely followed their pre-imputation, raw data counterparts. The parameter estimates and p-values in each of the twelve univariate linear regression models resulting from the imputed data analysis yielded very similar results to their corresponding raw data estimates and p-values. Standard errors were consistently lower in the imputed data results than in the raw data results in all but two scenarios. This was largely due to the information gained from auxiliary variables and the substantial increase in sample size in the imputed data sets. These differences were even more pronounced in the multiple linear regression comparison, in which the raw data results could not be trusted, with a sample size of $N = 10$ (due to very few participants not missing *any* data for each of the variables included in the regression model) as compared to the imputed data results ($N = 4{,}344$).

Conclusions: In the case of complex, mixed-type psychiatric data sets with large amounts of missing data, using MICE with RF models to multiply impute the missing values results in unbiased parameter estimates that closely agree with the raw data results, while also effectively reducing the standard error of parameter estimates and greatly increasing the analyzable sample size. Creating multiple completed data sets for clinical researchers to use for future analyses paves the way for additional research and studies to be conducted, and can ultimately improve clinicians' abilities to help those suffering from mood disorders.

Contents

List of Tables

List of Figures

List of Equations

1. Introduction and Background

1.1 Clinical Relevance

Mood disorders are one of the most complex issues that mental health professionals face today. Defined as a broad classification of all types of depression, mania, and bipolar disorders [26], mood disorders vary greatly from person-to-person in regard to both type and severity. Additionally, people of all ages, ethnicities, and genders can have mood disorders, which further complicates treatment and management options due the subjectivity of each individuals' specific combination of attributes [36]. The World Health Organization (WHO) estimates that worldwide, approximately 300 million people are affected by depression, and that nearly 60 million people suffer from bipolar affective disorder [62]. In the United States (U.S.) alone, the average annual cost of major depressive disorder and bipolar disorder is $100 billion and $150 billion, respectively [54], when considering both direct (healthcare spending, medications, etc.) and indirect (reduced labor supply, incarceration, etc.) costs. In addition to the general cost-related burden, mood disorders also often increase individuals' risk of developing comorbid conditions, such as heart disease, diabetes, and many other debilitating diseases. Thus, implementing a single, reliable treatment plan to each afflicted individual is both a complex and critically important task.

1.2 MoodNetwork

The Dauten Family Center for Bipolar Treatment Innovation at Massachusetts General Hospital in Boston, Massachusetts has introduced a modern approach aimed at assisting people across the globe with mood disorders. MN, which is one of the first of its kind, is an online,

patient-centered research community for individuals with mood disorders, along with their circle of support (i.e. family, friends, partners) [54]. MN is a Patient-Powered Research Network (PPRN), which is funded through the Patient-Centered Outcomes Research Institute (PCORI). Each PPRN shares the same central goal of improving the capacity that the U.S. has to conduct comparative effectiveness research by attempting to incorporate not only medical professionals' and stakeholders' points of view on care, but also the perspectives of the patients themselves [54]. To date, MN has enrolled over 5,000 individuals of varying ages, genders, and socioeconomic backgrounds. Each enrolled individual in this large and diverse sample contributes demographic information and is also able to complete any number of assessments and surveys, which range from mood evaluations to research priority questionnaires that help guide future research efforts [54]. Participants can elect to complete these assessments any number of times, which allows them to track their scores over time and monitor their personal trends.

Undeniably, there is a wealth of data present and there are likely key insights to be gained from these data, but there is one major limitation. Given the large number of participants enrolled in MN and given the fact that all surveys and assessments are voluntary, there is a substantial amount of missing data in the MN data set. Clinicians and statisticians alike are eager to investigate the data, but many statistical and data analysis techniques require complete data sets in order to achieve unbiased and accurate inferences. To further complicate things, some participants contribute multiple observations to MN – either by completing an assessment more than once, completing more than one assessment, or both. Additionally, there are both quantitative (assessment scores, age in years, etc.) and qualitative (sex, marital status, etc.) variables contributing to the missingness, and the large number of variables present results in a high-dimensional data structure. These data have yet to be analyzed in any capacity and with

online platform data rising in popularity in the clinical realm, developing a plan for handling this kind of data is of the utmost importance. All of these factors lead to a unique and variable data set that, if analyzed properly, has the potential to yield groundbreaking results.

1.3 Missing Data

Missing data is one of the most common and most complex issues associated with data analysis. This is due largely to the fact that most analyses rely on a complete (zero missing values) data set in order to be run [34]. Missing data can arise for a multitude of reasons, such as incomplete data entry by a researcher, improper data conversion from one software to another, or participant unwillingness to answer a question in a survey-based setting. Thus, careful assumptions must be made on the pattern of missingness before the data can be analyzed. In 1976, Donald Rubin proposed a system for classifying the different types of missing data [41].

The three types of missing data include missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) [41]. When data are MCAR, the probability of an observation (X_i) being missing does not depend on the value of X_i, and also does not depend on the value of any of the other observed variables in the data or unobserved variables not in the data. That is, for X_i:

$$P(missing|observed, unobserved) = P(missing)$$

If the data are MCAR, simply dropping the missing observations is a valid method of adjustment for the missingness due to the nature of the missingness being truly random. As such, more formal adjustments or imputation methods are not needed for analyzing the data, given that a sufficient number of observations are left in the data set after dropping the missing

observations so as not to reduce efficiency of estimates. The estimates themselves, though, remain asymptotically unbiased [41].

Oftentimes, data is instead classified as MAR rather than MCAR. If the data are MAR, the probability of an observation (X_i) being missing does not depend on the value of X_i or *unobserved* data, but does depend on the *observed* data. That is, the missingness of X_i can be explained after adjusting for one or more of the variables found in the data set. If the data are truly MAR (or MCAR), the missingness is considered ignorable and unbiased parameter estimates can still be produced without enacting a missing data model [13]. However, it is important to note that data can never be definitively be shown to be MAR, as values may be missing for unanticipated reasons.

Thus, often the safest assumption for missing data is MNAR. When the data are MNAR, the probability of an observation (X_i) being missing *does* depend on the unobserved value of X_i. That is, the probability of the observation being missing is related to factors which are not measured or captured by the data that is available to the researcher, and the factors that *are available* in the data are unable to explain or predict the missingness present. This proves to be problematic, and in this case the missingness is considered non-ignorable [13]. When the data are MNAR, specifying an appropriate missing data model is necessary in order to attempt to obtain any unbiased estimates of the outcome of interest.

1.4 Multiple Imputation

One of the most popular and statistically efficient methods for dealing with missing data is the process of multiple imputation (MI). Many studies and research support the use of MI, and consider it to be one of the best methods for handling missing data (either MCAR or MAR, but

not necessarily MNAR), due to the general flexibility of the approach, its ease-of-use with virtually any kind of data, and the fact that the MI procedure takes into account the uncertainty associated with the missing values [1, 13, 20, 30, 34, 42]. Furthermore, the MI procedure appropriately includes the random error that results from the imputation process itself, which leads to approximately unbiased parameter estimates when analyzing the data [1, 20]. This is a key distinction and advantage over single imputation approaches, such as the common method of mean imputation, which fail to account for the error induced by imputation (without specialized software [1]), which leads to biased parameter estimates.

The general idea behind MI is to first create several (m) imputed data sets under a suitable model that successfully incorporates variation, as opposed to creating only one imputed data set. Then, the desired analysis can be performed on each of the m data sets, and the parameter estimates from each of the m analyses can simply be averaged together to form a single point estimate. The standard errors associated with this process can be calculated using a formula developed by Rubin that takes into account both within-imputation variance and the between-imputation variance [42]. This process, which has been appropriately named "Rubin's rules", will be described in further detail in subsection 2.3 (Statistical Analyses & Comparisons) of the Methods section of this paper.

1.5 Random Forests

Machine learning has become one of the most cutting-edge areas of research in the past decade. The general idea behind machine learning algorithms is to make some sort of decision or classification by detecting patterns in data and generalizing from these patterns to predict future outcomes [9, 14]. Random forests (RFs) are just one of many different machine learning methods.

RFs are meta-learners, since they are a collection of individual learners called decision trees (DTs). To form an RF of DTs, the DTs are grown through a multi-step process. In the first step, two data sets are created from the original data by splitting the original data into a training set and a test set. Next, the data that will be used to grow each tree is determined by taking a different bootstrap (a random sample *with replacement*) sample from the training set, with each sample containing the same number of observations as the training set – likely containing duplicate observations due to the nature of the bootstrap. Each bootstrap sample is referred to as the *in-bag* (IB) data, and the IB data usually contains approximately two-thirds of the training data. The remaining, left-over data that was not randomly sampled into the IB data is referred to as the *out-of-bag* (OOB) data, and comprises the remaining one-third of the training data [6, 25].

Once the IB and OOB data are established, the DTs themselves are created. The standard way each individual DT learner is grown is by taking a random number of variables (attributes) from the IB data and treating them as questions (root nodes) that will have a binary ("yes" or "no") answer (leaf nodes) [28]. These binary splits are present for both qualitative and quantitative attributes, as the DT algorithm iteratively "looks for" the best split point using a method such as sum-of-squares regression or entropy for classification [6, 9]. Thus, a hierarchy of root nodes and leaf nodes develop, with a new root node stemming from the branch of each leaf node for each attribute that is assessed in the DT. The basic idea behind the RF procedure is to grow many, slightly different DTs, each of which make their own vote or prediction on what the outcome value ought to be. Each of the DTs are grown on a different bootstrap sample of the training data, thus the total number of trees will be equivalent to the total number of different IB data sets. Additionally, the DTs are grown without pre-setting the attributes to be used before growing the trees, signified by only considering a random set of attributes for each tree that is grown, which

appropriately accounts for any collinearity that may be present among the attributes [6, 25]. This conveniently results in each individual DT being is slightly different than the others amongst the total collection of all the DTs.

Next, the collection of all the DTs grown and each of their respective decisions (which comprises the RF) are considered, and the forest makes a final classification or prediction based on which of the individual DT classifications or predictions received the most votes in the case of qualitative variables, or by calculating the mean across all DTs in the case of quantitative variables [6, 18, 25, 52]. The performance of the RF algorithm can be assessed via the OOB error, which is calculated by finding the average misclassification over the entire RF using all of the OOB data. This OOB error also acts as a built-in validation system of the algorithm, which is attractive from a statistical point of view since the performance of the method can be assessed simply from the piece of the training data that was not used (OOB data), rather than relying on a test set. This is especially useful if there is not a test set to check performance against, which is often the case when data are missing [18].

1.6 Approach and Goals

There are numerous techniques that have been proposed to handle missing data using MI, including multiple imputation by chained equations (MICE), multiple imputation by classification and regression trees (CART), and Expectation-Maximization with Bootstrapping (EMB) [34]. Each method has its advantages and disadvantages, but a novel approach utilizing the machine learning technique of RFs in combination with MICE offers a surprisingly simple, non-parametric way to handle the mixed data types present in the MN data set [56]. The goal of this

paper is to apply this flexible MI method to a large, mixed-type, partially-longitudinal data set that has yet to be analyzed in any major capacity.

Once multiple completed data sets are generated using this MI method, a set of univariate linear regression models and a single, more complex multiple linear regression model will be fit to each individual imputed data set, as well as to the raw, unimputed data. Then, Rubin's rules [43] will be used to pool the imputed results together over the repeated analyses. This will ultimately serve to determine if there are significantly different conclusions drawn between the imputed data results and the raw data results, and if the resulting variances and standard errors from the imputations are low/agree with the corresponding errors from the raw data. Lastly, a set of complete, multiply imputed data sets utilizing this application will be stored so that they can be used for any kind of future analyses, in order to allow medical professionals to investigate any future research question of interest arising from the MN data.

2. Methods

2.1 Data

Before the data could undergo cleaning or imputation, a critical step in the data preparation process was to determine how to best subset the data in a clinically relevant fashion. Following consultations with key MN investigators and personnel, it was determined that basing the pre-imputation data on only certain mood trackers/assessments was the most practical course of action. This is because some mood trackers/assessments are more validated than others in the current literature, and the researchers were only interested in this smaller subset of mood trackers/assessments and the degree to which their measurements do or do not coincide. Furthermore, although data is available on individuals from multiple countries, the pre-

imputation data was pruned to include only participants from the U.S. This was done in order to properly impute values for the variable "region", which was created based on the U.S. Census Bureau's Census Divisions [57]. For the purposes of this paper, the analyses were restricted to U.S. participants only, but the imputation procedure could easily be extended to include international data in the future. However, this would likely be at the expense of losing information at the state level for U.S. participants, since the variable "state" would be undefined for international participants. It is also important to note that the two participants from Puerto Rico (PR) were excluded as a result of this decision, since PR does not fall under any of the Census Divisions as defined by the U.S. Census Bureau.

As was previously stated, there are over 5,000 total participants who have consented and enrolled in MN ($N = 5{,}211$) as of July 2018. After accounting for these initial decisions regarding which subset of data were to be used prior to imputation, a final sample size of $N = 4{,}344$ participants remained. This pre-imputation data set, which was based solely on demographic data, defined the base data set from which all subsequent assessment-included pre-imputation data sets were derived from. Each of these participants contribute demographic information, such as age, sex, marital status, etc. Each participant is also eligible to complete any number of mood trackers/assessments as many times as they wish. These additional assessments, which will hereafter be referred to as "instruments", attempt to measure and capture different aspects of mood state. Tables 1 and 2 below give a summary of the quantitative and qualitative variables, respectively.

Table 1: Descriptive Statistics and Missingness – Quantitative Variables

Variable	Type	Mean (SD)	Min	Max	# Miss	% Miss	Description
UID	Demographic	-	-	-	0	0.00%	Unique identification number (UID) for each participant.

Age	Demographic	43.21 (13.45)	18.00	92.00	1	0.02%	Age in years.
Number of people in household	Demographic	2.74 (1.44)	1.00	15.00	151	3.48%	Number of individuals currently living in participant's household.
ASRM Total Score	Instrument	4.16 (4.13)	0.00	17.00	4065	93.58%	Total score on the Altman Self-Rating Mania Scale (ASRM) [2], which measures mania.
DBSA Total Score	Instrument	-1.10 (1.56)	-4.00	4.00	3218	74.08%	Total score on the Depression and Bipolar Support Alliance (DBSA) Wellness Tracker [12], which measures both depression and mania on a bipolar scale.
QIDS-SR Total Score	Instrument	15.58 (5.19)	1.00	26.00	3684	84.81%	Total score on the Quick Inventory of Depressive Symptomatology Self-Report (QIDS-SR) [44], which measures depression.
SDS Total Score	Instrument	20.87 (7.17)	0.00	30.00	4245	97.72%	Total score on the Sheehan Disability Scale (SDS) [51], which measures disability/functional impairment.
BART Total Score	Instrument	23.52 (24.84)	0.00	75.00	3657	84.19%	Total score on the Balloon Analog Risk Task (BART) [31], which measures riskiness/risk-taking behavior tendency.
WHO-5 Total Score	Instrument	7.38 (5.12)	0.00	25.00	4229	97.35%	Total score on the World Health Organization Well-Being Index (WHO-5) [5], which measures quality of life/well-being.

Table 2: Descriptive Statistics and Missingness – Qualitative Variables

Variable	Type	Levels	N	%	# Miss	% Miss	Description
Depression	Demographic	No	124	2.90%	0	0.00%	Indicator of if participant has ever experienced depression.
		Yes	4220	97.10%			
Hypomania/Mania	Demographic	No	1071	24.70%	0	0.00%	Indicator of if participant has ever experienced hypomania or mania.

		Yes	3273	75.30%			
Friend/Family Member	Demographic	No	124	2.90%	535	12.32%	Indicator of if participant has a friend or family member who has experienced depression or bipolar disorder.
		Yes	4220	97.10%			
Sex	Demographic	Male	829	19.40%	65	1.50%	Sex of participant.
		Female	3394	79.30%			
		Ambiguous	20	0.50%			
		Other	36	0.80%			
Region*	Demographic	East North Central	553	13.20%	163	3.75%	U.S. Census Division region in which participant receives medical care.
		East South Central	159	3.80%			
		Mid-Atlantic	464	11.10%			
		Mountain	312	7.50%			
		Northeast	825	19.70%			
		Pacific	586	14.00%			
		South Atlantic	685	16.40%			
		West North Central	243	5.80%			
		West South Central	354	8.50%			
Hispanic	Demographic	No	3769	91.50%	225	5.18%	Indicator of if participant is Hispanic.
		Yes	350	8.50%			
Race	Demographic	Native American, American Indian, or Alaskan Native	45	1.10%	113	2.60%	Race of participant.
		Asian	112	2.60%			
		Black	150	3.50%			
		Native Hawaiian or Pacific Islander	9	0.20%			
		White	3639	86.00%			
		Other	276	6.50%			

Education	Demographic	8th grade or less	20	0.50%	106	2.44%	Highest level of education completed by participant.
		Some high school	114	2.70%			
		High school graduate or GED	491	11.60%			
		Some college or two-year degree	1554	36.70%			
		Four-year college graduate	970	22.90%			
		More than four-year college degree	1089	25.70%			
Marital Status	Demographic	Now married	1613	38.00%	96	2.21%	Marital status of participant.
		Living with partner or significant other	468	11.00%			
		Widowed	81	1.90%			
		Divorced	671	15.80%			
		Separated	172	4.00%			
		Never married	1243	29.30%			
Employment Status*	Demographic	Employed	2227	51.30%	0	0.00%	Indicator of if participant is employed, unemployed, or if they indicated another option (disabled, volunteer, etc.) without also indicating employed or unemployed.
		Unemployed	871	20.10%			
		Other/Unknown	1246	28.70%			
Currently Receiving Medical Care*	Demographic	No	3472	89.60%	467	10.75%	Indicator of if participant currently receives medical treatment at a hospital or medical facility.
		Yes	405	10.40%			

Table 2: *Variables marked with an asterisk (*) represent variables that were created/modified based off of original, raw MN data.*

An important observation to be made from these tables is that the percent missingness (given by the % Miss column) varies greatly from variable to variable and is extremely high (more than 90% missing) in some cases for instrument variables. Due to the nature of MN, in that some demographic questions and all instrument completions are optional, a "complete case" in these data would be considered relatively rare. For the purposes of the analyses in this paper, a participant's value for any given variable is considered missing if it is not reported, even for the

optional demographic questions and instruments. While this may seem concerning from a statistical standpoint, it is widely agreed upon in the literature that there is no absolute maximum percent missing data that would justify discarding variable(s), and the MICE method is able to account for large amounts of missing data by design [4, 15, 46, 47, 55, 59]. For example, even if a variable has > 90% missing values, if the overall sample size is sufficiently large, the number of observations without missing values may very well provide enough information about the highly missing variable. Assuming computations are not too intensive/time-consuming, van Buuren and Groothuis-Oudshoorn (2011) suggest setting the number of imputations (m) equal to the average percent of missing data across all variables in the data set [59]. In this paper, this is the approach taken, where the average percent missingness across all variables (excluding UID) is 30.3%, so $m = 30$ imputations are used.

Additionally, given that all instrument completions are optional for each participant, the number of unique, first completions (one completion per participant) of each instrument varies greatly. As a result, there is a wide range of possible completion patterns among all participants, with some participants contributing more information than others to the instrument variables in the full, wide data set. To give a general idea of the completion patterns present in the data, Table 3 below gives a summary of the number of first completions for each instrument in the data set, as well as the pairwise completions for each instrument (i.e., the number of participants who completed two instruments, for all possible combinations of two instruments). Also, to give a general idea of the amount of people who completed one instrument, two instruments, etc., Table 4 depicts the raw number and percentage for each possible number of instruments, with a minimum of zero and a maximum of six instruments completed.

Table 3: Unique First Completions and Pairwise Completions for Instruments

Instrument	ASRM	DBSA	QIDS-SR	SDS	WHO-5	BART
ASRM	**279**	242	234	37	233	20
DBSA	242	**1126**	519	90	620	75
QIDS-SR	234	519	**660**	75	483	49
SDS	37	90	75	**99**	82	23
WHO-5	233	620	483	82	**687**	52
BART	20	75	49	23	52	**115**

Table 3: *Bolded cells indicate the sample size for the total unique, first completions for each instrument.*

Table 4: Total Instrument Completeness

# Instruments Completed	# Participants	% Participants
0	2999	69.03%
1	552	12.71%
2	278	6.40%
3	264	6.08%
4	199	4.58%
5	42	0.97%
6	10	0.23%

2.2 The *mice* Package

2.2.1 Background and Justification of Use

The R package titled "*mice"* was first introduced by van Buuren and Groothuis-Oudshoorn in 2011 when they published their paper titled "mice: Multivariate Imputation by

Chained Equations in R" [59]. The package was aimed at providing a relatively simple, easy-to-use method for imputing mixed-type data, which is data that includes both qualitative and quantitative variables in its structure. The *mice* package, appropriately named after the MICE MI method, is sometimes alternatively referred to as the fully-conditional specification (FCS) method. There are two primary imputation methods that currently exist in the literature: FCS and joint modelling (JM). The main difference between the FCS and JM approaches is that JM requires the specification of a multivariate normal distribution for the missing data, as well as imputing from the missing datum's conditional distributions via Markov chain Monte Carlo (MCMC) techniques [8, 45], while FCS does not have these requirements. Instead, FCS specifies the multivariate imputation model variable-wise by defining a conditional density for each variable that has missing values [4, 59]. In other words, the FCS approach simply models the missing data conditional upon the other variables that are available in the data set.

Furthermore, due to the nature of the chained equations in the MICE procedure, each variable in the data set can be modeled based on its own unique distribution, which allows for much more pliable imputation model choices than JM would. The sole assumption of using MICE under most applications for imputation is for the data to be MAR. MICE can also be used under MNAR conditions [59], but proper modeling of the missing data is necessary, and Little warns that MNAR models are subject to a deficiency in identification in all cases [33]. It is also important to note that the exact regression methods used in the MICE procedure still operate under the same assumptions that would be made if a traditional regression were run without imputation. If the data truly are MNAR, any imputation method is likely to perform poorly since the pattern of missingness, by definition, cannot be explained by the variables that are present. For the purposes of this thesis, the missing data will be presumed to be MAR. This seems

plausible, since the missingness of instrument completions can likely be explained by demographic variables (which generally have very little missing data), and/or by the other instrument measures that are present in the data.

The authors present the MICE method as a superior alternative to JM MI methods, due to the aforementioned reduction in assumptions compared to JM, the fact that it is possible to specify models for which there is no known joint distribution, and the ease of implementation via statistical programming software using MICE [4, 59]. Results from multiple different studies and simulations suggest that MICE works quite well across different situations, including epidemiological data settings [35], data sets with different classes of MAR data (more missingness for larger values, more missingness for the center of a distribution, etc.) [58], and even in large survey-type data settings with large amounts of missing data [48]; an extremely similar situation to MN.

While it seems clear that MICE is a widely used and appropriate method for MI, there are a multitude of different ways the *mice* package can be implemented. Due to the flexibility of the MICE framework, in that the algorithm is a concatenation of several univariate models for the missing data, it is possible to specify a different model for each variable according to its own unique distribution. This is where the different applications of *mice* come into play, as the entire procedure is dependent on the overall pattern of model specifications for each variable in the data set. The model of choice for this paper was unanimous across all variables: using the method of RF to impute the missing values.

The justification for using the RF method for the imputation of all variables in the MN data set is simple: other FCS methods for imputation, such as linear regression for continuous variable modeling or logistic regression for binary variable modeling require that these models

are specified correctly, including interactions and non-linearities [49]. However, neglecting interactions or non-linear terms can lead to biased results. Moreover, using parametric regression models in the MICE procedure cannot be done in cases where there are more predictor variables than the number of observations without sufficient prior information [23], or when collinearity is present between variables. Using RFs to impute all of the missing data overcomes these challenges, since this method is non-parametric, can handle both qualitative and quantitative variable types, and accommodates non-linearities and interactions in the data, if any are present [49]. Burgette and Reiter corroborate these advantages following a simulation study comparing MICE using linear regression vs. MICE using DTs, hinting at the potential advantage of extending their ideas of MICE using DT's to MICE using RFs [7]. Thus, the MN data set, which can be described as a mixed-type data set arising from an observational study, proves to be particularly well suited for the use of RFs in the MICE procedure of the imputation process.

2.2.2 Data Setup & Algorithm

The general approach that van Buuren and Groothuis-Oudshoorn implement in the *mice* package involves a multi-step process. First, we define the following notation from van Buuren & Groothuis-Oudshoorn. Let Y_j with $(j = 1, \dots, p)$ be one of p incomplete variables, where $Y = (Y_1, \dots, Y_p)$. The observed and missing parts of Y_j are denoted by Y_j^{obs} and Y_j^{mis}, respectively, so $Y^{obs} = (Y_1^{obs}, \dots, Y_p^{obs})$ and $Y^{mis} = (Y_1^{mis}, \dots, Y_p^{mis})$ stand for the observed and missing data in Y. The number of imputations is equal to $m \geq 1$. The h^{th} imputed data sets are denoted as $Y^{(h)}$ where $h = 1, \dots, m$. Let $Y_{-j} = (Y_1, \dots, Y_{j-1}, Y_{j+1}, \dots, Y_p)$ denote the collection of the $p - 1$ variables in Y except Y_j. Let Q denote the quantity of scientific interest (e.g., a regression

coefficient). In practice, Q is often a multivariate vector. More generally, Q encompasses any model of scientific interest.

As an extension to the *mice* package in R, in 2014 Shah et al. [49] proposed a unique framework that would allow for RF models to be used as the imputation model of choice within the MICE procedure. Many of the most commonly used methods for imputation models in *mice* are not able to account for interactions or non-linearities in the data, but the introduction of this particular method provided a way to accomplish this. Furthermore, Shah et al. implemented this technique in two slightly different ways: one for quantitative variables and one for qualitative variables – which offers the advantage of flexibility for mixed-type data sets, while still preserving the same general methodology of an RF across both variable types. The authors compared their method, which can be used in either their own package (*CALIBERrfimpute*) or as a method within the base *mice* package, to the standard parametric *mice* package. Their results showed that using RFs within MICE using their technique provided narrower confidence intervals, less-biased parameter estimates, and was more efficient than standard parametric MICE methods [49].

Putting all of these components together results in the final, formal *mice* algorithm (using Shah et al.'s RFs method as the method of imputation), as outlined below in Table 5:

Table 5: The *mice* Algorithm Using the RF Method

Step 1. Begin by using a simple imputation approach, such as mean imputation, for all missing values in the data set. These initial imputations will act as "placeholders" for the soon-to-be MICE-imputed values.
Step 2. Once the order in which variables will be imputed is established, the placeholder imputed values for the first variable to be imputed, given by Y_1, are set back to missing.
Step 3. The non-missing values of Y_1, given by Y_1^{obs}, are regressed on all of the other variables from the data set that will be included in the appropriate imputation model (in this case, an RF model is used).

Step 4. **4a) Quantitative Variables:** After the imputation model is fit, the missing values of Y_1, given by Y_1^{mis}, are imputed by random draws from the independent normal distributions centered on the conditional means predicted by the RF imputation model from **Step 3,** with variance equal to the OOB error estimate. **4b) Qualitative Variables:** After the imputation model is fit, the missing values of Y_1, given by Y_1^{mis}, are imputed by taking the prediction from the terminal node of a randomly chosen DT from the RF imputation model from **Step 3**.
Step 5. Steps 2 – 4 are repeated for each variable that has missing data. Once a variable, Y_j, has undergone these steps, both Y_j^{obs} and the newly imputed values of Y_j, call them Y_j^{imp}, are used in the imputation model for all other variables $(Y_{j+1}, \dots, Y_p)$.
Step 6. Steps 2 – 5 are repeated for a pre-defined number of iterations, as a means of assuring the convergence of the imputed values – regardless of the order in which $(Y_j, \dots, Y_p)$ are imputed. The default number of iterations is ten.
Step 7. Steps 2 – 6 are repeated m times, where m is the desired number of fully-imputed data sets.

All of the variables present in the MN data were used in the imputation procedure except for UID (19 total variables), since it simply serves as an identification variable for each participant. Because an RF application of *mice* is used, it is not necessary to include all possible interactions of interest before undergoing imputation. Including all of the possible variables except for UID allows for the imputation model to be more general than any of the analysis models that could be fit as a result of the imputation, which is a key consideration when setting up a MICE procedure that also serves to reduce bias [4, 13, 21, 47]. Only the total scores for each of the six instruments, as opposed to all of the individual items of each instrument, were used in the imputation procedure. This choice was made partly due to the MN clinicians' preferences, and also because in the raw MN data, there was hardly any missing data at the individual item level

within each instrument [21]. It is also important to note that many of the imputed values for the quantitative variables were not whole numbers. Given that all of the quantitative variables in these data are only defined for integer values, these imputed non-integer values are certainly impossible to obtain in a real-world setting. Rounding the imputed values was considered, but it has been found that leaving imputed values as non-integer values reduces bias in parameter estimates, and generally leads to improved MI results [34, 59, 60]. Therefore, the imputed values were left as-is and were not rounded up or down to whole numbers.

There are several parameters used in the *mice* [59] algorithm that required specification before the procedure itself could be run. First, $m = 30$ imputed data sets were specified to be generated. Second, the method used for all qualitative variables was specified as *rfcat* [50] and the method used for all quantitative variables was specified as *rfcont* [50]. Third, the number of variables considered at each split (*mtry*) was left as the default, which is equal to one third of the number of predictors for qualitative variables, and is equal to the square root of the number of predictors for quantitative variables. Fourth, the UID variable was removed from predictions by using a custom *predictorMatrix* specification. Fifth, the order in which variables were imputed was not specified, so the variables were imputed in the same order as they appear in the data set. Sixth, specific restrictions on the minimum and maximum value of imputed values were specified for some quantitative variables using the *post* and *squeeze* functions in mice, due to the fact that there are only specific possible score ranges for the instrument variables, and other quantitative variables such as Age and Number in Household are only defined for positive values. Seventh, a random seed was set for reproducibility purposes. Lastly, $n = 20$ trees were specified to be grown in each RF to minimize bias in the estimates following MI, rather than the *mice* package's default of $n = 10$, per the recommendations of Shah et al [49].

2.3 Statistical Analyses & Comparisons

From a clinical perspective, it was determined that conducting a series of univariate linear regressions would be a relevant, informative way of both assessing the performance of the MI and of demonstrating a practical application using the multiply imputed data. This embodies the central purpose of using the RF method within the MICE procedure, in that the goal was to explore and compare the final point estimates and their associated errors after conducting the series of univariate linear regressions on the raw data results. More specifically, to cover the range of variable types (qualitative vs. quantitative), data types (demographic vs. instrumental), and variability in the completeness across different instruments (relative to other variables/data of the same type), twelve simple linear regression models were fit on each of the thirty data sets resulting from the implementation of the *mice* algorithm. Below, Table 6 summarizes the twelve regression models and the statistical rationale for choosing each combination of variable types, data types, and variable completeness.

Table 6: Univariate Linear Regression Models

Regression Model	Rationale
1. $E[QIDS\ Total\ Score] = \beta_0 + \beta_1 * Age$	**Outcome:** • Quantitative • High # of completions (N = 660) **Predictor:** • Demographic variable • Quantitative • Low # of imputed values (N = 1)
2. $E[QIDS\ Total\ Score] = \beta_0 + \beta_1 * Number\ in\ Household$	**Outcome:** • Quantitative • High # of completions (N = 660) **Predictor:** • Demographic variable • Quantitative • High # of imputed values (N = 151)
3. $E[QIDS\ Total\ Score] = \beta_0 + \beta_1 * I(Receiving\ Care)$	**Outcome:** • Quantitative

	• High # of completions (N = 660) **Predictor:** • Demographic variable • Qualitative • High # of imputed values (N = 467)
4. $E[QIDS\ Total\ Score] = \beta_0 + \beta_1 * I(School = 2) + \beta_2 * I(School = 3) + \beta_3 * I(School = 4) + \beta_4 * I(School = 5) + \beta_6 * I(School = 6)$	**Outcome:** • Quantitative • High # of completions (N = 660) **Predictor:** • Demographic variable • Qualitative • Low # of imputed values (N = 106)
5. $E[QIDS\ Total\ Score] = \beta_0 + \beta_1 * DBSA\ Total\ Score$	**Outcome:** • Quantitative • High # of completions (N = 660) **Predictor:** • Instrument (total score) variable • Quantitative • High # of completions (N = 1126)
6. $E[QIDS\ Total\ Score] = \beta_0 + \beta_1 * Sheehan\ Total\ Score$	**Outcome:** • Quantitative • High # of completions (N = 660) **Predictor:** • Instrument (total score) variable • Quantitative • Low # of completions (N = 99)
7. $E[ASRM\ Total\ Score] = \beta_0 + \beta_1 * Age$	**Outcome:** • Quantitative • Low # of completions (N = 279) **Predictor:** • Demographic variable • Quantitative • Low # of imputed values (N = 1)
8. $E[ASRM\ Total\ Score] = \beta_0 + \beta_1 * Number\ in\ Household$	**Outcome:** • Quantitative • Low # of completions (N = 279) **Predictor:** • Demographic variable • Quantitative • High # of imputed values (N = 151)
9. $E[ASRM\ Total\ Score] = \beta_0 + \beta_1 * I(Receiving\ Care)$	**Outcome:** • Quantitative • Low # of completions (N = 279) **Predictor:** • Demographic variable • Qualitative • High # of imputed values (N = 467)

10. $E[ASRM\ Total\ Score] = \beta_0 + \beta_1 * I(School = 2) + \beta_2 * I(School = 3) + \beta_3 * I(School = 4) + \beta_4 * I(School = 5) + \beta_6 * I(School = 6)$	**Outcome:** • Quantitative • Low # of completions (N = 279) **Predictor:** • Demographic variable • Qualitative • Low # of imputed values (N = 106)
11. $E[ASRM\ Total\ Score] = \beta_0 + \beta_1 * DBSA\ Total\ Score$	**Outcome:** • Quantitative • Low # of completions (N = 279) **Predictor:** • Instrument (total score) variable • Quantitative • High # of completions (N = 1126)
12. $E[ASRM\ Total\ Score] = \beta_0 + \beta_1 * Sheehan\ Total\ Score$	**Outcome:** • Quantitative • Low # of completions (N = 279) **Predictor:** • Instrument (total score) variable • Quantitative • Low # of completions (N = 99)

While simple univariate linear regressions are appropriate to explore and compare the imputed vs. raw data results, to get a sense of the full potential that this MI would have in the context of MN, a multiple linear regression was also fit on the imputed data. In many cases, multiple linear regression models cannot be fit on the raw MN data, due to the high percentage of missingness present in most covariate patterns [34]. Furthermore, even when a multiple linear regression can be run on the raw data, the sample size available for the regression will be extremely small, thus severely limiting the power and generalizability of the analysis, as well as likely overfitting the model itself [10]. For example, consider a multiple linear regression with DBSA total score as the outcome, where the DBSA instrument assesses both depression and mania on a bipolar scale. Although there are 1,126 unique first completions of the DBSA instrument, the sample size available for analysis drastically decreases by nearly 50% once the WHO-5 total score is also included as a predictor ($N_{DBSA+WHO} = 620$). The analyzable sample size shrinks even further to less than twenty observations when other relevant predictors, such as

the BART and ASRM total scores, Education, and Number in Household ($N_{DBSA+WHO+BART+ASRM+Edu+NumHouse} = 17$), are included. Moreover, the multiple linear regression completely fails/cannot be run if, for example, the Friend or Family (FoF) indicator variable is added to the regression formula. This is because each participant who is not missing data for the FoF variable has $FoF = 1$, and therefore the linear model cannot be fit if there is no variability in one of the predictor variables.

Given that MN researchers and/or other clinicians in the future may want to explore any number of different multiple linear regression options (or even logistic or ordinal regression), MI presents a perfect opportunity to make this possible by creating multiple complete data sets. To illustrate the advantages of analyzing the MI data over the raw MN data in the case of multiple linear regression, a plausible model was fit to both the imputed data and the raw data, as outlined by Equation 1 below.

Equation 1: Multiple Linear Regression Model

$$E[WHO\ Total\ Score] = \beta_0 + \beta_1 * Sheehan\ Total\ Score + \beta_2 * QIDS\ Total\ Score + \beta_3 * ASRM\ Total\ Score + \beta_4 * BART\ Total\ Score + \beta_5 * DBSA\ Total\ Score + \beta_6 * Age + \beta_7 * I(Sex = Female) + \beta_8 * I(Sex = Ambiguous) + \beta_9 * I(Sex = Other)$$

In order to appropriately conduct the MI procedure, the $m = 30$ imputed data sets resulting from the *mice* algorithm were analyzed according to Rubin's original MI framework [41, 42, 43]. This standard MI framework requires, by definition, the generation of multiple imputed data sets so that the uncertainty associated with each individual imputation can properly be accounted for. The random draws from Step 4 in Table 5 are the key components of this aspect of the MI procedure, and in combination with the bootstrap sampling associated with RFs, these random draws help to ensure that the imputations are "proper" [32, 49]. Thus, each of the above

univariate linear regression models, as well as the multiple linear regression model, were conducted on each of the $m = 30$ imputed data sets resulting from the *mice* run. Subsequently, each of the thirty parameter estimates were calculated using standard linear regression utilizing the ordinary least-squares method [27]. These parameter estimates and their associated standard errors were then pooled and calculated using Rubin's rules. Rubin's rules, as previously mentioned, are a series of equations that are used to appropriately pool the estimates resulting from the models fit on each of the imputed data sets generated in an MI procedure [43]. The first equation, which yields the overall pooled parameter estimate, is given by Equation 2 below.

Equation 2: The Pooled Parameter Estimate

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

Where $\hat{Q}_i(i = 1 \ldots m)$ is each of the $m = 30$ parameter estimates, and $\bar{Q}$ is the pooled parameter estimate. The pooled parameter estimate is simply the mean of each of the parameter estimates.

In order to calculate the combined variance, and therefore the combined standard error, both the within-imputation variance and the between-imputation variance need to be properly accounted for. First, the within-imputation variance is given by Equation 3 below.

Equation 3: Within-Imputation Variance

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

Where $U_i(i = 1 \ldots m)$ is each of the $m = 30$ variances associated with each of the $m = 30$ parameter estimates, $\hat{Q}_i(i = 1 \ldots m)$, and $\bar{U}$ is the average within-imputation variance. The

average within-imputation variance is simply the mean of each of the $m = 30$ variances. Next, the between-imputation variance is given by Equation 4 below.

Equation 4: Between-Imputation Variance

$$B = \frac{1}{m-1}\sum_{i=1}^{m}(\hat{Q}_i - \bar{Q})^2$$

Where B is the between-imputation variance. Putting both Equations 3 and 4 together, the total variance of $\bar{Q}$ is given by Equation 5 below.

Equation 5: Total Variance of the Pooled Parameter Estimate

$$T = \bar{U} + (1 + \frac{1}{m})B$$

Where T is the total variance. The total variance of the pooled parameter estimate is equivalent to the weighted sum of the within-imputation and between-imputation variances, and the total standard error can be calculated by simply taking the square root of this value.

Standard Wald-type confidence intervals and statistical tests for $\bar{Q}$ can be calculated using a Student's t approximation, given by Equation 6 below.

Equation 6: Reference Distribution for Statistical Tests

$$\frac{(Q - \bar{Q})}{\sqrt{T}} \sim t_v$$

Where the degrees of freedom (df), v, are given by Equation 7 below.

Equation 7: Degrees of Freedom for the Reference Distribution

$$v = (m-1)(1 + \frac{\bar{U}}{\left(1 + \frac{1}{m}\right)B})^2$$

And lastly, the estimate of information lost due to missingness for each estimated parameter is given by Equation 8 below.

Equation 8: Information Lost Due to Missingness

$$\lambda = \frac{B + \frac{B}{m}}{T}$$

Analyses were performed in R 3.4.1 [37], using the *tidyverse* [61] package for data cleaning/preparation and diagnostic plots, the *broom* [39] package for the extraction of omnibus p-values from qualitative predictor models with more than one category, the *xlsx* [17] package for importing the raw data files, the *lubridate* [22] package for the handling of dates/times, the *data.table* [16] package for the merging of data frames, the *psych* [38] package for generating descriptive statistics for quantitative variables, the *Hmisc* [24] package for generating descriptive statistics for qualitative variables, the *VIM* [27] package for generating missing value visualizations, the *gridExtra* [3] package for the creation of gridded graphics, the *CALIBERrfimpute* [50] package for setting the number of trees to be grown in each RF and for imputation, and the *mice* [59] package for imputation via MICE methodology, conduction of analyses, and for the pooling of results according to Rubin's rules.

3. Results

3.1 Missingness Pattern and Pre vs. Post-Imputation Distributions

Tables 1 & 2 from section 2.1 gave a summary of the raw number of missing values and the percent missingness for each variable. To give a visual representation of the pre-imputation data, Figure 1 below shows both the pattern of missing data over the entire data set and the proportion of missing values for each variable, sorted in order of descending missingness. In the

pattern plot, red boxes indicate missing data points while blue boxes indicate non-missing data points.

Figure 1: Missingness Pattern and Proportions

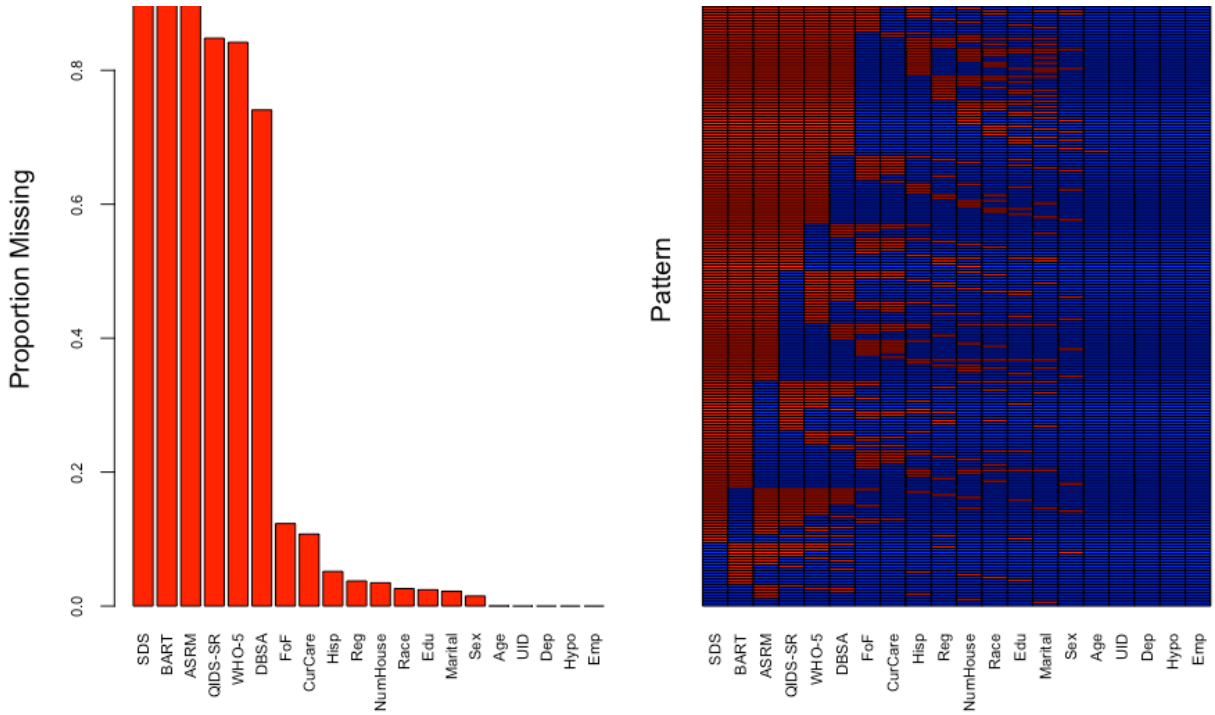


Figure 1: *In both plots, red color indicates missing data. In the left plot, the proportion of missing values for each variable is plotted in order of descending missingness. In the right plot, different patterns of missingness are displayed with non-missing data points represented by blue boxes.*

There is a steep drop off in percent missingness after the instrument variables. The DBSA total score is the least-missing instrument variable, with 74.08% missing data, and the next most-missing variable in the data set is the friend or family member indicator variable (FoF) at 12.32% missing. This signifies that overall, most of the missingness present in the MN data comes from the instrument variables and not the demographic variables, which is to be expected. SDS total score is the most-missing variable in the data set with 97.72% missing data, which gives a sense of the extreme missingness present for some of the instrument variables. From the missingness pattern plot, we are able to get a sense of the different combinations of missing variables across

all participants (observations). For example, there are very few observations that are only missing values for a few variables (given by mostly blue boxes for any given row), but there are also no completely missing observations in the data set (given by completely red rows). Furthermore, we are able to observe trends in the missingness across multiple variables. For example, for the rows that are *not* missing an SDS score, the same participants are oftentimes also *not* missing a DBSA score. This is useful information, because this tells us that participants who choose to fill out the disability-assessing instrument (SDS) also often choose to fill out a depression/mania-assessing instrument (DBSA), indicating a possible linkage between disability and depression and/or mania.

Following the MI procedure, the post-imputation distributions of each of the variables were expected to be similar to their pre-imputation distributions, if the MI procedure was performing well. Figures 2 & 3 depict both the pre-imputation (blue) and post-imputation (red) distributions of all of the quantitative variables (the age variable is not included here, because it contained only one missing value), as well as the qualitative variables, respectively.

Figure 6: Quantitative Variable Distributions – Pre vs. Post-Imputation

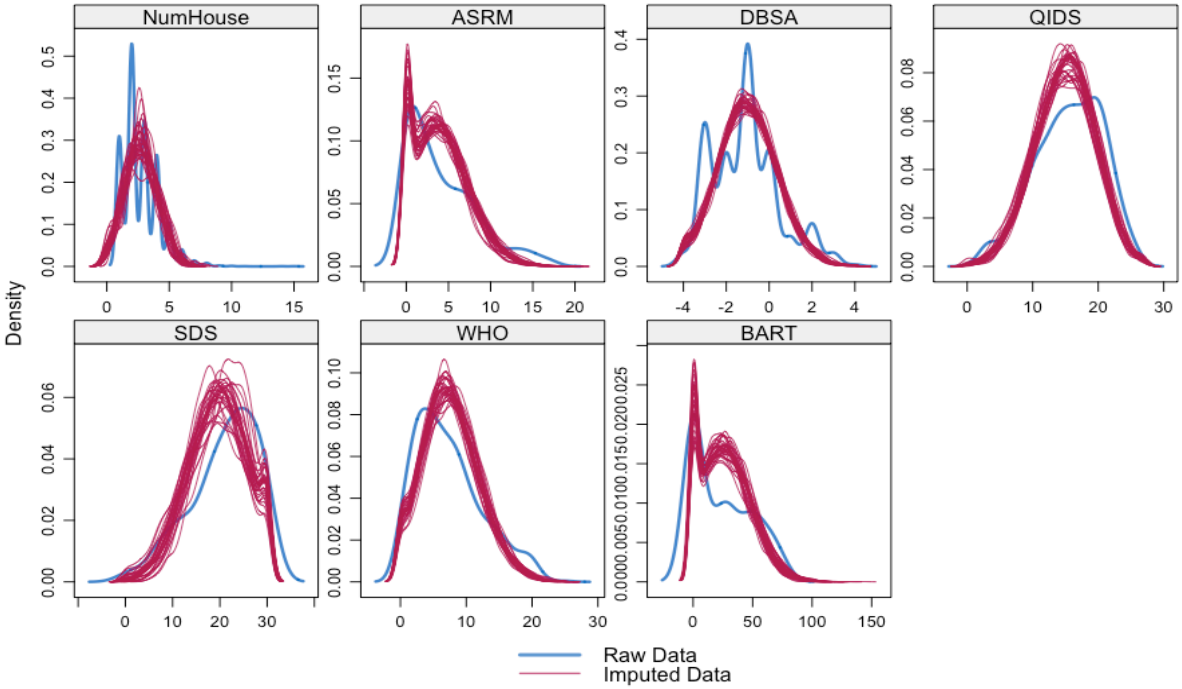
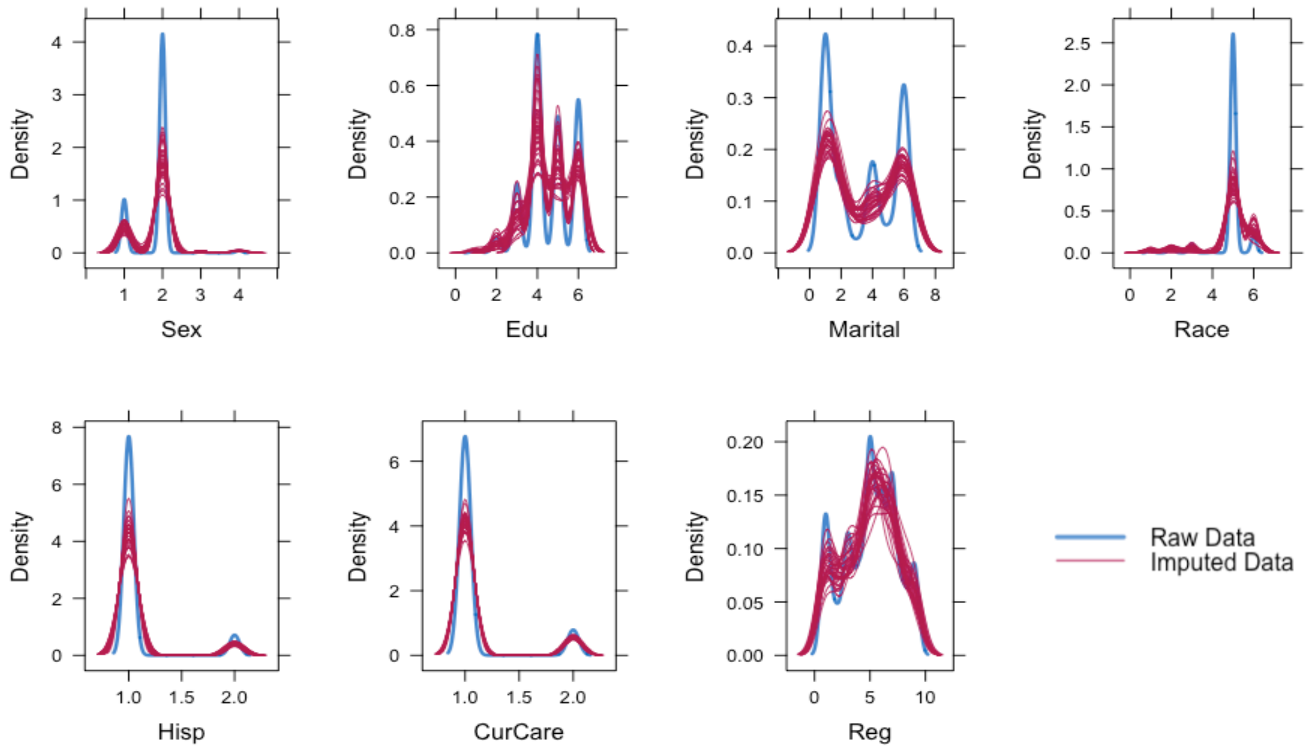


Figure 11: Qualitative Variable Distributions – Pre vs. Post-Imputation



For each of these variables, the peaks of their post-imputation densities are centered near the peak of their pre-imputation density. There are multiple red lines for the imputed data densities, because a single line is drawn for each imputed data set, and $m = 30$ imputed data sets were generated as a result of this analysis. Clearly, the pre-imputation distributions of these variables are complex, and tend not to follow a known distribution (normal, uniform, etc.). However, the post-imputation densities for some of the quantitative variables appear to closely follow a normal distribution. This is because these values were imputed by taking a random sample from a normal distribution using the RF predicted mean and variance equal to the OOB error. Therefore, although these post-imputation densities do not identically match their corresponding pre-imputation density, the fact that their post-imputation density is generally centered around the peak of their pre-imputation density indicates that the imputation procedure seemed to perform well. For the qualitative variables, their post-imputation densities align slightly more closely with their corresponding raw data densities. Furthermore, the shape of these densities are not bell-shaped like the quantitative post-imputation densities were, since the *mice* RF imputation method for qualitative variables did not randomly sample from the normal distribution, but instead simply sampled a terminal node from a randomly chosen DT.

3.2 Univariate Linear Regression Estimates

Each of the twelve linear regressions outlined in Table 6 were conducted on each of the $m = 30$ imputed data sets, and results were pooled according Rubin's rules (given by Equations 2 – 8). The regression estimates for each of these twelve regression formulas are compared between the pooled post-imputation results and the results fitting these same models on the raw, pre-imputation data in Tables 7 and 8 below for the QIDS-SR and ASRM outcomes, respectively.

Table 7: Univariate Regression Estimate Comparisons – QIDS-SR as Outcome

Model	Coef	Imputed Data						Raw Data				
		Est	SE	t	DF	P-val	λ	Est	SE	t	DF	P-val
1. QIDS ~ Age	β_0	17.225	0.427	40.299	58.235	< 0.001	0.690	18.374	0.635	28.944	-	< 0.001
	β_1	-0.052	0.011	-4.628	45.692	< 0.001	0.778	-0.067	0.014	-4.632	658	< 0.001
2. QIDS ~ NumHouse	β_0	14.357	0.387	37.135	38.506	< 0.001	0.843	14.880	0.436	34.161	-	< 0.001
	β_1	0.233	0.094	2.494	53.512	**0.016**[a]	0.720	0.281	0.148	1.896	638	**0.058**[a]
3. QIDS ~ CurCare	β_0	15.040	0.236	63.809	33.329	< 0.001	0.897	15.847	0.227	69.862	-	< 0.001
	β_1	-0.421	0.337	-1.249	100.163	0.214	0.525	-1.259	0.753	-1.672	560	0.095
4. QIDS ~ Edu	β_0	16.510	1.328	12.433	163.626	< 0.001	0.407	18.500	2.528	7.317	-	< 0.001
	β_1	0.017	1.500	0.011	132.232	< 0.001	0.455	0.088	2.810	0.031	644	< 0.001
	β_2	-0.640	1.404	-0.456	135.932		0.449	-1.630	2.601	-0.627		
	β_3	-0.753	1.342	-0.561	159.895		0.412	-2.242	2.550	-0.880		
	β_4	-1.719	1.330	-1.292	171.326		0.398	-2.886	2.561	-1.127		
	β_5	-2.996	1.311	-2.286	186.007		0.381	-4.747	2.558	-1.856		
5. QIDS ~ DBSA	β_0	13.984	0.278	50.242	32.297	< 0.001	0.908	14.277	0.293	48.785	-	< 0.001
	β_1	-0.985	0.132	-7.444	36.126	< 0.001	0.867	-1.033	0.149	-6.931	517	< 0.001
6. QIDS ~ SDS	β_0	7.625	**1.456**[b]	5.237	22.661	< 0.001	0.979	7.173	**1.268**[b]	5.655	-	< 0.001
	β_1	0.376	**0.070**[c]	5.359	22.764	< 0.001	0.979	0.455	**0.058**[c]	7.804	73	< 0.001

Table 7: *Highlighted, bolded cells indicate a key difference between the imputed data results and the raw data results. (a) The NumHouse predictor variable is statistically significant at the $\alpha = 0.05$ level in the imputed data results but is not in the raw data results. (b) The standard error estimate for the intercept of the SDS predictor variable is higher in the imputed data results than the raw data results. (c) The standard error estimate for the SDS predictor variable is higher in the imputed data results than the raw data results.*

Table 8: Univariate Regression Estimate Comparisons – ASRM as Outcome

Model	Coef	Imputed Data						Raw Data				
		Est	SE	t	DF	P-val	λ	Est	SE	t	DF	P-val
1. ASRM ~ Age	β_0	4.746	0.426	11.130	37.905	< 0.001	0.849	5.519	0.812	6.797	-	< 0.001
	β_1	-0.016	0.011	-1.504	34.618	0.142	0.883	-0.033	0.019	-1.755	277	0.080
2. ASRM ~ NumHouse	β_0	3.888	0.326	11.935	33.626	< 0.001	0.894	4.549	0.543	8.378	-	< 0.001
	β_1	**0.059**[a]	0.074	0.797	44.983	0.429	0.783	**-0.137**[a]	0.177	-0.776	271	0.439
3. ASRM ~ CurCare	β_0	4.073	0.248	16.421	27.244	< 0.001	0.956	4.393	0.281	15.626	-	< 0.001
	β_1	-0.222	0.357	-0.622	43.186	0.538	0.799	-1.275	1.059	-1.205	239	0.229
4. ASRM ~ Edu	β_0	4.233	0.987	4.290	123.827	**< 0.001**[b]	0.471	4.500	2.919	1.542	-	**0.124**[b]
	β_1	0.009	0.965	0.010	218.241	**0.023**[c]	0.350	-0.786	3.310	-0.237	267	**0.128**[c]
	β_2	0.425	1.004	0.424	124.868		0.469	0.557	3.001	0.186		
	β_3	0.041	0.948	0.043	154.579		0.420	0.252	2.944	0.086		
	β_4	-0.547	0.971	-0.563	139.290		0.443	-0.991	2.970	-0.334		
	β_5	-0.478	0.963	-0.496	144.750		0.434	-1.318	2.972	-0.444		
5. ASRM ~ DBSA	β_0	4.763	**0.323**[d]	14.766	25.378	< 0.001	0.968	4.832	**0.255**[d]	18.937	-	< 0.001
	β_1	0.696	**0.132**[e]	5.259	29.463	< 0.001	0.937	1.018	**0.131**[e]	7.794	240	< 0.001
6. ASRM ~ SDS	β_0	3.944	0.637	6.188	30.041	< 0.001	0.932	5.181	1.878	2.759	-	0.009
	β_1	**0.005**[f]	0.034	0.156	28.927	0.877	0.942	**-0.029**[f]	0.084	-0.348	35	0.730

Table 8: *Highlighted, bolded cells indicate a key difference between the imputed data results and the raw data results. (a) The estimate for the NumHouse predictor variable is positive in the imputed data results but is negative in the raw data results. (b) The intercept for the Edu predictor variable is statistically significant at the* $\alpha = 0.05$ *level in the imputed data results but is not in the raw data results. (c) The Edu predictor variable is statistically significant at the* $\alpha = 0.05$ *level in the imputed data results but is not in the raw data results. (d) The standard error estimate for the intercept of the DBSA predictor variable is higher in the imputed data results than the raw data results. (e) The standard error estimate for the DBSA predictor variable is higher in the imputed data results than the raw data results. (f) The estimate for the SDS predictor variable is positive in the imputed data results but is negative in the raw data results.*

For most models, the imputed results did not differ much from the raw data results. Coefficient estimates (Est) did tend to differ slightly in each comparison, but in most cases, were still indicating the same directionality (either a positive or negative relationship) between the outcome and predictor of interest. As seen in Table 8, the coefficient estimate for number of people in the household (NumHouse) predicting ASRM total score shifted from negative ($\beta = -0.137$) in the raw data results to positive ($\beta = 0.059$) in the imputed results. Similarly, the coefficient estimate for SDS total score predicting ASRM total score shifted from negative ($\beta = -0.029$) in the raw data results to positive ($\beta = 0.005$) in the imputed results.

Standard errors (SE) were consistently lower in the imputed results than in the raw data results in almost all cases. Given the increased sample size and the probable information gain from auxiliary variables due to the MI procedure, this is to be expected. As seen in Table 7, the SE for the intercept and the SDS total score predictor variable were slightly higher in the imputed data results ($SE = 1.546$ and $SE = 0.070$, respectively) than in the raw data results ($SE = 1.268$ and $SE = 0.058$). Similarly, in Table 8, the SE for the intercept and the DBSA total score predictor variable were slightly higher in the imputed data results ($SE = 0.323$ and $SE = 0.132$, respectively) than in the raw data results ($SE = 0.255$ and $SE = 0.131$).

The statistical significance (at the $\alpha = 0.05$ significance level) given by the p-values in both Table 7 and Table 8 (P-val) also remained mostly consistent between the imputed results and the raw data results, but there were a few notable changes. In Table 7, NumHouse was found

to be a statistically significant predictor of QIDS-SR total score in the imputed results ($p = 0.016$), but not in the raw data results ($p = 0.058$). In Table 8, the intercept for education level (Edu) was found to be statistically significant in the imputed results ($p < 0.001$), but not in the raw data results ($p = 0.124$). Additionally, Edu was found to be a statistically significant predictor of ASRM total score in the imputed results ($p = 0.023$), but not in the raw data results ($p = 0.128$).

For the models fit on the imputed data, an additional parameter that represents the proportion of the total variance that is attributable to the missing data (λ), was calculated. As indicated in Equation 8, λ represents the information lost due to missingness, which acts as a numerical representation of the quality of the imputation where higher values of λ indicate a poorer quality imputation. The λ values in the regression formulas containing predictors with very high missingness were, as expected, very high. For example, $\lambda = 0.979$ and $\lambda = 0.942$ in the regression formulas predicting QIDS-SR total score from SDS total score and ASRM total score from SDS total score, respectively. Whereas for a predictor with a much lower rate of missingness, such as age, $\lambda = 0.778$ and $\lambda = 0.883$ in the regression formulas predicting QIDS-SR total score from age and ASRM total score from age, respectively. Given the high rates of missingness for all instrument variables, and subsequently, for both outcome variables of interest in these analyses, λ is expected to be relatively high even in regression formulas containing nearly fully-observed predictor variables. In the two instances where the SE is higher in the imputed results than in the raw data results, $\lambda > 0.900$, indicating that a total variance that is attributable to the missing data above 90% may be indicative of substantial between-imputation variance.

3.3 Multiple Linear Regression Estimates

The multiple linear regression model outlined in Equation 1 was fit to both the raw MN data and the imputed data. In the same fashion as the twelve univariate linear regression models, the multiple linear regression was fit on each of the $m = 30$ imputed data sets, and the results were pooled using Rubin's rules. The regression estimates for the multiple regression formula are compared between the pooled post-imputation results and the results using the same formulas on the raw, pre-imputation data in Tables 9 below.

Table 9: Multiple Regression Estimate Comparisons

Coef	Imputed Data						Raw Data				
	Est	SE	t	DF	P-val	λ	Est	SE	t	DF	P-val
β_0 (Intercept)	16.137	0.828	19.490	36.344	**< 0.001[a]**	0.865	4.140	2.657	1.558	-	**0.260[a]**
β_1 (SDS)	**-0.113[b]**	0.055	-2.059	25.024	0.050	0.969	**0.973[b]**	0.188	5.178		0.035
β_2 (QIDS)	-0.429	0.049	-8.710	29.927	< 0.001	0.933	-0.958	0.157	-6.093		0.026
β_3 (ASRM)	0.110	0.043	2.576	37.527	0.014	0.853	0.807	0.092	8.768		0.013
β_4 (BART)	-0.009	0.009	-1.010	30.760	0.321	0.924	-0.030	0.017	-1.735		0.225
β_5 (DBSA)	**0.555[c]**	0.107	5.181	36.547	**< 0.001[d]**	0.863	**-0.933[c]**	0.275	-3.397	2	**0.077[d]**
β_6 (Age)	**0.015[e]**	0.009	1.558	39.494	0.127	0.833	**-0.044[e]**	0.039	-1.140		0.372
β_7 (Sex = Female)	-0.162	0.212	-0.769	68.103		0.638	-2.841	0.990	-2.870		
β_8 (Sex = Ambiguous)	-0.113	1.082	-0.104	95.712	0.687	0.537	-	-	-		0.103
β_9 (Sex = Other)	-0.163	0.765	-0.213	123.927		0.470	-	-	-		
Sample Size	N = 4344						N = 10				

Table 9: *Yellow highlighted, bolded cells indicate a key difference between the imputed data results and the raw data results. (a) The intercept for the model is statistically significant at the $\alpha = 0.05$ level in the imputed data results but is not in the raw data results. (b) The estimate for the SDS predictor variable is negative in the imputed data results but is positive in the raw data results. (c) The estimate for the DBSA predictor variable is positive in the imputed data results but is negative in the raw data results. (d) The DBSA predictor variable is statistically significant at the $\alpha = 0.05$ level in the imputed data results but is not in the raw data results. (e) The estimate for the Age predictor variable is positive in the imputed data results but is negative in the raw data results. Red highlighted, bolded cells indicate parameters that were unable to be estimated using the raw data – in this case, these were Sex variable categories.*

For this multiple linear regression model, there are many important differences between the imputed data results and the raw data results. As seen in Table 9, both the intercept ($p < 0.001$) and the DBSA total score predictor ($p < 0.001$) are statistically significant in the model resulting from the imputed data, but are not statistically significant ($p = 0.260$ and $p = 0.077$

for intercept and DBSA total score, respectively) in the model resulting from the raw data. Furthermore, there are major discrepancies in the directionality of the coefficient estimates between the imputed data model results and the raw data model results. The SDS total score coefficient estimate shifted from negative ($\beta = -0.113$) in the imputed results to positive ($\beta = 0.973$) in the raw data results, the DBSA total score coefficient shifted from positive ($\beta = 0.555$) in the imputed results to negative ($\beta = -0.933$) in the raw data results, and the coefficient for age shifted from positive ($\beta = 0.015$) in the imputed results to negative ($\beta = -0.044$) in the raw data results.

Other important results from comparing the multiple linear regression results include the size of the sample included in each analysis. For the imputed data results, $N = 4{,}344$ observations were used in each of the $m = 30$ runs of the regression, whereas for the raw data results, only $N = 10$ observations were used in the analysis. Another important distinction between the results of the two different models was the fact that the model fit on the raw data was unable to estimate coefficients for the "Ambiguous" and "Other" categories for the sex variable, since of the $N = 10$ total observations used in the analysis, none of those 10 participants fell into either of those categories. Using the raw data for analysis will likely often be subject to this pitfall, in that with such a small sample size, estimates for some less-common covariate patterns may not be possible, severely limiting the generalizability of the results.

4. Discussion

Using the MICE method in tandem with RFs to multiply impute missing data in large, mixed-type data sets such as MN provides a reliable, flexible way to gain useful, clinically relevant results from otherwise nearly unanalyzable data. Even for variables with large

percentages of missing data, the *mice* algorithm using RFs for each of the chained equations is able to impute the missing values of these highly missing variables with considerable accuracy, as the pre and post-distributions of these variables tended to align quite well. The univariate regression results comparing coefficient estimates, standard errors, p-values, and other quantities of interest indicate that similar results are obtained using the imputed data as compared to the raw "true" data, which is highly desirable when conducting an MI. And, the few differences that did arise between the imputed vs. raw data results for these regressions provide useful information regarding the need for both a larger sample and further efforts at getting participants to complete more assessments than they currently are.

Additionally, the two cases where the imputed results SEs were higher than the raw data SEs serve as an important reminder that using MI results is not always necessary nor is it always the best option. In these cases, there are still valuable insights to be gained; when the variable missingness and the total variance that is attributable to the missing data is too high, in addition to insufficient information gained from the auxiliary variables, a raw data analysis may be preferred over using the MI results. Alternatively, more imputed data sets may need to be generated in order to properly estimate the between-imputation variance. In any case, in most practical settings the clinician/researcher likely will not be interested in univariate results, and will instead be interested in multiple possible predictors in a given model. The multiple regression results clearly highlight the maximum potential benefit of using MI to generate multiple imputed data sets and analyze the data this way. The analyzable sample size gained from performing such a procedure is immense, and lets researchers investigate many more possible relationships between covariates that otherwise would be near impossible to do with highly missing data.

Using a RF model to multiply impute the missing data is only one of several possible options that could have been used on these data. Given the complexity of the MN data in that it contains both qualitative and quantitative variables, it made sense to consider a RF model-oriented approach. However, using other well-established methods within the MICE framework such as predictive mean matching for continuous variables, logistic regression for binary categorical variables, or linear discriminant analysis for factor variables are other viable options. Yet, many of these other methods cannot appropriately account for interactions or non-linearities in the data, and much more careful consideration would likely need to be taken if a non-RF method were to be used for imputation instead.

An alternative R package altogether, called *missForest*, was also considered for use in these analyses instead of using MICE with RFs. The *missForest* package has been shown to outperform MICE, CART, EMB and other imputation methods in multiple different data settings [53], but for MI purposes its justification for use is fundamentally flawed. The authors suggest that the nature of RFs inherently constitute a MI scheme, since as part of any RF, data are sampled with replacement at random in order to train the RF model. While the idea of randomly sampling data is similar to MI by Rubin's original standards, the *missForest* method produces only one completed data set after implementation and also does not randomly sample around the predictions for the missing values themselves, which are two blatant violations of the MI scheme. Ultimately, this ruled out the use of *missForest* for the purpose of this paper, and while *missForest* is a strong candidate for single-imputation purposes, it cannot (yet) be used for conducting a proper MI procedure.

There were some important limitations in this analysis that need to be considered. First, there was an extremely high amount of missing data for the instrument variables. This resulted in

high λ values, which indicates a lower-quality imputation than what might be hoped for. Relatedly, in the case where both the predictor variable and the outcome variable are highly missing (as seen in Table 8, Model 6), caution should be taken in generalizing these results towards the larger population. However, the regression estimates even with these high λ values and large amount of missing data for the instrument variables still yielded imputed data results that were similar to the raw data results in almost all cases. Future efforts towards encouraging MN participants to complete as many of the instruments as possible may help with this issue.

Additionally, nearly 70% of all MN participants had not completed *any* of the instruments. Thus, imputing instrument scores for these individuals without any within-subject information gain from instrument variables may not be advisable. Nevertheless, it does seem plausible that the demographic factors available in the MN data would be sufficient to impute instrument score values for these participants, due to the high quality of the imputed results. Further encouragement of participants to complete instruments would also likely help with this issue. Lastly, the variables chosen for this analysis were merely a subset of the full MN data available. It is possible that some information was lost in the exclusion of select demographic and/or instrument variables, and that the imputed data may provide a better representation of the "true" population if these omitted variables had been included.

These limitations point to numerous possible extensions of applying this MI scheme to the MN data. Future analyses could include all of the possible variables, even if they are considered less clinically relevant, to determine if this might result in higher quality imputations. Additionally, as van Buuren and Groothuis-Oudshoorn point out, parallel computation streams could also be implemented using another R package in combination with the mice package in order to reduce computation time [59]. Different imputation models (other than RFs) might also

be applied to these data to compare the resulting parameter estimates to these results, in an attempt to determine a "best" imputation model choice. A comparison of the statistical power between the raw data analyses and the imputed data analyses would also be a worthwhile investigation, requiring more rigorous calculations for the power of the imputed analyses. Lastly, the complex data structure of the MN data is not unique to this one data set – seeing as online support platforms and projects are becoming increasingly popular in psychiatric research, the MI method outlined in this paper provides a framework for analysis of many more data sets of a similar type to MN. Given the non-parametric, flexible nature of RFs and the statistical validity of MI, the challenges that missing data impose upon psychiatric data can be properly accounted for with relative ease using these two statistical techniques.

In conclusion, using MICE with RF models for the imputation of the missing data associated with MN yields multiple, complete data sets that are very similar to the raw, unimputed data. Using proper statistical techniques for the pooling of results across the multiple imputed data sets gives accurate parameter estimates with SEs almost always lower than that of the unimputed data, due largely to the drastically advantageous increase in analyzable sample size and information gain from auxiliary variables. Providing clinicians with these multiple complete data sets paves the way for numerous future studies and analyses that otherwise might not be possible with the complete cases data. Additionally, the code that has been written as a product of conducting this analysis provides a general, relatively simple framework for future statisticians to follow so that these future analyses can be possible; even for data sets other than the MN data. All of this is with the hope in mind that advancing the ability to analyze and draw valid inference from psychiatric data will enable clinicians and statisticians alike to work

harmoniously to improve the health of individuals across the world suffering from mood and other mental health disorders.

5. References

[1] Allison, P. D. (2000). Multiple imputation for missing data: A cautionary tale. *Sociological Methods & Research*, 28(3); 301-309.

[2] Altman, E. G., Hedeker, D., Peterson, J. L., & Davis, J. M. (1997). The Altman Self-Rating Mania Scale. *Biological Psychiatry*, 42(10); 948-955.

[3] Auguie, Baptiste (2017). gridExtra: Miscellaneous Functions for "Grid" Graphics. R package version 2.3. https://CRAN.R-project.org/package=gridExtra.

[4] Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: what is it and how does it work?. *International Journal of Methods in Psychiatric Research*, 20(1); 40-9.

[5] Bech, P. (2004). Measuring the dimension of psychological general well-being by the WHO-5. *Quality of Life Newsletter*, 15-16.

[6] Breiman, L. (2001). "Random forests", *Mach. Learn,* 45; 5-32).

[7] Burgette L.F., Reiter J.P. (2010). Multiple imputation for missing data via sequential regression trees. *Am J Epidemiol*, 172(9); 1070-1076.

[8] Catellier D.J. et al. (2005). Imputation of missing data when measuring physical activity by accelerometry. *Med Sci Sports Exerc*, 37(11); S555-62.

[9] Chang, Robert (2014). Machine Learning: a Probabilistic Perspective, *CHANCE*, 27(2); 62-63.

[10] Cohen, J. (1988). Statistical power analysis for the behavioral sciences. (2).

[11] Collins L.M,. Schafer J.L., Kam C.M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, (6); 330-351.

[12] Depression and Bipolar Support Alliance. DBSA wellness tracker. http://www.dbsalliance.org/site/PageServer?pagename=wellness_tracker.

[13] Dibal, Nicholas Pindar, et al. (2017). Challenges and Implications of Missing Data on the Validity of Inferences and Options for Choosing the Right Strategy in Handling Them. *International Journal of Statistical Distributions and Applications*, 3(4); 87-94., doi:10.11648/j.ijsd.20170304.15.

[14] Domingos, Pedro (2012). A Few Useful Things to Know about Machine Learning. *Communications of the ACM,* 55(10); 78-87.

[15] Dong, Y., & Peng, C. Y. (2013). Principled missing data methods for researchers. *SpringerPlus*, 2(1); 222. doi:10.1186/2193-1801-2-222.

[16] Dowle, Matt & Srinivasan, Arun (2017). data.table: Extension of `data.frame`. R package version 1.10.4-3. https://CRAN.R-project.org/package=data.table.

[17] Dragulescu, Adrian A. (2014). xlsx: Read, write, format Excel 2007 and Excel 97/2000/XP/2003 files. R package version 0.5.7. https://CRAN.R-project.org/package=xlsx.

[18] Livingston, F. (2005). Implementation of Breiman's random forest machine learning algorithm. *Machine Learning Journal Paper*.

[19] Fleurence R. et al. (2013). How the Patient-Centered Outcomes Research Institute is engaging patients and others in shaping its research agenda. *Health Aff*., (32); 393-400.

[20] Graham, J.W., Olchowski, A. E., Gilreath, T. D. (2007). "How many imputations are really needed: Some practical clarifications of multiple imputation theory". *Prevention Science*, (8); 206-213.

[21] Graham, J.W. (2009). Missing data analysis: making it work in the real world. *Annual Review of Psychology*, (60); 549-576.

[22] Grolemund, Garrett & Wickham, Hadley (2011). Dates and Times Made Easy with lubridate. *Journal of Statistical Software*, 40(3); 1-25. http://www.jstatsoft.org/v40/i03/.

[23] Hardt, J., Herke, M., Leonhart, R. (2012). Auxiliary variables in multiple imputation in regression with missing X: a warning against including too many in small sample research. *BMC Med ResMethodol*, 12(1); 184.

[24] Harrell Jr., Frank E., with contributions from Charles Dupont and many others (2017). Hmisc: Harrell Miscellaneous. R package version 4.0-3. https://CRAN.R-project.org/package=Hmisc.

[25] James, Gareth et al. (2017). Chapter 8: Tree-Based Methods. An Introduction to Statistical Learning with Applications in R. *Springer*, 316-319.

[26] Johns Hopkins Medicine (2011). Overview of Mood Disorders. https://www.hopkinsmedicine.org/healthlibrary/conditions/mental_health_disorders/overview_of_mood_disorders_85, P00759.

[27] Kenney, J. F. & Keeping, E. S. (1962). Chapter 15: Linear Regression and Correlation. Mathematics of Statistics, Pt. 1. (3); 252-285.

[28] Kingsford, C., & Salzberg, S. L. (2008). What are decision trees?. *Nature Biotechnology*, 26(9); 1011-1013. http://doi.org/10.1038/nbt0908-1011.

[29] Kowarik, Alexander & Templ, Matthias (2016). Imputation with the R Package VIM. *Journal of Statistical Software,* 74(7); 1-16. doi:10.18637/jss.v074.i07.

[30] Lee, Jin Hyuk & Huber Jr., John (2011). Multiple imputation with large proportions of missing data: How much is too much?. *United Kingdom Stata Users' Group Meetings*, (23).

[31] Lejuez, C. W. et al. (2002). Evaluation of a behavioral measure of risk taking: The Balloon Analogue Risk Task (BART). *Journal of Experimental Psychology: Applied*, 8(2); 75-84.

[32] Little, R.J.A., & Rubin, D. B. (2002). Statistical analysis with missing data. (2).

[33] Little, R.J.A. (2009). Comments on: Missing Data Methods in Longitudinal Studies: A Review. *Test*, (18); 47-50.

[34] Molenberghs, G., Fitzmaurice, G., Kenward, M. G., Tsiatis, A., Verbeke, G. (Eds.) (2015). *Handbook of Missing Data Methodology*. Boca Raton, FL: Chapman & Hall/CRC.

[35] Moons, K.G.M., Donders, R., Stijnen, T., Harrell, F.J. (2006). Using the Outcome for Imputation of Missing Predictor Values was Preferred. *Journal of Clinical Epidemiology*, 59(10); 1092-1101.

[36] Ostacoli, L. et al. (2013). Age of Onset of Mood Disorders and Complexity of Personality Traits. *ISRN Psychiatry*, 1(7). doi:10.1155/2013/246358.

[37] R Core Team (2017). "R: A language and environment for statistical computing". *R Foundation for Statistical Computing*, Vienna, Austria. URL https://www.R-project.org/.

[38] Revelle, W. (2017). psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA, https://CRAN.R-project.org/package=psych, Version = 1.7.8.

[39] Robinson, David (2017). broom: Convert Statistical Analysis Objects into Tidy Data Frames. R package version 0.4.3. https://CRAN.R-project.org/package=broom

[40] Rodwell, L., Lee, K. J., Romaniuk, H., Carlin, J. B. (2014). Comparison of methods for imputing limited-range variables: a simulation study. *BMC Medical Research Methodology*, (14); 57. http://doi.org/10.1186/1471-2288-14-57.

[41] Rubin, D. B. (1976). Inference and missing data. *Biometrika,* 63(3); 581-592.

[42] Rubin, D. B. (1987). Multiple imputation for non-response in Surveys.

[43] Rubin, D.B. (1988). Multiple Imputation for Data-Base Construction. *Compstat*, 389-400. Physica-Verlag HD, doi:10.1007/978-3-642-46900-8_53.

[44] Rush, A.J. et al. (2003). The 16-Item Quick Inventory of Depressive Symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): a psychometric evaluation in patients with chronic major depression. *Biol. Psychiatry,* (54); 573-583.

[45] Schafer, J.L. (1997). Analysis of incomplete multivariate data. Chapman & Hall/CRC.

[46] Schafer, J.L., & Graham, J.W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7(2); 147-177.

[47] Schafer, J.L. (2003). Multiple imputation in multivariate problems when the imputation and analysis models differ. *Statistica Neerlandica*, (57); 19-35.

[48] Schunk, D. (2008). A Markov Chain Monte Carlo Algorithm for Multiple Imputation in Large Surveys. *Advances in Statistical Analysis*, 92(1); 101-114.

[49] Shah, A., Bartlett, J., Carpenter, J., Nicholas, O., Hemingway, H. (2014). Comparison of random forest and parametric imputation models for imputing missing data using MICE: A CALIBER study. *American Journal of Epidemiology*, 179(6); 764-744. doi:10.1093/aje/kwt312.

[50] Shah, Anoop (2018). CALIBERrfimpute: Imputation in MICE using Random Forest. R package version 1.0-1.

[51] Sheehan, D. V. (1983). Sheehan disability scale. *Handbook of Psychiatric Measures*, (2); 100-2.

[52] Statnikov A., Wang, L., Aliferis, C. (2008). A comprehensive comparison of Random Forests and support vector machines for microarray-based cancer classification. *BMC Bioinformatics*, 9(319). doi:10.1186/1471-2105-9-319.

[53] Stekhoven, Daniel J. & Bühlmann, Peter. (2012). MissForest – non-parametric missing value imputation for mixed-type data. *Bioinformatics*. 28(1); 112-118. https://doi.org/10.1093/bioinformatics/btr597

[54] Sylvia, L. G. et al. (2018). MoodNetwork. *Medical Care*, 1-5. doi:10.1097/mlr.0000000000000789.

[55] Tabachnick B.G. & Fidell, L.S. (2012). Using multivariate statistics. (6). Needham Heights, MA: Allyn & Bacon.

[56] Tang, Fei (2017). Random Forest Missing Data Approaches. *Open Access Dissertations*. (1852). https://scholarlyrepository.miami.edu/oa_dissertations/1852.

[57] United States Census Bureau, Geography Division (2015). Census Regions and Divisions of the United States. https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf.

[58] van Buuren, S., Brand, J.P.L., Groothuis-Oudshoorn, K., Rubin, D.B. (2006). Fully Conditional Specification in Multivariate Imputation. *Journal of Statistical Computation and Simulation*, 76(12); 1049-1064.

[59] van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45(3); 1-67. doi:http://dx.doi.org/10.18637/jss.v045.i03.

[60] von Hippel, P.T. (2013). Should a Normal Imputation Model be Modified to Impute Skewed Variables? *Sociol Method Res*, 42(1); 105-138. doi:10.1177/0049124112464866.

[61] Wickham, Hadley (2017). tidyverse: Easily Install and Load the 'Tidyverse'. R package version 1.2.1. https://CRAN.R-project.org/package=tidyverse.

[62] World Health Organization (2018). Mental disorders. http://www.who.int/news-room/fact-sheets/detail/mental-disorders.