Exploratory Analyses for Missing Data in Meta-Analyses

Introduction

Systematic reviews of substance abuse research hold great promise for examining what makes different substance abuse interventions effective. Methodological tools such as meta-regression can formally test relationships between how effective an intervention is and how or on whom it is implemented. However, such tools must contend with the real-world difficulties of modern research syntheses, including the fact that it is often impossible to extract all relevant information from the literature to conduct such analyses.

The fact that not every study reports the information required to run a meta-regression means that many meta-analyses run into a missing data problem. Issues with missing data are not new. There is a large literature on methods for handling missing data in primary studies, as well as work on related issues in meta-analysis. This literature highlights the ways that missingness can bias an analysis, examines conditions under which these biases can be corrected, and proposes various statistical procedures to adjust for bias.

Diagnosing missing data issues remains an important aspect of any statistical analysis of incomplete data (i.e., data with missing values). Understanding what and how much data is missing, and how problematic that is crucial for contextualizing the results of a meta-analysis. As well, a key assumption of analyses involving missing data is that the analyst has some idea about why data is missing. While much of the literature has focused on the implications of that assumption, considerably less attention is paid to approaches to examining it in a dataset.

Recent research has suggested analysts can better understand missingness in their data through exploratory analyses, including visual and numerical summaries akin to classical exploratory data analyses. These explorations, which occur before conducting the formal meta-analysis, can shed greater light on key issues relevant to missingness. Tools for doing so are only now emerging, but these tools have yet to gain broader traction in quantitative disciplines. Nor has this approach seemingly made its way into modern meta-analyses, where missing data is a common problem.

This tutorial discusses tools for exploring and diagnosing missing data problems in meta-analysis. The following section clarifies the types of missing data for which these tools are appropriate. We then describe principles of missing data that can guide exploratory analyses. Finally, we demonstrate some of these tools on data from a meta-analysis on substance abuse interventions for adolescents. In addition, our supplementary material expands on these results with a vignette that contains additional visualizations and executable code.

Missing Data and Meta-Analysis

Because a meta-analysis involves a series of primary studies, "missing data" in meta-analysis can be seen as a broad term that could refer to several different scenarios. For instance, data could be missing on individual participants within studies, including their outcomes in the study or other characteristics (e.g., their age, race, prior substance use). "Missing data" could also refer when information cannot be extracted from a completed study by a meta-analyst. This might occur if a study fails to report enough detail for analysts to back out effect estimates, standard errors, or study-level characteristics. Finally, entire studies or effects may be missing from a meta-analytic dataset. This might occur if effects (or entire studies) are not reported or published. There is empirical evidence that statistically significant results are more likely to be published and hence wind up in a meta-analysis, which can induce *publication bias*, a well-known problem in the field. The studies or effects that are not reported, and thus are not included in a meta-analysis, can be considered missing data.

Precisely how to examine, diagnose, and adjust for missing data will be different depending on what scenario we mean when we say "missing data." For instance, meta-analysts have used *funnel plots* to examine if their systematic review is missing studies or effects due to publication bias. Our focus will be on the second scenario, where information cannot be extracted from some studies. This is a common problem in meta-analysis and one that that can limit the accuracy of any statistical inferences.

Assume we have data on k effect estimates and p variables (including the estimate itself). This can be summarized and stored in a $k \times p$ table where rows correspond to effect estimates and columns correspond to variables concerning those estimates. One column would contain the effect estimates themselves, and another would contain the standard error or estimation error variance of those estimates. The remaining $p - 2$ columns could contain effect- or study-level covariates, including summary demographics (e.g., the percent of a study's sample that were minorities), treatment type (e.g., behavioral therapy versus pharmacological interventions), or dosage/duration of an intervention. Some of the cells in this table may be missing values, and the analyses presented in this article provide ways to summarize and examine patterns of missingness.

Data

A prime example of this type of missingness can be seen in data from Tanner-Smith et al. (2016), who examined the effects of substance abuse interventions for adolescents on subsequent substance use. These data were extracted from 61 randomized trials and quasi-experiments, and include $k = 95$ different effect size estimates.

These estimates include contrasts between a given treatment condition and a control condition within a study, or between two different treatment conditions in the same study. These data will be used to illustrate key concepts of missingness and some useful tools to exploring missingness in this tutorial.

Tanner-Smith et al. found a range of intervention types and venues that have been studied on individuals who use different substances and who differ in a variety of ways. Some interventions in their data focus on cognitive behavioral therapy (CBT), family therapy, or pharmacological therapy. Some interventions are in-patient, and others are out-patient. Individuals in studies might present using marijuana, which is most common among adolescents, or alcohol or opioids, and they may come from wealthy families or poor families. Finally, some effects reported in studies contrasted a given intervention with some control condition, while others contrasted two alternative interventions or implementations.

To explore relevant relationships, Tanner-Smith et al. extracted a considerable amount of information from the studies they found. Their raw data included some $p = 46$ variables per study. In addition to estimated effects and their standard errors, they documented the types of interventions being contrasted, as well as their intensity and context. This included where interventions occurred, and how much time subjects spent in the intervention. For instance, if a study contrasted two interventions that involved behavioral therapy, Tanner-Smith et al. documented how many hours per week subjects in each intervention (referred to in this article as *groups*) spent in therapy. They also documented the demographics of subjects in the studies, such as the percentage of subjects who were minorities, as well as the substances that subjects reported using.

Tanner-Smith et al. then fit a series of meta-regression models to their data in order to examine how treatment effects varied according to the type of therapies and individuals studied. They found that assertive continuing care (ACC), behavioral therapy, CBT, motivational enhancement therapy (MET), and family therapy tended to be more effective than generic "practice as usual" interventions that often involved referrals to community services. However, they did not find strong relationships between the characteristics of adolescents in the studies and the effectiveness of interventions (net of intervention type).

A complicating factor in conducting these analyses was that some of the data were missing. Not every study reported the requisite information for extracting covariates for every effect size. For instance, not all studies reported how many hours per week subjects spend in therapy or the racial or socioeconomic makeup of their subject pool. As a result, not all effect estimates had information about the types of individuals in the study or the intensity of the interventions. It was often the case that one or two of the fields in their data table were missing for any given effect estimate. Thus, when it came time to run meta-regressions, Tanner et al. were faced with a decision about how to address effects for which they had missing covariates.

Principles of Missing Data

Tanner-Smith et al. ultimately opted for a sophisticated statistical procedure called the expectation-maximization (EM) algorithm to estimate their models, which has been an important tool for analyzing data with missing values. However, that was not their only option. A common approach in meta-analysis is a *complete-case* analysis that excludes effects for which any of the relevant covariates in the meta-regression model are missing.

An alternative to complete-case analysis that has gained broad use in various fields is to impute (i.e., fill in) missing values. Pigott (2019) discusses the use of single-value imputations in meta-analysis. A more common approach to missingness in primary studies is to use multiple imputations for a missing value, which can better reflect the uncertainty introduced by filling in unknown values. Often imputations are based on predictive models that can better inform what values we might have observed had a given field not been missing. These predictive models typically leverage information from other variables in the data.

Analyses involving incomplete data will be impacted by which variables are missing in a dataset and how frequently they are missing, as well as relationships between variables. In this section we provide an overview of principles of missing data that apply to meta-analysis, and list some potential statistical approaches to handling missing data.

Notation

As noted above, we can describe the datasets commonly used in meta-analysis as tables with k rows and p columns. Each row corresponds to an effect size, and each column corresponds to some variable related to that effect. These are the types of tables used in most standard meta-analysis software, including Comprehensive Meta-Analysis, the `metafor` library, or OpenMeta[Analyst]. The structure of these tables is shown in the matrix below. In the matrix, T_i denote the effect size estimates and σ_i are their standard errors. The X_{ij} refer to additional variables collected that pertain to a given effect size and that might be used in an analysis.

$$\begin{bmatrix} T_1 & \sigma_1 & X_{11} & ... & X_{1,p-2} \\ T_2 & \sigma_2 & X_{21} & ... & \mathbf{NA} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ T_k & \sigma_k & \mathbf{NA} & ... & X_{k,p-2} \end{bmatrix}$$

Missing values in the matrix are denoted as **NA**. Most missing data literature augments the traditional table above with a table $\mathbf{R}$ of response indicators R_{ij}. These indicators take a value of $R_{ij} = 1$ if entry i, j in the table is observed, and $R_{ij} = 0$ if it is missing. For instance, $R_{21} = 1$ because the first element of row 2 (T_2) is observed, but $R_{2p} = 0$ because the last element of row 2 is missing (NA).

In the matrix above, there are both *observed* and *unobserved* data. The observed data comprises all of the entries in the table above for which $R_{ij} = 1$. For instance, the observed data would include the effect estimates T_i and standard errors σ_i. The unobserved data comprises all of the entries for which $R_{ij} = 0$, including $X_{2,p-2}$ and $X_{k,1}$. Note that the entire matrix above, and hence the incomplete data on which we conduct an analysis, is simply the union of the observed and unobserved data.

It is often of interest how much data is missing. This could refer to several different quantitites. For instance, we could be interested in the total amount of cells with missing values:

$$\sum_{i=1}^{k} \sum_{j=1}^{p} \frac{1 - R_{ij}}{kp}$$

We might also be interested in how many observations are missing a given variable (i.e., how much of each column is missing):

$$\sum_{i=1}^{k} \frac{1 - R_{ij}}{k}$$

While such percentages are somewhat intuitive and can be used in exploratory analyses of missingness, they are not the only numerical summary of missingness. In a meta-analysis, models are often estimated by weighting effect size estimates in a way that those estimates that are most precisely estimated (i.e., that have the smallest standard errors) receive the most weight. This means that the precision with which we estimate important quantities in a meta-analysis, including meta-regression coefficients, will therefore depend on the precision of the studies included in an analysis. That is, we will have better estimates of a meta-regression model if the precision of each effect $w_i = 1/\sigma_i^2$ is large. Missing a variable for an effect estimate with a large standard error (and thus low precision w_i) can potentially be less detrimental than missing the same variable for an effect estimate with a small standard error (i.e., a large w_i). When standard errors of effect sizes are fully observed, an alternative way to quantify the extent of missingness is with a precision-weighted percentage:

$$\sum_{i=1}^{k} \frac{w_i(1-R_{ij})}{\sum_{i=1}^{k} w_i}$$

This describes the percentage of the total precision of effects for which a covariate is missing. If this is greater than the raw percentage, that would indicate that the missingness problem may be more acute because larger studies might be missing important covariates.

Missingness Mechanisms

A key assumption that underpins how to analyze a dataset with missing values involves why those values are missing, typically referred to as the missingness *mechanism*. Rubin (1976) classified three different possible types of mechanisms based on the probability that a value is missing. Rubin noted that data could be missing completely at random (MCAR), which means that the probability that a given value is missing is independent of all of the observed or unobserved data. This can be expressed as $P[R|\text{observed data}, \text{unobserved data}] = P[R]$, and intuitively means the fact that a given value is missing is unrelated to anything; it is as if it were deleted on the basis of a coin flip.

Values could be missing at random (MAR), which occurs if the probability that a value is missing depends only on the observed data. This can be expressed as $P[R|\text{observed data}, \text{unobserved data}] = P[R||\text{observed data}]$. Note that this differs from MCAR in that missingness might be related to observed values. For instance, suppose studies with larger standard errors are less likely to report the racial composition of their samples. Then, assuming the standard errors are observed, this might constitute MAR. It would not constitute MCAR, because missingness is related to an observed value, the standard error of an effect estimate.

Finally, data are said to be missing not at random (MNAR) if the probability that a value is missing depends on the unobserved data in some way. This differs from MAR in that missingness depends on unobserved data. As an example, suppose that studies with larger standard errors and a greater proportion of minorities are less likely to report the racial composition of their samples. Then missingness of racial categories in the data would depend on an observed value (the standard error), but also the racial composition that could itself be missing.

Missingness Patterns

In addition to the mechanism, it is often useful to understand which variables are missing together from the same rows. For instance, some rows in the Tanner-Smith et al. data might be missing the hours of therapy per week for one of the groups, while other row may be missing the hours of therapy per week *and* the percentage of study participants who were minorities. In other words, different rows may exhibit different *missingness patterns*. Understanding these patterns can give some insight into missingness mechanisms, but it can also help identify variables that might be more or less useful in dealing with issues that arise from missingness. For instance, it will be difficult to build imputation models for a variable if it is frequently missing alongside several other variables in the data. It would also seem useful to examine how missingness in one variable is related to observed values of other variables. If missingness in one column (e.g., therapy hours per week) occurs more frequently with low values of another column (e.g., the effect estimate), that might be cause to re-assess any MCAR assumptions.

Exploratory Analyses

The tools discussed in the remainder of this article facilitate exploratory analyses of missingness in a meta-analytic dataset. Exploratory missingness analyses (EMA) are difficult for several reasons. First, they are not always wholly conclusive. It will often be difficult to draw ironclad inferences about missingness mechanisms based on exploratory analyses. Even proposed tests for missingness mechanisms can have misleading results. Instead, EMA can provide support for or help generate theories that explain missingness in ways that are consistent with the mechanisms described above. Some of these theories may require consultation with data curators and other individuals who extracted information from the studies reviewed.

Second, there is no single visualization or set of metrics guaranteed to provide a complete picture of missingness for all datasets. A plot that is tremendously useful for one dataset may be of less interest in others.

Third, EMA differs from traditional exploratory data analyses because the focus on the unobserved data. Many software tools, including most graphics software actually deletes observations with missing values, which would eliminate the information EMA seek to understand. Thus, a new and emerging suite of tools that differ from traditional exploratory data analyses are required. Moreover, as with the case with much general-use statistical software, adaptations of these tools may be necessary to tailor EMA for meta-analytic datasets.

In the following sections, we show and discuss different types of visualizations and numerical summaries relevant to EMA. These are rooted in an approach to understand the scope and correlates of missingness problems in a dataset described in the previous section. However, they are not exhaustive, and so as part of the supplementary material to this tutorial, we have included a vignette that presents and describes alternative visualization and numerical summaries of missingness. Both the demonstration presented in this article and the supplementary vignette are implemented in the `R` software language and draw heavily on the `visdat` and `naniar` libraries with some custom extensions developed specifically for meta-analysis.

Aggregation Plots

Aggregation plots can be a useful starting point when exploring missingness in a meta-analytic dataset. They visualize the entire dataset as a "heatmap" that indicates which values are missing from which rows. Two examples of aggregation plots are given in Figure 1 below. The plots in Figure 1 are laid out exactly like the data: the columns correspond to variables in the data, and rows correspond to effect sizes. Figure 1A colors each column according to the type of variable (e.g., numeric, categorical, etc.) they represent, and gray spots indicate that a value in a given row and column is missing. Figure 1B does not indicate the type of variables in the data, but instead shows the total number of cells that contain missing values. That is, out of all of the $k \times p$ cells in the data, 6.2% are missing values.

Aggregation plots provide a high-level picture of missingness in a dataset. They can show which columns are complete, such as the columns corresponding to the effect size estimates and standard errors, as well as which columns or groups of columns have a lot of missingness. In particularly, Figure 1 shows that there are groups of variables regarding one of the treatment arms (group 2) in studies that are missing fairly frequently.

Univariate Explorations

While aggregation plots can provide a good overview, typically we want more detailed information about how many observations are missing a given variable. Variable missing plots can provide a convenient way to explore this. These plots display the overall missingness in each column of a dataset and present the results in order of which column has the most missingness.

Figure 2 shows two approaches to making these plots. Figure 2A shows the raw count missing values in each column, while Figure 2B shows the percent of each column missing. For example, Figure 2 shows that the hours per week that the contrast group spent in their assigned condition (`g2hrsperweek`) is missing for almost half of the effects in the data.

From variable missing plots, it is often easy to identify variables that might be driving any missing

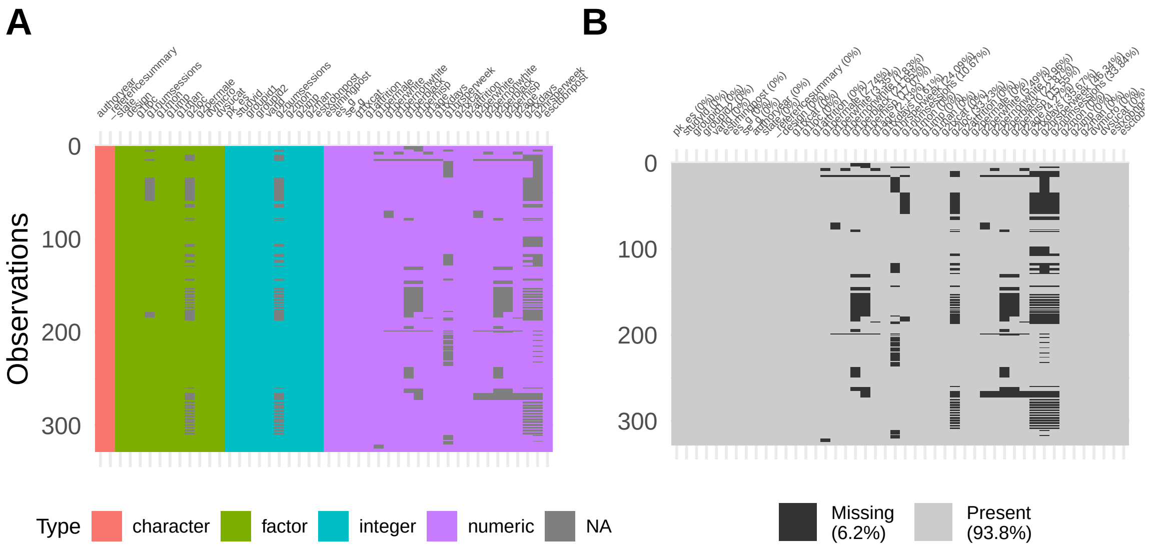

Figure 1: *These plots indicate the severity of missingness in the adolescent substance abuse intervention data. Each row in the plot corresponds to a row in the data, and each column corresponds to a variable collected in the data. Missing cells in the data are indicated by a gray dash in plot (A) and a black dash in plot (b). Plot (A) also colors columns according to the type of variables they store. Plot (B) indicates that 6.2% of the total dataset (i.e., 6.3% of all cells in the table) are missing values.*

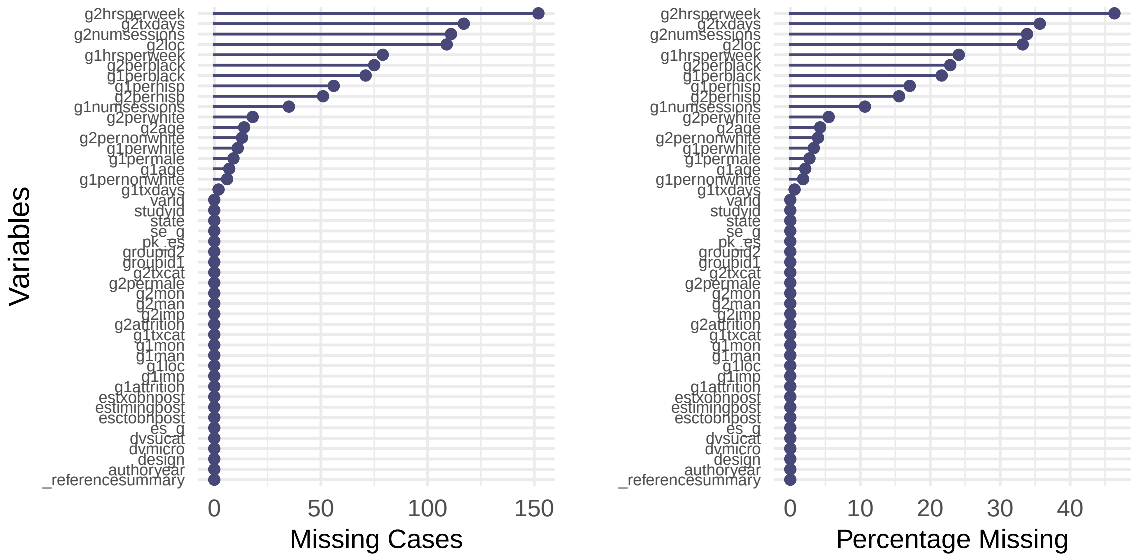

Figure 2: *Graphical summaries of missingness in variables, ordered by missingness, for the Substance Abuse data. There are 10 variables with at least 10% of missing cases. This visualization becomes relevant when deciding which variable to include in the analysis.*

data problems, and they can quantify the extent to which a given column has missing values on the scale of raw percentages. Previously, we argued that precision-weighted percentages may be more informative in describing the extent of missingness in a meta-analysis. A comparison of raw percentages and precision-weighted percentages is given below:

Variable	# Missing	% Missing	Weighted % Missing
g2hrsperweek	152	46.3	45.2
g2txdays	117	35.7	32.6
g2numsessions	111	33.8	30.2
g2loc	109	33.2	30.0
g1hrsperweek	79	24.1	36.8
g2perblack	75	22.9	17.5
g1perblack	71	21.6	16.1
g1perhisp	56	17.1	12.7
g2perhisp	51	15.5	12.4
g1numsessions	35	10.7	6.9
g2perwhite	18	5.5	4.6
g2age	14	4.3	3.0
g2pernonwhite	13	4.0	2.3
g1perwhite	11	3.4	3.1
g1permale	9	2.7	1.3
g1age	7	2.1	1.5
g1pernonwhite	6	1.8	0.8
g1txdays	2	0.6	0.1

Exploring Patterns of Missingness

Different combinations of missingness across cases can be visualized using an “upset plot” (Conway et al. 2017) with the `gg_miss_upset()` function in the `naniar` package; thus providing the number of times certain variables go missing together.

We explore the combinations among variables with higher percentage of missing (see Figure 3 below).

Relating Missingness to Observed Values

Variable missingness in cases over some other factor variable is visualized with a “heatmap” style using the `gg_miss_fct()` function from the `naniar`package (see Figure 4 below). This provides information regarding any relationship between observed values and missingness condition.

To explore the relationship of variables presenting a large percentage of missing data with effect size and error variances, the original data is transform in order to create a data structure that keeps track of missing values (Tierney, 2018). Using the `as_shadow` and `bind_shadow` functions from the `naniar` package a data frame with the same set of columns, but with the column names added a `suffix_NA`, is bound to the original dataset.

Later, the distribution of effect size and error variance are visualized when some covariates are missing, and when they are not using the ggplot function. Figure 5 shows three scenarios with different relevant covariates.

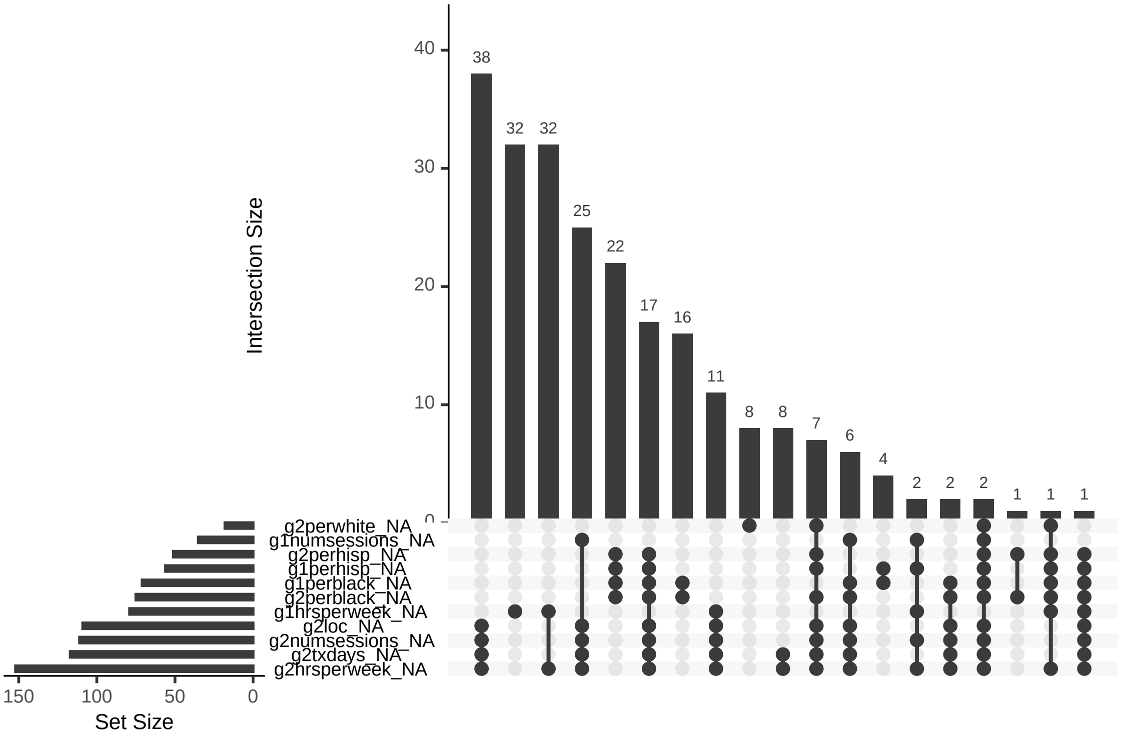

Figure 3: *Details those variables that are missing together. For instance, there are a large number of cases where Group 2 Level of Care, Number of Sessions, Treatment Contact (hours per week) and Duration of Treatment (days) are missing together. This simple exploration provides valuable information for imputation.*

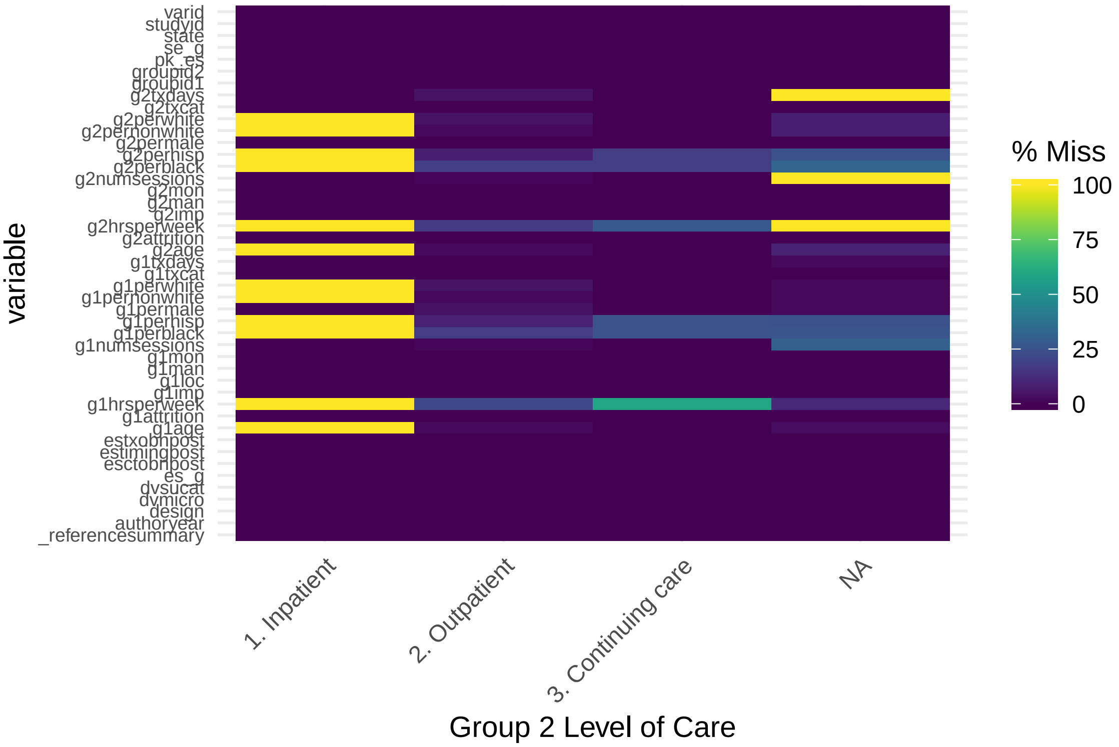

Figure 4: *Highlights the number of missings in each column, broken down by a factor variable, in this case the Level of Care for group 2. The inpatient category has 100% of missing values in at least 12 different variables, suggesting that this category could impose a problem when fitting a regression model.*

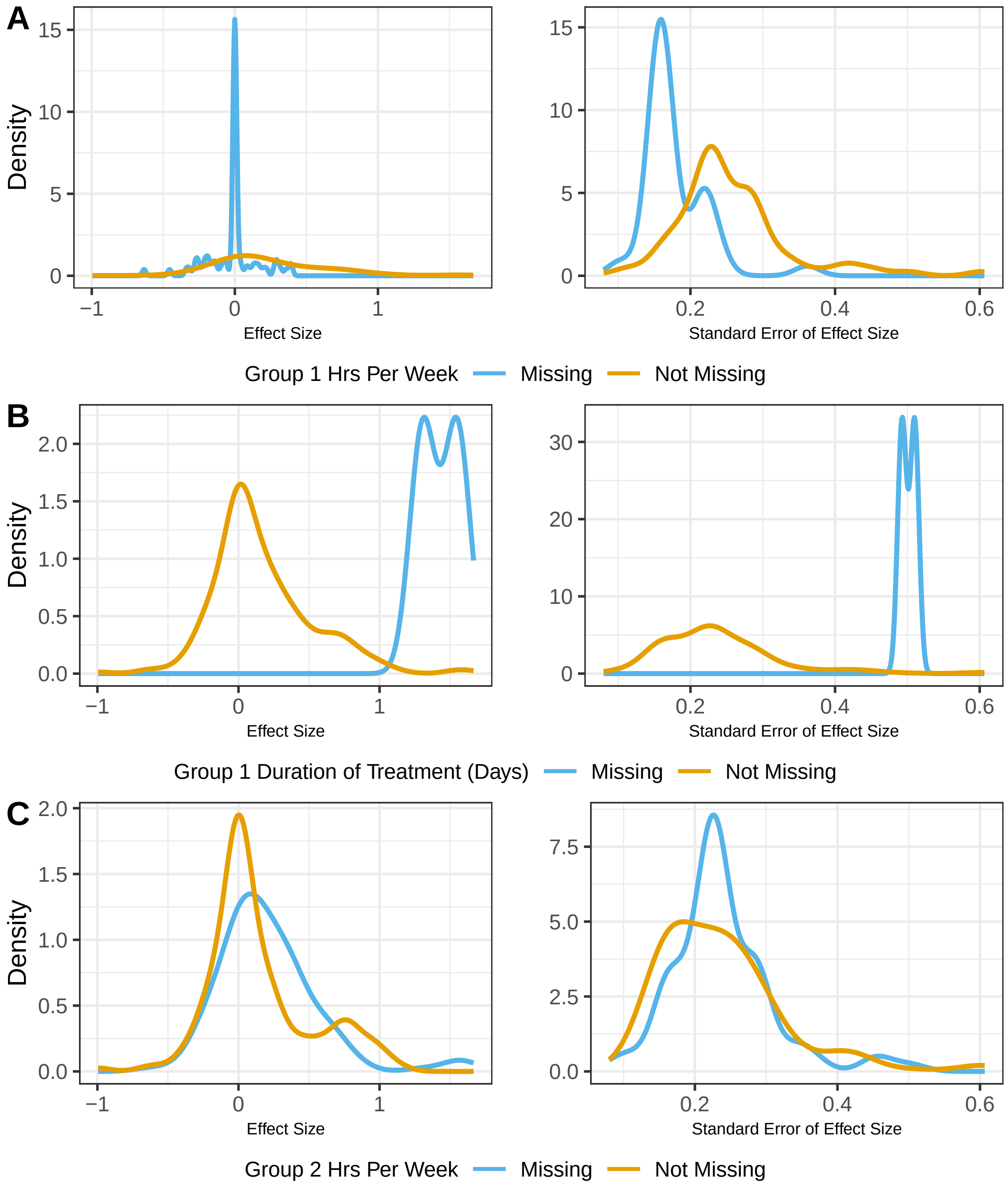

Figure 5: *Plot (A) shows that the covariance Duration of Treatment (days) for Group 1 is mostly missing for larger effect size values. Further, the effect size has larger standard error, when this covariate is missing. Plot (B) illustrates a case where the effect size tends to be closer to zero when a particular covariate is missing. Specifically, when Treatment Contact (hours per week) for group 1 is missing, both the effect size and its standard errors tend to be smaller than when the covariate is present.Plot (C) shows that both, the effect size and its standard errors, have a similar distribution either when the covariate Treatment Contact (hours per week) for group 2 is present or not.*