Exploratory Analyses for Missing Data in Meta-Analyses

Introduction

Systematic reviews of substance abuse literature hold great promise for unpacking correlates of effective substance abuse interventions. Methodological tools such as meta-regression can formally test relationships between the type or implementation of an intervention and how effective it is. However, such tools must contend with the real-world difficulties of modern research syntheses, including the fact that it is often impossible to extract relevant information from the literature.

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 some 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.

A key assumption of most missing data methods is that the analyst has some idea about which of their data are missing and why. However, 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. In fact, most work on summarizing missingness in a dataset and examining its sources arises in literature on graphical summaries of data.

This is inherently an exploratory data analysis (EDA), wherein the analyst seeks to identify patterns of missing variables in their data and what those may be correlated with. As such, there is no single procedure or silver bullet for a given dataset. Instead, analysts...

Missing Data and Meta-Analysis

In the context of meta-analysis, "missing data" is a broad term that can be used to describe several different types of scenarios. For instance, data could be missing on individual participants within studies, including their outcomes in the study or other characterisitcs (e.g., their age, race, prior substance use). "Missing data" could also refer to scenarios where 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 pharacological 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

As part of this tutorial we use data analyzed by Tanner-Smith et al. (2016) on the effects of substance abuse inverventions for adolescents on subsequent substance use. These data were extracted from 61 randomized trials and quasi-experiments, and include 95 different effects of or contrasts between interventions. These effects include contrasts between a given treatment condition and a control condition within a study, or between two different treatement conditions in the same study.

There are 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. For instance, interventions might focus on cognitive behavioral therapy (CBT), family therapy, or pharmacological therapy. Interventions could be in- or out-patient. Individuals in studies might present using marijuana, which is most common among adolescents, or alcohol or opioids. They may come from wealthy families or poor families. Thus, Tanner-Smith et al. fit a series of meta-regression models 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 characterisitcs 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, [INSERT EXAMPLE]. As a result, not all effect estimate had information about the types of individuals in the study or the [INSERT OTHER COVARIATE]. 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 no covariates.

Tanner-Smith et al. ultimately opted for a sophisticated statistical procedure called the expectation-maximization (EM) algorithm to estimate their models. However, that was not their only option. They could have simply omitted effects with missing covariates from their analysis, imputed values that were missing, or augmented their meta-regression models so that they included a model for the missing covariates.

Principles of Missing Data

Analyses involving incomplete data will be affected both by how much data is missing, and why it is missing.

Define MAR/MNAR/MCAR.

Describe why we would want to do an exploratory analysis

Visualizations

First, any potential biases are related to the amount of missing data. When a greater amount of data are missing and excluded from an analysis, then any potential biases can be larger. Conversely, if only a small amount of data is missing, then any potential biases will be small. Second, any corrections one might make will depend on and be limited by which variables are missing and how frequently. Strategies that impute missing values for variable X tend to perform better if imputations can make use of important related variables. If those related variables are also likely to be missing when X is missing, this can limit how "good" imputations are.

This section examines different visualizations using `naniar` (Tierney, 2018), `visdat` (Tierney, 2017) and `ggplot2` (Wickham, 2009) R's packages. We demonstrate how visualizations typically used with datasets outside of meta-analysis can be adapted to the realities of meta-analytic data and contribute to understanding a dataset structure. Three types of visualization of missing data are discussed: whole-data plot; bivariate plots; and comparison with effect size and error variance plots.

1. Whole-data plots

Aggregation plots are useful tools for identifying the number of missing in each variable and case. Overall missingness is visualized with a "heatmap" style using `vis_dat()` and `vis_miss()` functions from the `visdat` package (see Figure 1 below).

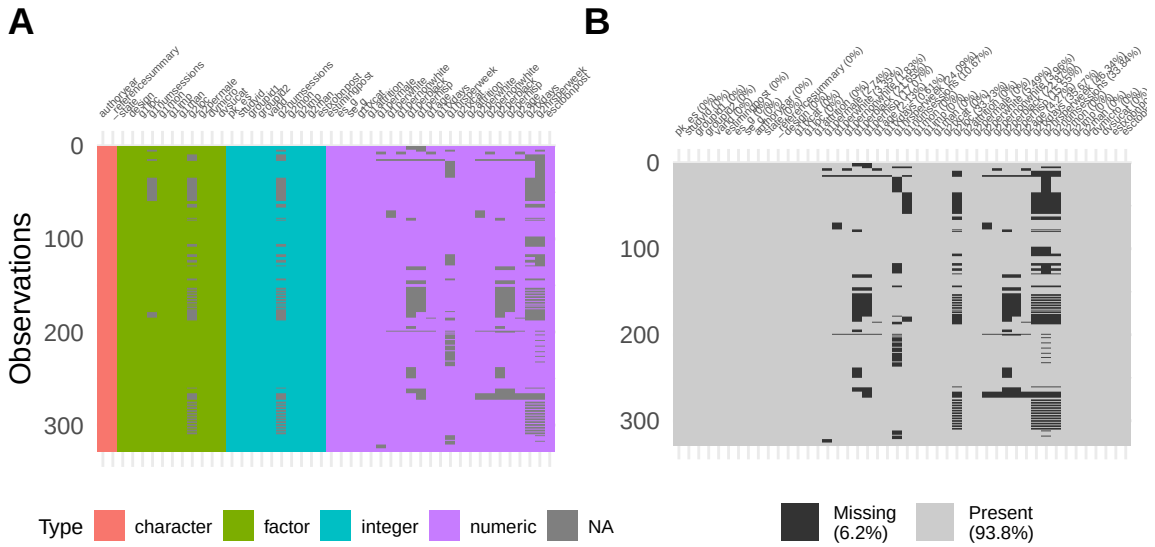

Figure 1: *Demonstrate how severe the level of data loss is within the Substance Abuse data. Plot (A) highlights variables with missing data. While none of the effect estimates are missing, it is clear that missing values appear in other 18 variables of interest, which, as shown in plot (B), represent 6.2% of the total dataset.*

Similarly, function `gg_miss_var()` from naniar package provides an approach to visualizing overall missingness by identifying variables with a greater number or percentage of missing cases, ordering by missingness (see Figure 2 below).

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).

2. Bivariate plots

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.

3. Comparison with effect size and error variances plots

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.

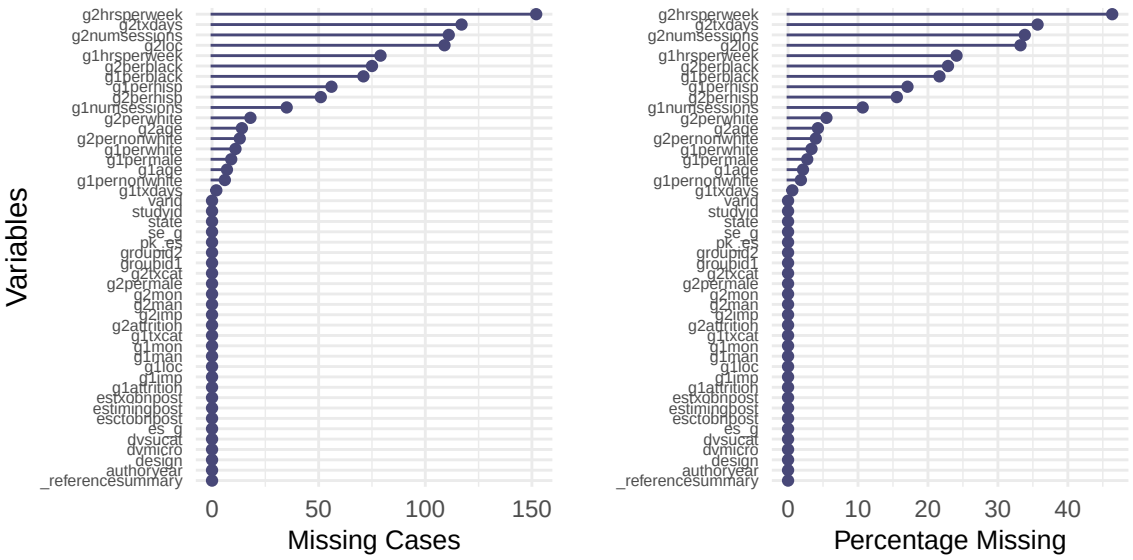

Figure 2: *Graphical summaries of missingness in variables, ordered by missingness, for the Substance Abuse data. There are 10 variables with at least 10% of missing cases. This visualization becomes relevant when deciding which variable to include in the analysis.*

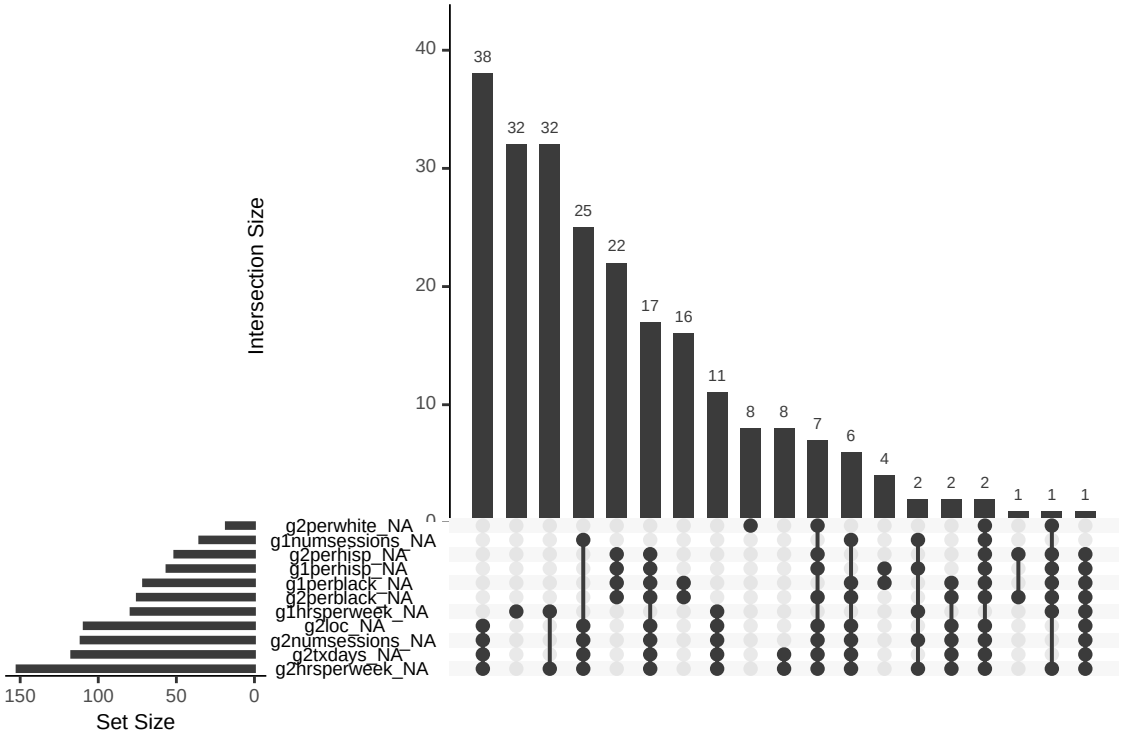

Figure 3: *Details those variables that are missing together. For instance, there are a large number of cases where Group 2 Level of Care, Number of Sessions, Treatment Contact (hours per week) and Duration of Treatment (days) are missing together. This simple exploration provides valuable information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

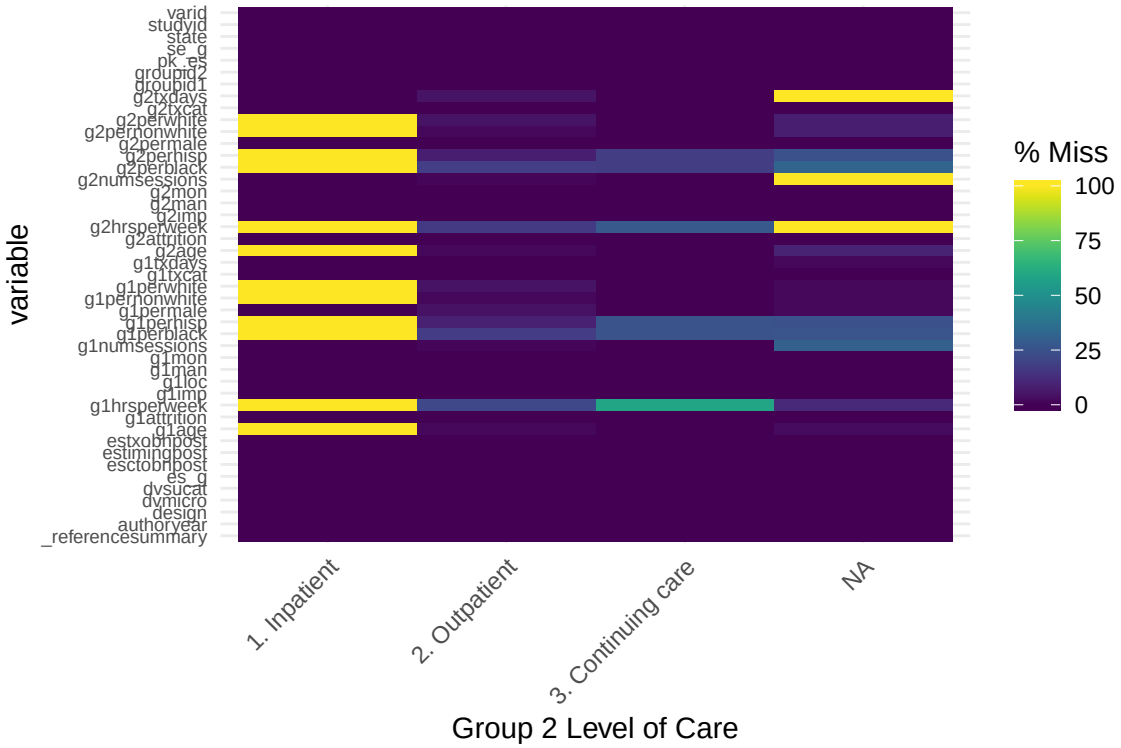

Figure 4: *Highlights the number of missings in each column, broken down by a factor variable, in this case the Level of Care for group 2. The inpatient category has 100% of missing values in at least 12 different variables, suggesting that this category could impose a problem when fitting a regression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

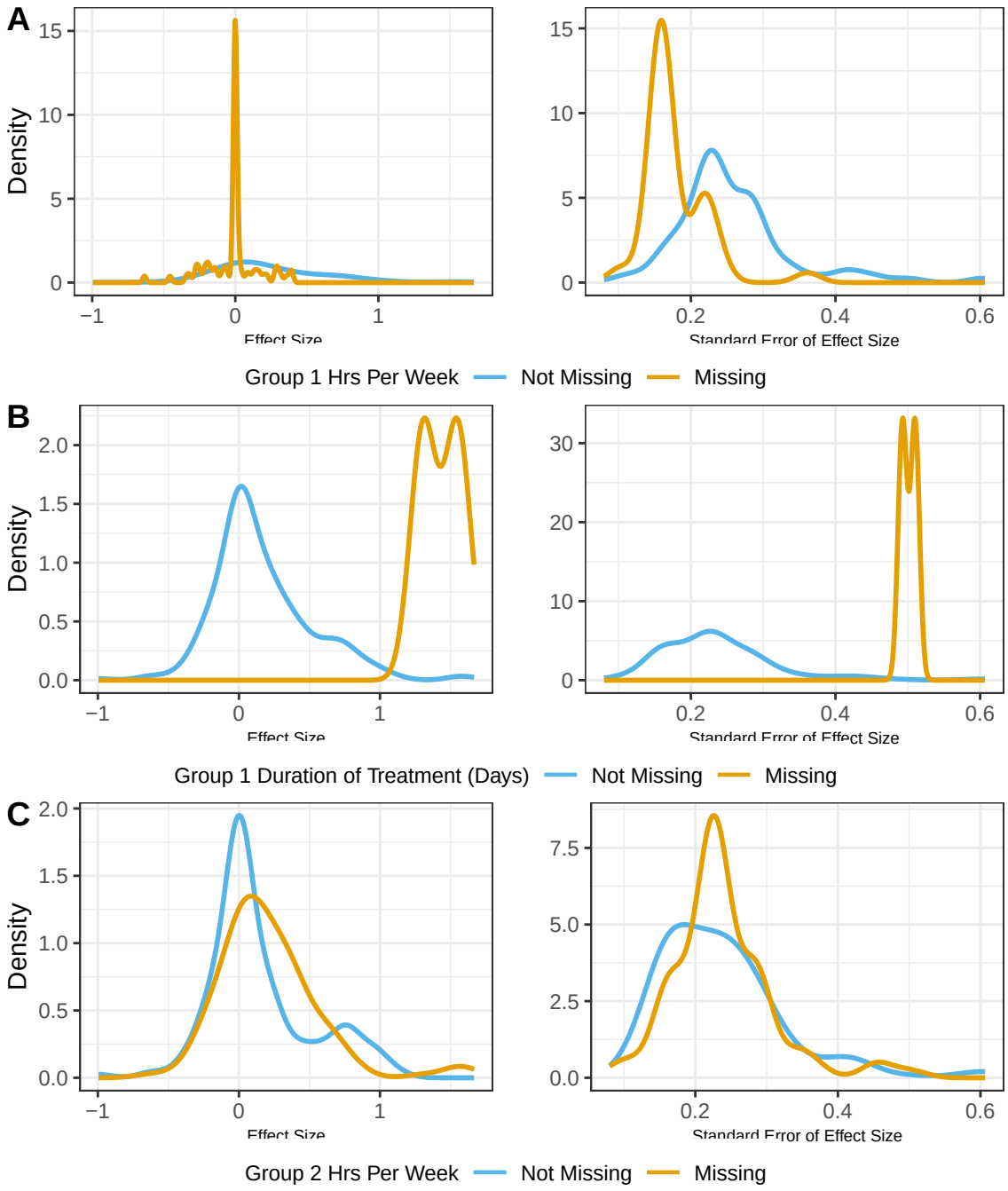

Figure 5: *Plot (A) shows that the covariance Duration of Treatment (days) for Group 1 is mostly missing for larger effect size values. Further, the effect size has larger standard error, when this covariate is missing. Plot (B) illustrates a case where the effect size tends to be closer to zero when 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.*