

Exploratory Analyses for Missing Data in Meta-Analyses

Introduction

Systematic reviews of substance abuse research hold great promise for examining what makes potential interventions effective (White *et al.*, 2010; Tanner-Smith, Wilson, & Lipsey, 2013; Tanner-Smith *et al.*, 2016; Newbury-Birch *et al.*, 2018; Ramsey *et al.*, 2019; Yuvaraj *et al.*, 2019). Methodological tools such as meta-regression can formally test relationships between an intervention’s impact and how or on whom it is studied (Hedges & Olkin, 1985; Cooper, Hedges, & Valentine, 2019; Tipton, Pustejovsky, & Ahmadi, 2019a). 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 face missing data problems (Pigott, 2019). Issues with missing data are not new. There is a vast literature on methods for handling missing data in primary studies, as well as work on related issues in meta-analysis (Rubin, 1976; Pigott, 2001a, 2019; Schafer & Graham, 2002; Little & Rubin, 2002; Graham, 2009; van Buuren, 2018). 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 or accurately compute uncertainty.

Diagnosing missing data issues remains an important aspect of any analysis of incomplete data (i.e., data with missing values). Understanding which and how much data is missing, and how problematic that can be, is crucial in determining how to proceed in a meta-analysis and how to contextualize the results. A key assumption of many analysis methods for incomplete data is that the analyst has some idea about *why* data is missing. Much of the literature on missing data has focused on the implications of that assumption (Little & Rubin, 2002; Pigott, 2019). Outside of some statistical tests (e.g., Little, 1988), considerably less attention is paid to how to form and examine theories about missingness (Tierney & Cook, 2018).

Various researchers have suggested analysts can better understand missingness in their data through exploratory analyses, including visual and numerical summaries (Buja, Cook, & Swayne, 1996; Cheng, Cook, & Hofmann, 2015) akin to classical exploratory data analyses (Tukey, 1962). These explorations, which occur before running a formal meta-analysis, can shed greater light on key issues relevant to missingness. Tools for doing so are only now emerging in statistics, but these tools have yet to gain broader traction in quantitative disciplines (Tierney, 2017; Tierney & Cook, 2018). Nor has this approach made its way into meta-analysis, where missing data is a common problem.

This tutorial examines exploratory analysis methods for diagnosing missing data problems in a meta-analysis, including data visualizations and numerical summaries. The following section clarifies the types of missing data for which these methods are appropriate. We then describe principles of missing data that can guide exploratory analyses. Finally, we demonstrate an exploration of missing data on a meta-analysis of substance abuse interventions for adolescents. Additional examples and executable code are available as part of the supplementary material for this article.

Missing Data in a Meta-Analysis

Because a meta-analysis involves an ensemble of *effects* (i.e., intervention impacts) reported by primary studies, *missing data* or *missingness* in meta-analysis could refer to at least three 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) (e.g., Higgins, White, & Wood, 2008). Missingness could also refer to information that could not be extracted from a completed study by a meta-analyst (Pigott, 2001a). This may occur if a study fails to report enough detail for analysts to

back out effect estimates, standard errors, or study- and effect-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 (Rosenthal, 1979). 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 (Hedges, 1984; Rothstein, Sutton, & Borenstein, 2005). The studies or effects that are not reported, and thus are not included in a meta-analysis, can be seen as 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 assess whether their systematic review is missing studies or effects due to publication bias (Light & Pillemer, 1984; Egger *et al.*, 1997). Our focus will be on the second scenario, where information cannot be extracted from some studies. This is a common problem in meta-analysis that can limit the accuracy of any statistical inferences (Pigott, 2019; Tipton, Pustejovsky, & Ahmadi, 2019b).

Notation

Assume a literature search reveals k effect estimates and we collect data on p variables regarding each effect (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. These tables are used by most standard meta-analysis software, including Comprehensive Meta-Analysis, the `metafor` library, or OpenMetaAnalyst (Viechtbauer, 2010; Borenstein *et al.*, 2012; Trikalinos, 2012).

The structure of such tables is shown in the matrix \mathbf{D} 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.

$$\mathbf{D} = \begin{bmatrix} T_1 & \sigma_1 & X_{11} & \dots & X_{1,p-2} \\ T_2 & \sigma_2 & X_{21} & \dots & \mathbf{NA} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ T_k & \sigma_k & \mathbf{NA} & \dots & X_{k,p-2} \end{bmatrix} \quad \mathbf{R} = \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & 1 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 0 & \dots & 1 \end{bmatrix}$$

Missing data in this context would refer to individual cells in the table that are missing values, and are denoted in \mathbf{D} as **NA**. To gain insight about missingness, we can augment the traditional dataset \mathbf{D} with a matrix \mathbf{R} , which is comprised of indicators R_{ij} for whether a cell has a missing value. These indicators take a value of $R_{ij} = 1$ if entry i, j in the dataset \mathbf{D} 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**).

Note that there are both *observed* and *unobserved* data. The observed data comprises all of the entries in \mathbf{D} 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 of \mathbf{D} for which $R_{ij} = 0$, including $X_{2,p-2}$ and $X_{k,1}$.

Data

A prime example of this type of missingness can be seen in data from Tanner-Smith *et al.* (2016), who examined the impacts 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 data will be used to illustrate key concepts of missingness and some useful tools to exploring missingness in this tutorial.

Tanner-Smith *et al.* identified a range of intervention types that have been studied in different venues and on different types of adolescent substance users. Some interventions 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, alcohol, or opioids, and they may come from wealthy families or poor families. While each effect involves the difference between two groups of study participants (referred to here as *Group 1* and *Group 2*), some reported effects contrasted a given intervention with some control or “business as usual” condition, while others contrasted two alternative interventions or implementations. Finally, Tanner-Smith et al. were able to extract effect estimates at different time points after intake. In all, their raw data totaled some $k = 328$ effect estimates and $p = 43$ variables for each effect.

In addition to estimated effects and their standard errors, Tanner-Smith et al. 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, Tanner-Smith et al. documented how many hours per week subjects in each group spent in receiving treatment. 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 impacts 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 dataset 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 the information that was missing.

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 incomplete data (Dempster, Laird, & Rubin, 1977; Graham, Cumsille, & Elek-Fisk, 2003). The EM algorithm has also been studied as a useful approach to estimation when missing covariates in a statistical model, which was primarily the issue facing Tanner-Smith et al. (Ibrahim, 1990; Ibrahim, Lipsitz, & Chen, 1999).

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 (Pigott, 2001a, 2001b, 2019; Tipton *et al.*, 2019b). An alternative to complete-case analysis that has gained broad use in various fields is to impute (i.e., fill in) missing values (Rubin, 1987; Pigott, 2001b, 2019). 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 (van Buuren, 2018).

Analyses involving incomplete data will be impacted by which variables are missing in a dataset, how frequently they are missing, as well as relationships between variables. Some understanding of these issues will guide decisions about which analytic approach may be appropriate. This section provides an overview of principles of missing data that apply to meta-analysis, and lists some potential statistical approaches to handling missing data. These principles can be used to guide exploratory analyses of missingness.

Quantifying the Amount of Missingness

One issue with missingness involves how much data is missing. This could refer to several different quantities. First, we could be interested in the total fraction of cells with missing values:

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

Second, we can compute the proportion of effects missing any variables:

$$\sum_{i=1}^k \frac{1 - \prod_{j=1}^p R_{ij}}{k} \quad (2)$$

Third, we may wish to know the percentage of effects are missing a given variable (i.e., how much of each column is missing):

$$\sum_{i=1}^k \frac{1 - R_{ij}}{k} \quad (3)$$

While such quantities 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, parameters are often estimated by weighting effect size estimates in such a way that the most precisely estimated effect estimates (i.e., that have the smallest standard errors) receive the most weight (Borenstein, 2009; Cooper *et al.*, 2019). 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 $1/\sigma_i^2$ is large (Hedges & Olkin, 1985). Missing a variable for an effect estimate with a large standard error (and thus low precision) can potentially be less detrimental than missing the same variable for an effect estimate with a small standard error.

Thus, when standard errors of effect sizes are fully observed, an alternative way to quantify the extent of missingness is with a precision-weighted percentage. The weighted percentage

$$\sum_{i=1}^k \frac{(1 - \prod_{j=1}^p R_{ij}) / \sigma_i^2}{\sum_{i=1}^k 1 / \sigma_i^2} \quad (4)$$

gives the fraction of information in a meta-analysis associated with effects that are missing any data. The quantity

$$\sum_{i=1}^k \frac{(1 - R_{ij}) / \sigma_i^2}{\sum_{i=1}^k 1 / \sigma_i^2} \quad (5)$$

is the percentage of information associated with effects missing a specific variable. If the weighted percentage (4) is greater than the raw percentage in (2) or if (5) is greater than (3), that would indicate that a missingness problem may be more acute because data is missing from larger studies.

Missingness Mechanisms

A key assumption that underpins analyses of incomplete data involves why those values are missing, typically referred to as the missingness *mechanism*. Rubin (1976) classified three different possible types of mechanisms. These mechanisms relate the probability that a value is missing to the observed and unobserved data; that is, they relate \mathbf{R} to the observed and unobserved data.

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 $P[R|\text{observed data, unobserved data}] = P[R]$, and means the probability that a given value is missing is unrelated to anything observed or unobserved.

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 $P[R|\text{observed data, unobserved data}] =$

$P[R|\text{observed data}]$. Note that this differs from MCAR in that missingness might be related to observed values. For instance, suppose studies with smaller sample sizes (and hence larger standard errors) are less likely to report the racial composition of their samples. Then, assuming the standard errors are observed, this could be consistent with MAR. It would violate an assumption of 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 unobserved data in some way. This differs from MAR in that missingness depends on unobserved, rather than just observed 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.

It is worth noting that various researchers have proposed hypothesis tests of some of these assumptions. For instance, Little (1988) describes a test for MCAR, while other authors have posited tests whether data are MAR versus MNAR (Baker, Rosenberger, & Dersimonian, 1992; Diggle & Kenward, 1994; Molenberghs, Kenward, & Lesaffre, 1997; Troxel, Harrington, & Lipsitz, 2002). However these tests may not be appropriate for data typically used in a meta-regression. While Little’s test assumes that all variables are continuous and normally distributed, meta-regressions frequently involve categorical covariates and the standard errors of effect estimates are likely not normally distributed. Much of the literature on MAR tests compares specific models for dropout in longitudinal studies, which is almost never an issue for the meta-analyst (Molenberghs *et al.*, 2008; Rhoads, 2012).

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 are missing the hours of therapy per week for one of the groups, while other rows are missing the hours of therapy per week *and* the percentage of study participants who were minorities. In other words, different rows in the data exhibit different *missingness patterns*. Missingness patterns can be thought of examining relationships within the matrix \mathbf{R} .

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 (van Buuren, 2018). For instance, suppose an analyst decides to deal with missingness by imputing missing values. Typically, imputations can be improved by using predictive models that leverage other variables in the dataset. However, if a variable to be imputed is frequently missing alongside many other variables in the data, this can affect the accuracy of the imputation model.

Missing Data Analysis Methods

There has been a large amount of research into methods for analyzing incomplete data (see Graham, 2012; Little & Rubin, 2002; van Buuren, 2018). Pigott (2019) provides a comprehensive overview of methods for handling missing data in a meta-analysis. A common approach in meta-analysis is to conduct a complete-case analysis that excludes effects that are missing any variables (Tipton *et al.*, 2019b). Meta-analysts also make regular use of *available-case* analyses, which attempts to use all observed data, even rows that may be missing variables. Typically, this will involve running regression models that involve one or two covariates at a time, and is sometimes referred to as “shifting units of analysis” (Cooper, 1998). As an example, one might regress the intervention effects on the intervention type in one model, and then on intervention duration in a second model using Tanner-Smith’s *et al.* data. A related approach is the EM algorithm, which makes use of all the observed data when estimating parameters in a meta-regression using an iterative procedure to obtain maximum likelihood estimates. Finally, imputing missing values has become increasingly common for statistical analyses of incomplete data in many fields. The standard approach is to use a method call multiple imputation, where missing fields are filled in with several values that the missing field might have contained had it not been missing. This creates several “complete” datasets, each of which are analyzed and the results of those analyses are then pooled (Rubin, 1987; Little & Rubin, 2002).

Methods for analyzing incomplete data rely on assumptions about the missingness mechanism. Complete-case analyses may be appropriate assuming MCAR, but not necessarily if data are MAR (Pig-

ott, 2001b; Little & Rubin, 2002). More sophisticated approaches, such as multiple imputation or the EM algorithm, are typically implemented in software in a way that assumes data are MAR but not MNAR (Schafer & Graham, 2002; Graham, 2009; van Buuren, 2018). When data are MNAR, the mechanism that produces missingness will typically need to be explicitly modeled in any analysis.

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) combine numerical and visual summaries in order to better understand the extent and sources of missingness in a dataset (Buja *et al.*, 1996; Cheng *et al.*, 2015; Tierney & Cook, 2018). EMA differ from traditional exploratory data analyses (EDA) because they focus on missingness indicators in **R**, as well as the relationship between those indicators and the observed data in **D**. Many software tools, including most graphics software used to conduct a standard EDA actually delete observations with missing values, which would eliminate information about missingness (Tierney, 2017).

While the following sections present an example of an EMA, it is worth noting two aspects about EMA to better contextualize this process. First, it will often be difficult to draw very strong inferences about missingness mechanisms based on exploratory analyses. Even proposed tests for missingness mechanisms can have misleading results (Molenberghs *et al.*, 2008; Rhoads, 2012; Seaman *et al.*, 2013). 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 for others. Any EMA must rely on knowledge of how data were collected and extracted, and can help leverage that knowledge to examine and propose theories about missingness.

In the following sections, we present and discuss an example EMA of Tanner-Smith’s *et al.* data on substance abuse interventions for adolescents. To simplify presentation, we focus on $p = 20$ variables relevant to the analyses conducted by Tanner-Smith *et al.* This example serves to highlight some potential techniques and tools, but it is not exhaustive, and so as part of the supplementary material to this tutorial, we have included a vignette that presents and describes alternative visualizations 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 (Tierney, 2017; Tierney & Cook, 2018). Executable code is included with the supplementary materials.

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. Figure 1 shows an aggregation plot for the Tanner-Smith *et al.* data. The plot is laid out exactly like the data: The columns correspond to variables in the data, and rows correspond to effect sizes. Dark areas correspond to cells that are missing values.

Aggregation plots provide a high-level picture of missingness in a dataset. They can indicate which columns are complete, such as the columns corresponding to the effect size estimates, standard errors, or study designs. They also show which columns or groups of columns contain many missing values. In particular, Figure 1 appears to show three general kinds of missingness patterns. First, studies are missing information on the treatment intensity (hours per week and duration) for Group 1 or Group 2. Note that occasionally this information is missing for both groups, as with the rows near the top of the plot. Second, studies are missing information on the demographic makeup (percent of the group that is white, black, Hispanic, or male) for Group 1 and Group 2 simultaneously. Finally, for a number of rows in the middle of the data, it appears that studies are missing information both on Group 2’s treatment intensity and demographics.

Figure 1 also displays some numerical summaries regarding the extent of missingness in the

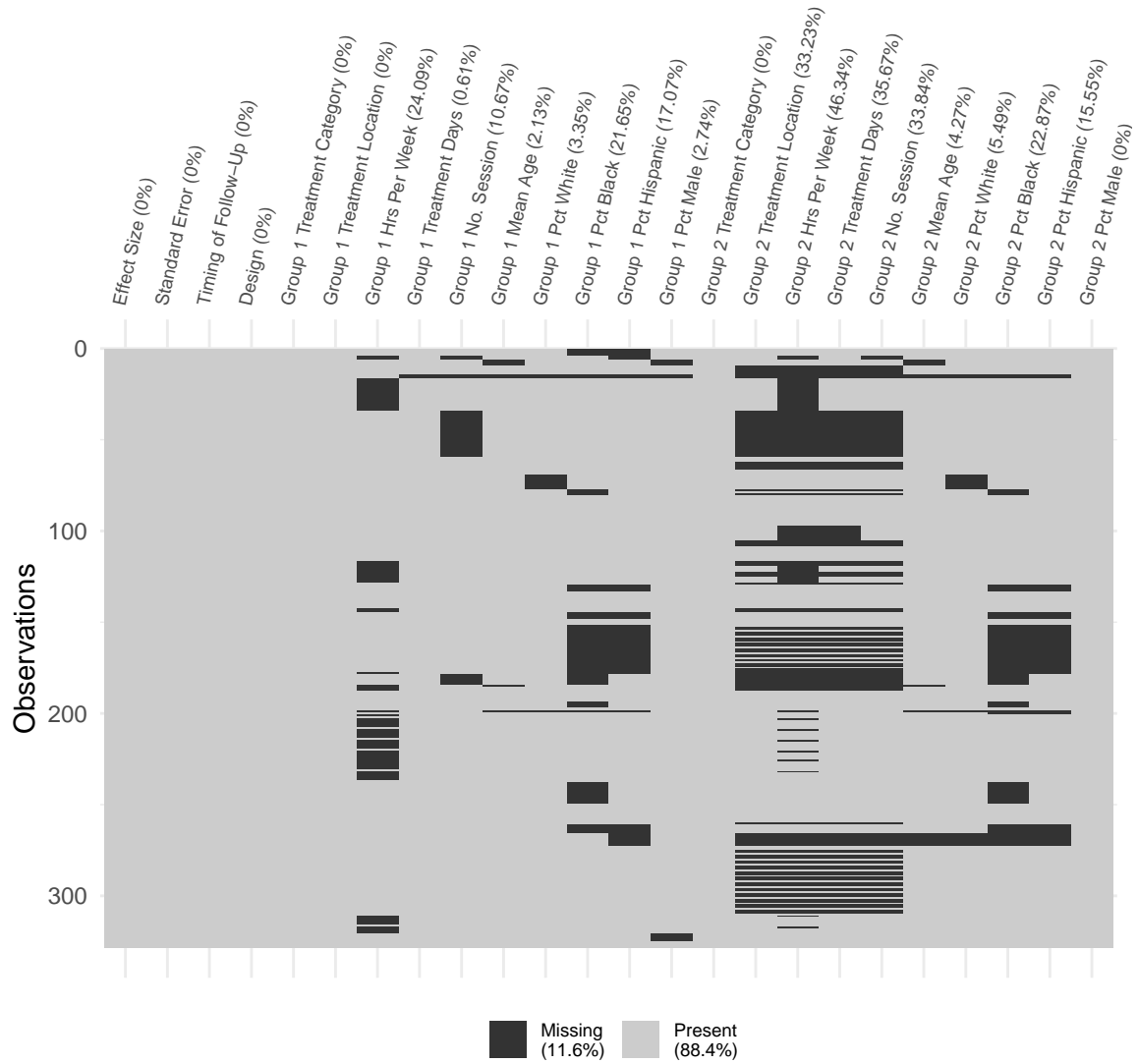


Figure 1: This plot indicates the severity of missingness in the adolescent substance abuse intervention data. Each row in the plot corresponds to a row in the data, and each column corresponds to a variable collected in the data. Missing cells in the data are indicated by a dark dash in plot. The legend shows the percent of cells in the data that contain missing values. The column labels show the percent of rows missing each variable in the data.

data. In the legend, we see that over 11% of all cells are missing values in the table. Figure 1 also reports the percent of each column that is missing. Overall, 73.8% of rows are missing at least one value, and effects that are missing any covariate make up roughly 74.4% of the total precision in the data. Thus, a complete-case analysis of all variables would require dropping nearly 27% of the rows in the data.

Univariate Explorations

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

From variable missing plots, it is often easy to identify variables that might be driving any missing data problems, and they can quantify the extent to which a given column has missing values on the scale of raw percentages.

Figure 2 shows a variable missing plot for the Tanner-Smith et al. data. Each variable is listed on the y -axis, and the line extending along the x -axis indicates the fraction of rows for which a variable is missing. Figure 2 suggests that the hours per week that Group 2 spent in their assigned treatment is missing for almost half of the effects in the data, while the hours per week that Group 1 spent in treatment was missing for far fewer rows—roughly half as many. Further, the percentage of each group that is black or Hispanic is missing more frequently than is the percentage of each group that is white. That is, studies reported the breakdown of white/non-white adolescents in their studies more frequently than they reported the percentage of adolescents identifying as a specific non-white race.

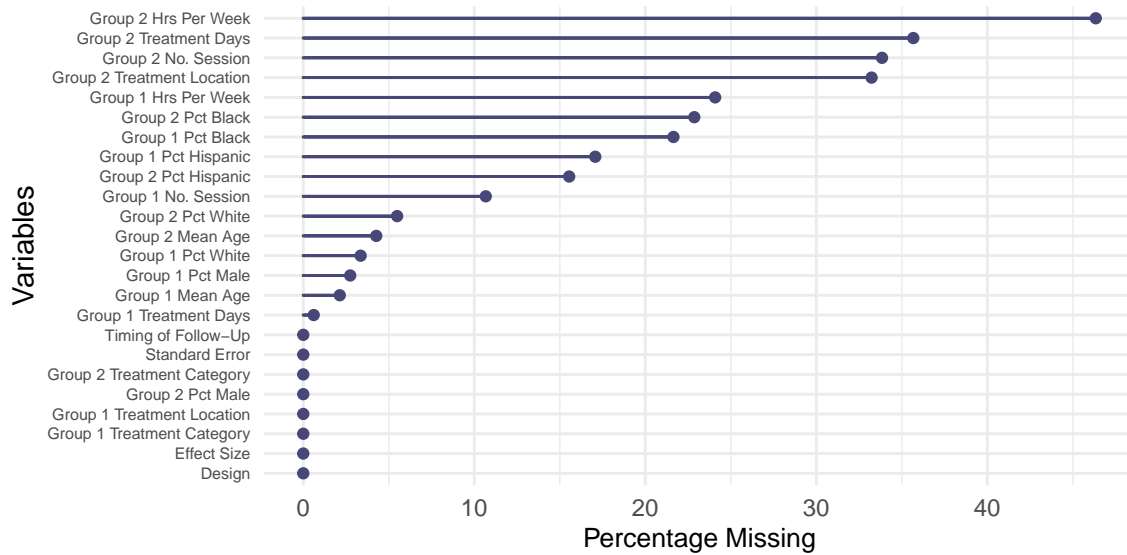


Figure 2: *This plot summarizes missingness in variables, ordered by missingness, in the adolescent substance abuse intervention data. Indicating that there are 10 variables with at least 10% of missing cases. This kind of visualization becomes relevant when deciding which variable to include in the analysis.*

Previously, we argued that precision-weighted percentages may be more informative in describing the extent of missingness in a meta-analysis. Raw percentages and precision-weighted percentages are presented in Table 1. We would discourage interpreting the size of the differences between the raw and weighted percentages, however comparing those columns can identify variables missing from larger studies, which typically receive more weight in a meta-analysis. For example, the raw percentage column and Figure 2 would suggest missingness in the hours per week that Group 1 spent in treatment (24% missing) may be much less acute than missingness in the hours per week Group 2 spent in treatment (over 46% missing). But the weighted percentage indicates that the effects for which Group 1's hours per week variable is missing make up nearly 37% of the total precision of effect estimates. Hence, excluding those effects in a complete-case or available-case analysis would reduce how accurately the relationship between Group

1's treatment intensity and the intervention's impact can be assessed. This reduction in accuracy would likely be greater what is indicated by the raw percentages.

Table 1: *This table displays the total number, percentage, and precision-weighted percentage of effect sizes that are missing a given variable.*

Variable	# Missing	% Missing	Wt. % Missing
Group 2 Hrs Per Week	152	46.3	45.2
Group 1 Hrs Per Week	79	24.1	36.8
Group 2 Treatment Days	117	35.7	32.6
Group 2 No. Session	111	33.8	30.2
Group 2 Treatment Location	109	33.2	30.0
Group 2 Pct Black	75	22.9	17.5
Group 1 Pct Black	71	21.6	16.1
Group 1 Pct Hispanic	56	17.1	12.7
Group 2 Pct Hispanic	51	15.5	12.4
Group 1 No. Session	35	10.7	6.9
Group 2 Pct White	18	5.5	4.6
Group 1 Pct White	11	3.4	3.1
Group 2 Mean Age	14	4.3	3.0
Group 1 Mean Age	7	2.1	1.5
Group 1 Pct Male	9	2.7	1.3
Group 1 Treatment Days	2	0.6	0.1

Exploring Patterns of Missingness

When diagnosing potential missing data problems, it also matters which variables are missing together. Certain patterns of missingness can be indicative of the missingness mechanism, and patterns can also point to potential issues for analytic methods with incomplete data, such as multiple imputation, available case analysis, or the EM algorithm.

While the aggregation plot gives some idea about patterns, an *upset plot* can provide greater insight into how frequently those patterns occur (Conway, Lex, & Gehlenborg, 2017). Figure 3 exemplifies an upset plot. The bottom of Figure 3 presents different variables and indicates how many rows each of those variables is missing from. The dots along the bottom panel indicate different patterns of missingness, which means that a given set of variables are missing from the same row(s). The bars in the top panel of Figure 3 show the frequency with which these patterns occur.

The bottom four variables listed in Figure 3 involve the treatment duration and intensity for Group 2. Four of the top five variables listed involve the racial composition of the studies. Judging by the bottom panel in the figure, there are few different patterns worth noting. The first pattern, in which much of the information about the treatment condition in Group 2 is missing. However, that pattern also occurs in conjunction with other patterns. For instance, the sixth pattern involves rows that are missing information about Group 2's treatment, as well as the racial composition of the study.

In the top panel, we see how frequently each pattern occurs. The first pattern, which largely contains information about Group 2's treatment duration and intensity, occurs for 38 rows in the data. However, because that pattern is part of other patterns, it occurs frequently in the data. Judging from Figure 3, many rows are missing information about Group 2's treatment, and several of those rows are also missing other variables, including variables that describe the demographics of the study participants.

Relating Missingness to Observed Values

Explorations of missingness can also examine whether missingness in one variable is related to observed values in another variable. This can highlight potential biases in the observed data. It can also clarify

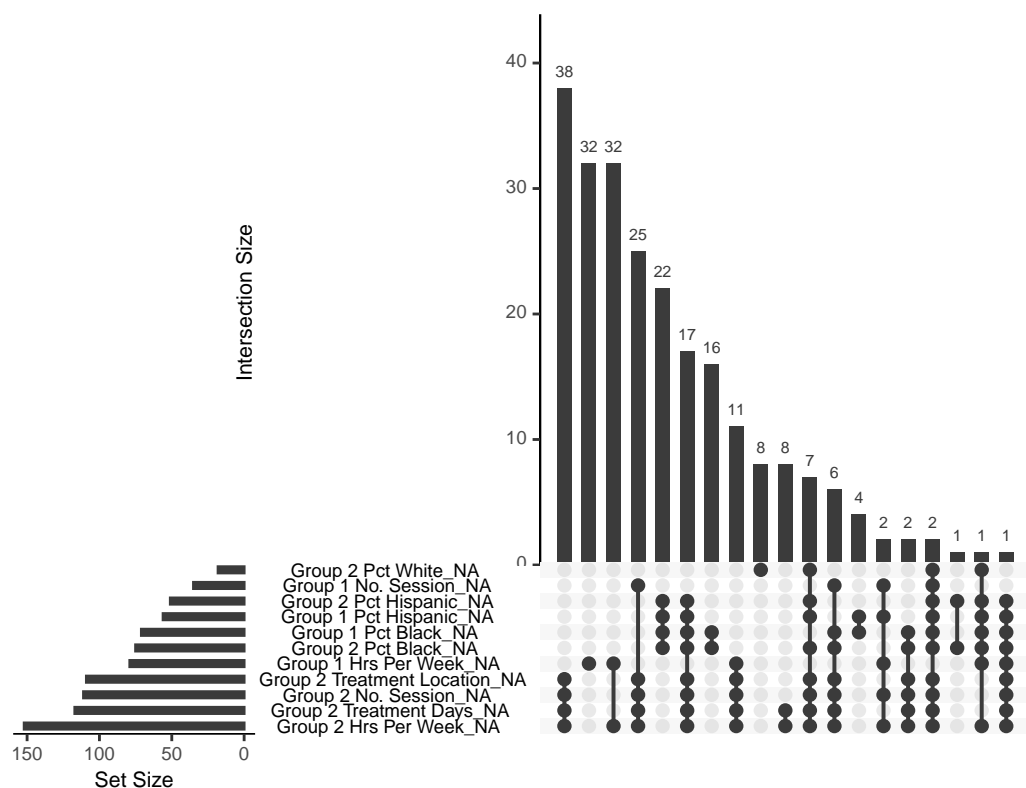


Figure 3: This plot details those variables that are missing together. For instance, there are a large number of cases where group 2 level of care, number of sessions, treatment contact (hours per week) and duration of treatment (days) are missing together. This simple exploration provides valuable information for imputation.

whether data appear to meet the MCAR assumption. The idea behind this is that MCAR assumes that the probability that a value is missing is independent of both observed and unobserved values. However, if missingness in one column of the data is correlated with observed values in another column, that would be an indication that data are not MCAR. A similar logic underpins the test for MCAR proposed by (Little, 1988).

Since covariates in a meta-regression are often categorical, visualizations like the *heatmap* in Figure 4 can be useful. Figure 4 plots the missingness rate for each variable as it relates to Group 1’s treatment location. The figure is grouped into tiles. The columns of the figure correspond to the different types of care provided to Group 1, shown along the bottom of the graph: inpatient, outpatient, or continuing care. The rows of the figure correspond to other variables in the dataset. Each tile is shaded according to how frequently those variables are missing for each level of care provided.

Figure 4 shows that missingness in other variables is more common for effects where Group 1 received inpatient or continuing care treatment. In particular, studies where Group 1 received outpatient treatment were more likely to report racial demographics, information about Group 1’s treatment intensity (in hours per week) and duration (in days and number of sessions). Likewise, studies where Group 1 received outpatient care were more likely to report characteristics of Group 2’s treatment type, duration, and intensity. We should note, however, that the *Inpatient* column should be interpreted with some caution, as only five (5) effects in the raw data involve Group 1 receiving inpatient treatment. Still, the differences in missingness for effects where Group 1 received outpatient versus continuing care interventions suggest that missingness in several variables is related to the venue of treatment.

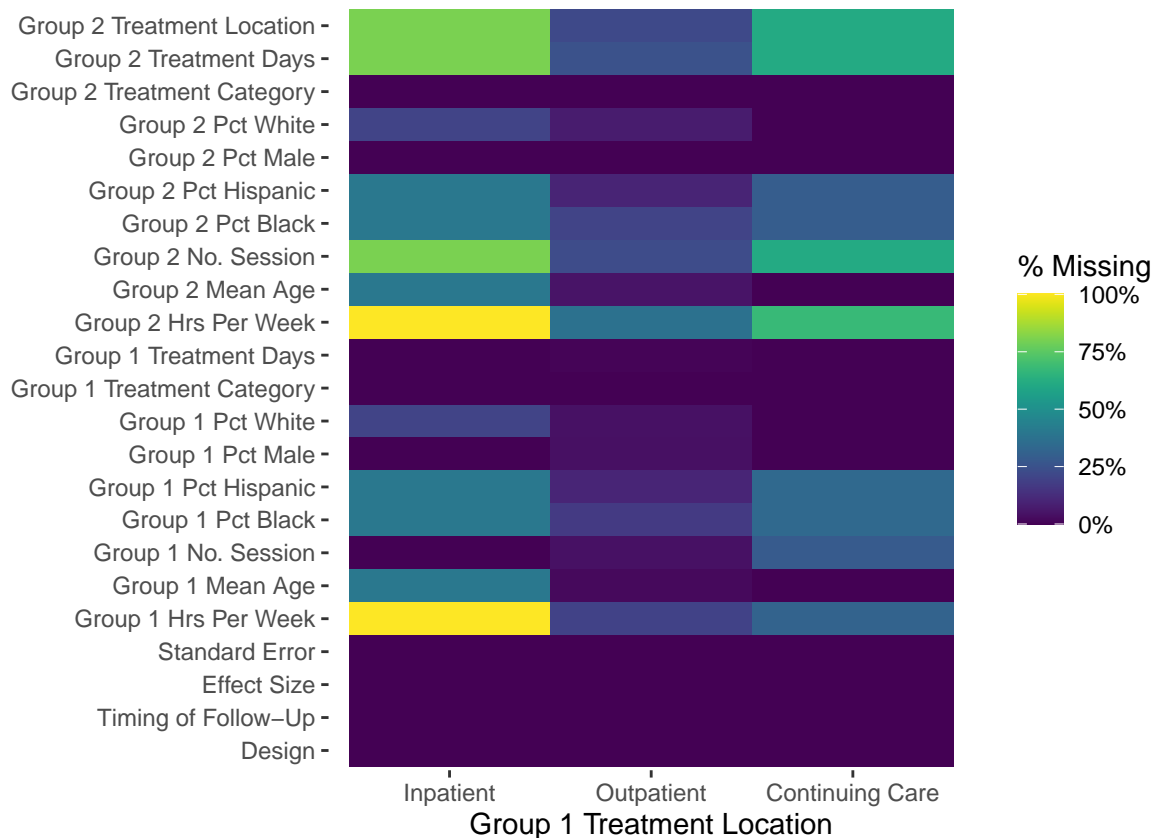


Figure 4: This plot shows the rate of missingness for each variable as a function of Group 1’s treatment location. Each column is broken down by where Group 1 received treatment (inpatient, outpatient, and continuing care). Each row represents another variable in the data. Tiles are shaded according to the fraction of rows in the data for which each variable is missing for a given level of Group 1 treatment location.

Similar patterns emerge in Figure 5, which shows that missingness in several variables is also cor-

related with the venue of Group 2's treatment. Note that Figure 5 includes an *NA* column, which indicates that Group 2's treatment location is itself missing. As with Figure 4, the *Inpatient* column is difficult to really interpret as it represents only one (1) effect in the raw data. However, comparing the *Outpatient* and *Continuing Care* columns reveals that studies where Group 2 received outpatient care were more likely to report other variables, particularly those related to treatment intensity (in hours per week) and racial composition.

The *NA* column in Figure 5 also confirms a notable missingness pattern identified in the upset plot. The *NA* column corresponds to rows in the data where Group 2's treatment location is missing. Among those rows, we see that other information about Group 2's treatment, including duration in days and number of sessions, are almost always missing as well. This would suggest that some studies simply fail to provide any information that could be extracted on the intensity and duration of Group 2's treatment.

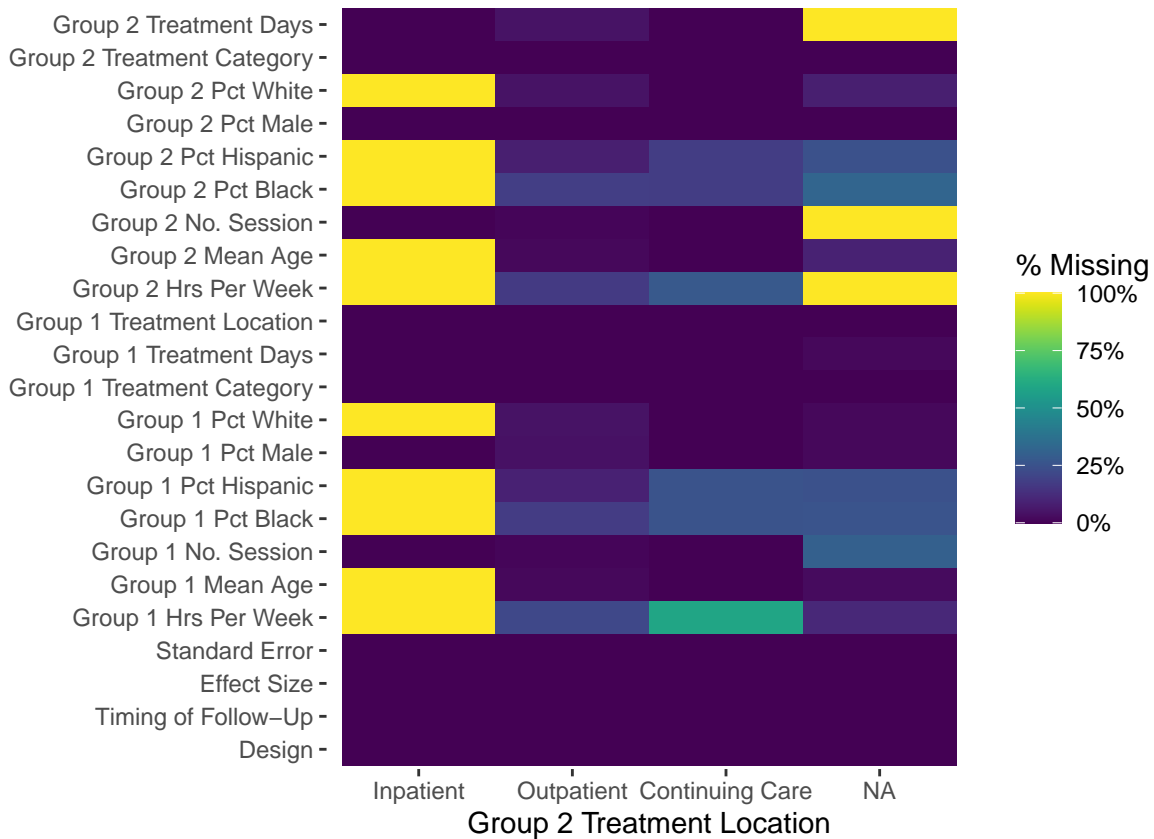


Figure 5: This plot shows the rate of missingness for each variable as a function of Group 2's treatment location. Each column is broken down by where Group 2 received treatment (inpatient, outpatient, and continuing care). Each row represents another variable in the data. Tiles are shaded according to the fraction of rows in the data for which each variable is missing for a given level of Group 2 treatment location.

Two crucial variables that are in nearly all meta-analyses are the effect size estimates and their standard errors σ_i (or variances σ_i^2). If missingness in a covariate is correlated with effect sizes or standard errors, this is likely to impact the analytic results. Therefore, it will typically be a good idea to explore such relationships.

Because both effect estimates and standard errors are continuous, *comparative density plots* can be useful tools. Comparative density plots present the distribution of effect estimates and standard errors among effects for which a covariate is missing versus when the covariate is observed. Figure 6 displays several density plot pairs for different variables. For each pair of plots we see the distribution of effect estimates (left) and standard errors (right), each colored according to whether a given covariate is missing.

Figure 6A, shows the relationship between the missingness in Group 1's treatment intensity (in hours per week) and the distribution of effect estimates and standard errors. Effects are on the scale of Cohen's d . From Figure 6A, we can see that effect estimates for which Group 1's treatment intensity is missing tend to be slightly smaller than the effect estimates for which Group 1's treatment intensity is observed. Effect estimates tend to have smaller standard errors when Group 1's treatment intensity is missing than when it is reported. This is consistent with the weighted percentages reported in Table 1, which found that missingness in Group 1's treatment intensity occurred with studies with greater precision.

Contrast that with Figure 6B, which focuses on whether Group 1's treatment duration (in days) is missing or not. Figure 6B shows that effect estimates for which Group 1's treatment duration is missing tend to be larger and have larger standard errors than effects for which duration of treatment is reported. Plots A and B suggest that missingness of information about treatment dosage will be related to the size of effects found and how precisely those effects were estimated.

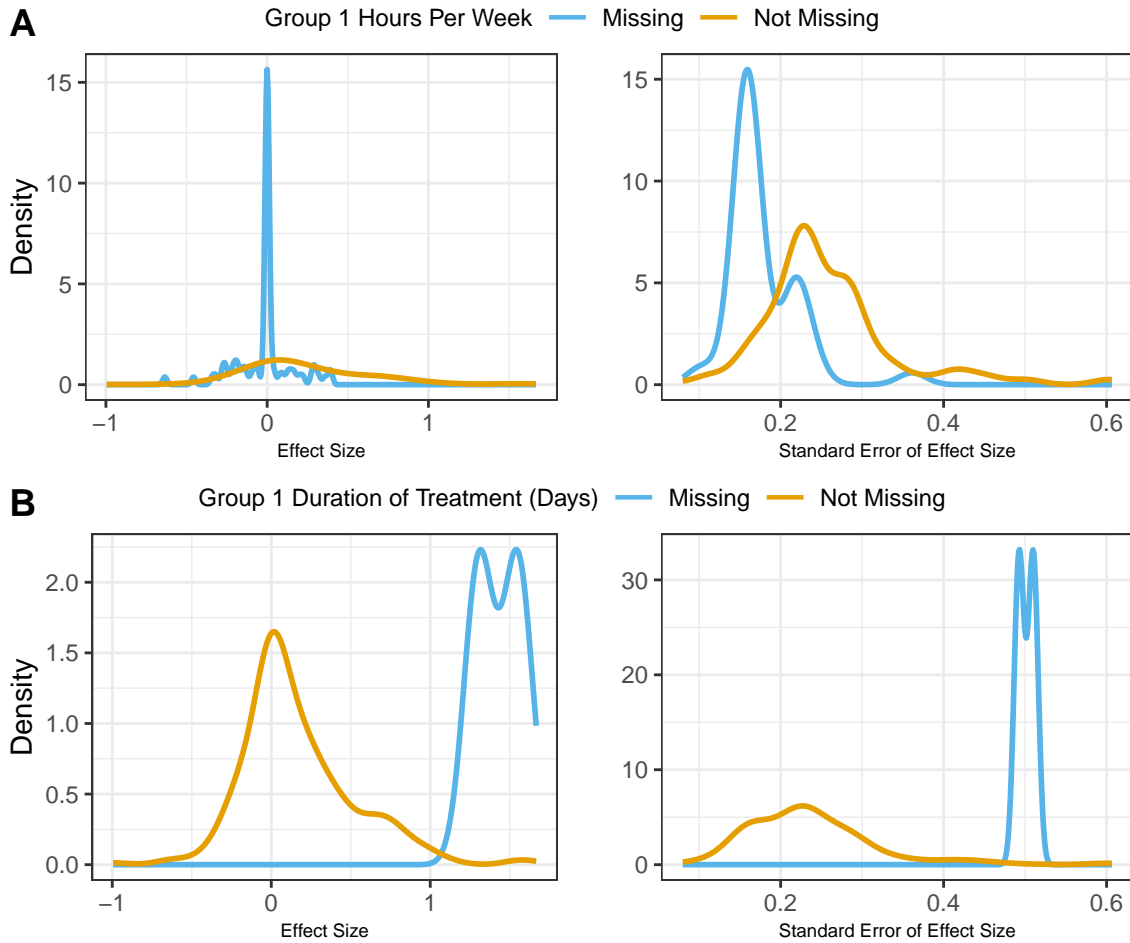


Figure 6: This figure compares the distribution of effect size estimates and standard errors for when various covariates regarding Group 1 are observed versus missing. Plot (A) compares the distribution effect size estimates and standard errors for when Group 1 treatment intensity (hours per week) is missing versus observed. Plot (B) compares the distributions for when Group 1 treatment duration (in days) is observed versus missing.

Figure 7 shows an analogous set of plots, only it compares the effect size estimate and standard error distributions for when Group 2's treatment intensity (in hours per week) and duration (in days) are missing or observed. Figure 7A shows that when Group 2's treatment intensity is observed, effect sizes and standard errors are slightly smaller, though not drastically different than when Group 2's treatment intensity is missing. As well, Figure 7B suggests that effect estimates and standard errors are slightly

smaller, though not particularly different, when Group 2's treatment duration is observed compared to when it is missing.

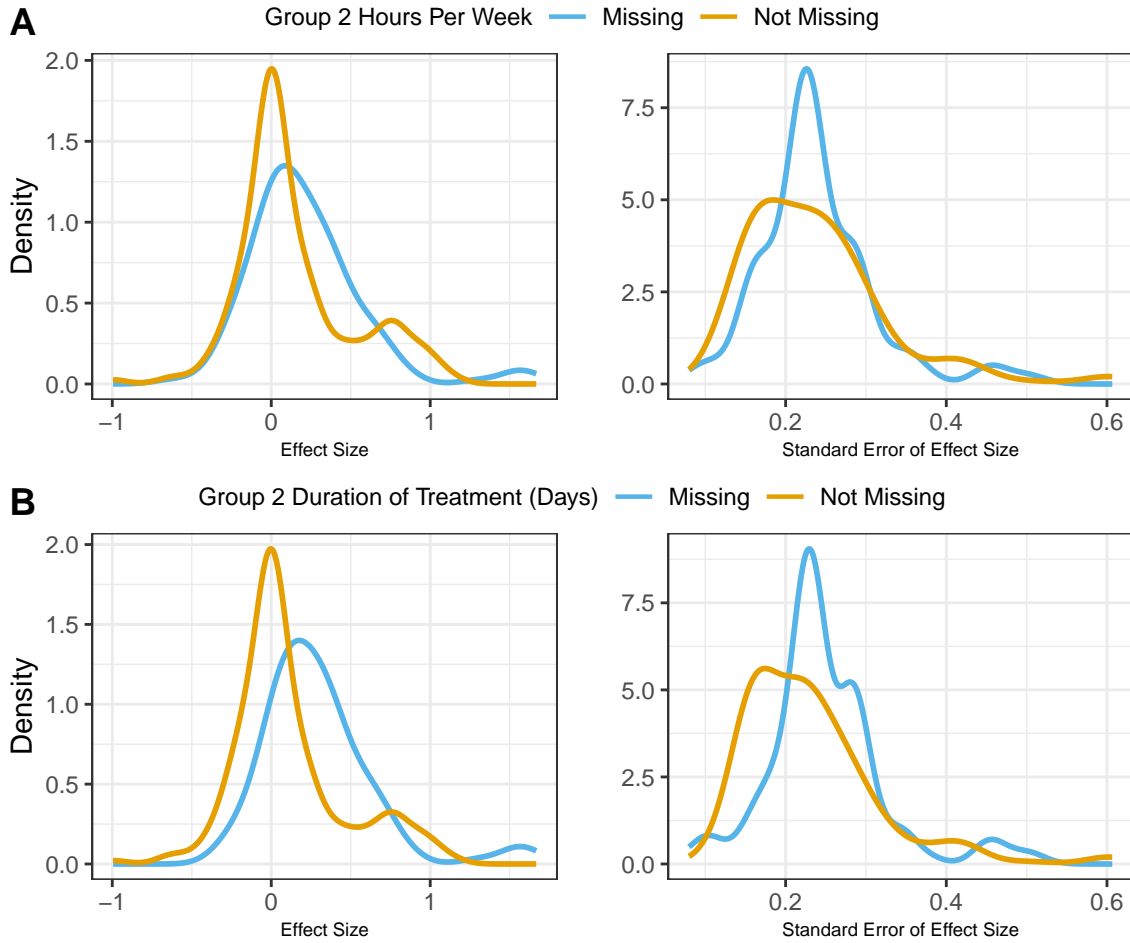


Figure 7: This figure compares the distribution of effect size estimates and standard errors for when various covariates regarding Group 2 are observed versus missing. Plot (A) compares the distribution effect size estimates and standard errors for when Group 2 treatment intensity (hours per week) is missing versus observed. Plot (B) compares the distributions for when Group 2 treatment duration (in days) is observed versus missing.

Comparing Figures 6 and 7 reveals that missingness in variables pertaining to Group 1's treatment has a stronger relationship the effect size estimates and standard errors than does missingness in variables pertaining to Group 2's treatment. Omitting effects for which Group 1's treatment duration or intensity are missing would seemingly have a stronger impact on an analysis.

Discussion

Missing data is and will continue to be an issue with most meta-analyses, and that can affect what we can learn about substance abuse interventions from research syntheses. While there are various potential approaches to handling missing data in meta-analysis, most of those approaches assume that the missingness mechanism is known to the analyst. That is not always the case, and so this article argued that an exploratory analysis of missingness might help analysts better understand and diagnose their missing data problems. It also outlined and demonstrated some tools that can support exploratory analyses into the scale and correlates of missingness in a meta-analytic dataset. These tools proved to be useful as a first step to understanding why data is missing.

These tools were applied to data on a large meta-analysis conducted by Tanner-Smith *et al.* (2016) on substance abuse interventions for adolescents. We found 73.8% of the effect sizes were missing at least one of their corresponding covariates. This was driven by some variables that were missing frequently (e.g., Group 2 hours of treatment per week). Our analysis also revealed that missingness in some variables may be more severe than was obvious from first glance (e.g., Group 1 hours of treatment per week). Variables quantifying the intensity and duration of treatment in a study were frequently missing together. Finally, we identified some variables whose missingness appears to be related to the size and standard errors of effect size estimates (e.g., Group 1 hours of treatment per week and treatment duration in days), which suggests that missingness was not MCAR.

How to proceed from an EMA will depend on what is known about the data collection and missingness. Based on our analysis, we would be cautious of using analysis methods that assume data are MCAR, such as complete-case analysis or shifting units of analysis. Tanner-Smith *et al.* used the EM algorithm to estimate their meta-regression models, which assumes data are MAR. This is consistent with our findings, and it is a common assumption made in analyses of incomplete data.

The analyses presented, while suggestive, do have several limitations. First, data curators and analysts who extract data for a meta-analysis can and perhaps should play a larger role in EMA. While our post-hoc analysis had limited input from these individuals, they will almost certainly have some insight about what made particular variables difficult to extract from the literature, and why that may have occurred.

Second, it will be impossible to distinguish between data that are MAR or MNAR using the methods demonstrated in this article. This is because MNAR assumes that missingness is related to data that is not actually observed. Thus, in order to study or confirm whether data are MNAR would require some knowledge of the unobserved data. Analogous limitations have been noted in tests for whether data are MAR or MNAR (Molenberghs *et al.*, 2008; Rhoads, 2012).

The methodology discussed in this tutorial could be used to create different visualizations than were presented in this paper. Our complementary material develops on these results with a vignette that contains further visualizations and executable code implemented in the R computing language. Moreover, even though the data on substance abuse interventions for adolescents has a particular structure with information at the effect size level for each study, the tools exposed in this tutorial can be easily applied to other dataset structures.

References

- Baker SG, Rosenberger WF, Dersimonian R. (1992) Closed-form estimates for missing counts in two-way contingency tables. *Statist Med* **11**: 643–657.
- Borenstein M. (2009) *Introduction to meta-analysis*. Chichester, U.K.: John Wiley & Sons.
- Borenstein M, Hedges LV, Higgins JP, Rothstein H. (2012) Comprehensive Meta-Analysis Version 3.0. [https://www.meta-analysis.com/downloads/Meta-Analysis Manual V3.pdf](https://www.meta-analysis.com/downloads/Meta-Analysis%20Manual%20V3.pdf).
- Buja A, Cook D, Swayne DF. (1996) Interactive high-dimensional data visualization. *Journal of Computational and Graphical Statistics* **5**: 78.
- van Buuren S. (2018) *Flexible Imputation of Missing Data, Second Edition*, 2nd edn. Second edition. | Boca Raton, Florida : CRC Press, [2019] |: Chapman and Hall/CRC.
- Cheng X, Cook D, Hofmann H. (2015) Visually exploring missing values in multivariable data using a graphical user interface. *J Stat Soft* **68**.
- Conway JR, Lex A, Gehlenborg N. (2017) UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* **33**: 2938–2940.
- Cooper HM. (1998) *Synthesizing research: A guide for literature reviews* 3rd ed. Thousand Oaks, Calif: Sage Publications.
- Cooper HM, Hedges LV, Valentine JC (eds). (2019) *Handbook of research synthesis and meta-analysis* 3rd edition. New York: Russell Sage Foundation.
- Dempster AP, Laird NM, Rubin DB. (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* **39**: 1–22.
- Diggle P, Kenward MG. (1994) Informative drop-out in longitudinal data analysis. *Applied Statistics* **43**: 49.
- Egger M, Smith GD, Schneider M, Minder C. (1997) Bias in meta-analysis detected by a simple, graphical test. *BMJ* **315**: 629–634.
- Graham JW. (2009) Missing data analysis: Making it work in the real world. *Annu Rev Psychol* **60**: 549–576.
- Graham JW. (2012) *Missing Data*. New York, NY: Springer New York.
- Graham JW, Cumsille PE, Elek-Fisk E. (2003) Methods for handling missing data. In I.B. Weiner (ed), *Handbook of Psychology*, p. wei0204. Hoboken, NJ, USA: John Wiley & Sons, Inc.
- Hedges LV. (1984) Estimation of effect size under nonrandom sampling: The effects of censoring studies yielding statistically insignificant mean differences. *Journal of Educational Statistics* **9**: 61.
- Hedges LV, Olkin I. (1985) *Statistical methods for meta-analysis*. Orlando: Academic Press.
- Higgins JP, White IR, Wood AM. (2008) Imputation methods for missing outcome data in meta-analysis of clinical trials. *Clinical Trials* **5**: 225–239.
- Ibrahim JG. (1990) Incomplete data in generalized linear models. *Journal of the American Statistical Association* **85**: 765–769.
- Ibrahim JG, Lipsitz SR, Chen M-H. (1999) Missing covariates in generalized linear models when the missing data mechanism is non-ignorable. *J Royal Statistical Soc B* **61**: 173–190.
- Light RJ, Pillemer DB. (1984) *Summing up: the science of reviewing research*. Cambridge, Mass: Harvard

University Press.

- Little RJA. (1988) A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association* **83**: 1198–1202.
- Little RJA, Rubin DB. (2002) *Statistical Analysis with Missing Data*. Hoboken, NJ, USA: John Wiley & Sons, Inc.
- Molenberghs G, Beunckens C, Sotito C, Kenward MG. (2008) Every missingness not at random model has a missingness at random counterpart with equal fit. *J Royal Statistical Soc B* **70**: 371–388.
- Molenberghs G, Kenward MG, Lesaffre E. (1997) The analysis of longitudinal ordinal data with nonrandom drop-out. *Biometrika* **84**: 33–44.
- Newbury-Birch D, Ferguson J, Landale S, et al. (2018) A systematic review of the efficacy of alcohol interventions for incarcerated people. *Alcohol and Alcoholism* **53**: 412–425.
- Pigott TD. (2001a) Missing predictors in models of effect size. *Eval Health Prof* **24**: 277–307.
- Pigott TD. (2001b) A review of methods for missing data. *Educational Research and Evaluation* **7**: 353–383.
- Pigott TD. (2019) Handling missing data. In Harris Cooper, Larry V. Hedges & Jeffrey C. Valentine (eds), *The Handbook for Research Synthesis and Meta-analysis*, 3rd edn. New York: Russell Sage.
- Ramsey AT, Satterfield JM, Gerke DR, Proctor EK. (2019) Technology-based alcohol interventions in primary care: Systematic review. *J Med Internet Res* **21**: e10859.
- Rhoads CH. (2012) Problems with tests of the missingness mechanism in quantitative policy studies. *Statistics, Politics, and Policy* **3**.
- Rosenthal R. (1979) The file drawer problem and tolerance for null results. *Psychological Bulletin* **86**: 638–641.
- Rothstein H, Sutton AJ, Borenstein M (eds). (2005) *Publication bias in meta-analysis: prevention, assessment and adjustments*. Chichester, England ; Hoboken, NJ: Wiley.
- Rubin DB. (1976) Inference and missing data. *Biometrika* **63**: 581–592.
- Rubin DB. (1987) *Multiple imputation for nonresponse in surveys*. New York: Wiley.
- Schafer JL, Graham JW. (2002) Missing data: our view of the state of the art. *Psychol Methods* **7**: 147–177.
- Seaman S, Galati J, Jackson D, Carlin J. (2013) What is meant by ‘missing at random’? *Statist Sci* **28**: 257–268.
- Tanner-Smith EE, Steinka-Fry KT, Kettrey HH, Lipsey MW. (2016) Adolescent substance use treatment effectiveness: A systematic review and meta-analysis. Office of Justice Programs.
- Tanner-Smith EE, Wilson SJ, Lipsey MW. (2013) The comparative effectiveness of outpatient treatment for adolescent substance abuse: A meta-analysis. *Journal of Substance Abuse Treatment* **44**: 145–158.
- Tierney NJ. (2017) visdat: Visualising whole data frames. *JOSS* **2**: 355.
- Tierney NJ, Cook DH. (2018) Expanding tidy data principles to facilitate missing data exploration, visualization and assessment of imputations. *arXiv:180902264 [stat]*.
- Tipton E, Pustejovsky JE, Ahmadi H. (2019a) A history of meta-regression: Technical, conceptual, and practical developments between 1974 and 2018. *Res Syn Meth* **10**: 161–179.

- Tipton E, Pustejovsky JE, Ahmadi H. (2019b) Current practices in meta-regression in psychology, education, and medicine. *Res Syn Meth* **10**: 180–194.
- Trikalinos T. (2012) OpenMetaAnalyst: Powerful open-source software for meta-analysis. <https://effectivehealthcare.ahrq.gov/products/open-meta-analyst/abstract>.
- Troxel AB, Harrington DP, Lipsitz SR. (2002) Analysis of longitudinal data with non-ignorable non-monotone missing values. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **47**: 425–438.
- Tukey JW. (1962) The future of data analysis. *Ann Math Statist* **33**: 1–67.
- Viechtbauer W. (2010) Conducting meta-analyses in R with the metafor package. *J Stat Soft* **36**.
- White A, Kavanagh D, Stallman H, et al. (2010) Online alcohol Interventions: A systematic review. *J Med Internet Res* **12**: e62.
- Yuvaraj K, Eliyas SK, Gokul S, Manikandanesan S. (2019) Effectiveness of workplace intervention for reducing alcohol consumption: A systematic review and meta-analysis. *Alcohol and Alcoholism* **54**: 264–271.