# PHP 2550: Worksheet 4

**Due: September 27th at 11:59pm**

## Reading Recap

1. Recap Chapters 3-6 of the "The 9 Pitfalls of Data Science" in three points and choose your favorite example. Each point could be summarizing a key takeaway, something that surprised you, or something that you want to remember.

   **Answer:**

   - In pitfall 4, worshiping computer, the example of "Seeing the World Through Pixels" surprises me a lot. This example talks about how differently computers and human "see" an object. When a picture of school bus was shown, the DNN (deep neural network) identified it as an airplane instead of a school bus. This occurs since, unlike humans, computers would not understand the picture but analyzes the pixel patterns. Human always identify the object by understanding the context and meaning, even the picture is blocked or unclear, we rely on our experiences to identify it. I choose this example since it helped me understand something I had been confused about. When trying to log in to some websites, I have encountered tasks asking me to recognize traffic lights, buses, or taxis and they said this task can prevent computer from logging in. I used to believe this should be a easy task for "smart computers", but after reading this example, I understand the reason for that and I realize how important and sophisticated human perception really is. In addition, I would like to remember this tip and avoid overly trust computers.
   - In pitfall 5, Torturing Data, the example of "Aspartame doesn't cause cancer" does catch my attention. I usually drink diet coke to avoid suger intake and I was concerned about this topic previously when there are people saying that Aspartame could cause cancer. A lot of studies have shown that aspartame does not cause cancer, and this result is confirmed by both the FDA and European Food Safety Authority. However, because of fears inflamed by hoax emails, people continued to test it. Because of over-testing, there a few occasional anomal cases popped up, suggesting possible links between brain tumors and aspartame. This p-hacking case shows that over-testing or over-analyzing is a way to torturing data and this would

increase the probability of finding meaningless patterns which are not the result we would like to have. This is also a tip I would like to remind myself.

- In pitall 6, fooling yourself, the example of "Wishful Thinking" attracts my attention since I also have experiences of overestimating my math test scores during high school and this scenario is common in our class. This example talks about how people can overestimate their ability that the 100 high school students predicted a score of 75 for a math test on average while their mean score is actually 60. This discrepancy occurs due to bias of optimism or an underestimation of the math test's difficulty, leading to wrong decision making or conclusion. Our human nature can lead us to be overly optimistic which is a way of fooling ourselves, letting our subjective biases and desires influence our judgement. This is also a tip I would remind myself in the future.

2. The questions below relate to the article Personalized Research on Diet in Ulcerative Colitis and Crohn's Disease: A Series of N-of-1 Diet Trials.

   - Summarize the paper in 1-2 paragraphs using your own words (a good way to help this is to not look at the paper when writing your response). Use the six questions from last week to help guide your summary
     - What do the author(s) want to know (motivation)?
     - What did they do (approach/methods)?
     - Why was it done that way (context within the field)?
     - What do the results show (figures and data tables)?
     - How did the author(s) interpret the results (interpretation/discussion)?
     - What should be done next (discussion/own reflection)?

     **Answer:** As the evidence regarding specific carbohydrate diet (SCD) for inflammatory bowel disease (IBD) remains limited, the aim of this study was to investigate the effect of two different diets, SCD and modified SCD (MSCD), on IBD symptoms and inflammation in children and adolescents with IBD. To accomplish this, an N-of-1 trial study design was implemented. First, 54 individuals who met the inclusion criteria began a usual diet (US) for at least one week. Then, they were randomized to begin either SCD or MSCD using a centralized, stratified, block randomization approach with a 1:1 allocation ratio. To avoid confounding secular trends with the actual treatment effects, multiple crossovers were conducted, in which participants alternated between SCD and MSCD for 4 periods, each 8 weeks long.

     Analyses were conducted at both the individual and aggregate levels. The individual-level figures and results shown in the table indicate that there were no clinically meaningful differences between SCD and MSCD for the participants.

Changes in IBD symptoms in SCD versus UD varied by individual. Bayesian generalized linear mixed models with non-informative priors were used to aggregate the results across individuals. Results showed that the probability of a clinically meaningful difference in IBD symptoms between SCD and MSCD was less than 1% in the pooled estimation sample. The probability that SCD was better than UD in reducing IBD symptoms was 62%, and the probability that MSCD was better than UD was 45%. The authors concluded that most individuals showed no differences between the SCD and MSCD diets, and this was also true at the aggregate level.

Some drawbacks and limitations were discussed in the paper. Although the N-of-1 trials were designed to allow for a robust comparison between SCD and MSCD, they were not optimized for a robust comparison between these two intervention diets and UD due to the lack of randomization and information on the UD. In addition, the authors noted that N-of-1 trials are more suitable for studies where interventions take effect and wear off quickly, which was not the case in this study. Therefore, they suggested applying the N-of-1 trial methodology to interventions that have a quicker cause-effect relationship in future gastrointestinal disease studies.

- Reflecting on the replication crisis and documentation discussion from class, how easy would it be to try to replicate this study? Explain your response.

  **Answer:** This study would be somewhat difficult to replicate. In this study, the author provides detailed explanation of their study design and method, and they detailed introduce their analysis process. They includes a well-structured explanation of their use of bayesian generalized linear models, and they provide many supplementary materials including detailed information about their used equations and underlying logics of their approach. Although these materials are quite helpful, without providing the data and code script of their analysis, it might require significant efforts to replicate this study.

  Since the original dataset is not provided, people who would like to reproduce this study need to recollect data which would introduce variability to the replication. Differences among sample characteristics, data collection protocols, and settings would influence the reproducibility.

  In addition, without providing the exact code, although they provide detailed explanation of their methods, it is difficult to reproduce the exact analysis. The use of package, software version, and coding logistic would lead to various outcomes.

- How was the data presented in the paper? What parts of an exploratory analysis were included?

  **Answer:** Summary tables and figures were used to present the data in this paper. The paper firstly exhibits a comprehensive summary table of all baseline characteristics such as age, gender, disease type, medication, and disease duration, grouped by the completeness of the patients' participation in the study (full completers, early

completers, and withdrawals). The table provide readers insights of comparisons of patients' characteristics across completeness groups using summary statistics and p-values. By including this table, the author presented the key characteristics of the study participants, and allowed the reader to assess the generalisability of the findings and identify potential confounding factors or biases.

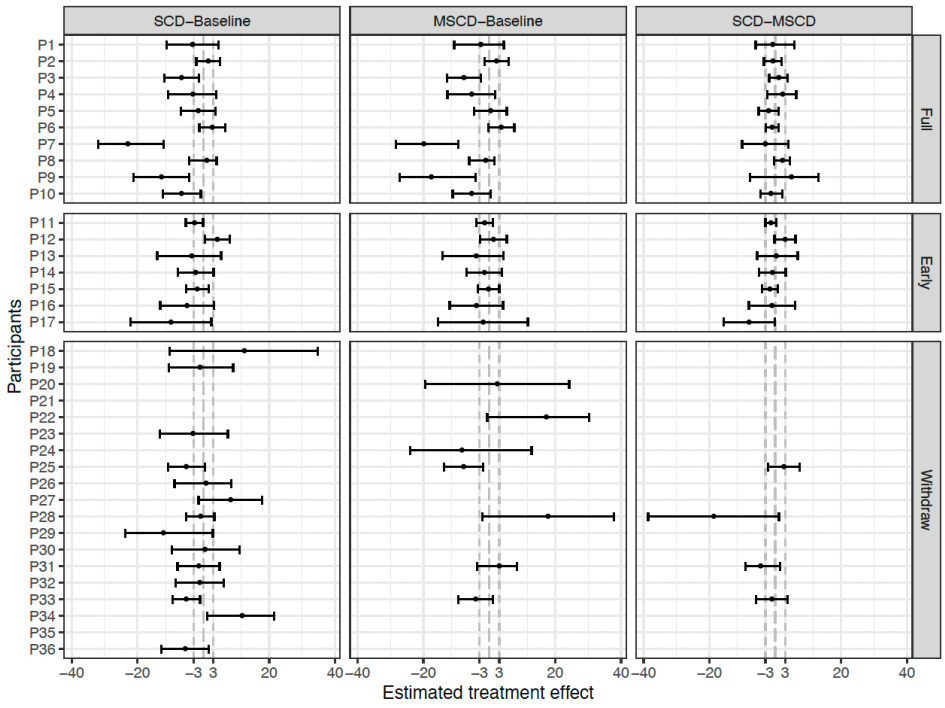
In addition, the author presents tables of the withdrawal and early completed patients with their reasons for dropping out or early completion, along with their basic demographic and clinical information. This table could help to analyze missing patterns, understanding the logic of dropping our rate and reasons.

In addition, in terms of findings, the paper presents changes in IBD symptoms and fecal calprotectin levels between different diets through figures, which helps readers understand the underlying treatment effects. They also summarize individual results of symptom scales, changes, and their overall diet adherence presented in the table, exploring how individuals responded differently to the diet intervention. The author also included the posterior estimates as a result of aggregate analysis.

The mixture of summary tables and graphs provides a detailed introduction of the data, introducing both patients' characteristics and outcomes across groups. Additionally, the combination of grouped and individual outcome results helps us to understand both the overall and personalized responses to the dietary interventions.

- Take a look at the visual below that did not make into the paper and was replaced with Figure 1 in the paper. How is the data presented in each visual? Why do you think the authors went with Figure 1?

  **Answer:** Both figures attempt to show treatment effects for three subpopulations and three comparative groups of diets. However, the below graph shows the estimated treatment effects with 95% confidence intervals demonstrating the uncertainty of the estimates, whereas Figure 1 in the paper shows the probability of improvement in IBD symptoms in different colors. In contrast, Figure 1 is more readable because it helps the reader distinguish the difference in outcome between diets more intuitively. For example, instead of looking to see if the 95% confidence interval includes 0, the reader can tell if the first diet is better than the second by looking to see if black is the dominant color in the stacked bars. This improvement of readability might be the reason that the author switched to Figure 1.




3. The following questions relate to the reading "An Introduction to Modern Missing Data Analyses".

   - Explain the three missing data mechanisms introduced Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR) and give an example of each. Describe a setting in which it would be difficult to tell which setting is appropriate for our data.

     **Answer:**

     1. MCAR: MCAR, missing completely at random, is when the missing is completely random and it does not depend on any variables in the data. This means that the probability of missing data on a variable is unrelated to any observed or unobserved data. An example for MCAR is in a heart disease study, the technician accidently loses some lab samples of some participants. Here, the missingness is not related to any participants' health related or other observed and unobserved variables, making it completely random.

     2. MAR: MAR, missing at random, is when the missingness is related to observed variables, but not on the missing variable itself. In this situation, the missingness can be explained by other observed variables, but not the missing variable itself. An example of MAR is, in a study of depression, male participants are less likely to complete the survey questions asking about depression severity

compared to the female participants, leading to missingness. This missingness is not related to the severity of participants' severity of depression, but is related to their gender which is observed.

3. MNAR: MNAR, missing not at random, is when the probability of missing data is related to the incompleted variables itself. An example for MNAR is that people with higher incomes may be more likely to leave the income question blank due to privacy concerns in a survey. Here, the missing of income response is related to participants' income level but not related to other observed variables.

In a longitudinal depression study, people complete surveys regularly through time and some participants might drop out from the study. This missingness might have various causes which might be difficult to distinguish among these three types. For example, if the participants move to other places where the reason for the moving is not related to any variables in our studies, this missingness is random and it should be MCAR. Or, if the participants are too old to make regular appointment and complete the survey, the missing data is related to their age, but not related to their severity of depression. This case is MAR. Or, if the participants' depression severity becomes much worse and they drop out since they are not willing to share their worse conditions, this missingness is MNAR where it is related to the missing variable itself. In this case, it would be difficult to tell which setting is appropriate for our data because in reality, researchers might not have clear information about the drop our reason and drop out might occur due to mix of reasons mentioned earlier.

- In this class, we will focus on exploring missing data, thinking through what could have contributed to missing data, and introduce using multiple imputation as a possible tool to use. Write a 3-5 sentence explanation of multiple imputation and then describe settings when multiple imputation might not be appropriate to use.

  **Answer:** Multiple imputation is a modern statistical technique to handle missing data. The basic procedure for multiple imputation includes three steps: the imputation phase, the analysis phase, and the pooling phase. In the imputation phase, we generate a specified number of datasets with different estimates of missing values. Then, we perform analyses separately using the same methods and steps (with no change due to missingness) on each dataset generated during the imputation phase, yielding multiple sets of estimates of each parameter and standard error. Last, in the pooling phase, we combine these multiple sets of results from each dataset's analysis into a single set of results.

  Multiple imputation might not be appropriate to use when MNAR occurs, that is, when the missingness is related to the missing variable itself. This is because multiple imputation assumes the missingness in the data can be explained by observed data. The implementation of multiple imputation on MNAR data would lead to

biases. Or, if our data has small sample size, it would be challenged to do reliable imputation using limited observations. Moreover, if the proportion of missingness is extremely high, our imputation might also be unreliable due to lack of information which would introduce uncertainty.

## Data Exploration and Visualization

The data (schmid_data.csv and schmid_codebook.xslx) in this question comes from a previous PhD qualifying exam. The data is a national data set of demographic, diagnostic, and respiratory parameters of infants with severe bronchopulmonary dysplasia (sBPD) admitted to collaborative NICUs and with known respiratory support parameters at 36 weeks postmenstrual age (PMA). This data was used to develop a regression model to predict the composite outcome of tracheostomy/death to guide the indication criteria and timing of tracheostomy placement. For our purposes, we will focus on exploratory analysis of this data and you do not need to do any modeling.

Conduct an exploratory analysis of this data (3-5 pages). To guide yourself, think of at least three questions you want to answer using your EDA. You should also look at patterns of missing data. As part of your analysis, create a summary table and include at least two visualizations. When creating your visuals, you should think about what you want the reader to learn from your visual, making your visual clear and effective, and the data-to-ink ratio (i.e. how much information your visual contains and whether the visual is more effective than putting the same information in text). Your writing accompanying your analysis can be short and informal.

**Answer:**

Baseline characteristics (birth, demographics, medications, and clinical measurements) are listed in Table 1. The outcomes we are interested in are binary variables indicating whether the infant received tracheotomy prior to discharge or whether the infant died. According to Table 1, we observe a large proportion of missing values for certain variables, such as complete prenatal steroids, maternal chorioamnionitis, hospital discharge. For example, 42% of the tracheotomy outcome 'Trach' and 39% of the death outcome 'Death' lacked complete prenatal steroid information. This pattern suggests that the exclusion of these variables needs to be considered when conducting multiple imputations.

To further explore the missing patterns in the dataset, we created Figure 1. Missing observations are shown in gray, with more gray shaded areas indicating more missing values. We can see from the figure that missing values increase over time, with more values missing at 44 weeks than at 36 weeks. This may be due to the fact that healthier infants are more likely to be discharged from the hospital and the researchers were unable to collect information on them after discharge.

To explore the association between outcomes and the medical centers where the data were collected, we created a new categorical variable that combined information on both outcomes into four categories: had tracheotomy and died (TD), did not have tracheotomy and died (NTD), had tracheotomy without death (TND), and had no tracheotomy without death (NTND). As we can see in Figure 2, more than half of the infants did not have a tracheotomy and did not die, except in centers 1, 12, and 21. Especially in center 21, all infants had tracheotomy without death. This implies that there may be potential differences in the characteristics and settings of infants admitted to the centers.

The longitudinal structure of the data also allows us to understand the distribution of infant weight over time, and how this may affect our new outcome of 'tracheal intubation/death'. As shown in Figure 3, the distribution of infant weights is approximately normally distributed, with the distribution becoming more spread out over time. Looking at the stacked bars in the figure, we find that the light blue bars (NTND) consistently have the largest proportion, suggesting that the majority of infants did not require tracheotomy and survived during the study period. Notably, comparing the red bars (TD) between the three time points, we found that more newborns required tacheometry at baseline compared to 36 and 44 weeks but did not survive the study period compared.

Table 1: Summary Statistics by Death and Tracheostomy

| | **Trachoestomy** | | **Death** | |
|---|---|---|---|---|
| **Characteristic** | **No**, N = 853[1] | **Yes**, N = 146[1] | **No**, N = 943[1] | **Yes**, N = 54[1] |
| **Gender** | | | | |
| Female | 348 (41%) | 60 (41%) | 390 (41%) | 17 (31%) |
| Male | 501 (59%) | 86 (59%) | 549 (58%) | 37 (69%) |
| Missing | 4 (0.5%) | 0 (0%) | 4 (0.4%) | 0 (0%) |
| **Ethnicity** | | | | |
| Hispanic or Latino | 66 (7.7%) | 8 (5.5%) | 71 (7.5%) | 3 (5.6%) |
| Not Hispanic or Latino | 740 (87%) | 128 (88%) | 818 (87%) | 48 (89%) |
| Missing | 47 (5.5%) | 10 (6.8%) | 54 (5.7%) | 3 (5.6%) |
| **Birth Weight** | 814 (295) | 761 (303) | 811 (288) | 721 (406) |
| **Gestational Age** | 26 (2) | 26 (2) | 26 (2) | 26 (2) |
| **Birth Length** | 33 (4) | 32 (4) | 33 (4) | 30 (4) |
| Unknown | 48 | 30 | 71 | 7 |
| **Birth Head Circumference** | 23.22 (2.71) | 22.99 (3.07) | 23.24 (2.70) | 22.34 (3.54) |
| Unknown | 46 | 31 | 72 | 5 |
| **Delivery Method** | | | | |
| Vaginal delivery | 254 (30%) | 31 (21%) | 274 (29%) | 10 (19%) |
| Cesarean section | 597 (70%) | 114 (78%) | 666 (71%) | 44 (81%) |
| Missing | 2 (0.2%) | 1 (0.7%) | 3 (0.3%) | 0 (0%) |
| **Prenatal Corticosteroids** | | | | |

| | | | | |
|---|---|---|---|---|
| Yes | 715 (84%) | 123 (84%) | 792 (84%) | 44 (81%) |
| No | 118 (14%) | 8 (5.5%) | 121 (13%) | 5 (9.3%) |
| Missing | 20 (2.3%) | 15 (10%) | 30 (3.2%) | 5 (9.3%) |
| **Small for gestational age** | | | | |
| SGA | 161 (19%) | 42 (29%) | 177 (19%) | 26 (48%) |
| Not SGA | 681 (80%) | 100 (68%) | 751 (80%) | 28 (52%) |
| Missing | 11 (1.3%) | 4 (2.7%) | 15 (1.6%) | 0 (0%) |
| **Complete Prenatal Steroids** | | | | |
| Yes | 527 (62%) | 86 (59%) | 577 (61%) | 34 (63%) |
| No | 166 (19%) | 27 (18%) | 184 (20%) | 9 (17%) |
| Missing | 160 (19%) | 33 (23%) | 182 (19%) | 11 (20%) |
| **Maternal Chorioamnionitis** | | | | |
| Yes | 138 (16%) | 22 (15%) | 153 (16%) | 7 (13%) |
| No | 665 (78%) | 112 (77%) | 732 (78%) | 43 (80%) |
| Missing | 50 (5.9%) | 12 (8.2%) | 58 (6.2%) | 4 (7.4%) |
| **Hospital Discharge Gestational Age** | 49 (24) | 80 (30) | 52 (27) | 59 (24) |
| Unknown | 86 | 38 | 116 | 7 |

[1]Mean (SD) for continuous; n (%) for categorical

Figure 1: Heat Map across All variables

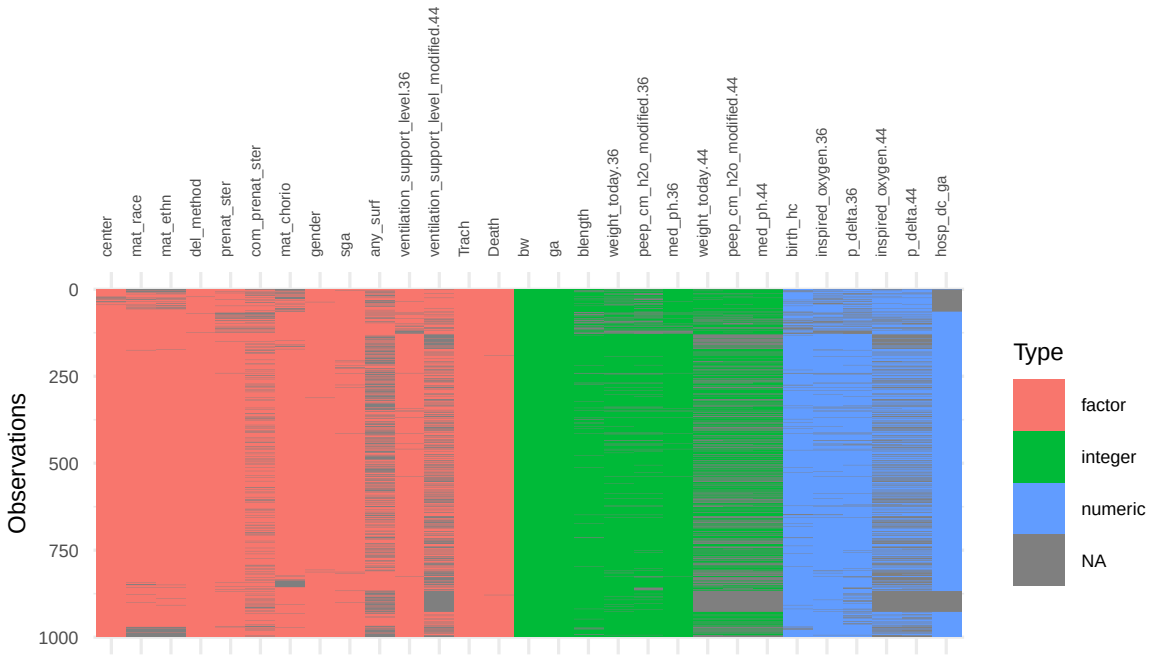


Figure 2: Outcome Proportion by Center

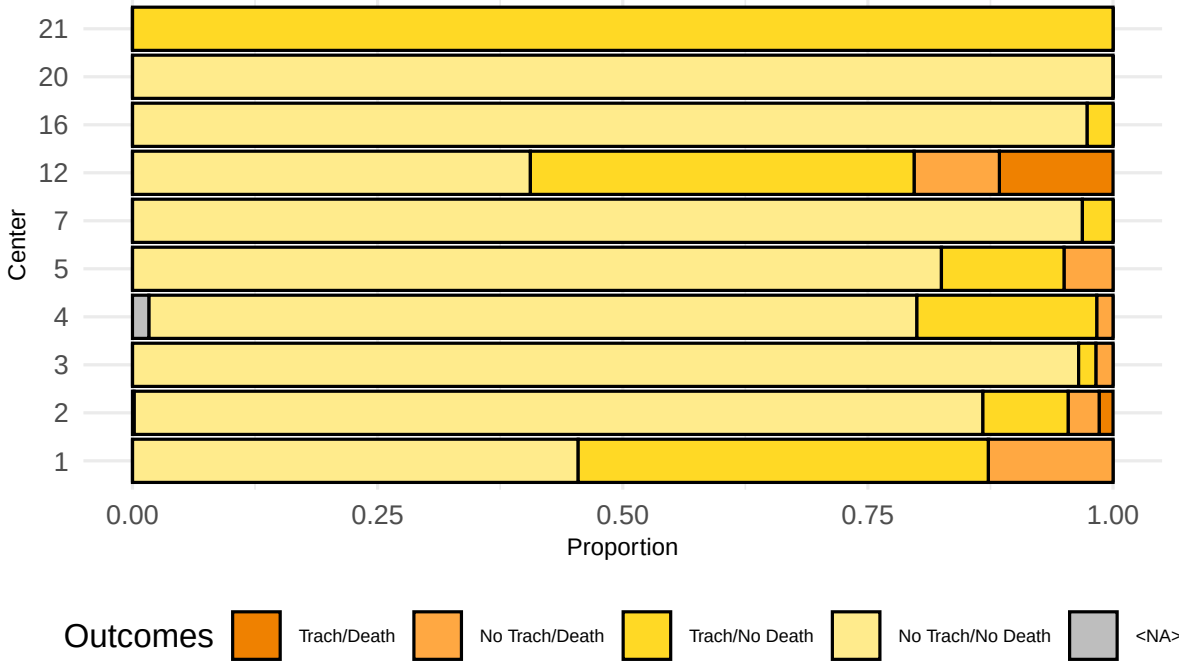
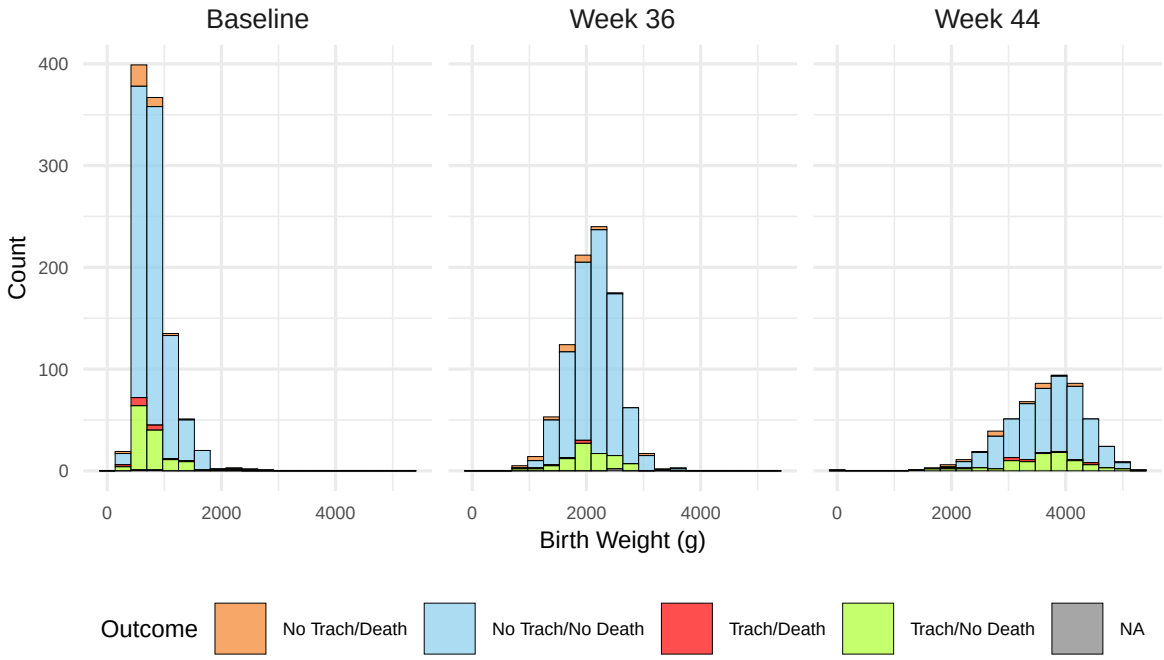



Figure 3: Birth Weight Distribution by Tracheostomy/Death Outcome



## Missing Data and Imputation with MICE

Load in the data called `pain` from the `HDSinRdata` package and read the documentation. The data contains information from patient-reported pain assessments at baseline and at a 3-month follow-up.

1. First, describe the patterns of missing data observed in the data set overall. How would you present this information in an exploratory analysis?

   **Answer:** I would present the missing data pattern using heat map. Base on the data structure, I plot three separate heat maps focusing on the body region variables, baseline characteristics, and the follow-up variable, respectively.

   In our search for missing patterns, we first noticed that the entire row of 11749 was empty. Our guess is that there was a manual error in the coding of the dataset, and thus we consider excluding this row when discussing the missing patterns. From the heat map of the body region variables, we can see that all individuals have completed data since there is no visible grey area on the plot, indicating no missing values. The baseline characteristics, including both continuous and categorical, exhibits some missingness among certain columns. Categorical variables, such as `PAT_RACE` and `MEDICAID_BIN`, have small portion of missing data, while continuous variables like `BMI` and `GH_PHYSICAL_SCORE` have larger proportion of missingness. To be noticed, the follow-up variable in the third heat map exhibits a significantly large portion of missingness, with over 50% of value missing, which might be a crucial issue.

Figure 4: Heat Map for Body Region Variables

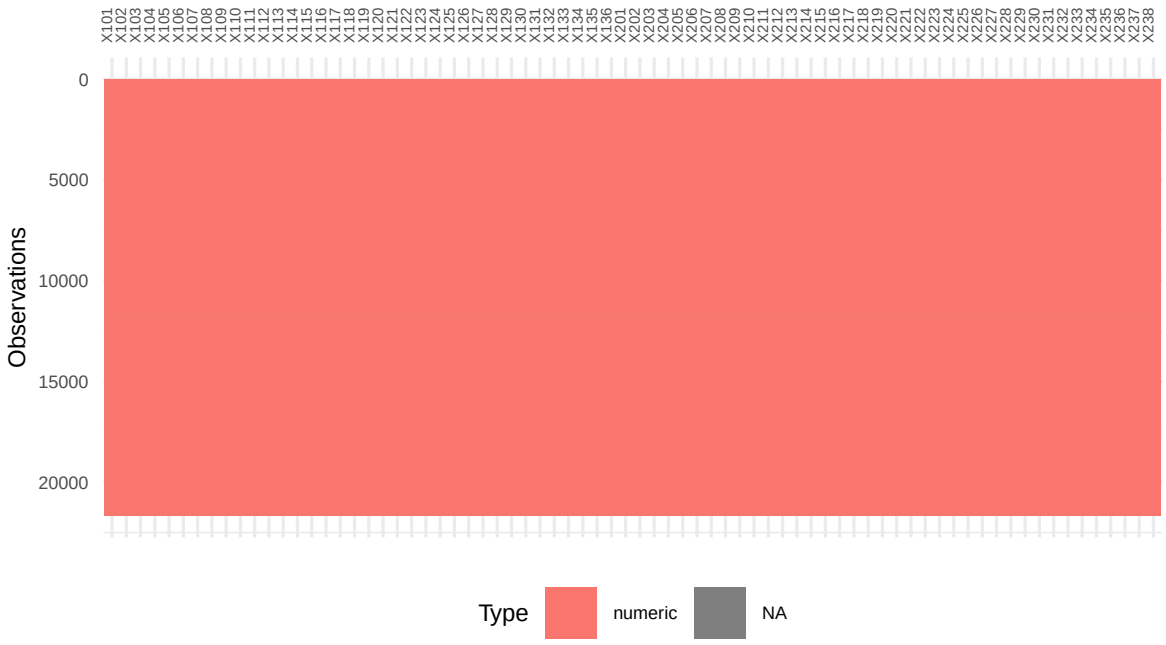



Figure 5: Heat Map for Baseline Characteristics

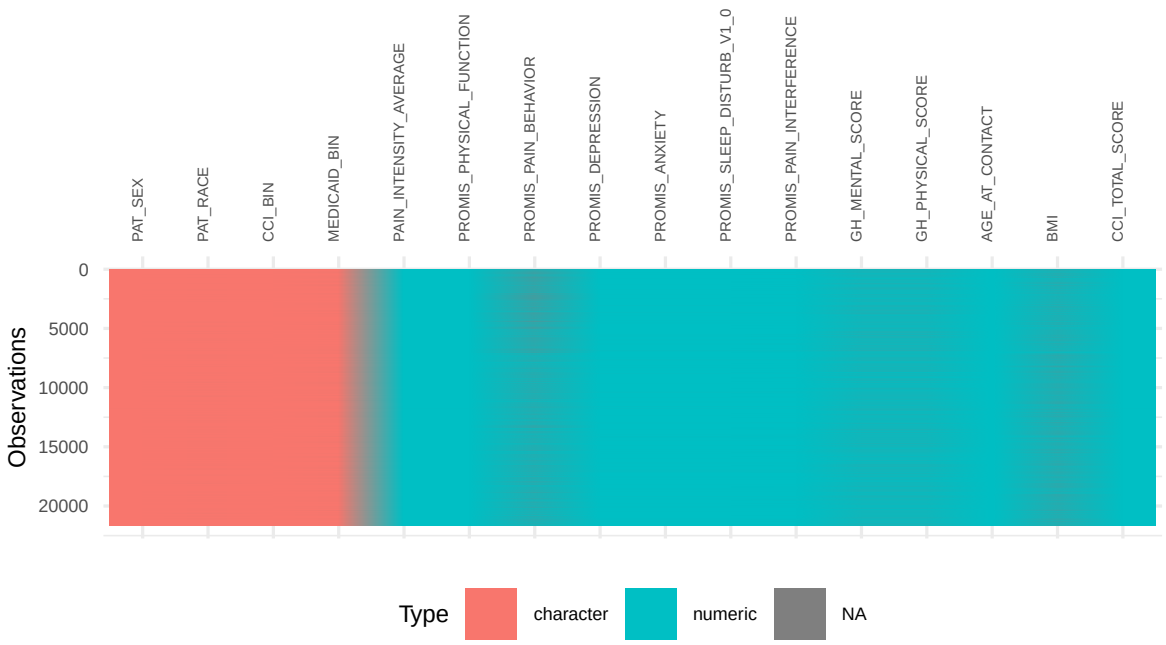
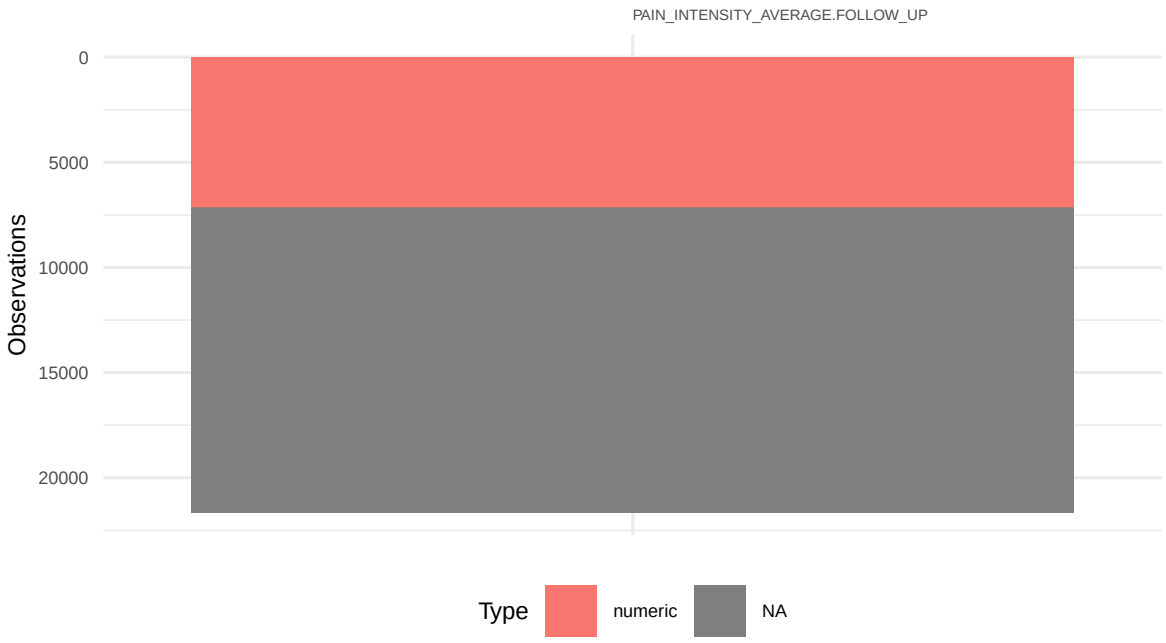


Figure 6: Heat Map for Follow−up Variable




2. If we were interested in analyzing the change in pain over time, it would be important to think about the missing data due to loss to follow-up. Compare the baseline characteristics between those with and without follow-up information. Comment on your results and discuss whether you think the data is MCAR, MAR, or MNAR.

   **Answer:** To compare the baseline characteristics between those with and without follow-up information, we create a table summary of all baseline characteristics by follow-up presence. Here, participants without follow-up had significantly higher pain intensity score (7.00 vs. 6.00 with $p < 0.01$) compared to participants with follow-up. Moreover, in several PROMIS measures, patients with and without follow-up exhibits small but statistically significant differences with smaller p-value. For those demographic variables, age, BMI, gender, and race, significant difference between groups present as well. For example, younger people or patients with lower BMI value are more likely not to have the follow-up information. Moreover, observing the MEDICAID_BIN summary, medicaid recipients are more likely to skip the follow-up.

   We think this data is more likely to be MAR where the follow-up variable missingness is relates to observed baseline characteristics variables, but not depend on the unobserved data. Since many baseline characteristics variables exhibits significant difference between people with and without follow-up information as we mentioned earlier, these characteristics seem associated with the probability of missingness which fits the assumption of MAR.

Table 2: Summary Statistics by Follow-up Presence

| **Characteristic** | **Follow-up**, N = 7,138[1] | **No Follow-up**, N = 14,521[1] | **p-value**[2] |
|---|---|---|---|
| PAIN_INTENSITY_AVERAGE | 6.00 (5.00, 8.00) | 7.00 (5.00, 8.00) | <0.001 |
| PROMIS_PHYSICAL_FUNCTION | 35 (31, 39) | 35 (30, 39) | 0.045 |
| PROMIS_PAIN_BEHAVIOR | 61.6 (59.0, 63.4) | 61.7 (59.5, 63.4) | <0.001 |
| PROMIS_DEPRESSION | 55 (48, 62) | 55 (48, 62) | 0.009 |
| PROMIS_ANXIETY | 56 (49, 63) | 58 (50, 63) | <0.001 |
| PROMIS_SLEEP_DISTURB_V1_0 | 59 (54, 65) | 60 (54, 65) | <0.001 |
| PROMIS_PAIN_INTERFERENCE | 67 (62, 70) | 67 (63, 72) | <0.001 |
| GH_MENTAL_SCORE | 44 (36, 51) | 44 (36, 51) | 0.5 |
| GH_PHYSICAL_SCORE | 35 (30, 40) | 35 (30, 40) | 0.028 |
| AGE_AT_CONTACT | 58 (48, 67) | 57 (45, 68) | 0.001 |
| BMI | 30 (25, 35) | 29 (25, 34) | <0.001 |
| CCI_TOTAL_SCORE | | | 0.6 |
| 0 | 6,405 (90%) | 13,119 (90%) | |
| 1 | 606 (8.5%) | 1,149 (7.9%) | |
| 2 | 104 (1.5%) | 196 (1.3%) | |
| 3 | 19 (0.3%) | 45 (0.3%) | |
| 4 | 4 (<0.1%) | 10 (<0.1%) | |
| 5 | 0 (0%) | 1 (<0.1%) | |
| PAT_SEX | | | <0.001 |
| female | 4,431 (62%) | 8,671 (60%) | |
| male | 2,707 (38%) | 5,849 (40%) | |
| PAT_RACE | | | <0.001 |
| ALASKA NATIVE | 1 (<0.1%) | 1 (<0.1%) | |
| AMERICAN INDIAN | 29 (0.4%) | 29 (0.2%) | |
| BLACK | 934 (13%) | 2,295 (16%) | |
| CHINESE | 4 (<0.1%) | 17 (0.1%) | |
| DECLINED | 28 (0.4%) | 93 (0.6%) | |
| FILIPINO | 3 (<0.1%) | 3 (<0.1%) | |
| GUAM/CHAMORRO | 0 (0%) | 1 (<0.1%) | |
| HAWAIIAN | 1 (<0.1%) | 0 (0%) | |
| INDIAN (ASIAN) | 13 (0.2%) | 36 (0.2%) | |
| JAPANESE | 1 (<0.1%) | 8 (<0.1%) | |
| KOREAN | 3 (<0.1%) | 7 (<0.1%) | |
| NOT SPECIFIED | 2 (<0.1%) | 2 (<0.1%) | |
| OTHER | 0 (0%) | 1 (<0.1%) | |
| OTHER ASIAN | 8 (0.1%) | 39 (0.3%) | |
| OTHER PACIFIC ISLANDER | 1 (<0.1%) | 11 (<0.1%) | |
| VIETNAMESE | 3 (<0.1%) | 3 (<0.1%) | |
| WHITE | 6,080 (86%) | 11,860 (82%) | |

| | | | |
|---|---|---|---|
| CCI_BIN | | | 0.2 |
| Any comorbidity | 733 (10%) | 1,401 (9.6%) | |
| No comorbidity | 6,405 (90%) | 13,119 (90%) | |
| MEDICAID_BIN | | | <0.001 |
| Missing | 92 (1.3%) | 209 (1.4%) | |
| no | 5,708 (80%) | 11,049 (76%) | |
| yes | 1,338 (19%) | 3,263 (22%) | |

[1]Median (IQR); n (%)
[2]Welch Two Sample t-test; Fisher's Exact Test for Count Data with simulated p-value (based on 2000 replicates); Fisher's exact test

3. Now suppose we are interested in mental and physical function at baseline and how that is associated with pain intensity. We will use the `mice` package to perform multiple imputation on this data. First, drop the variable `PAIN_INTENSITY_AVERAGE.FOLLOW_UP` and the body map variables `X101:X238`.

```
# Drop PAIN_INTENSITY_AVERAGE.FOLLOW_U and X101:X238
pain_mod <- pain[, -c(1:75)] %>%
  select(-PAIN_INTENSITY_AVERAGE.FOLLOW_UP)

pain_mod$PAT_SEX <- as.factor(pain_mod$PAT_SEX)
pain_mod$PAT_RACE <- as.factor(pain_mod$PAT_RACE)
pain_mod$CCI_BIN <- as.factor(pain_mod$CCI_BIN)
pain_mod$MEDICAID_BIN <- as.factor(pain_mod$MEDICAID_BIN)
```

4. Use the `mice()` function to perform multiple imputation to create five imputed data sets and save the result as `pain_mice`. You should read the documentation on this function to understand how it is implementing this and what arguments you might want to set. What is the structure of the returned object?

```
# pain_mice <- mice(pain_mod, m = 5, maxit = 5, seed = (2550))
# saveRDS(pain_mice, file = "mice_result.rds")
pain_mice <- readRDS("mice_result.rds")
```

**Answer:** The returned object contains the following parts:

- **data:** The original dataset with missing data.
- **imp:** A list of imputed values for variables with missingness.
- **m:** The number of imputed dataset which is 5 here.

- **where:** A matrix presents the location of missing values in our original dataset where missing values are coded as TRUE.
- **blocks:** A list of variable names. Variables in one block would be imputed together.
- **call:** The command used to run the mice() function, including information of data, number of iterations, random seed, and any set parameters.
- **nmis:** A numeric vector presents number of missing values for each variable.
- **method:** The imputation method used.
- **predictorMatrix:** A numeric matrix specifying which variables are used as predictors in the imputation models.
- **visitSequence:** The order variables are imputed in mice function.
- **formulas:** A list of formulas used for imputation of each variable.
- **post:** Post-processing rules applied to the imputed data.
- **blots:** Contains information of additional processing for certain variables during imputation.
- **ignore:** A vector specifying whether specific rows were ignored during imputation. False represents the observation was included in the process while TURE represents it is ignored.
- **seed:** The random seed set in parameter.
- **iteration:** Number of interation for the imputation.
- **lastSeedValue:** This component stores the final state of the random number generator after all imputations.
- **chainMean:** The mean of the imputed values for each variable at each iteration and for each imputation chain.
- **chainVar:** The variance of the imputed values for each variable at each iteration and for each imputation chain.
- **loggedEvents**: A data frame that records unusual events occurred during imputation, like warnings and errors.
- **version:** The version of mice package used for the imputation.
- **date:** The date when the imputation performed.

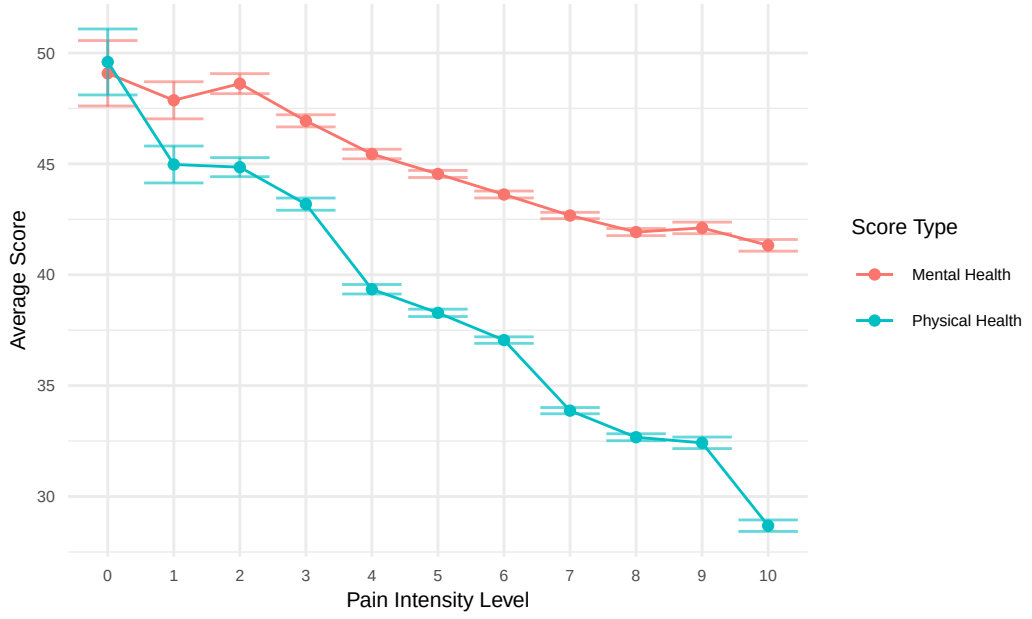
5. For one imputed data set, find the average mental and physical health score for each possible pain intensity level (0-10). To access the first imputed data set you can use the code below.
   `mice::complete(pain_mice,1)`

Table 3: Average Summary of Pain Intensity Score

| Pain Intensity Level | Average Mental Health Score | Average Physical Health Score |
|---|---|---|
| 0 | 49.12553 | 49.43191 |
| 1 | 47.97703 | 45.07432 |
| 2 | 48.68802 | 44.81221 |
| 3 | 46.97733 | 43.19800 |
| 4 | 45.47257 | 39.30918 |
| 5 | 44.58407 | 38.27742 |
| 6 | 43.67522 | 37.04332 |
| 7 | 42.71223 | 33.86519 |
| 8 | 41.94154 | 32.67854 |
| 9 | 42.12265 | 32.38978 |
| 10 | 41.34510 | 28.63168 |

6. Repeat this for the other four data sets and then use Rubin's rules to plot the results. https://bookdown.org/mwheymans/bookmi/rubins-rules.html

Figure 7: Pooled Average Mental and Physical Health Scores by Pain Intensity Levels with SE



## Appendix

```
library(tidyverse)
library(ggplot2)
library(visdat)
library(gt)
schmid_data <- read.csv("schmid_data.csv")
library(tidyr)
library(gtsummary)
library(ggpubr)

schmid_trach <- schmid_data %>%
  select(gender, mat_ethn, bw, ga, blength, birth_hc, del_method,
         prenat_ster, sga, com_prenat_ster,mat_chorio,hosp_dc_ga,Trach) %>%
  mutate(Trach = recode(Trach, `0` = "No", `1` ="Yes")) %>%
  mutate(
    # mat_race = case_when(mat_race == 1 ~ "American Indian or Alaskan Native",
    #                      mat_race == 2 ~ "Asian",
    #                      mat_race == 3 ~ "Black or African American",
    #                      mat_race == 4 ~ "Native Hawaiian or Other Pacific Islande",
    #                      mat_race == 5 ~ "White",
    #                      mat_race == 6 ~ "Other",
    #                      TRUE  ~ "Missing")
    gender = case_when(gender == "Female" ~  "Female",
                       gender == "Male" ~  "Male",
                      TRUE ~ "Missing"),
    mat_ethn = case_when(mat_ethn == 1 ~ "Hispanic or Latino",
                         mat_ethn == 2 ~ "Not Hispanic or Latino",
                         TRUE ~ "Missing"),
    del_method = case_when(del_method==1 ~ 'Vaginal delivery',
                           del_method==2 ~ 'Cesarean section',
                           TRUE ~ "Missing"),
    sga = case_when(sga=="SGA" ~ "SGA",
                    sga=="Not SGA" ~ "Not SGA",
                    TRUE ~ "Missing"),
    prenat_ster = case_when(prenat_ster == "Yes" ~ "Yes",
                            prenat_ster == "No" ~ "No",
                            TRUE ~ "Missing"),
    com_prenat_ster = case_when(com_prenat_ster == "Yes" ~ "Yes",
                            com_prenat_ster == "No" ~ "No",
                            TRUE ~ "Missing"),
```

```
    mat_chorio = case_when(mat_chorio == "Yes" ~ "Yes",
                           mat_chorio == "No" ~ "No",
                           TRUE ~ "Missing")) %>%
  mutate(mat_ethn = factor(mat_ethn, levels = c('Hispanic or Latino',
                                                'Not Hispanic or Latino', 'Missing')),
         del_method = factor(del_method,
                             levels = c('Vaginal delivery',
                                        'Cesarean section', 'Missing')),
         sga = factor(sga, levels = c('SGA', 'Not SGA', 'Missing')),
         prenat_ster = factor(prenat_ster, levels = c('Yes', 'No', 'Missing')),
         com_prenat_ster = factor(com_prenat_ster, levels = c('Yes', 'No', 'Missing')),
         mat_chorio = factor(mat_chorio, levels = c('Yes', 'No', 'Missing'))) %>%
  tbl_summary(by=Trach,
              label = list(gender ~ "Gender",
                           #mat_race ~ "Race",
                           mat_ethn ~ "Ethnicity",
                           bw ~ "Birth Weight",
                           ga ~ " Gestational Age",
                           blength ~ "Birth Length",
                           birth_hc ~ "Birth Head Circumference",
                           del_method ~ "Delivery Method",
                           prenat_ster ~ "Prenatal Corticosteroids",
                           sga ~ "Small for gestational age",
                           com_prenat_ster ~ "Complete Prenatal Steroids",
                           mat_chorio ~ "Maternal Chorioamnionitis",
                           hosp_dc_ga ~ "Hospital Discharge Gestational Age"
                           ),
              statistic = all_continuous() ~ "{mean} ({sd})") %>%
  modify_spanning_header(update =  all_stat_cols() ~  "**Trachoestomy**") %>%
  modify_footnote(update = all_stat_cols() ~ "Mean (SD) for continuous; n (%) for categori
  bold_labels()


schmid_death <- schmid_data %>%
  select(gender, mat_ethn, bw, ga, blength, birth_hc, del_method,
         prenat_ster, sga, com_prenat_ster,mat_chorio,hosp_dc_ga,Death) %>%
  mutate(Death = recode(Death, `0` = "No", `1` ="Yes")) %>%
  mutate(
    # mat_race = case_when(mat_race == 1 ~ "American Indian or Alaskan Native",
    #                      mat_race == 2 ~ "Asian",
    #                      mat_race == 3 ~ "Black or African American",
```

```
  #                     mat_race == 4 ~ "Native Hawaiian or Other Pacific Islande",
  #                     mat_race == 5 ~ "White",
  #                     mat_race == 6 ~ "Other",
  #                     TRUE  ~ "Missing")
  gender = case_when(gender == "Female" ~  "Female",
                     gender == "Male" ~  "Male",
                    TRUE ~ "Missing"),
  mat_ethn = case_when(mat_ethn == 1 ~ "Hispanic or Latino",
                       mat_ethn == 2 ~ "Not Hispanic or Latino",
                       TRUE ~ "Missing"),
  del_method = case_when(del_method==1 ~ 'Vaginal delivery',
                         del_method==2 ~ 'Cesarean section',
                         TRUE ~ "Missing"),
  sga = case_when(sga=="SGA" ~ "SGA",
                  sga=="Not SGA" ~ "Not SGA",
                  TRUE ~ "Missing"),
  prenat_ster = case_when(prenat_ster == "Yes" ~ "Yes",
                          prenat_ster == "No" ~ "No",
                          TRUE ~ "Missing"),
  com_prenat_ster = case_when(com_prenat_ster == "Yes" ~ "Yes",
                          com_prenat_ster == "No" ~ "No",
                          TRUE ~ "Missing"),
  mat_chorio = case_when(mat_chorio == "Yes" ~ "Yes",
                          mat_chorio == "No" ~ "No",
                          TRUE ~ "Missing")) %>%
mutate(mat_ethn = factor(mat_ethn, levels = c('Hispanic or Latino',
                                              'Not Hispanic or Latino', 'Missing')),
       del_method = factor(del_method, levels = c('Vaginal delivery', 'Cesarean section'
                                                  'Missing')),
       sga = factor(sga, levels = c('SGA', 'Not SGA', 'Missing')),
       prenat_ster = factor(prenat_ster, levels = c('Yes', 'No', 'Missing')),
       com_prenat_ster = factor(com_prenat_ster, levels = c('Yes', 'No', 'Missing')),
       mat_chorio = factor(mat_chorio, levels = c('Yes', 'No', 'Missing'))) %>%
tbl_summary(by=Death,
            label = list(gender ~ "Gender",
                         #mat_race ~ "Race",
                         mat_ethn ~ "Ethnicity",
                         bw ~ "Birth Weight",
                         ga ~ " Gestational Age",
                         blength ~ "Birth Length",
                         birth_hc ~ "Birth Head Circumference",
```

```
                            del_method ~ "Delivery Method",
                            prenat_ster ~ "Prenatal Corticosteroids",
                            sga ~ "Small for gestational age",
                            com_prenat_ster ~ "Complete Prenatal Steroids",
                            mat_chorio ~ "Maternal Chorioamnionitis",
                            hosp_dc_ga ~ "Hospital Discharge Gestational Age"
                            ),
              statistic = all_continuous() ~ "{mean} ({sd})") %>%
  modify_spanning_header(update =  all_stat_cols() ~  "**Death**") %>%
  modify_footnote(update = all_stat_cols() ~ "Mean (SD) for continuous; n (%) for categori

schimd_tbl <- tbl_merge(list(schmid_trach, schmid_death),
                        tab_spanner = c("**Trachoestomy**", "**Death**"))

schimd_tbl %>%
  as_gt() %>%
  cols_align(align = "center", columns = everything()) %>%
  tab_options(
    table.width = pct(120),
    table.align = "center"
  ) %>%
  tab_header(
    title = md("Table 1: Summary Statistics by Death and Tracheostomy")
  ) %>%
  tab_style(
    style = cell_text(size = px(10)),  # Adjust the size as needed
    locations = cells_title(groups = "title")
  )
schmid_data_m <- schmid_data
schmid_data_m$center <- as.factor(schmid_data_m$center)
schmid_data_m$mat_race <- as.factor(schmid_data_m$mat_race)
schmid_data_m$mat_ethn <- as.factor(schmid_data_m$mat_ethn)
schmid_data_m$del_method <- as.factor(schmid_data_m$del_method)
schmid_data_m$prenat_ster <- as.factor(schmid_data_m$prenat_ster)
schmid_data_m$com_prenat_ster <- as.factor(schmid_data_m$com_prenat_ster)
schmid_data_m$mat_chorio <- as.factor(schmid_data_m$mat_chorio)
schmid_data_m$gender <- as.factor(schmid_data_m$gender)
schmid_data_m$any_surf <- as.factor(schmid_data_m$any_surf)
schmid_data_m$ventilation_support_level.36 <- as.factor(schmid_data_m$ventilation_support_
schmid_data_m$ventilation_support_level_modified.44 <- as.factor(schmid_data_m$ventilation
schmid_data_m$Trach <- as.factor(schmid_data_m$Trach)
```

```r
schmid_data_m$Death <- as.factor(schmid_data_m$Death)
schmid_data_m$sga <- as.factor(schmid_data_m$sga)

vis_dat(schmid_data_m[, -1]) +
  theme(axis.text.x = element_text(angle = 90, size = 5)) +
  ggtitle("Figure 1: Heat Map across All variables") +
  theme(legend.text = element_text(size = 6),
        legend.title = element_text(size = 8),
        axis.text = element_text(size = 6),
        plot.title = element_text(size = 10),
        axis.title = element_text(size = 8)
        )
schmid_data <- schmid_data %>%
  mutate(Outcome = case_when(Trach == 1 & Death == "Yes" ~ "Trach/Death",
                             Trach == 1 & Death == "No" ~ "Trach/No Death",
                             Trach == 0 & Death == "Yes" ~ "No Trach/Death",
                             Trach == 0 & Death == "No" ~ "No Trach/No Death"))
schmid_count <- schmid_data %>%
  filter(!is.na(center)) %>%
  group_by(center) %>%
  count(Outcome) %>%
  spread(Outcome, n, fill = 0)
 schmid_count <-  schmid_count %>%
   gather(`Trach/Death`,`No Trach/Death`,`No Trach/No Death`,`Trach/No Death`,`<NA>`,
          key = "Outcomes", value = "Count") %>%
   mutate(Outcomes = factor(Outcomes,
                            levels = c("Trach/Death", "No Trach/Death", "Trach/No Death",
                                       "No Trach/No Death", "<NA>"))) %>%
   group_by(center) %>%
   mutate(Proportion = Count/sum(Count))

ggplot(schmid_count, aes(x = Proportion, y = factor(center), fill = Outcomes)) +
  geom_bar(stat = "identity", position = "stack", color = "black") +
  labs(title = "Figure 2: Outcome Proportion by Center", x = "Proportion", y = "Center") +
  theme_minimal() +
  theme(legend.position = "bottom") +
  scale_fill_manual(values = c("Trach/Death" = "#EF8100",
                               "No Trach/Death" = "#FFA842",
                               "Trach/No Death" = "#FFD925",
                               "No Trach/No Death" = "#FFEB8C",
                               "<NA>" = "gray")) +
```

```
  theme(legend.text = element_text(size = 6),
        plot.title = element_text(size = 10),
        axis.title = element_text(size = 8)
        )
schmid_data$ventilation_support_level.36 <-
  as.factor(schmid_data$ventilation_support_level.36)
schmid_data$ventilation_support_level_modified.44 <-
  as.factor(schmid_data$ventilation_support_level_modified.44)

schmid_data$Trach <- as.factor(schmid_data$Trach)
schmid_data$Death <- as.factor(schmid_data$Death)

schmid_long <- schmid_data %>%
  rename(ventilation_support_level.44 = 'ventilation_support_level_modified.44') %>%
  pivot_longer(cols = c('ventilation_support_level.36',
                        'ventilation_support_level.44',
                        'p_delta.36',
                        'p_delta.44'),
               names_to = c(".value", "time"),
               names_pattern = "(ventilation_support_level|p_delta).(\\d+)")
schmid_long$time <- as.factor(schmid_long$time)
schmid_long$Trach <- as.factor(schmid_long$Trach)
schmid_long$ventilation_support_level <- as.factor(schmid_long$ventilation_support_level)
weight_df <- schmid_data %>%
  pivot_longer(cols = c("bw", "weight_today.36", "weight_today.44"),
               names_to = "Time",
               values_to = "Weight")

weight_df <- weight_df %>%
  mutate(Time = case_when(Time == "bw" ~ "Baseline",
                          Time == "weight_today.36" ~ "Week 36",
                          Time == "weight_today.44" ~ "Week 44",
                          TRUE ~ NA))
options(repr.plot.width = 10, repr.plot.height = 10)

weight_df %>%
  group_by(Time) %>%
  ggplot(aes(x = Weight, fill = Outcome)) +
  geom_histogram(bins = 20, position = "stack", color = "black", alpha = 0.7, size = 0.1)
  labs(title = "Figure 3: Birth Weight Distribution by Tracheostomy/Death Outcome",
       x = "Birth Weight (g)", y = "Count") +
```

```
  facet_wrap(~Time, ncol = 3) +
  theme_minimal() +
  theme(legend.position = "bottom") +
  scale_fill_manual(values = c("#F48126", "skyblue", "red", "greenyellow")) +
  theme(legend.text = element_text(size = 6),
        legend.title = element_text(size = 8),
        axis.text = element_text(size = 6),
        plot.title = element_text(size = 10),
        axis.title = element_text(size = 8)
        )

library(HDSinRdata)
library(mice)
library(dplyr)
library(ggplot2)
library(kableExtra)
library(visdat)
library(gtsummary)
data(pain)

vis_dat(pain[, 2:75], warn_large_data = FALSE) +
  theme(axis.text.x = element_text(angle = 90, size = 5)) +
  theme(legend.position = "bottom") +
  ggtitle("Figure 4: Heat Map for Body Region Variables") +
  theme(legend.text = element_text(size = 6),
        legend.title = element_text(size = 8),
        axis.text = element_text(size = 6),
        plot.title = element_text(size = 10),
        axis.title = element_text(size = 8)
        )

vis_dat(pain[, c(76:87, 89:92)]) +
  theme(axis.text.x = element_text(angle = 90, size = 5)) +
  theme(legend.position = "bottom") +
  ggtitle("Figure 5: Heat Map for Baseline Characteristics") +
  theme(legend.text = element_text(size = 6),
        legend.title = element_text(size = 8),
        axis.text = element_text(size = 6),
        plot.title = element_text(size = 10),
        axis.title = element_text(size = 8)
        )
```

```r
vis_dat(pain[, 88]) +
  theme(axis.text.x = element_text(angle = 0, size = 5)) +
  theme(legend.position = "bottom") +
  ggtitle("Figure 6: Heat Map for Follow-up Variable") +
  theme(legend.text = element_text(size = 6),
        legend.title = element_text(size = 8),
        axis.text = element_text(size = 6),
        plot.title = element_text(size = 10),
        axis.title = element_text(size = 8)
        )
pain_followup <- pain %>%
  mutate(follow_up_status = ifelse(is.na(PAIN_INTENSITY_AVERAGE.FOLLOW_UP),
                                   "No Follow-up", "Follow-up")) %>%
  mutate(MEDICAID_BIN = ifelse(is.na(MEDICAID_BIN), "Missing", MEDICAID_BIN))

tbl <- pain_followup %>%
  select(follow_up_status, colnames(pain)[c(76:87, 89:92)]) %>%
  tbl_summary(
    by = follow_up_status,
    missing = "no"
  ) %>%
  add_p(
    test = list(
      all_categorical() ~ "fisher.test",
      all_continuous() ~ "t.test"
    ),
    test.args = list(
      all_categorical() ~ list(simulate.p.value = TRUE)
    )
  ) %>%
  modify_table_body(~ .x %>% dplyr::mutate_at(vars(label), as.character))

tbl %>%
  as_gt() %>%
  tab_options(
    table.font.size = px(6)
  ) %>%
  tab_header(
    title = md("Table 2: Summary Statistics by Follow-up Presence")
  )
# Drop PAIN_INTENSITY_AVERAGE.FOLLOW_U and X101:X238
```

```r
pain_mod <- pain[, -c(1:75)] %>%
  select(-PAIN_INTENSITY_AVERAGE.FOLLOW_UP)

pain_mod$PAT_SEX <- as.factor(pain_mod$PAT_SEX)
pain_mod$PAT_RACE <- as.factor(pain_mod$PAT_RACE)
pain_mod$CCI_BIN <- as.factor(pain_mod$CCI_BIN)
pain_mod$MEDICAID_BIN <- as.factor(pain_mod$MEDICAID_BIN)
# pain_mice <- mice(pain_mod, m = 5, maxit = 5, seed = (2550))
# saveRDS(pain_mice, file = "mice_result.rds")
pain_mice <- readRDS("mice_result.rds")
score_list <- list()

for (i in 1:5) {
  mice_imp <- mice::complete(pain_mice, i)

  score <- mice_imp %>%
  group_by(PAIN_INTENSITY_AVERAGE) %>%
  summarize(
avg_mental_health = mean(GH_MENTAL_SCORE),
se_mental_health = sd(GH_MENTAL_SCORE)/sqrt(n()),
avg_physical_health = mean(GH_PHYSICAL_SCORE),
se_physical_health = sd(GH_PHYSICAL_SCORE)/sqrt(n())
  )

  score_list[[i]] <-  score
}

score1_avg <- score_list[[1]][, c("PAIN_INTENSITY_AVERAGE", "avg_mental_health",
                                  "avg_physical_health")]
colnames(score1_avg) <- c("Pain Intensity Level", "Average Mental Health Score", "Average
knitr::kable(score1_avg,
         caption = "Average Summary of Pain Intensity Score") %>%
  kable_styling(latex_options = "HOLD_position")
mice_pool <- do.call(rbind, score_list)

score_pool <- mice_pool %>%
  group_by(PAIN_INTENSITY_AVERAGE) %>%
  summarize(
avg_mental_health_pool = mean(avg_mental_health),

se_mental_health_pool = sqrt(mean(se_mental_health^2)
```

```r
+ (sum((avg_mental_health - mean(avg_mental_health))^2))/4
+ (sum((avg_mental_health - mean(avg_mental_health))^2))/20),

avg_physical_health_pool = mean(avg_physical_health),

se_physical_health_pool = sqrt(mean(se_mental_health^2)
+ (sum((avg_physical_health - mean(avg_physical_health))^2))/4
+ (sum((avg_physical_health - mean(avg_physical_health))^2))/20)
)

ggplot(score_pool, aes(x = factor(PAIN_INTENSITY_AVERAGE))) +
  geom_point(aes(y = avg_mental_health_pool, color = "Mental Health")) +
  geom_line(aes(y = avg_mental_health_pool, color = "Mental Health"), group = 1) +
  geom_errorbar(aes(ymin = avg_mental_health_pool - se_mental_health_pool,
                ymax = avg_mental_health_pool + se_mental_health_pool,
                color = "Mental Health",), alpha = 0.6) +
  geom_point(aes(y = avg_physical_health_pool, color = "Physical Health")) +
  geom_line(aes(y = avg_physical_health_pool, color = "Physical Health"), group = 1) +
  geom_errorbar(aes(ymin = avg_physical_health_pool - se_physical_health_pool,
                ymax = avg_physical_health_pool + se_physical_health_pool,
                color = "Physical Health"), alpha = 0.6) +
  labs(x = "Pain Intensity Level", y = "Average Score",
   title = "Figure 7: Pooled Average Mental and Physical Health Scores by Pain Intensity L
   color = "Score Type") +
  theme_minimal() +
  theme(legend.text = element_text(size = 6),
    legend.title = element_text(size = 8),
    axis.text = element_text(size = 6),
    plot.title = element_text(size = 8),
    axis.title = element_text(size = 8)
    )
```