Estimating Treatment Effects in Longitudinal Clinical Trials with Missing Data

A Preprint

Amy Browne
School of Mathematical and Statistical Sciences
University of Galway

a.browne47@universityofgalway.ie

Tsz Mang Yip
School of Mathematical and Statistical Sciences
University of Galway

k.yip2@universityofgalway.ie

June 24, 2025

Abstract

Motivation Missing data is a pervasive issue in longitudinal clinical trials, potentially introducing bias and reducing statistical power. This project tested different missing data handling techniques, using two real-world longitudinal randomized clinical trial (RCT) datasets. **Results** No substantial differences were observed between the methods across both datasets. **Supplementary Information** Code is available at `https://github.com/amydebrun/Missing-Data`.

Keywords Missing Data · Longitudinal Data · Randomized Clinical Trial · Real World Data Analysis · Sensitivity Analysis · Multiple Imputation · Missingness Mechanisms

1 Introduction

1.1 Missing data in longitudinal study

Longitudinal Study Design

Longitudinal studies are research designs in which data are collected from the same participants at multiple time points. They provide valuable insights into how variables change over time—often over several years or decades. While this design is powerful for understanding causal relationships and long-term trends, it presents significant challenges in retaining participants across all waves of data collection.

Longitudinal trials require consistent data collection methods at predefined intervals. Researchers must work to minimize participant dropout or non-response. However, due to the extended duration and complexity of these studies, missing data is almost inevitable.

Similarly, clinical trials which evaluate the effects of biomedical or behavioral interventions are also vulnerable to missing data. Participants may miss follow-ups due to illness or other commitments. This loss of data can compromise the study's validity. Ignoring missing data introduces bias, as we're disregarding potentially informative cases. For instance, if participants drop out due to adverse effects from the treatment, the missing data is not random—this introduces attrition bias, which can distort estimates of the treatment's effect.

Removing incomplete cases reduces the sample size, which in turn decreases statistical power, limiting the ability to detect valid treatment effects.

In statistical software like R, missing values are represented as NA among the observed values. In longitudinal data, rows of data belonging to individuals at a certain time-point can be missing. When fitting models like linear regression, R automatically excludes any subjects with missing values in either predictor or outcome variables. While convenient, this method can lead to selection bias if the missingness is systematic.

1.2 Project Objective

The objective of this project is to estimate treatment effects using real-world data, with a primary focus on addressing issues that come with missing data in clinical trials. We will briefly discuss different assumptions about missing data mechanisms—such as Missing Completely at Random, Missing at Random, and Missing Not at Random and investigate their implications for statistical analysis.

While a range of methods are available to handle missing data, they all vary in levels of complexity and assumptions. This project aims to explore these methods, understand the reasons behind their differing performance, and implement them on real-world longitudinal clinical data to compare their ability to estimate treatment effect under real-world conditions.

After initial exploration of the different methods, we also decided to adjust the estimands through data wrangling our chosen datasets to compare how the estimate differs when the estimand is either categorical or continuous and to compare the effects using linear regression or linear mixed effects models. Additionally, we will conduct sensitivity analyses to investigate how treatment effect estimates change under different assumptions about the missing data.

1.3 Missingness Mechanisms

Missing data mechanisms are important to consider when choosing which sort of missing data handling method to use. There are three mechanisms which missing data can follow:

- Missing Completely At Random (MCAR)
- Missing At Random (MAR)
- Missing Not At Random (MNAR)

Although they may appear similar at first glance, continuing to handle missing data without considering these mechanisms may still result in biased estimates and inaccurate conclusions. Missing data mechanisms appear to be mentioned first by Donald Rubin in 1976. We can formally define these mechanisms using the following notation;

- R represents a missing data indicator which when $R = 1$ indicates observed data and $R = 0$ indicates unobserved
- X represents data that is always observed
- Y represents data that is potentially missing.

Missing Completely At Random

The formal definition of MCAR data is:

$$P(R = 1 \mid Y, X) = P(R = 1)$$

The probability of the data is observed given observed data and missing data is the same as the probability of being observed without the given data. This mechanism is considered the easiest to deal with as it does not bias the result although data is rarely MCAR. This can occur due to system failure and some data is deleted accidentally, or else there is issues with the treatment system and data cannot be recorded. The DAG below visualises an example of this mechanism based on a randomised trial with randomisation variable L. Missing indicator R_Y, which denotes whether the outcome is observed ($R_Y = 1$) or missing ($R_Y = 0$), is independent from the observed baseline variables X_1 and X_2, which are pre-randomised variables, and outcome variable Y (which potentially contains missing data).

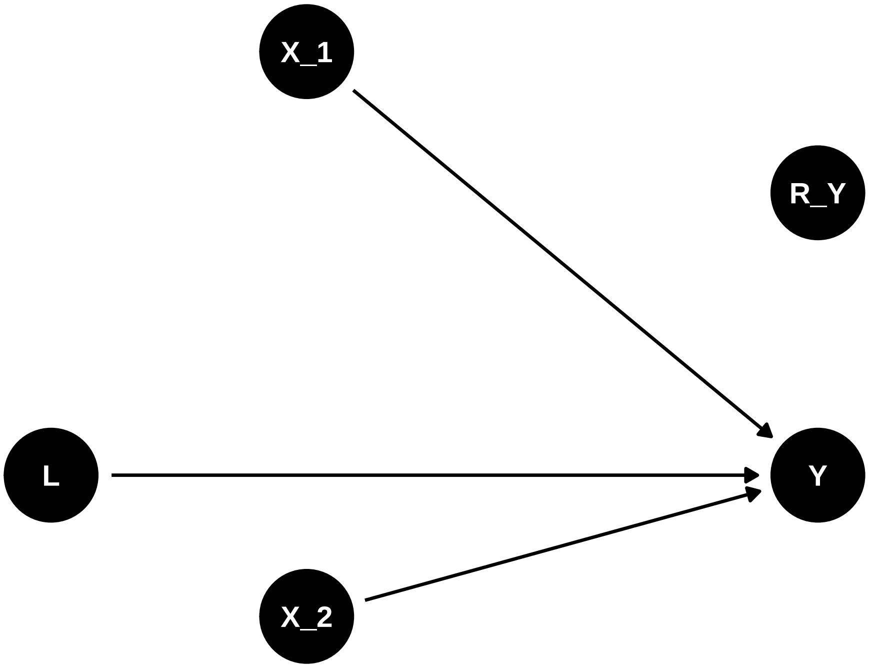


Figure 1: Missing Completely At Random

Missing At Random

$P(R = 1 \mid Y, X) = P(R = 1 \mid X)$

The probability of data being observed given observed and unobserved data is the same as the probability being observed given the observed data. In short, the missing data dependends on the observed data. We can say that the missing data indicator is conditional on another variable.

MAR is a more realistic mechanism than MCAR and requires more intensive handling methods. Below shows 2 DAGs that shows a general example of a randomised treatment L on response variable Y, which potentially contains missing data. X_1 and X_2 are baseline variables, recorded pre-randomisation. The first DAG shows the MAR mechanism conditioned on both baseline variables; missingness indicator R_Y is influenced by baseline variables X_1 and X_2. The second DAG is similar, but shows the data is MAR conditional on randomised treatment L and baseline variable X_1. X_2 no longer influences the missingness indicator R_Y, but R_Y is still conditioned on L and X_1, showing the MAR assumption still holds given these variables.

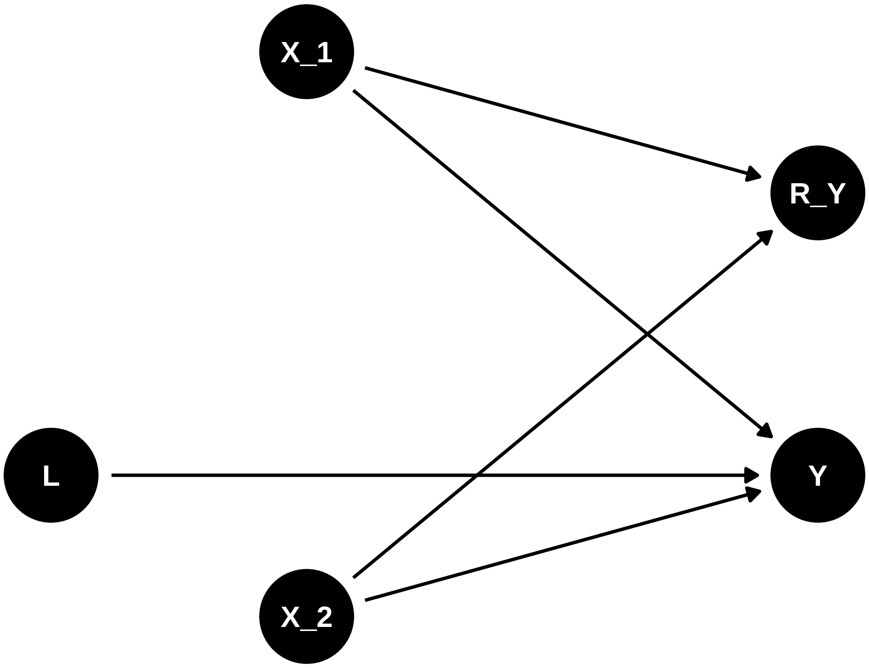


Figure 2: Missing at Random conditional on L, X_1, and X_2

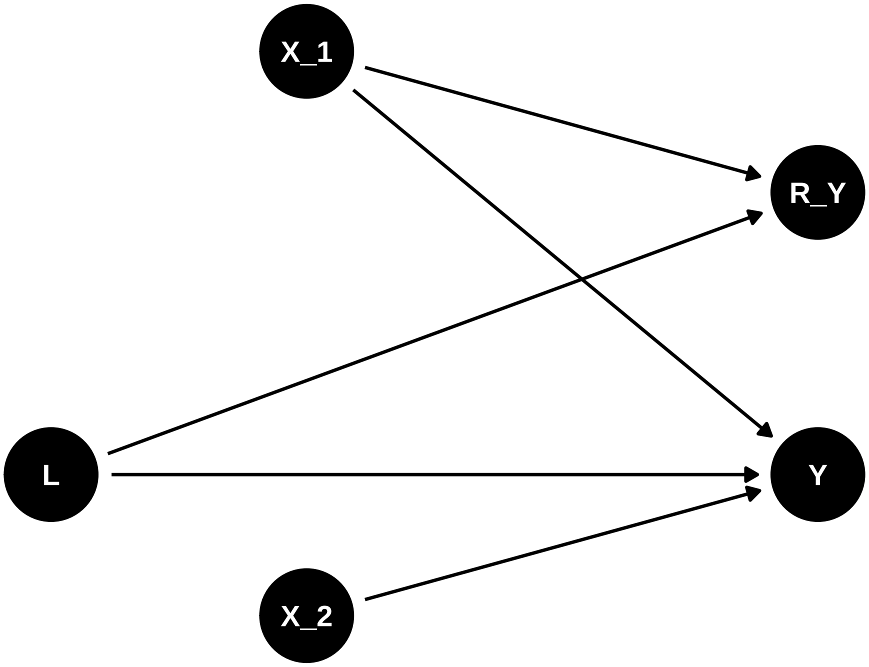


Figure 3: Missing at Random conditional on L and X_1

Missing Not At Random

$P(R = 1 \mid Y, X) \neq P(R = 1 \mid X)$

The probability of data being observed depends both on the observed data and the missing data. This mechanism is the most difficult to deal with as it relates to the unobserved data, so producing valid results is a challenge. Certain participants in a general health study may avoid answering questions truthfully about smoking habits or their diet in order to make themselves more appealing. Sensitivity analysis is an option to determine the treatment effect when assuming different mechanisms. The 2 following DAGs show this mechanism using the above scenarios. In the first DAGs structure, the potentially missing outcome variable Y directly influences R_Y. The outcome Y is directly influenced by the randomised treatment L and baseline variables X_1 and X_2, but the missingness indicator is influenced by the missing outcome data. The second DAG shows MNAR conditional on X_2 only as it influences the missingness indicator R_Y along with the missing outcome.

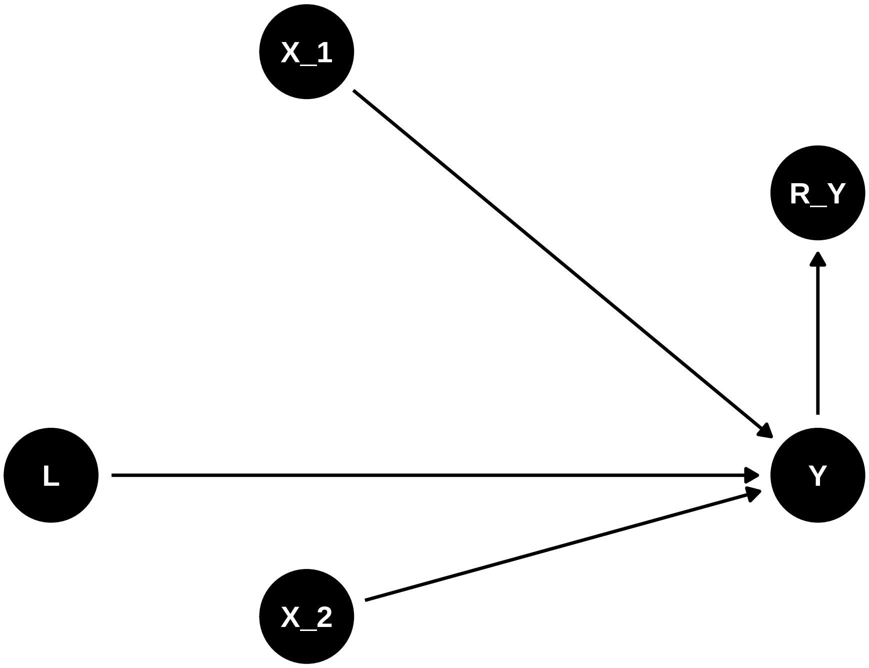


Figure 4: Missing Not at Random conditional on L, X_1 and X_2

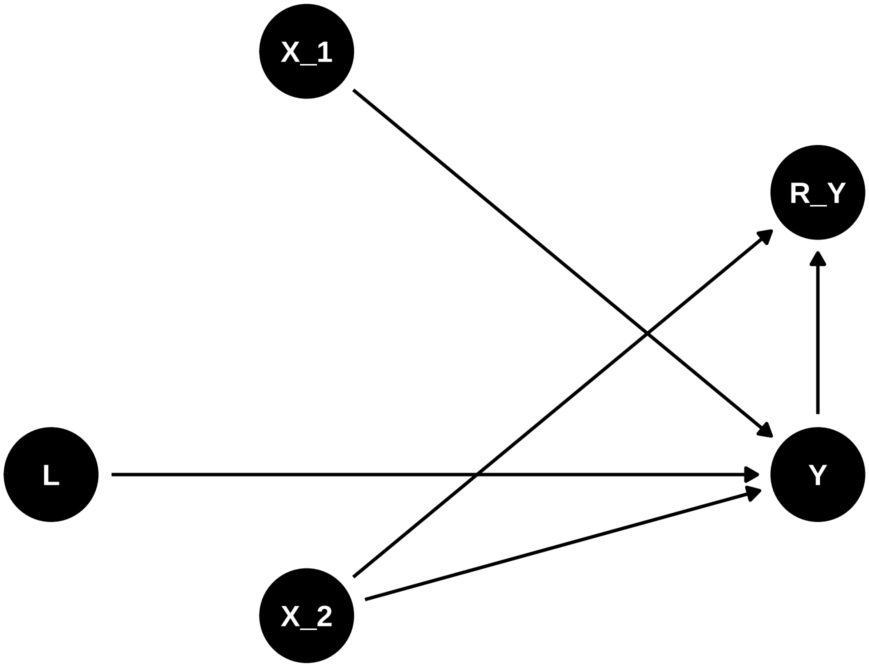


Figure 5: Missing not at Random conditional on L, X_2

Although these considerations are important, the mechanism underlying missing data is fundamentally untestable. Since the true values of missing observations are never observed, we cannot definitively determine which missingness assumption holds in practice. There is limited hypothesis test available to see whether data is MCAR (*Little's MCAR test*). It is a likelihood-ratio test that groups missing data patterns, estimates the expected means under a multivariate normal model using maximum likelihood, and computes a chi-square statistic to assess model fit. The decision lies under the hypotheses;

- H_0: Data is MCAR
- H_1: Data is MAR or MNAR

If the p-value resulting from the chi-square distribution is greater than 0.05, we fail to reject the null hypothesis. It is limited as it does not necessarily confirm MCAR but provides no evidence to reject that the data MCAR. If the test has a significant p-value, it suggests that the data is either MAR or MNAR, but its unclear which one. This test is also vulnerable to Type 1 error (false positive result) and Type 2 error (false negative result) as the test may not detect MCAR violations with its low statistical power, even if the MCAR assumption holds.

1.4 Different missing data patterns

A missing data pattern describes the pattern of missing data among the observed data. As described by Little & Rubin (2014), it's important to remark the missing data patterns of the data set prior to data handling as some handling methods are intended for certain classified patterns. The following table summarises some missing data patterns.

Table 1: Missing Data Pattern

Pattern	Description
Univariate	Missing values in a single variable
Multivariate	Missing values present in multiple variables
Monotonic	Variables Y_j can be ordered such that if Y_j is missing, all subsequent variables are also missing
General	Missing values have no structure and are scattered throughout data

1.5 Literature

While reviewing the literature on missing data, it became evident that despite the use of established handling methods, there remains room for improvement in how missing data are addressed in clinical study reports and in the selection of appropriate methods.

In *2014, Powney et al.* reviewed 100 longitudinal clinical trials conducted between 2005 and 2012 and found that only 44 reported an adequate method to handle missing data, while 30 used potentially inadequate methods.

Spineli et al. (2015) examined 190 systematic reviews published after 2009 to investigate how missing data were reported in randomized controlled trials. Although 175 of these reviews mentioned missing data, only 61 discussed its implications.

Hunt et al. (2021) reviewed 62 pharmacoepidemiologic multi-database studies from 2018 to 2019 to assess missing data reporting and handling. Thirty-five studies reported missing data, but only 19 described methods for handling it. The most popular approach was complete case analysis (CCA) used in 13 studies, followed by multiple imputation (MI) in 2 studies.

Complete case analysis was also the predominant method in a systematic review of 229 observational studies from the United Network for Organ Sharing (UNOS) database *(Baker et al., 2025)*. Forty-one studies used CCA, 22 reported MI, and notably, 31 studies removed covariates due to missing values—an action that could introduce further bias depending on the covariates' relationship with other predictors.

2 Clinical Trial Data

We have sourced two different data sets to perform our missingness analysis on. They both have different sample sizes and different levels of missing data which is an advantage as we can conduct analysis in different settings.

2.1 Acupuncture Data

Our first dataset contains results from a clinical trial determining the effects of acupuncture therapy on chronic headache in primary care *(Vickers et al 2004)*. Vickers released the dataset from the study in 2006 and is publicly available to download in excel format from `https://pmc.ncbi.nlm.nih.gov/articles/PMC1489946/`. The study design for the acupuncture trial is a longitudinal randomised controlled trial with two follow up time points. 401 participants were gathered from general practices in England and Wales who suffered from migraines. The main objective was to determine the effects of acupuncture use on headache severity, health status, absence days and GP and therapist visits in comparison to avoiding acupuncture against a control intervention whih offered standard care. Headache and medication use was recorded at baseline and additional factors were recorded post-randomisation such as such sick days and daily activities. The results of the trial showed a 34% decrease in headache severity score at the final time point in comparison to the control group with a 16% decrease. The acupuncture therapy group also showed a 37% decrease in medication use, which compares to 23% in the control group.

2.1.1 Data Overview

13 variables were recorded during the acupuncture trial.

Table 2: Acupuncture trial variables

Variable	Description
id	patient ID code
age	Age
sex	sex; female (1) vs. male (0)
migraine	diagnosis ; migraine (1) vs. tension-type (0)
chronicity	number of years of headache disorder
acupuncturist	acupuncturist id code
practiceid	gp practice id
group	treatment group; acupuncture (1) vs. control (0)
pk1	headache severity score baseline
pk2	headache severity score 3 month
pk5	headache severity score 1 year
f1	headache frequency baseline
f2	headache frequency 3 month
f5	headache frequency 1 year

Variables `pk2` and `pk5` are the headache severity scores at 3 months and 12 months.

Table 3 are the summary statistics of the acupuncture trial data. These consist of 2 columns; column 0 showing control group statistics and column 1 showing treatment group statistics. The continous variables show the mean value (standard deviation) and the categorical variables show the count(percentage). Note pk2 and pk5 are post randomisation variables and contain missing data.

Table 3: Acupuncture data overview

Characteristic	**0** N = 196	**1** N = 205
id	470 (209)	472 (206)
age	45 (11)	46 (11)
sex	165(84%)	172(84%)
migraine	183(93%)	194(95%)
chronicity	22 (13)	21 (14)
acupuncturist	5.97 (2.82)	6.00 (2.80)
practice_id	24 (11)	25 (12)
pk1	27 (17)	26 (15)
pk2	24 (18)	19 (16)
Missing	43	32
pk5	22 (17)	16 (14)
Missing	56	44
f1	16 (7)	16 (7)
f2	11 (9)	11 (8)
f5	10 (9)	9 (8)

[1] Mean (SD); n(%)

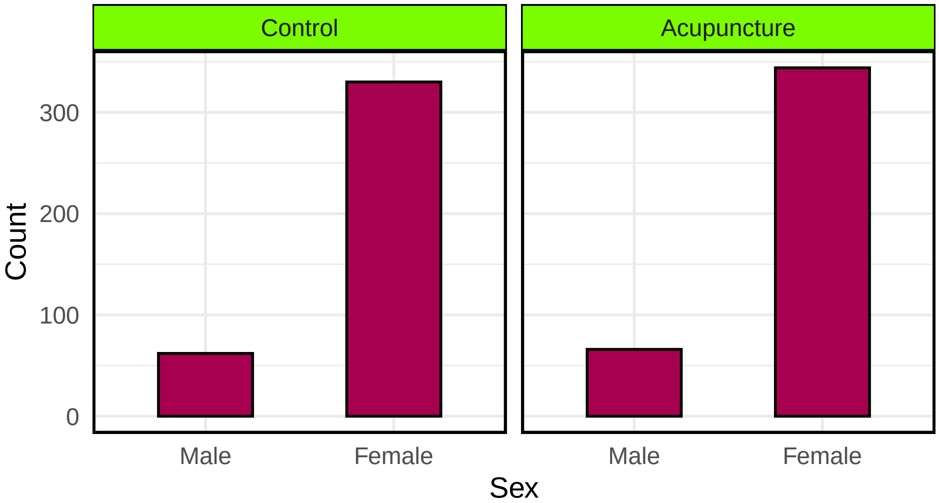


Figure 6: The distribution of sex in both treatment groups of the acupuncture trial data. The participants are predominantly female.

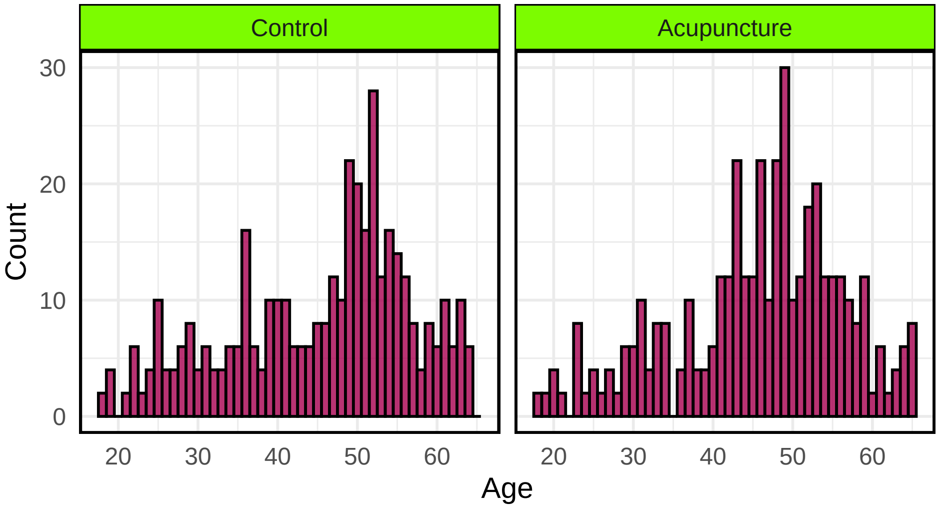


Figure 7: The distribution of age in both treatment groups. Both distributions are negatively skewed, predominantly middle aged and older.

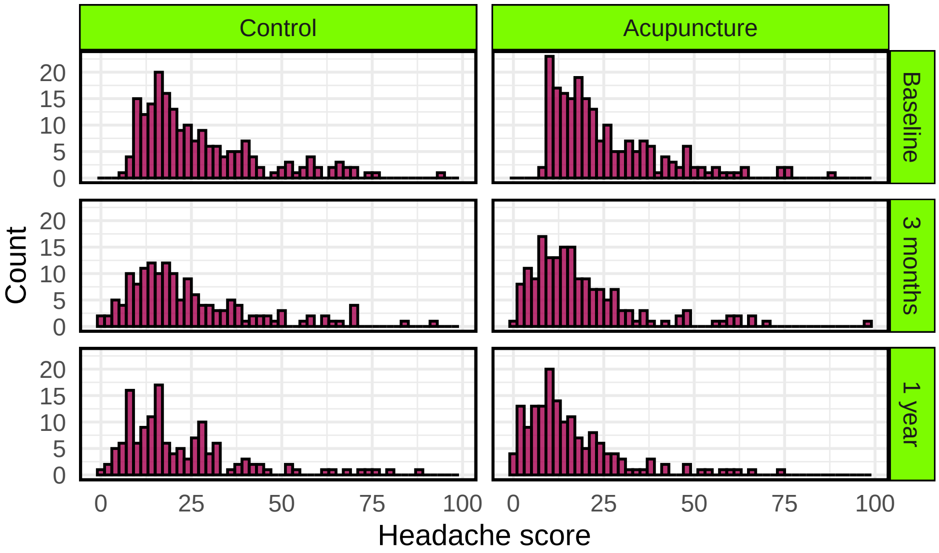


Figure 8: These plots show the distribution of the headache severity score at the three measured time-points. Over the three time-points we can see a generally positively skewed distribution with most headache severity scores ranging between 0 and 35. There appears to be minimal difference in headache severity score over the treatment group, and the control group appears to have a lower headache severity score overall. The 3 month and 12 month time-point distributions are incomplete due to the missing values.

2.2 VITAL data

The Vitamin D and Omega-3 trial (VITAL) was a nationwide clinical trial sponsored by Brigham and Women's Hospital in Boston to determine the effects of daily doses of 2000 IU vitamin D and 1 gram of omega-3 fatty acids (fishoil) on preventing cancer and cardiovascular disease (CVD). This was a 2x2 factorial designed longitudinal randomised controlled trial in which 25,871 adults were randomised to 4 possible combinations of treatment; fish oil, vitamin D, both treatments, an a control group with no treatment. The results of the vitamin D group said that it did not lower cancer development significantly neither did it cause any meaningful reuction in cardiovascular disease. As for the fish-oil intervention lowered risk of heart attack by 28% and risk of heart attack fatality by 50%, but no reduction in stroke or non-heart disease cardiovascular fatalities. Fish oil did not reduce occurence in breast or prostate cancer or cancer-deaths. For our research, we are analysing the results of an ancillary study which uses data from VITAL. MacFarlane et al (2020) conducted a study using VITAL on the therapeutic effects of vitamin D and fishoil on osteoarthritic knee pain. 1,398 participants in the study was randomised to 4 possible treatment combinations. Participants were randomised to receive either omega-3 oils or vitamin D, both treatments, or randomised to the placebo group. A knee pain scale grading index Western Ontario and McMaster Osteoarthritis Index (WOMAC) is used to record knee pain score at baseline and annually for 4 time-points post randomisation. Results showed that there was no significant change in knee pain score between both treatment groups and the control group. For this research project, we assumed that that the participants were only randomised to either receive the fish oil treatment, vitamin D treatment or the placebo. We did not include the interaction between vitamin D and fish oil in our models. The below table shows the sample size per treatment group.

	Vitamin D	**Placebo**
Fish Oil	N = 342 *(not analysed)*	N = 353
Placebo	N = 332	N = 371

2.2.1 Data Overview

There are x variables recorded in the VITAL data, the following table refers to a subset of the variables that we frequently use.

Table 5: VITAL study variables

Variable	Description
`Subject_ID`	Patient ID code
age	Age of patient
bmi	Body mass index of patient
sex	Sex of patient
vitdactive	1=vitamin D, 0=no vitamin D
fishoilactive	1= fish oil, 0= no fish oil
`pain_base`	Knee pain at baseline
`pain_yrX`	Knee pain X years post randomisation
`stiffness_base`	Knee stiffness at baseline
`stiffness_yrX`	Knee stiffness X years post randomisation
`function_base`	Knee function at baseline
`function_yrX`	Knee function X years post randomisation
kneepainfreq	Frequency of knee pain

Table 6 shows the summary statistics of partial VITAL data. It provides the mean value (standard deviation) of the continuous variables and the count(%) for the categorical variables. This is divided per treatment combination group. Note here that the VITAL data contains missing baseline (i.e. `pain_base`) data as well as post-randomisation data.

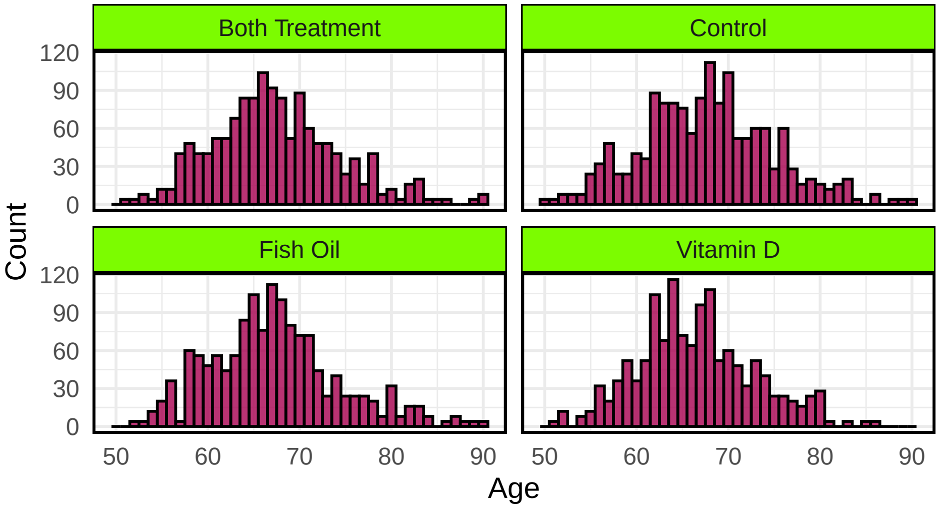


Figure 9: The distribution of age in each treatment group, which appear to be normal/ positive skewed distributions.

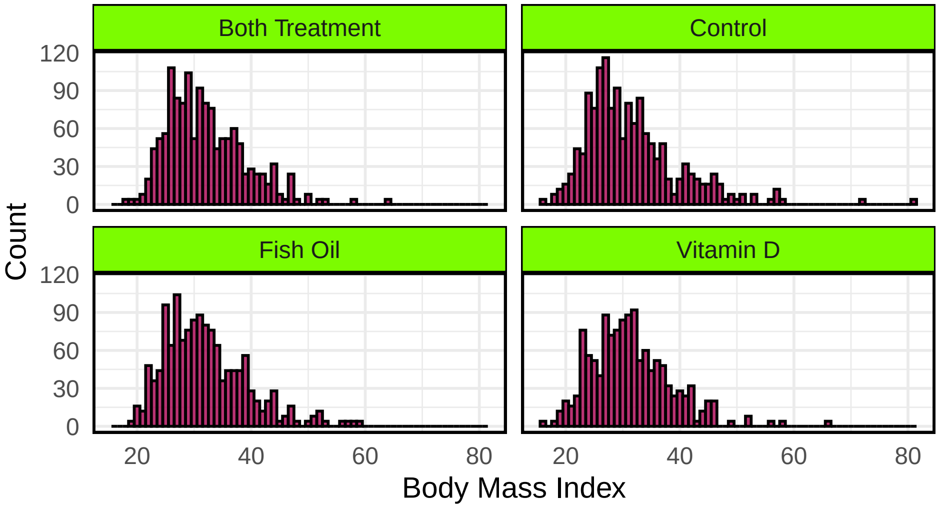


Figure 10: The distribution of body mass index in each treatment group, these appear to be positively skewed with a BMI between 20 and 40. This shows a range between healthy weighted and obese participants.

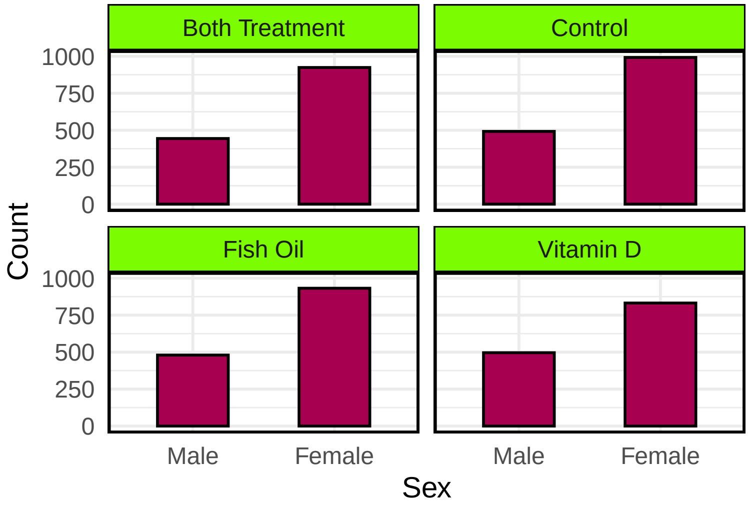


Figure 11: The distribution of males and females in the VITAL data. Similar to the acupuncture trial, the participants are predominantly female.

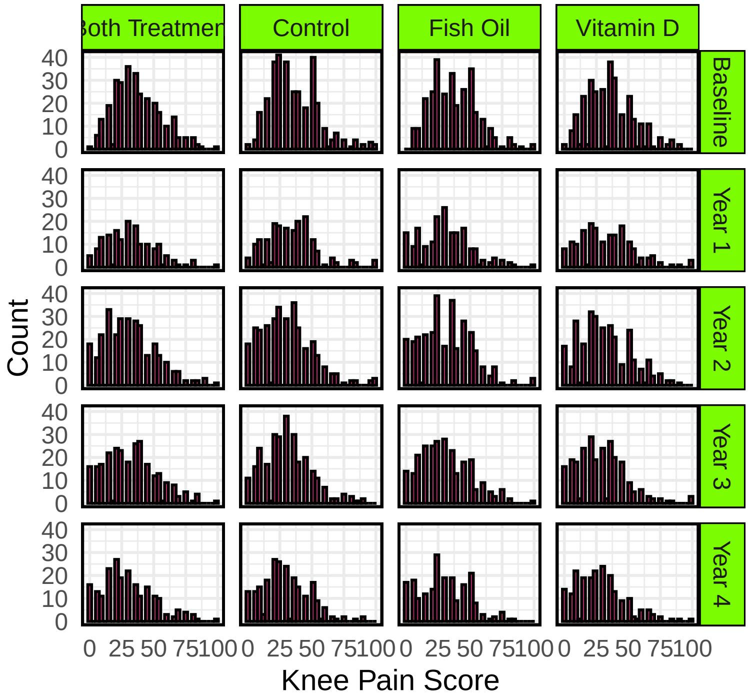


Figure 12: The distribution of the knee pain from baseline until 4 years post randomisation. In all treatment groups it appears that the distributions follow a positive/ near normal distribution and skews more positive as time goes on. This occurs in all groups. Each time-point has an incomplete distribution due to missing data.

2.3 Missing Data Summary

Acupuncture Data

The summary statistics of the acupuncture trial data shows that all baseline variables have been observed. It appears two post-randomisation outcome variables contain missing data. This pattern would be described as a multivariate monotonic pattern as the percentage of missing data increases at each follow up time point, which is common in longitudinal studies. We can visualise this in the plot below.

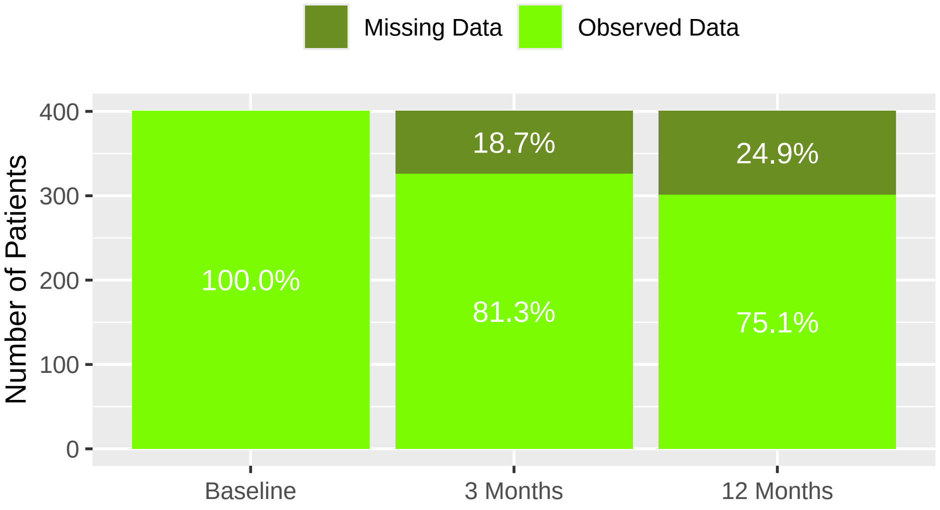


Figure 13: Proportion of missing data in outcome variables *pk*2 and *pk*5. This follows a monotonic missing data pattern.

If we separate the acupuncture trial data into the treatment group and control group, we can observe that there is a higher percentage of missing data in the control group.

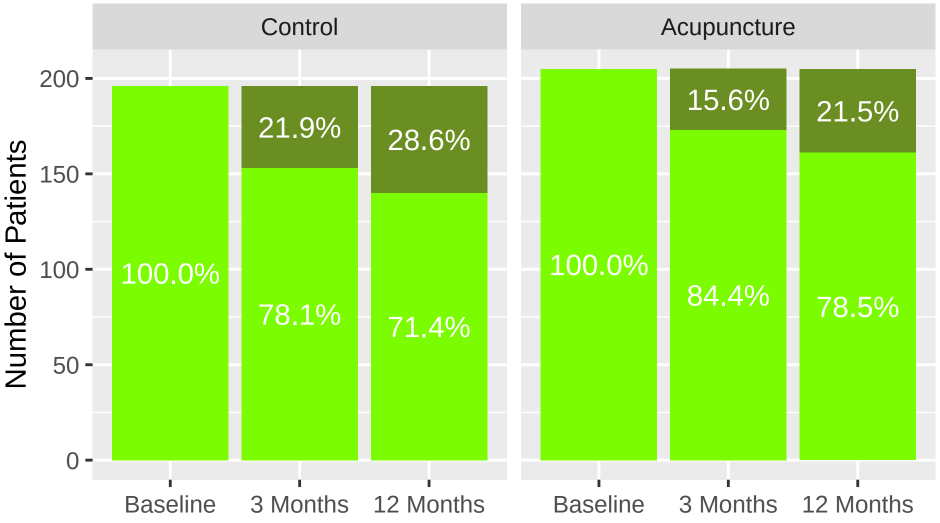


Figure 14: Higher percentage of missing data (dark green) in control group

If we investigate the missing data pattern in more detail, we can see that the pattern is similar in both groups. In the plot below, we can see that the pattern in which the three headache severity scores $pk1$, $pk2$ and $pk5$ are observed (1,1,1) followed by the pattern in which both post-randomised headache severity scores contain missing data (1,0,0).

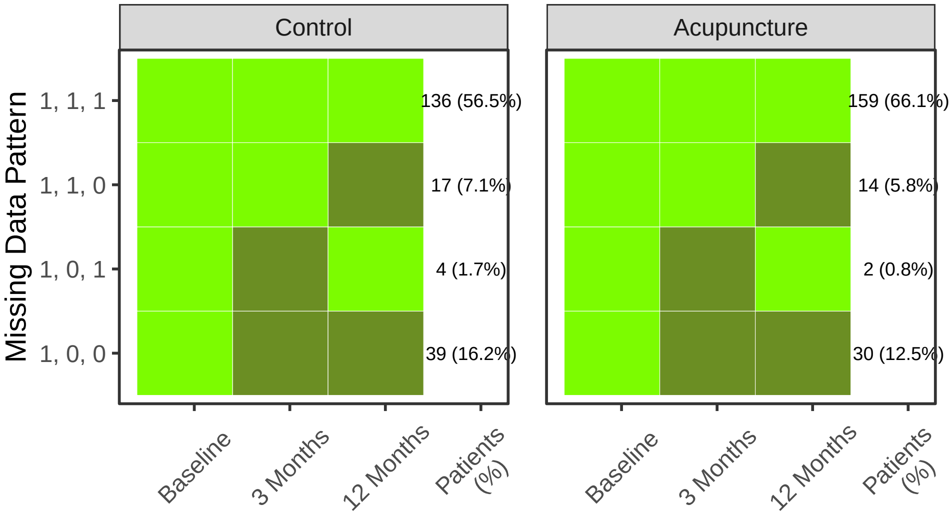


Figure 15: Missing data pattern is similar in both groups (1=observed, 0=missing)

VITAL data

The summary statistics of the VITAL data showed, unlike the acupuncture trial data, there are also baseline variables that contain missing data as well as the post randomised variables. A non-monotonic multivariate pattern appears to show here as there is a great proportion of missing data in the knee pain score at the first year post randomisation, which drastically decreases in year 2, then steadily increases again.

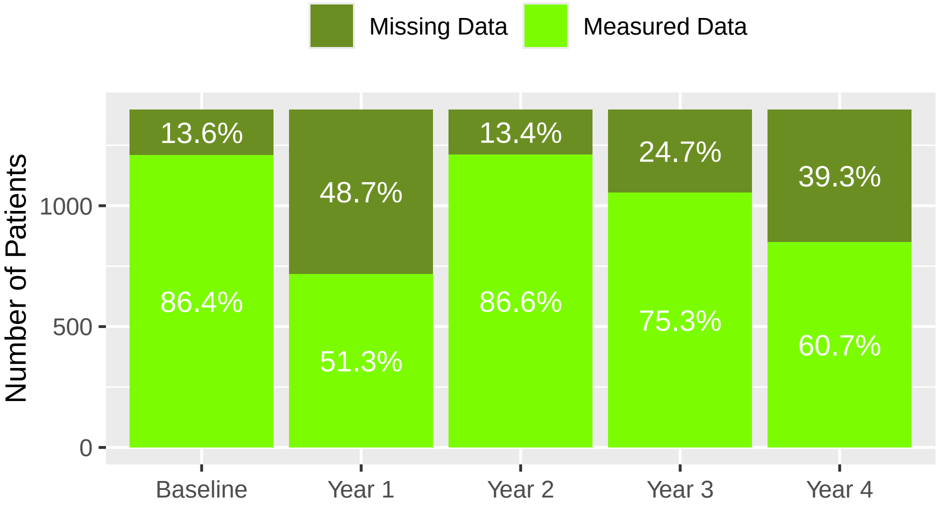


Figure 16: Non-monotonic pattern across out knee pain scores in VITAL data.

Recall that the VITAL data set contains 4 different combinations of treatment. Separating this out in the plot below we can see the percentage of missing data in each knee pain time point variable in each group. The proportion appears to be similar in each group with year 2 post randomisation containing the most missing data.

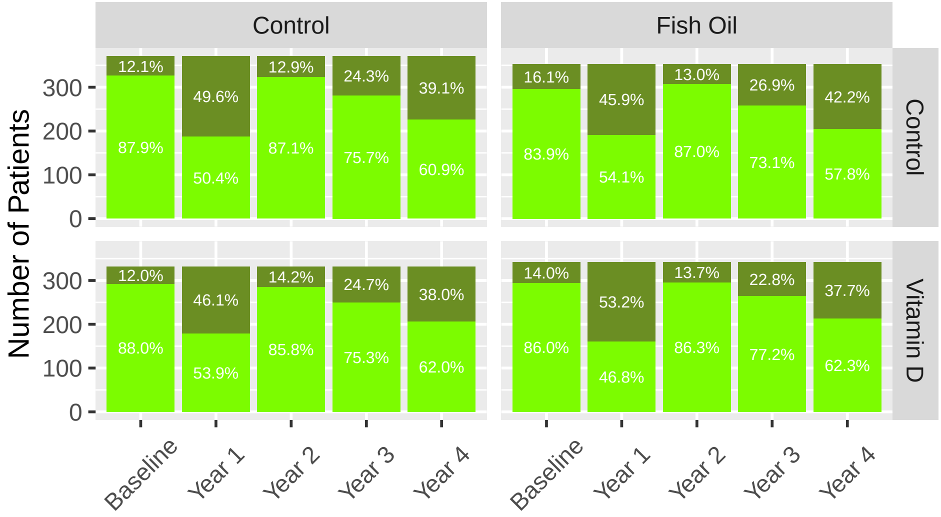


Figure 17: Percentage of missing data across baseline and 4 follow-up time points. Non-monotonic pattern appears similar in 4 randomised groups.

Figure 18 shows the possible missing data patterns in the 4 combinatory groups, they are mostly similar with little anomalies. As this dataset has more time-points recorded than the acupuncture trial data, and more treatment groups, there are more possible missing data patterns to observe. The most common data pattern still stands as (1,1,1,1,1) in which data is observed at all time-points (shown in light green blocks in the first row of each group).

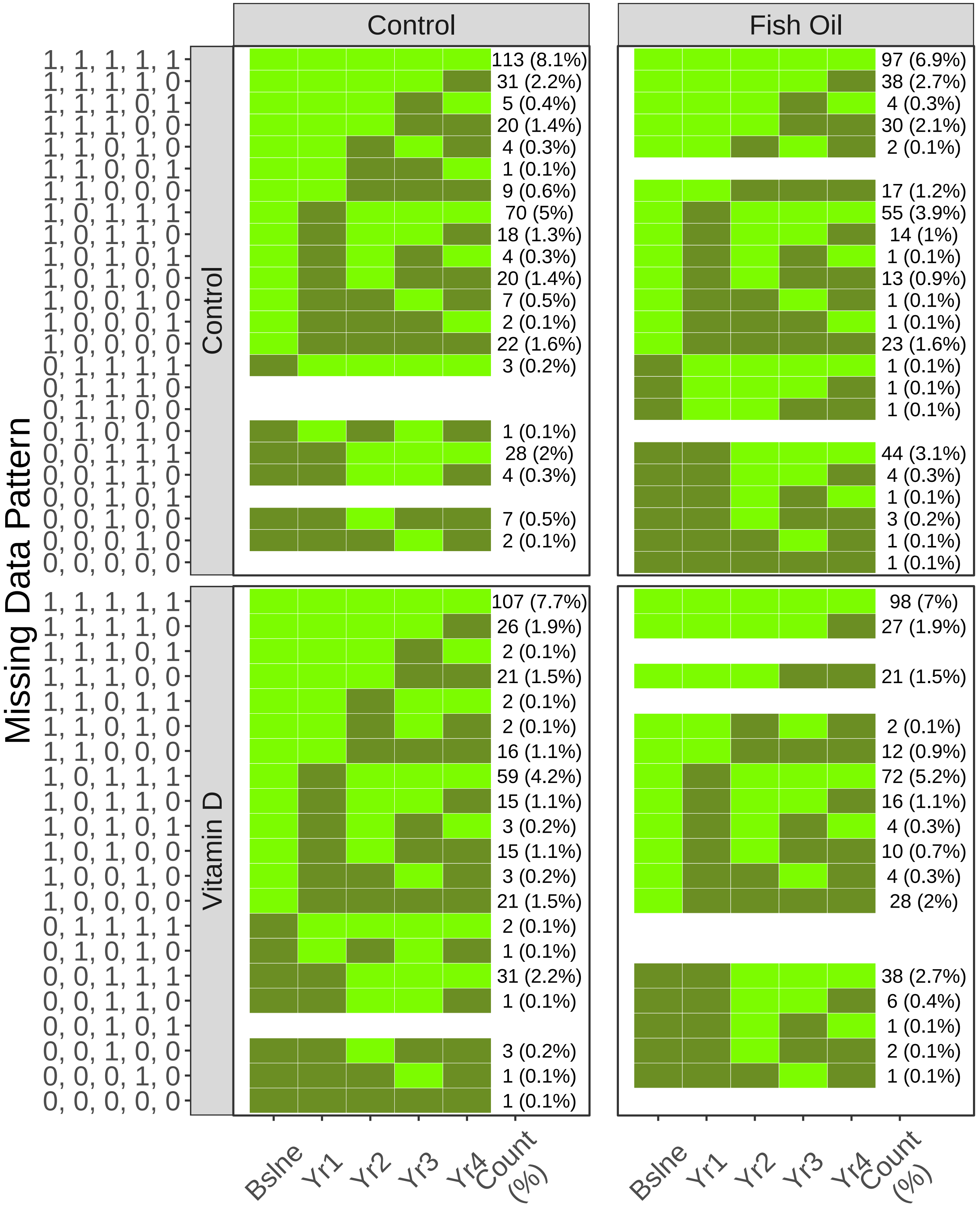


Figure 18: Missing data patterns across 4 treatment groups in VITAL data.

3 Methods

3.1 Anchoring methods

We began our analysis with basic methods to serve as a comparison benchmark for more advanced techniques. As previously mentioned, these included complete case analysis, LOCF, and the mean observation method. Each of these approaches was used to anchor the results and provide a reference point for evaluating the performance of more robust models.

3.1.1 Complete case analysis

For both datasets, given that our estimand is the mean difference in pain scores at the end of the study, we fit a model regressing the final pain score on treatment group and baseline pain score. CCA can yield valid inferences when the data are MCAR, or when the outcome is MAR conditional on the variables included in the model. In our analysis, we adjust for treatment group and baseline pain score. To ensure comparability across methods and to reduce computational burden, we did not include additional covariates in any of the analyses.

We applied the following models for each dataset:

```r
# For acupuncture data set
acu_CAA <- lm(pk5 ~ group + pk1 , data = acu_wide)
# For VITAL data set
vital_complete <- lm(pain_yr4 ~ fishoilactive + vitdactive + pain_base, data = vital_wide)
```

One of the main drawbacks of using CCA is the loss of power due to reduced sample size, along with the risk of introducing bias. In our implementation, the lm() function in R automatically excludes any individual with missing values in the model variables. As a result, we are left with only 301 out of 401 observations in the acupuncture dataset and 698 out of 1390 in the VITAL dataset—corresponding to a loss of approximately 25% and 50% of the sample size, respectively.

We have not included many covariates in our models, so bias could exist even with complete data due to omitted variables. Now, suppose we attempted to adjust for all relevant baseline covariates—missing data would still pose a problem. For instance, in the VITAL dataset, several baseline covariates have missing values. Excluding observations with missing data in these variables could still induce bias, particularly if the missingness is related to the outcome.

In the acupuncture dataset, although there is no missingness in baseline variables, CCA has further limitations. While intermediate outcomes such as the 3-month pain score or post-randomisation variables like pain frequency may contain valuable information, they are not included in our models. This is not due to a limitation of CCA or LME itself, but because including post-randomisation variables would violate the goal of estimating the total effect of randomisation. However, these variables can still contribute indirectly through MI. Rather than adjusting for them in the model, MI treats repeated and related measures as part of a multivariate distribution within each individual, allowing partially observed variables to help impute missing values without compromising the target estimand.

In rare cases—such as when the sole interest is in estimating the mean difference between groups at the end of the study, and there are no important non-baseline covariates complete case analysis can still yield valid results. Thus when apply complete case analysis, we are assuming: the outcome is MAR, conditional on the observed baseline predictors included in the model.

This condition sometimes referred to as “conditionally MCAR”. While it is weaker than the strict MCAR assumption, it is still quite strong. It implies that:

- Missingness depends only on the variables included in the model;
- The model is correctly specified;
- The outcome variable is not involved in the missingness mechanism—consistent with the MAR framework.

Under these assumptions, CCA can provide unbiased estimates. However, they are rarely fully met in practice, especially in complex longitudinal studies where many relevant variables and time points are involved.

This can be expressed in the mathematics form, assuming outcome of interest Y MAR depends on baseline covaraites X

$$Pr(R_i = 1 \mid Y_i, X_i) = Pr(R_i = 1 \mid X_i)$$

Thus, we can infer under MAR, the distribution of outcome within covaraites is the same in the observed data, the unobserved data, and the population.

$$\begin{aligned}
&Pr(Y_i \mid R_i = 1, X_i) \\
&= \frac{Pr(R_i = r, Y_i, X_i)}{Pr(R_i = 1, X_i)} \\
&= \frac{Pr(R_i = 1 \mid Y_i, X_i) Pr(Y_i, X_i)}{Pr(R_i = 1 \mid X_i) Pr(X_i)} \\
&Assuming \quad \Pr(R_i = 1 \mid Y_i, X_i) = Pr(R_i = 1 \mid X_i) \\
&= \frac{Pr(Y_i, X_i)}{Pr(X_i)} \\
&= Pr(Y_i \mid X_i)
\end{aligned}$$

The key issue is that we aim to include as much relevant information as possible in the set of predictors X to make the MAR assumption more plausible and robust. As discussed above, more statistically principled methods—such as multiple imputation or mixed-effects models—allow us to incorporate a broader set of variables and handle missingness more effectively. However, it is important to note that the MAR assumption is fundamentally untestable. Therefore, to assess the robustness of our conclusions, we must perform sensitivity analyses, which are presented in later sections.

3.1.2 Last observation carry forward

LOCF is a simple imputation method that fills in missing values based on the last observed outcome for each subject. It relies on assumptions unrelated to the missing data mechanism, meaning it does not model the reason for missingness. Instead, LOCF assumes that the outcome remains stable after dropout, which may be plausible in some clinical contexts, such as when the treatment effect has plateaued or when patients are expected to remain in a steady state.

However, this assumption is not convincing in either of the datasets analyzed here. In both the Acupuncture and VITAL trials, pain score are unlikely to be stabilized throughout, and the outcomes display trends over time, making the LOCF assumption potentially misleading and unsuitable for accurate estimation of treatment effects.

3.1.3 Mean observation method

Mean imputation is another simple method that fills in missing values by replacing them with the mean of the observed data at each timepoint. Like LOCF, it is based on assumptions unrelated to the missing data mechanism and does not attempt to model why the data are missing.

This method implicitly assumes that the mean of the observed values accurately represents the population mean, which can be problematic. In fact, this can be viewed as an even more radical assumption than assuming data are MCAR, since it ignores potential bias introduced by the missingness and oversimplifies individual variation.

As with LOCF, this assumption is not convincing in either of the data sets analyzed here. In both the Acupuncture and VITAL trials, pain scores vary over time, and individual trajectories differ. Replacing missing values with the average observed value at each time point disregards this variation and may lead to underestimated variability and biased treatment effect estimates.

3.2 Changing imputation methods

To address missing data in a statistically principled manner, it is helpful to distinguish between two components: the imputation step and the modeling step. In our anchoring methods, we either did not perform imputation—as in CCA—or relied on single imputation techniques, such as LOCF or mean imputation.

While methods like LME can utilize most available observed data and yield unbiased estimates under the MAR assumption without imputation, they cannot incorporate other post-randomization variables that are not part of the model (e.g., pain frequency in the acupuncture data set). In contrast, MI allows the inclusion of such auxiliary information during the imputation process.

In most realistic scenarios, doing no imputation or applying single imputation methods is likely to introduce bias and underestimate uncertainty, especially when missingness is related to unobserved outcomes. Thus, more flexible approaches—such as MI combined with appropriate modeling—are generally preferred for robust inference in the presence of missing data.

3.2.1 Multiple imputation

A key limitation of single imputation methods is that they treat imputed values as if they were observed data when fitting the substantive model. This approach fails to reflect the inherent uncertainty associated with missing values—it assumes we have perfectly recovered the unobserved data, which is never the case in practice.

In reality, the best we can do is to estimate the distribution of the missing data conditional on the observed data, under specific assumptions about the missingness mechanism (e.g., MAR). Importantly, this distribution is not adequately captured by a single draw, as is done in single imputation. Instead, to account for uncertainty, we must generate multiple draws from this distribution, resulting in multiple imputed datasets. This is the foundation of Multiple Imputation (MI).

As we will demonstrate in the following sections, the draws used in MI must be (at least approximately) Bayesian for Rubin's variance formula to yield valid inference. By fitting the substantive model separately to each of the K completed datasets, we obtain multiple estimates that, when pooled, incorporate both within-imputation variability and between-imputation variability. This approach not only addresses bias caused by missing data but also appropriately inflates standard errors to reflect the uncertainty introduced by imputation. Rubin's rules provide a general framework for combining these estimates to obtain point estimates, variances, confidence intervals, and statistical tests.

We also briefly introduce the different methods available for performing multiple imputation (Carpenter et al. 2023). In cases with a monotonic missing data pattern, sequential regression offers a straightforward solution. For arbitrary or multivariate missingness, more flexible approaches like joint modeling or Fully Conditional Specification (FCS)—also known as multiple imputation by chained equations (MICE)—are required.

In our project, we chose the FCS approach, given its flexibility and ease of use when dealing with datasets like Acupuncture and VITAL that have complex missingness patterns across multiple variables. A brief comparison of joint modeling versus FCS is included below to justify this decision.

While we used five imputations (K = 5) for our MI procedure in this analysis, we also discuss the considerations involved in choosing the number of imputations. These include factors such as the proportion of missing data, computational cost, and the desired precision of standard errors and confidence intervals.

Rubin's rules Once multiple imputed datasets are generated, Rubin's rules (*(Rubin1987) Multiple Imputation for Nonresponse in Surveys* 1987) are used to combine the parameter estimates and associated uncertainty across the K imputed datasets. The steps are as follows:

1. Impute K complete datasets, each containing different plausible values for the missing data. (More details in later section, we will focus on the following steps for now)

2. Fit the substantive model (e.g., a linear model or mixed-effects model) to each imputed dataset, obtaining Parameter estimate $\hat{\beta}_k$ and Variance estimate $\hat{\sigma}_k^2$ for each k=1,2,. . . ,K

3. Compute the pooled estimate $\hat{\beta}_{MI}$ and its total variance $\hat{V}_{MI}$:

$$\hat{\beta}_{MI} = \frac{1}{K}\sum_{k=1}^{K}\hat{\beta}_k$$
$$\hat{V}_{MI} = \hat{W} + (1+\frac{1}{K})\hat{B}$$
$$\hat{W} = \frac{1}{K}\sum_{k=1}^{K}\hat{\sigma}_k^2$$
$$\hat{B} = \frac{1}{K-1}\sum_{k=1}^{K}(\hat{\beta}_k - \hat{\beta_{MI}})^2$$

4. To test a null hypothesis $\hat{\beta}_{MI} = \beta^0$ use a t-statistic with ν degrees of freedom. This allows us to construct confidence intervals and perform hypothesis testing, reflecting the additional uncertainty due to imputation.

$$T = \frac{\hat{\beta}_{MI} - \beta^0}{\sqrt{\hat{V}_{MI}}}$$
$$\nu = (K-1)[1+\frac{\hat{W}}{(1+1/K)\hat{B}}]^2$$

Sequential regression MI In many longitudinal studies, the missing data pattern is approximately monotonic, particularly when dropout is due to participant withdrawal, a common situation in clinical research. In such cases, later measurements are often missing while earlier ones are observed, which permits the use of sequential regression for imputation under the assumption of MAR.

To justify this, consider the joint distribution of the outcome vector for individual i can be factorized as:

$$f(Y_{i,1}, Y_{i,2}, ..., Y_{i,p}) = f(Y_{i,p} \mid Y_{i,1}, ..., Y_{i,p-1}) * f(Y_{i,p-1} \mid Y_{i,1}, ..., Y_{i,p-2}) * ... * f(Y_{i,2} \mid Y_{i,1}) * f(Y_{i,1})$$

Under a monotonic missingness pattern, for each missing value $Y_{i,j}$ all preceding values $Y_{i,1}, ..., Y_{i,j-1}$ are observed. If we also assume MAR, then each of these conditional distributions can be estimated directly from the observed data. This provides a principled basis for sequential regression imputation, where each variable is regressed on the variables preceding it in order, and missing values are imputed based on those conditional models.

Suppose we have $i = 1, ..., n$ individual with $j = 1, ..., p$ variables. When the missing data pattern is monotonic, sequential regression provides an efficient and valid approach to imputation under the MAR assumption. The procedure follows these steps:

1. specify the model

$$Y_{i,j} = (1, Y_{i,1}, ..., Y_{i,j-1})^T * \beta_j + e_{i,j},\ e_{i,j}^{i.i.d.} \sim N(0, \sigma_j^2)$$

2. Under the monotonic pattern, for every missing $Y_{i,j}$ we assume that $Y_{i,1}, ..., Y_{i,j-1}$ are fully observed. We fit the regression model using ordinary least squares (OLS) to obtain estimates $\hat{\beta}_j, \hat{\sigma}_j^2$
3. To incorporate uncertainty, we draw new parameters β_j^*, σ_j^{*2} from their posterior distributions:

$$\sigma_j^{*2} = \frac{\hat{\sigma}_j^2(n_j - j)}{z}$$
$$\beta^* \sim N(\hat{\beta}, \sigma_j^{*2} A_j)$$
$$A_j = (\sum_{i=1}^{n_j} x_{i,j} x_{i,j}^T)^{-1}$$

4. For individuals with missing $Y_{i,j}$ generate imputations from the model: $Y_{i,j} = (1, Y_{i,1}, ..., Y_{i,j-1})\beta^* + e^*_{i,j}$, $e^*_{i,j} \sim N(0, \sigma_j^{*2})$
5. Repeat for $Y_{i,j+1}$ until complete

This sequential regression method is applicable to datasets with a monotonic missingness pattern, such as the Acupuncture dataset, where individuals tend to drop out in a consistent, time-ordered manner. However, it is not suitable for datasets with non-monotonic missingness, such as VITAL, where some individuals may have missing values at intermediate time points but return for later follow-ups. In such cases, more flexible approaches like Joint Modeling or FCS are required to properly handle the complex, arbitrary missing data structure.

Joint modelling When the missing data pattern is non-monotonic, as in the VITAL dataset, sequential regression is no longer appropriate. Instead, we can apply joint modeling, which makes no assumption about the missingness pattern but assumes the missing data mechanism is ignorable (typically, Missing At Random).

Under joint modeling, we assume that the complete multivariate outcome vector follows a multivariate normal distribution:

$$\begin{aligned}
&Y \sim N(\beta,\ \Omega)\\
&Y = (Y_{i,1}, Y_{i,2}, ..., Y_{i,p})^T\\
&\beta = (\beta_{0,1}, \beta_{0,2}, ..., \beta_{0,p})^T\\
&\Omega \text{ is the covariance matrix}
\end{aligned}$$

To estimate the parameters β and Ω and impute missing values, Gibbs sampling is one of the approaches. This algorithm draws each parameter in turn, conditional on the current values of all other parameters and the data.

To get priors to start Gibbs sampling. We begin with initial estimates β^0 and Ω^0, computed from the observed data. For each variable with missing values, we also generate an initial imputation Y_M^0 by sampling from the observed values of that variable with replacement. This allows us to calculate initial statistics such as $\overline{Y}^0$ and and the sample covariance matrix S^0 (Note it also used as prior sample covariance matrix S^P in each iteration)

Then, for each iteration r the following steps are performed:

1. Draw the precision matrix (inverse covariance):$\Omega^{-1,r} \sim W(n + \nu, (S_p^{-1} + S^{r-1})^{-1})$
2. Draw the mean vector:$\beta^r \sim N(\bar{Y}^{r-1}, n^{-1}\Omega^r)$
3. Impute missing values: $Y_M^r \sim f(Y_M \mid \beta^r, \Omega^r, Y_O)$
4. Update the mean and covariance estimates: $\bar{Y}^r$ the mean of the combined imputed and observed data, and S^r the sum of squares and cross-products from the combined data

After a sufficient number of burn-in iterations, we repeat this process K times to generate K imputed datasets. These are then analyzed using the substantive model of interest, and Rubin's rules are applied to pool the results, yielding valid parameter estimates and standard errors that account for the uncertainty due to missing data.

Full conditional specification Fully Conditional Specification (FCS), also known as multiple imputation by chained equations (MICE), is an extension of sequential regression imputation that relaxes the requirement that all covariate values used in the regressions be fully observed.

Importantly, when the missingness pattern is monotonic, FCS becomes equivalent to the sequential regression method discussed earlier. However, its key advantage is that it remains valid under non-monotonic missingness, making it suitable for more general data structures like the VITAL dataset.

The term "full conditional specification" refers to the fact that each variable is imputed from its full conditional distribution, given all other variables. This allows for more flexible modeling of multivariate missingness.

The general procedure involves the following steps:

1. Reorder the variables so that the overall missingness pattern is as close to monotonic as possible. This can improve stability and convergence in the imputation process.
2. Initialize missing values by filling in initial guesses—often by drawing, with replacement, from the observed values of each variable.
3. For each variable Y_j
    - Regress the observed part of Y_j on all other variables (including those with imputed values).
    - Use the fitted model to impute the missing values in Y_j, treating the other variables as given.
4. Repeat step 3 for all variables with missing data to complete one cycle.
5. Perform multiple cycles until convergence, and then repeat the entire process K times to generate K imputed datasets.

These datasets can then be analyzed with the substantive model, and the results pooled using Rubin's rules. FCS is widely used in practice due to its flexibility and implementation in tools like the mice package in R.

FCS VS joint modelling Generally, under the assumption of a multivariate normal distribution, the joint distribution uniquely determines the full set of conditional distributions, and vice versa. This means that, in theory, joint modeling and fully conditional specification (FCS) are mathematically compatible representations of the same underlying structure—provided all models are correctly specified.

In practice, however, joint modeling using a Gibbs sampler is often considered a more efficient algorithm. It also has the advantage of allowing the inclusion of prior information, which can be particularly useful when data are sparse or when integrating external knowledge into the model. Additionally, joint modeling methods can incorporate ridge parameters to stabilize the estimation of the covariance matrix, which becomes important when the number of variables is large relative to the sample size.

However, these concerns do not apply in our project, as our datasets have moderate dimensionality and sufficient sample size. Therefore, we opted to use Fully Conditional Specification (FCS) for multiple imputation.

We adopted the FCS approach using the mice package in R to perform multiple imputation throughout this project. One advantage of FCS is its ease of implementation, as it does not require explicit specification of a joint model or prior distributions, and it typically converges in fewer iterations in practice. Moreover, the balanced longitudinal design of both datasets—where follow-up measurements occur at consistent time points across individuals—makes FCS especially suitable. This structure allows for sequential imputation of partially observed variables in a consistent order across patients. In contrast, if the data were collected at irregular or patient-specific time points, or under a more complex design such as a cluster randomised trial, the FCS approach would be more difficult to apply reliably.

Finally, within the MICE framework, different imputation methods can be specified for different variable types—such as predictive mean matching or sampling from observed values. We will explore the impact of these options later in the analysis.

How to choose imputation number K Another practical consideration in multiple imputation (MI) is the choice of the number of imputations, denoted by K. While early applications often used K=5, more recent work emphasizes that the optimal number depends on the degree of missing information in the data.

A key parameter in determining K is γ_0, which represents the fraction of missing information for the parameter of interest. Unfortunately, γ_0 is typically unknown in advance, including in our project.

To address this, (White, Royston, and Wood 2011) proposed a simple and conservative strategy: using the proportion of complete cases in the dataset as a proxy for $1 - \gamma_0$, thereby estimating γ_0 conservatively. This approach allows for an informed yet practical choice of K, especially when precise calculation of the missing information is not feasible.

In our analysis, we used K=5 imputations as a baseline, while recognizing that more imputations may be needed in cases with higher levels of missingness or if more precise estimates of standard errors are required. We will explore the implication of changing K in the following sessions.

Choose 3-5 imputations

The classic advice for multiple imputation is to use a low number of imputations, typically between 3 and 5, when the proportion of missing information is moderate. As discussed in (*(Rubin1987) Multiple Imputation* *for Nonresponse in Surveys* 1987), the argument for choosing a small K is based on the total variance estimate:$T_K = (1 + \frac{\gamma 0}{K})T_\infty$

where T_K is the variance with K imputations, T_∞ is the asymptotic variance as $K \to \infty$ and γ_0 is the fraction of missing information. Since γ_0 is typically unknown, this formula helps illustrate the trade-off between the number of imputations and efficiency.

There is often limited benefit in increasing K beyond 5. For instance, if $\gamma 0 = 30\%$, using K = 5 result in $T_m = 1.06T_\infty$, indicating only a 6% inflation in variance compared to the ideal case.

In this project, we chose to use K=5 imputations for multiple imputation, following the classical recommendation to use a low number of imputations when the proportion of missing information is moderate.

Chosse >20 imputations

While early guidelines recommended using a low number of imputations (typically K=3-5) for moderate missingness, more recent research argue that increasing K beyond this range can yield important gains in statistical efficiency.

- (Royston 2004) suggested that to constrain the coefficient of variation of $ln(t_\nu \sqrt{T})$ to below 0.05—effectively keeping the width of confidence intervals within about 10% uncertainty, a minimum of K>20 is required. Here, t_ν is the quantile from a t-distribution with ν degrees of freedom, reflecting the uncertainty from finite imputations, and T is the total variance combining within- and between-imputation components.
- (Graham, Olchowski, and Gilreath 2007) argued that to achieve statistical power within 1% of the theoretical maximum, researchers should use at least K=20.
- (Bodner 2008) examined how the number of imputations relates to the fraction of missing information γ_0, and its effect on p-values and confidence intervals. He recommended increasing $K = (3, 6, 12, 24, 59, 114, 258)$ for $\gamma 0 = (0.1, 0.3, 0.5, 0.7, 0.9)$ accordingly.

In some situations—such as when estimating variance components or when dealing with highly uncertain estimands—using a very high number of imputations (e.g.,K=200) may be warranted to approximate the full posterior distribution.

The main drawback of increasing K is that it leads to longer computational time. However, this is generally the only limitation, and it becomes manageable with modern computing resources. Moreover, starting with a high number of imputations gives greater flexibility: we can always test the stability or sensitivity of our results by re-analyzing a subset of the imputations (e.g., comparing the performance at K=5,10,20) without needing to re-run the entire imputation process.

Thus, while K=3-5 is often sufficient under moderate missingness and when the focus is on point estimates, using a larger K can improve robustness in more demanding settings, with minimal trade-offs beyond processing time.

$K \approx 100\lambda$

A widely cited rule of thumb proposed by (White, Royston, and Wood 2011) recommends choosing the number of imputations based on the fraction of incomplete cases in the dataset, denoted as λ. Specifically, they suggest setting:$K \approx 100\lambda$

This rule has become a de facto standard, particularly in medical research, due to its simplicity and strong theoretical support. The key idea is that the number of imputations should roughly match the percentage of individuals with any missing data.

- The Monte Carlo error of the pooled point estimate $\hat{\beta}$ is approximately 10% of its standard error.
- The Monte Carlo error of the test statistic $\hat{\beta}/SE_{\hat{\beta}}$ is roughly 0.1.
- The Monte Carlo error of a p-value is approximately 0.01 when the true p-value is 0.05.

These error bounds are typically acceptable in applied research, ensuring stable estimates and valid inference without requiring an excessive number of imputations.

One challenge with applying this rule arises in high-dimensional settings, where the number of variables is large. In such cases, it is common for a large proportion of individuals to have at least one missing value, which can push λ close to 1. To address this, it is reasonable to use the overall missing rate (i.e., total proportion of missing cells in the dataset) as a conservative proxy for λ when needed.

This rule provides a useful upper bound for choosing K, especially when balancing the goals of statistical precision and computational efficiency.

3.3 Changing substantive model

Another important decision point in the missing data handling process is the choice of the substantive model—the model used for the final analysis after imputation. One natural alternative to standard multiple linear regression is the linear mixed-effects model (LME).

As discussed earlier, LME has the advantage of leveraging all available observed data, even in the presence of missingness. In fact, under certain conditions, LME can yield valid inferences without requiring multiple imputation, as long as the missingness mechanism is ignorable (e.g., MAR) and auxiliary variables are not essential.

Moreover, even with fully observed data—either originally complete or completed through imputation—LME remains a superior choice in many longitudinal settings because it explicitly accounts for within-subject correlation from repeated measurements. This leads to more accurate estimates and better statistical efficiency than standard linear models, which ignore the data's hierarchical structure.

In this project, we focus on multiple linear regression and LME, both of which assume a linear relationship between covariates and the outcome. However, when selecting a substantive model, it's important to evaluate this assumption. In cases where linearity is questionable, alternative models—such as polynomial regression or spline-based models—may provide a better fit and should be considered.

An additional benefit of using LME is that it allows for flexible modeling of time. Specifically, it enables us to treat time as a continuous variable, rather than as a categorical factor, which can improve interpretability and statistical power. This modeling choice will be explored further in a later section.

3.4 Data Wrangling

Both the Acupuncture and VITAL datasets originate in wide format, where each subject's repeated measurements are stored in separate columns. While this format is convenient for some operations, it is not directly compatible with LME, which require the data to be transformed into long format—with one row per observation per time point.

In our project, both datasets have a balanced longitudinal design, with follow-up measurements taken at regular, pre-specified intervals. This is not always the case in real-world studies, where irregular or missing time points can complicate analysis. The balanced structure in our data allows for a more straightforward imputation of partially observed follow-up outcomes using a multivariate distribution, making the MI process more reliable and easier to implement.

To address these differing requirements, the following general approach was used for both data sets:

1. Perform MI in wide format using the `mice()` function, treating the repeated measures as separate but related variables.
2. Convert the imputed wide data sets to long format, which allows time to be included as a continuous variable in the LME model.
3. Fit LME models on the long-format data to model outcome trajectories over time while appropriately accounting for missing values and repeated measures.

This strategy allows us to use multiple imputation methods compatible with the multivariate normal assumption across time points, while still leveraging the advantages of LME models in the analysis stage.

The balanced longitudinal design of both studies further supports this approach: each subject was planned to have the same number of repeated measurements at roughly equal intervals (monthly in the Acupuncture study and yearly in the VITAL study). This regular measurement structure allows partially observed follow-up outcomes to be imputed more reliably using a multivariate framework.

Below is a simple schematic to illustrate the reshaping process. Here, `Outcome1/2/3` represent repeated outcomes across three timepoints, and `x` indicates a missing value:

Original data (wide format):

ID	Outcome1	Outcome2	Outcome3
1001	2.5	3.5	x
1002	4.5	x	6.3
1003	3.3	6.2	8.1

Completed data after multiple imputation (still wide format):

ID	Outcome1	Outcome2	Outcome3
1001	2.5	3.5	4.5
1002	4.5	5.2	6.3
1003	3.3	6.2	8.1

Transformed data (long format):

ID	Time	Outcome
1001	1	2.5
1001	2	3.5
1001	3	4.5
1002	1	4.5
1002	2	5.2
1002	3	6.3
1003	1	3.3
1003	2	6.2
1003	3	8.1

3.5 Examing using forest plot

To visually compare how our estimand (treatment effect) changes across different missing data methods, we present results using a series of forest plots. These plots allow us to directly assess the impact of each imputation or modeling strategy on the estimated effect and its confidence interval.

The same process was applied to both the Acupuncture and VITAL datasets. For the VITAL trial, which follows a 2×2 factorial design, we simplify the analysis—as previously discussed—by treating fish oil and vitamin D as if they were tested in two separate randomized controlled trials. This means we ignore potential interaction effects between the two treatments and estimate their effects independently, which allows for a more straightforward comparison across methods.

3.5.1 Change imputation method and substantive model

We evaluated the impact of missing data by systematically varying both the imputation method and the substantive model used in the analysis. By doing that, we try to isolate the effects of changing either the imputation strategy or the analysis model, and highlights the practical impact of each decision on the resulting treatment effect estimates.

First, as described earlier, we applied three anchoring methods: Complete Case Analysis (CAA), Last Observation Carried Forward (LOCF), and mean imputation. Each was followed by multiple linear regression to estimate the treatment effect.

Next, we performed Multiple Imputation (MI) using the predictive mean matching method with 5 imputations, again analyzing the imputed datasets using multiple linear regression—maintaining consistency with the anchoring models for fair comparison.

Then, we changed the substantive model to a linear mixed-effects model (LME), without performing any imputation. This approach uses all available observed data and accounts for repeated measurements, offering a more efficient use of the data under the MAR assumption.

Finally, we combined both improvements: we performed MI using predictive mean matching with 5 imputations, and analyzed the completed datasets using the LME model.

3.5.2 Change estimand

Assuming our primary objective remains to estimate the mean difference in pain scores at the final timepoint for both datasets, we now explore how this estimand can be more efficiently estimated by incorporating interim outcome data.

In the previous section, we applied six different methods, two of which used Linear Mixed-Effects (LME) models as the substantive model. However, in those implementations, time was treated as a categorical variable. This approach limited our ability to take full advantage of the repeated measures structure, as it did not model the underlying trajectory of pain over time.

To address this, we revise the estimand by treating time as a continuous variable. This allows us to model individual pain trajectories across the study period, thereby leveraging all available interim outcome data. As a result, we improve the precision of the estimated treatment effect at the final timepoint—even in the presence of missing data.

Original Estimand (Categoricl Time) *Acupuncture data set*

The goal of the original acupuncture study was to estimate the effect of acupuncture therapy versus general care on chronic headache severity. The estimand of interest is the mean difference in headache scores at 12 months, conditional on baseline severity. This can be expressed with the following linear regression model:

$$pk5_i = \beta_0 + \beta_1 group_i + \beta_2 pk1_i + \varepsilon_i$$

where

- $pk5_i$: headache pain score of individual i at 12 months.
- $pk1_i$: adjusted baseline headache pain score of individual i
- $group_i$: binary indicator of treatment assignment (1 = acupuncture, 0 = control)
- β_0: intercept, representing the expected 12-month pain score for a control participant with zero baseline pain
- β_1: treatment effect of acupuncture relative to control; this is the estimand of interest
- β_2: effect of baseline headache severity on the 12-month outcome
- ε_i: residual error term for individual i

VITAL data set

In the VITAL study, the primary aim was to evaluate the effects of vitamin D and fish oil supplementation on knee pain four years after randomisation. The estimand of interest is the mean difference in knee pain scores at 4 years, conditional on baseline pain. This is modelled using the following linear regression:

$$painyear4_i = \beta_0 + \beta_1 fishoilactive_i + \beta_2 vitdactive_i + \beta_3 painbase_i + \varepsilon_i$$

where

- $pain_y r4_i$: knee pain score for individual i at 4 years post-randomization
- $fishoilactive_i$: binary indicator for fish oil treatment (1 = active, 0 = placebo)
- $vitdactive_i$: binary indicator for vitamin D treatment (1 = active, 0 = placebo)
- $painbase_i$: baseline knee pain score for individual i
- β_0: intercept, representing the expected 4-year pain score for a participant receiving neither treatment and with zero baseline pain
- β_1: effect of fish oil supplementation; this is the estimand of interest in the fish oil analysis
- β_2: effect of vitamin D supplementation; this is the estimand of interest in the vitamin D analysis
- β_3: effect of baseline pain on the 4-year outcome
- ε: residual error term for individual i

Changed Estimand (Continuous Time) To better utilise repeated measures in both datasets, we revise the estimand using a Linear Mixed-Effects (LME) model with time treated as a continuous variable.

To ensure that the intercept represents the estimated treatment effect specifically at the end of follow-up, we re-center the time variable: - In the acupuncture study, we define $time_c = time - 12$ (in months), so that time zero corresponds to month 12, which is the end of study. - In the VITAL study, we define $time_{contin} = time - 4$ (in years), so that time zero corresponds to year 4, which is the end of study.

With this centering, the model intercept represents the mean pain score at the end of the study, and the coefficient for the treatment group becomes the estimand of interest.

Acupuncture data set

The revised LME model for the acupuncture dataset is specified as below. Note we did not allowing random slop as there are only 2 data collecting time points here:

$$pkscore_{ij} = \beta_0 + \beta_1 group_i + \beta_2 time_{c,ij} + \beta_3 (group_i \times time_{c,ij}) + \beta_4 pk1_i + b_{0i} + \varepsilon_{ij}$$

where

- $pkscore_{ij}$: headache pain score for individual i at time j
- $group_i$: binary indicator of treatment assignment (1 = acupuncture, 0 = control)
- $time_{c,ij}$: time in months since baseline, centered at 12 months ($time_c = time - 12$)
- $pk1_i$: adjusted baseline headache pain score of individual i
- β_0: intercept, representing the expected pain score at 12 months (end of study) for the control group with zero baseline pain
- β_1: treatment effect of acupuncture relative to control at 12 months; this is the estimand of interest
- β_2: average rate of change in pain over time in the control group
- β_3: treatment-time interaction effect
- β_4: effect of baseline headache pain
- b_{0i}: random intercept for individual i
- ε_{ij}: residual error term for individual i at time j

VITAL data set

The revised LME model for the VITAL dataset is specified as:

$$pain_{ij} = \beta_0 + \beta_1 fishoilactive_i + \beta_2 vitdactive_i + \beta_3 time_{contin,ij} + \beta_4 (fishoilactive_i \times time_{contin,ij}) + \beta_5 (vitdactive_i \times time_{contin,ij}$$

where

- $pain_{ij}$: knee pain score for individual i at time j
- $fishoilactive_i$: binary indicator for fish oil treatment (1 = active, 0 = placebo)
- $vitdactive_i$: binary indicator for vitamin D treatment (1 = active, 0 = placebo)
- $time_{contin,ij}$: time in years since baseline, centered at 4 years ($time_{contin} = time - 4$)
- $painbase_i$: baseline knee pain score for individual i
- β_0: intercept, representing the expected pain score at 4 years (end of study) for a participant receiving neither treatment and with zero baseline pain
- β_1: effect of fish oil at 4 years; this is the estimand of interest in the fish oil analysis
- β_2: effect of vitamin D at 4 years; this is the estimand of interest in the vitamin D analysis
- β_3: average rate of change in pain over time in the placebo group
- β_4: fish oil-time interaction effect
- β_5: vitamin D-time interaction effect
- β_6: effect of baseline knee pain
- b_{0i}: random intercept for individual i
- b_{1i}: random slope for time for individual i
- ε_{ij}: residual error term for individual i at time j

3.5.3 Change FSC methods

Even within the FCS (Fully Conditional Specification) framework, there are multiple imputation strategies we can explore. We begin with deterministic prediction (method="norm.predict"), which imputes missing values using predicted values from a linear regression model. Since no uncertainty is added, this is not technically multiple imputation—each missing value is replaced by a single fixed estimate.

To account for uncertainty in the outcome, we adds random noise to the regression prediction (method="norm.nob"). Further, to reflect uncertainty in the model parameters, norm draws both regression coefficients and residual variance from their posterior distributions, providing a fully Bayesian imputation (method="norm").

Alternatively, instead of model-based predictions, we can impute using observed values. predictive mean matching (method="pmm") selects donor values from observed data that are closest in predicted value to the missing case. Weighted predictive mean matching (method="midastouch") extends this by weighting the distance between predicted values, offering a more robust variant of PMM.

Or even, if we prefer a non-parametric imputation approach, we can use the methods like random sampling (method="sample"), which imputes missing values by randomly sampling from the observed values of the same variable. This method does not rely on any model or predictor variables, and instead preserves the marginal distribution of the variable.

3.5.4 Change imputation numbers

As discussed above, there are various arguments for choosing a higher number of imputations based on the fraction of missing information and the desired precision of inference. To explore the impact of increasing K, we adjusted the number of imputations accordingly in both datasets.

For the Acupuncture dataset, where approximately 25% of cases have missing data, we increased K from the default 5 to 20 and 25. For the VITAL dataset, which has around 50% missingness, we increased K from 5 to 20 and then to 50.

3.5.5 Sensitivity analysis

As mentioned in our introduction, missing data mechanisms are unverifiable. While it is most common to assume that data are MAR, in reality, the mechanism could fall under MNAR. We conducted a sensitivity analysis to assess the robustness of the estimated treatment effect under alternative missingness assumptions. Sensitivity analysis is recommended by the National Research Council (2010) as a best practice for handling missing data; however, it is rarely conducted in practice. This may reflect an underappreciation of the impact of missing data in many studies, as well as the complexity and specialised expertise required to perform such analyses.

A systematic review by Fiero et al. (2016) examined 86 cluster randomised trials published between 2013 and 2014. Among the 80 trials that reported missing outcome data, only 14 (18%) reported conducting a sensitivity analysis.

In our study, we implemented multiple imputation with δ -adjustment as our sensitivity analysis approach. In this method, a constant shift δ is added to the imputed values after the initial imputation step. This enables us to explore a range of plausible scenarios by re-estimating the treatment effect under different assumptions about the missing data mechanism. By applying these shift parameters, we can model situations in which participants with missing data are assumed to have systematically better or worse outcomes than those observed, thereby testing the stability of our conclusions when relaxing the MAR assumption conditional on other variables.

The shift can be broken down in to:

$$\delta = \delta_{\text{MAR}} + \delta_{\text{MNAR}}$$

where; - δ_{MAR}: mean difference caused by the predictors in the imputation models - δ_{MNAR}: mean difference caused by an additional non-ignorable part of the imputation model.

Essentially, δ is the mean difference between the imputed values and the true values that we did not observe.

Deciding which δ values are suitable is subjective. There are two δ paramters we are considering for the treatment arm and cotrol arm. We need to include $\delta = 0$ as that assumes the data is MAR which acts as our baseline, and we add the non-zero δ values to test the sensitivity to the MNAR assumption. We decided to base our δ values on the standard deviations of the outcome variables of interest. The standard deviation of the headache severity score at the final time point in the treatment group is *13.71828* and *17.01089* in the control group. The standard deviation of the knee pain score at the final time point in the fishoil group is *19.13955*, *19.24856* in the vitamin D group, and *18.29479* in the control group. These statistics are based on the data in the original wide format. We also extracted the standard deviation of the outcome variables in the long format in which the estimand is continous. We decided an increasing and decreasing multiples of these values should suffice;

δ = -0.5*SD, -0.25*SD, 0 , 0.25*SD, 0.5*SD$

Applying these to both datasets, our chosen δ values are;

Original Estimand

- $\delta_{\text{Acupuncture}-\text{original}} = -6.859139, -3.429569, 0, 3.429569, 6.859139$
- $\delta_{\text{Acu}-\text{control}-\text{original}} = -8.505446, -4.252723, 0, 4.252723, 8.505446$
- $\delta_{\text{VITAL}-\text{fishoil}-\text{original}} = -9.569774, -4.784887, 0, 4.784887, 9.569774$
- $\delta_{\text{VITAL}-\text{vitd}-\text{original}} = -9.624281, -4.812141, 0, 4.812141, 9.624281$
- $\delta_{\text{VITAL}-\text{control}-\text{original}} = -9.147397, -4.573698, 0, 4.573698, 9.147397$

Changed Estimand

- $\delta_{\text{Acupuncture}-\text{changed}} = -7.397829, -3.698915, 0, 3.698915, 7.397829$
- $\delta_{\text{Acu}-\text{control}-\text{changed}} = -8.660196, -4.330098, 0, 4.330098, 8.660196$
- $\delta_{\text{VITAL}-\text{fishoil}-\text{changed}} = -9.783807, -4.891904, 0, 4.891904, 9.783807$
- $\delta_{\text{VITAL}-\text{vitd}-\text{changed}} = -9.967777, -4.983889, 0, 4.983889, 9.967777$
- $\delta_{\text{VITAL}-\text{control}-\text{changed}} = -9.646493, -4.823247, 0, 4.823247, 9.646493$

Sensitivity Analysis; the process

This section outlines the application of the δ-adjustment sensitivity analysis. We begin with the acupuncture trial and its original categorical estimand, specifying the predictor variables for multiple imputation as the treatment group and baseline headache severity score (pk1). These are used to generate a predictor matrix for imputation. An initial imputation run extracts the post object, which allows manual modification of the imputation process. For each value in the vector $\delta_{\text{Acupuncture}-\text{original}}$, we adjust the imputed values of the final headache severity score (pk5) by adding the corresponding δ. A linear regression model, $pk5_i = \beta_0 + \beta_1 \text{group}_i + \beta_2 pk1_i + \varepsilon_i$, is fitted to each imputed dataset, and the estimates are pooled. We extract the coefficient for the treatment group, our estimand of interest, along with its standard error, confidence interval, and p-value.

We then repeat this process using only the placebo group (group == 0) to isolate control group effects, generating the acu_wide_placebo dataset. Here, $\delta_{\text{Acu}-\text{control}-\text{original}}$ values are added to the imputed pk5 values, and the pooled regression estimates are filtered for the intercept term, representing the average outcome in the control group.

A similar approach is applied to the VITAL trial data in wide format, using pain_base, vitdactive, and fishoilactive as predictors, and pain_yr4 as the outcome. Following imputation and δ-adjustment, we fit the model $painyr4_i = \beta_0 + \beta_1 \text{fishoilactive}i + \beta_2 \text{vitdactive}i + \beta_3 \text{pain\_base}i + \varepsilon_i$, pool the estimates, and extract coefficients for the treatment groups (fishoilactive, vitdactive). For the control group, we filter the dataset to include only individuals with fishoilactive == 0 and vitdactive == 0, forming vital_wide_placebo. Here, $\delta\text{VITAL} - \text{fishoil} - \text{original}$, $\delta\text{VITAL} - \text{vitd} - \text{original}$, and $\delta\text{VITAL} - \text{control} - \text{original}$ are added to the respective imputed values, and pooled results are similarly obtained.

We performed sensitivity analyses in a similar fashion to assess the robustness of the treatment effect under our alternative estimand. This process was more computationally intensive, as it required performing multiple imputation on the wide-format data and subsequently transforming it into long format.

For the acupuncture trial, we added $\delta_{\text{Acupuncture-changed}}$ values to the imputed data and then used a custom function to convert the data into long format. Due to the repeated measures structure, we employed a linear mixed-effects model to estimate the treatment effect prior to pooling. The substantive model was:

$$pkscore_{ij} = \beta_0 + \beta_1 group_i + \beta_2 time_{c,ij} + \beta_3 (group_i time_{c,ij}) + \beta_4 pk1_i + b_{0i} + \varepsilon_{ij}$$

For the VITAL study, the imputed wide-format data was reshaped into long format using a separate function. The treatment effect was estimated using the following mixed-effects model:

$$\text{pain}_{ij} = \beta_0 + \beta_1 \text{fishoilactive}_i \text{time\_contin}_j + \beta_2 \text{vitdactive}_i \text{time\_contin}_j + \beta_3 \text{pain\_base}_i + b_{0i} + \varepsilon_{ij}$$

As for the control group, we only include the baseline knee pain score and time as predictors, so our model here was:

$$\text{pain}_{ij} = \beta_0 + \beta_1 \text{time\_contin}_{ij} + \beta_2 \text{pain\_base}_i + b_{0i} + \varepsilon_{ij}$$

4 Result

4.1 Comparing different methods

We begin by comparing our anchoring methods with more statistically principled approaches.

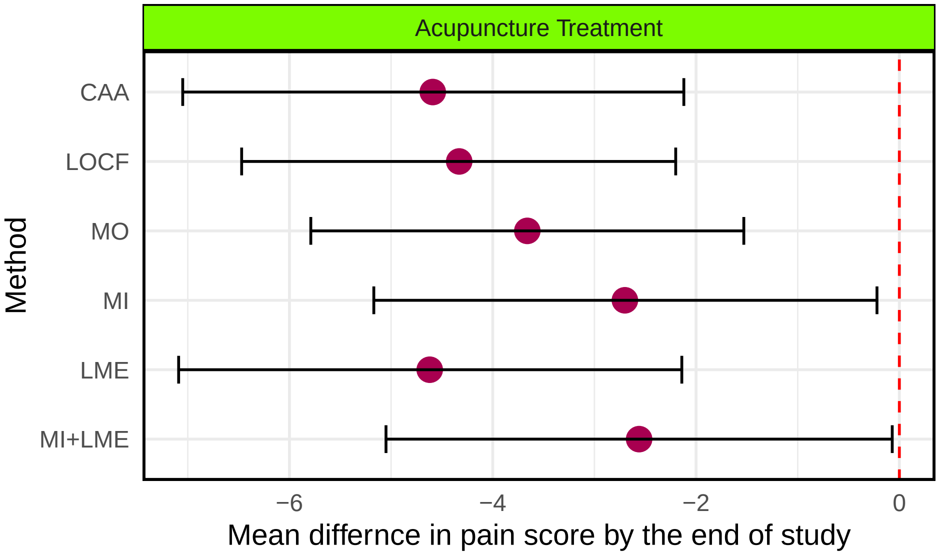


Figure 19: Acupuncture data set analysed with different methods; CCA = complete case analysis; LOCR = last observation carry forward; MO = mean observation; MI = multiple imputation with linear regression; LME = linear mixed-effects model without imputation; MI+LME = multiple imputation followed by linear mixed-effects model.

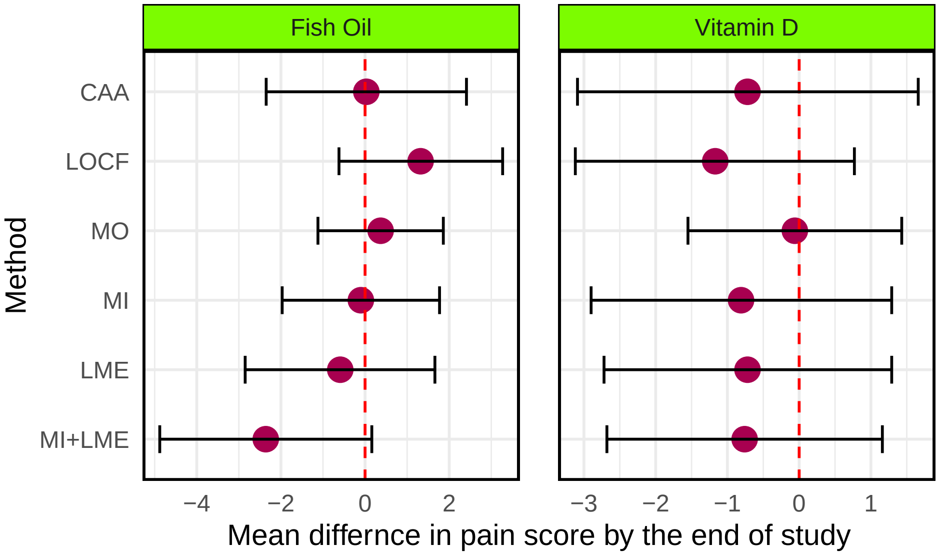


Figure 20: VITAL data set analysed with different methods; CCA = complete case analysis; LOCR = last observation carry forward; MO = mean observation; MI = multiple imputation with linear regression; LME = linear mixed-effects model without imputation; MI+LME = multiple imputation followed by linear mixed-effects model.

4.2 Changing estimand

In this section, we explore how changing the estimand (how time is modeled) to see how that affects our conclusions.

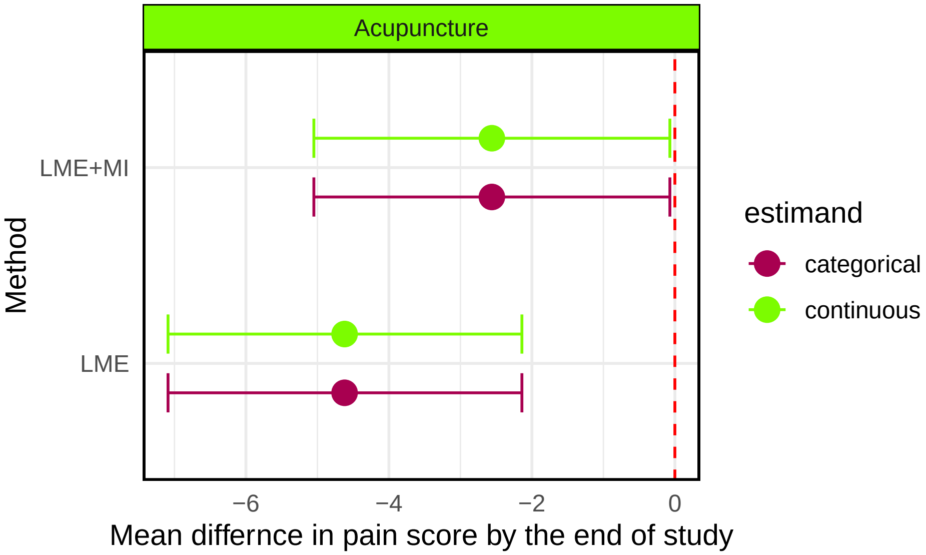


Figure 21: Acupuncture data set analysed with different estimands; LME: linear mix model without imputations; MI+LME: multiple imputation and linear miex model; Purple line (Original Estimand) estimand: The mean difference of pain score at the end of study conditional on baseline pain score, treating time as categorical variable; Green line (Changed Estimand) estimand: The mean difference of pain score at the end of study conditional on baseline pain score, treating time as continuous variable.

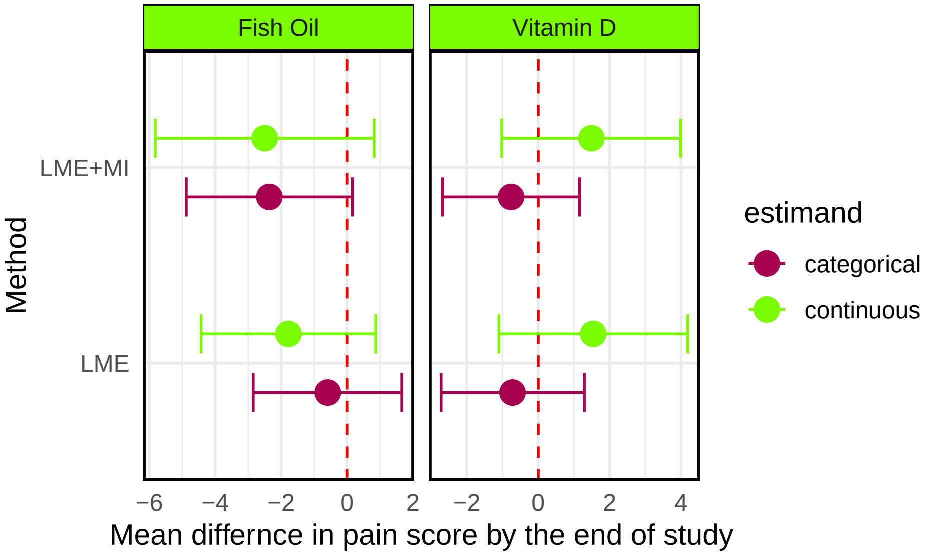


Figure 22: VITAL data set analysed with different estimands; LME: linear mix model without imputations; MI+LME: multiple imputation and linear miex model; Purple line (Original Estimand) estimand: The mean difference of pain score at the end of study conditional on baseline pain score, treating time as categorical variable; Green line (Changed Estimand) estimand: The mean difference of pain score at the end of study conditional on baseline pain score, treating time as continuous variable.

4.3 Changing imputation methods and imputation numbers

In this session. We keep our estimand treating time as a categorical variable and using multiple linear regression as substantive model.

We first compared different imputation methods within the FCS framework, using 5 imputations.

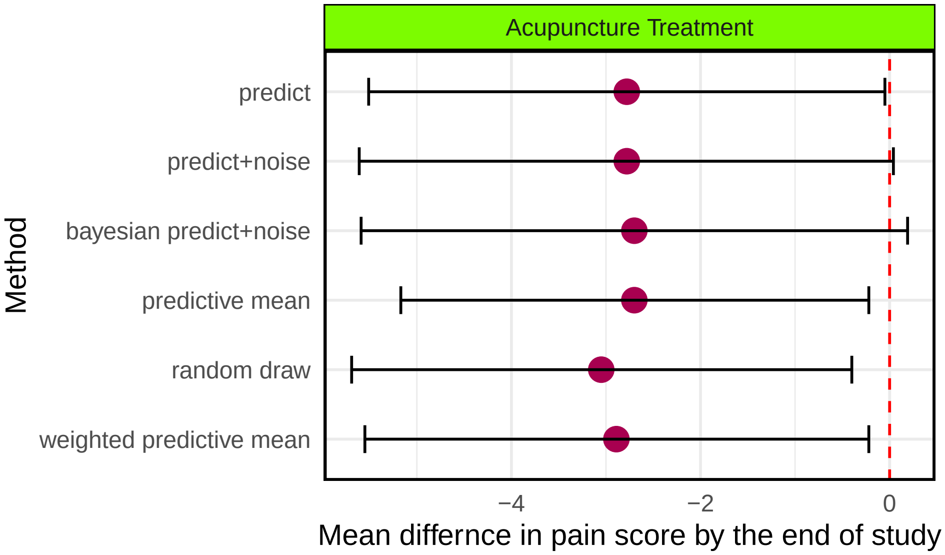


Figure 23: Acupuncture data set analysed with different FCS methods

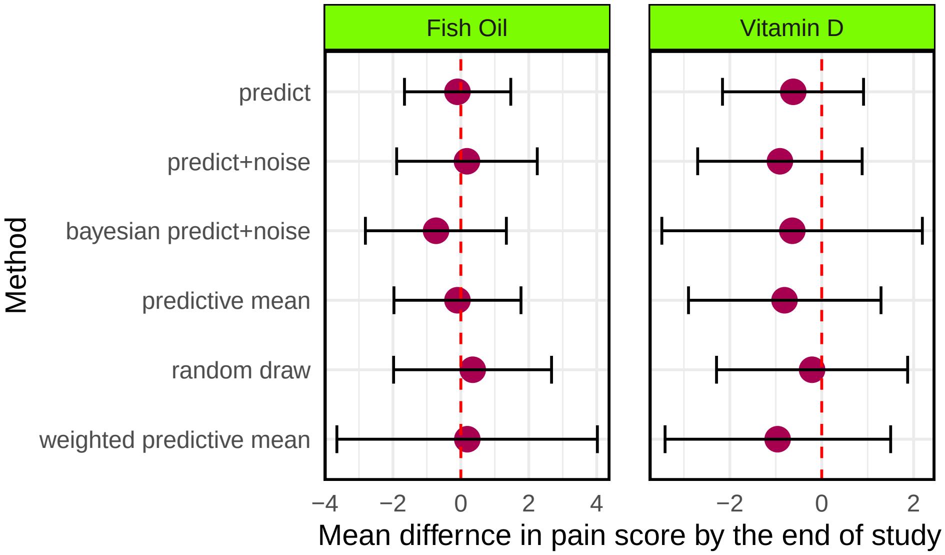


Figure 24: VITAL data set analysed with different FCS methods

Next, we assess the impact of changing the number of imputations, using predictive mean method. As introduced in the previous session, we used $K = 5$, $K = 20$, and $K = 100 * \lambda$ (Which is 25 and 50 for Acupuncture and VITAL data sets accordingly).

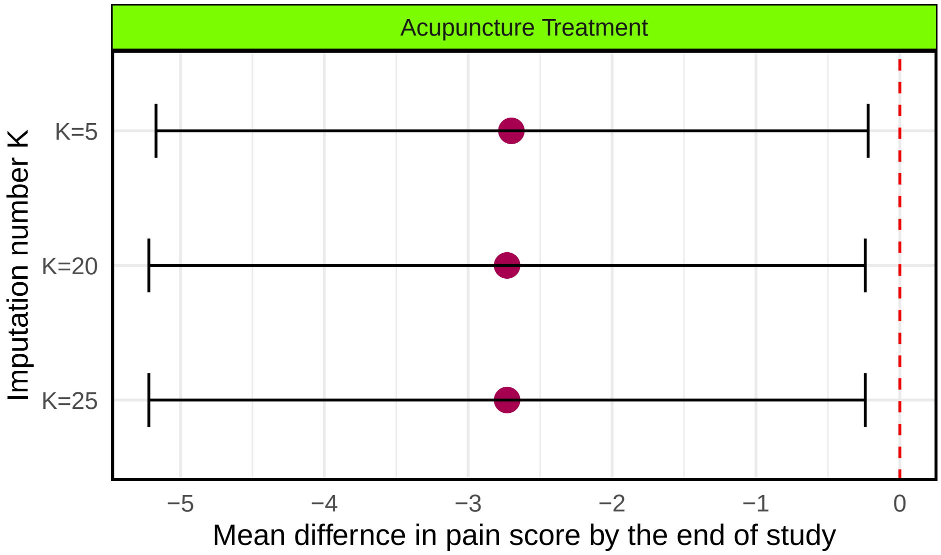


Figure 25: Acupuncture data set analysed with different imputation numbers

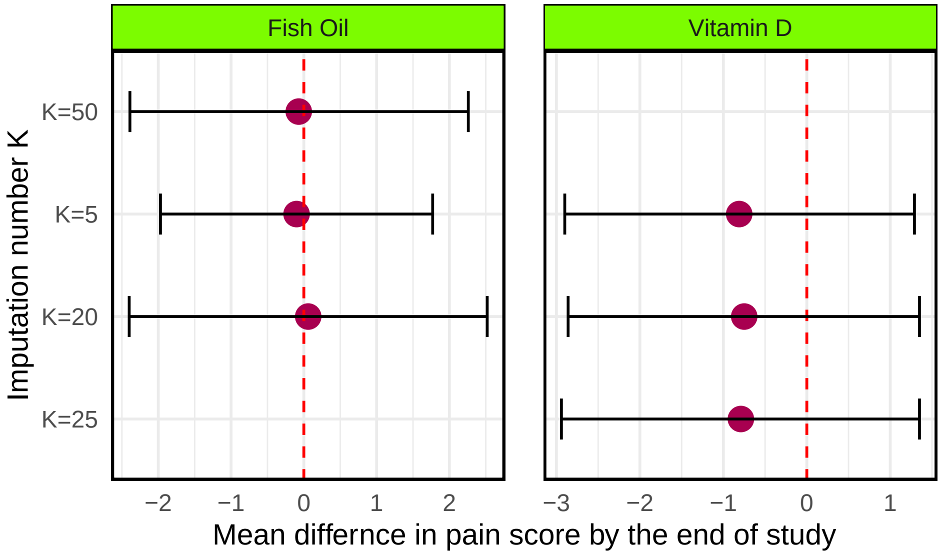


Figure 26: VITAL data set analysed with different imputation numbers

4.4 Sensitivity analysis

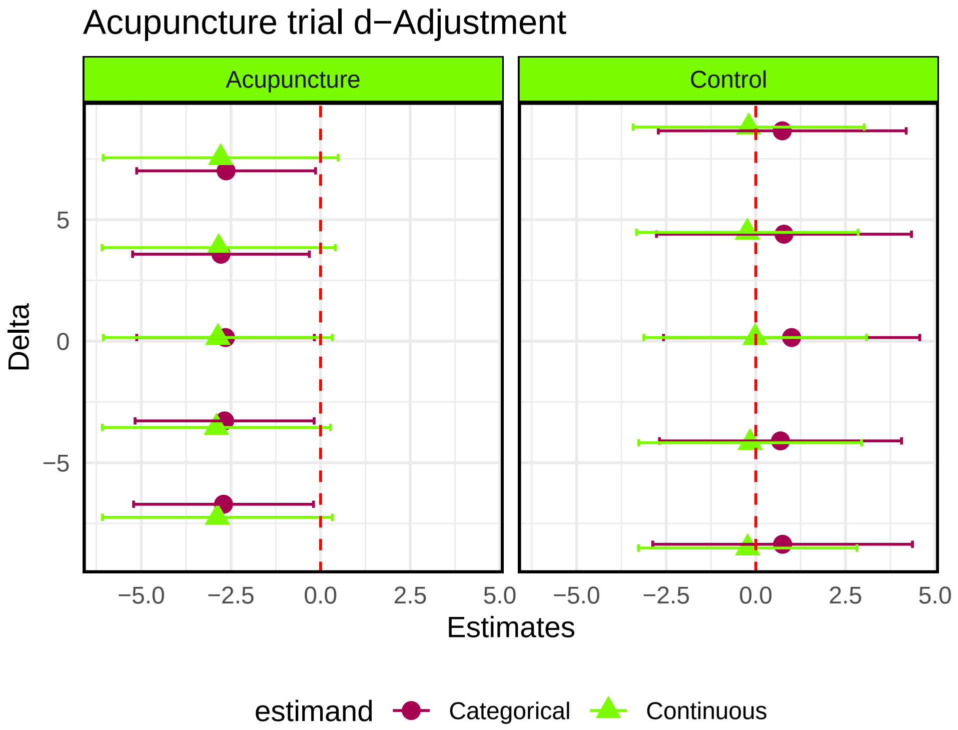


Figure 27: Sensitive analysis for Acupuncture data set

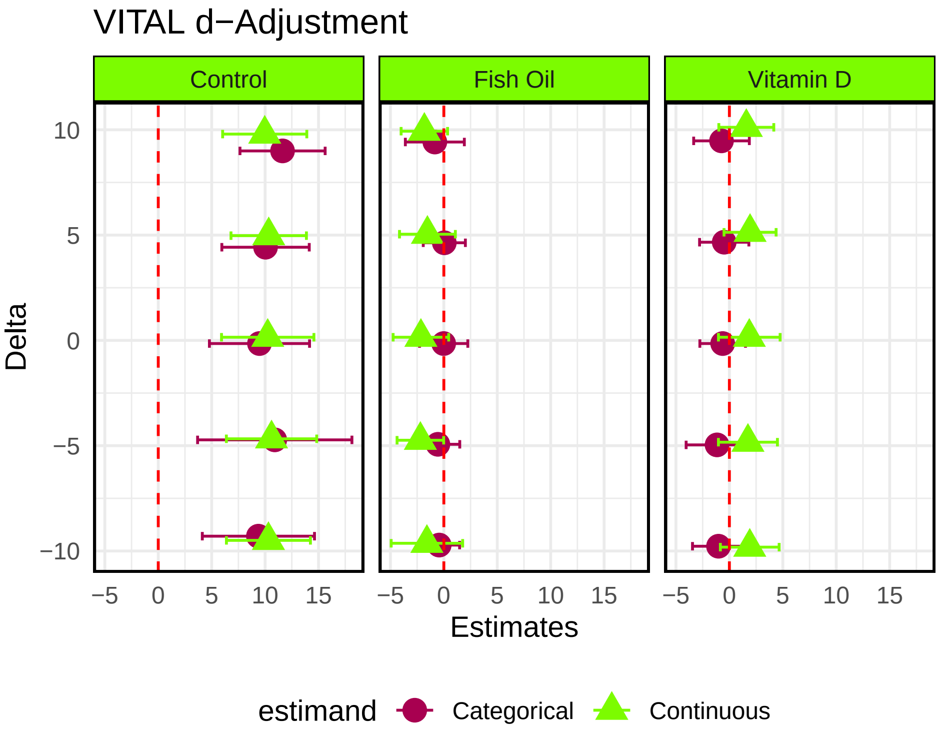


Figure 28: Sensitive analysis for VITAL data set

The forest plot aims to illustrate the robustness of the estimated treatment effects under varying missing data assumptions using δ adjustment sensitivity analysis. In figure 27 which shows the acupuncture trial results, we can observe the conditional expectation of headache severity at the final time point of 12 months given the baseline headache severity and acupuncture treatment under different δ shifts. Similarly we can also observe the conditional expectation of headache severity at the final time point in the control group, given baseline headache severity only and no treatment.

Figure 28 shows the VITAL results after applying the sensitivity analysis. In the fish oil panel, we observe the conditional expectation of knee pain score at 4 years given baseline knee pain score and fish oil treatment under the different δ shifts. The vitamin D panel shows similar conditional expectation, but the control group shows the conditional expectation of knee pain score given baseline knee pain score and no treatment.

The error bars represent the 95% confidence intervals.

The δ parameter shifts the imputed values to simulate scenarios where the missing data are systematically better or worse than the observed ones. Positive δ values imply better unobserved outcomes while negative values imply worse. The consistency of estimates across the δ values suggests the treatment effect is robust to departures from the MAR assumption.

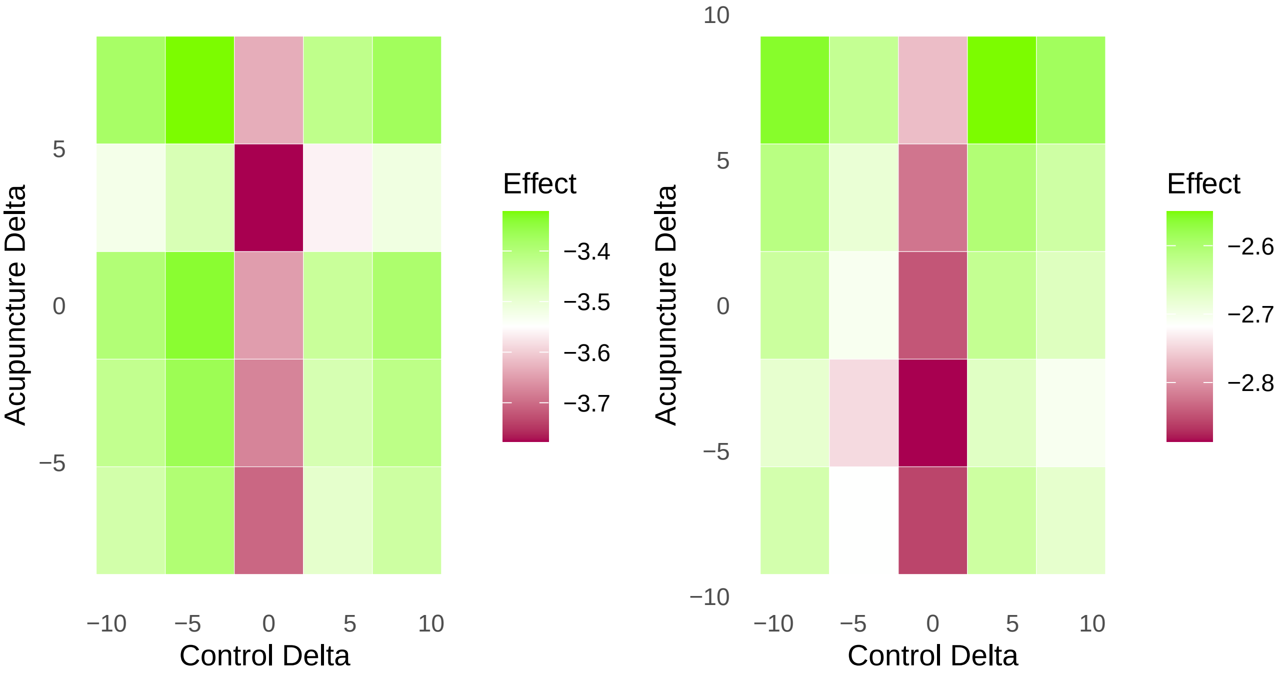
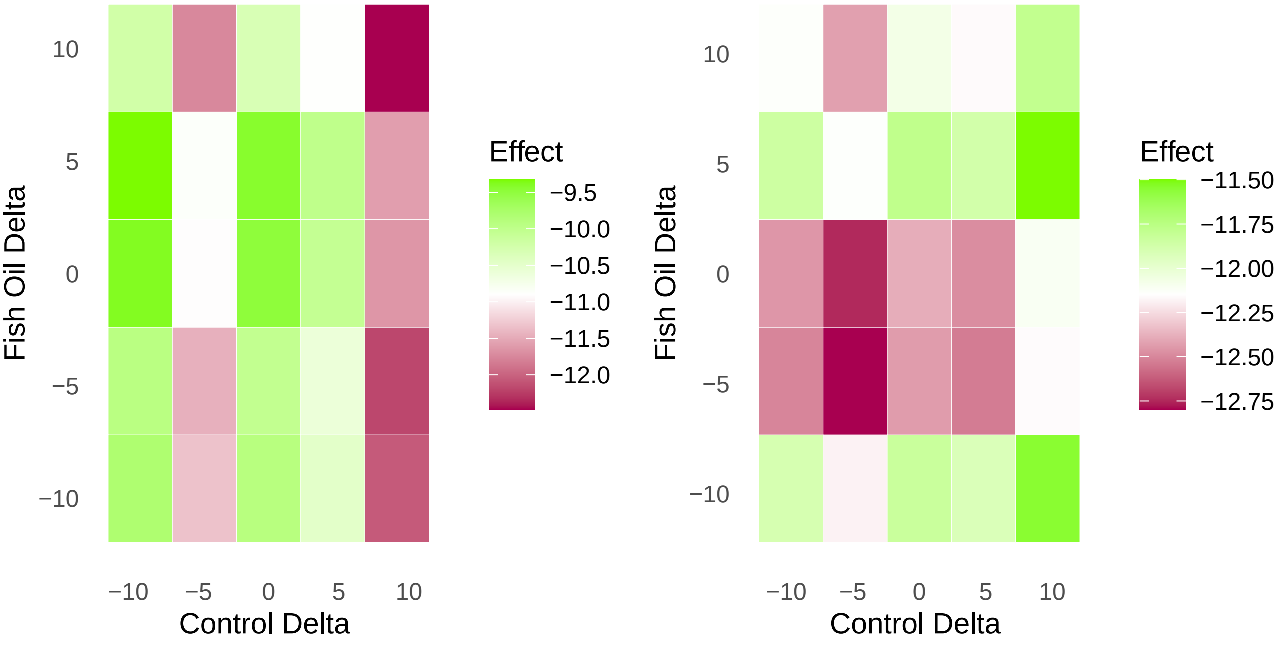


Figure 29: Original & changed estimand in acupuncture trial



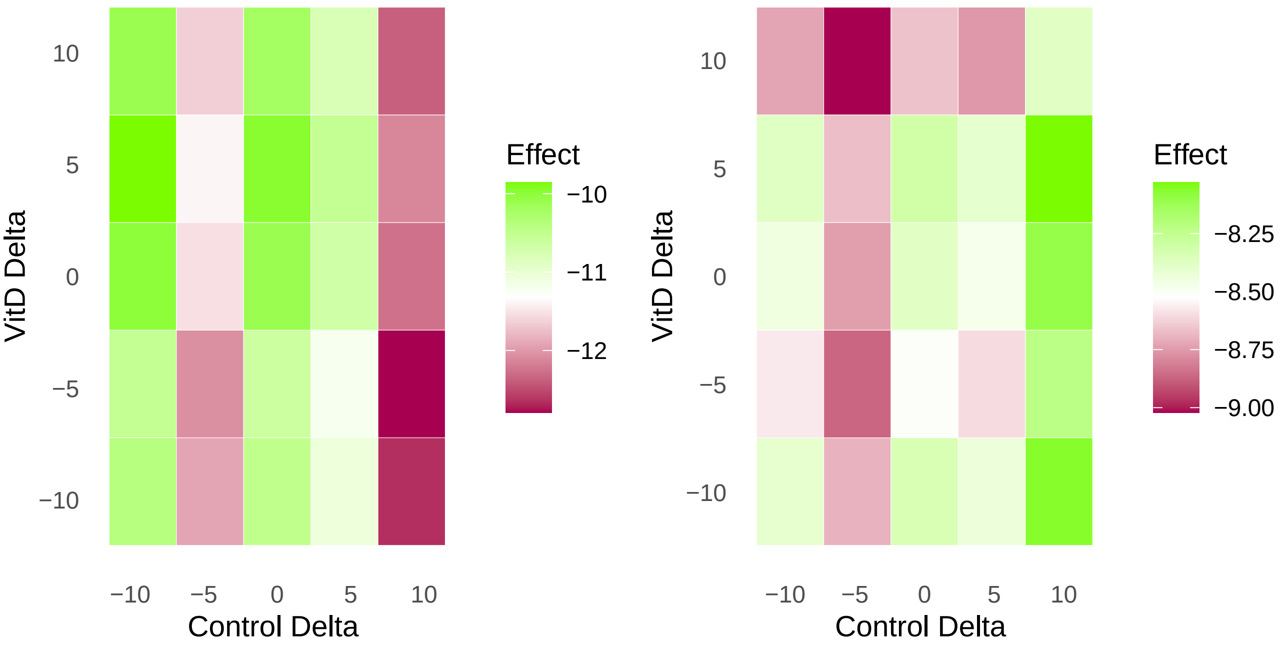


Figure 30: Original & changed estimand in vitamin D group

We created heatmaps plotting the δ values for both the control group and treatment group to visualise the change in treatment effect when varying the δ adjustments under MNAR. We are comparing both the original estimand (heatmap on the left handside) and the changed estimand (heatmaps on the right handside) The aim here is to show the robustness of the initial estimated treatment effect when assumed under a different mechanism. A similar range of colours would be interpretable as a strong estimate as there was not much change in the estimate after adding the δ shifts.

5 Discussion

Across all analyses, no meaningful difference was observed in the final treatment effect estimates across the various methods and settings explored.

However, several points are worth noting:

- In Figures 1 and 2, we observe that methods involving multiple imputation (MI) show more variation in the estimated treatment effects. This is expected, as MI leverages all available observed data, while the other four methods (CAA, LOCF, MO, and LME without imputation) use only a subset—typically including only group and baseline pain score.
- When we changed the estimand to treat time as a continuous variable, the impact was minimal in the Acupuncture study, which is unsurprising given that it includes only two follow-up time points. In such cases, modeling time continuously offers limited additional information.

It is important to emphasize that the goal of this project is not to revise prior conclusions from the original studies, but rather to gain practical insights by applying and comparing different methods for handling missing data.

The fact that our results are broadly consistent across methods should not be taken to imply that simpler approaches like CCA are equivalent to more principled methods. Even in a similar future study, the same results may not hold, especially under different missing data patterns or assumptions.

This project has highlighted that there are many valid approaches to handling missing data, and that these can be combined in different ways depending on the context. It is crucial to understand which methods are most appropriate under which conditions.

For example:

- LME is a powerful tool for analyzing longitudinal data and can yield unbiased results even without imputation—if the assumptions are met. However, in the Acupuncture dataset, its utility is limited due to the small number of timepoints and its inability to incorporate other post-randomisation variables like pain frequency.
- In the VITAL dataset, although there are more timepoints, baseline pain scores are sometimes missing. As a result, when fitting LME without imputation, individuals with missing baseline values are excluded. This exclusion becomes more severe as more covariates with missing data are included, potentially leading to substantial loss of information. While it is theoretically possible to incorporate post-randomisation variables into an LME model by extending the covariance structure, this approach is complex and less flexible. In contrast, using MI provides a more straightforward way to incorporate auxiliary post-randomisation information during the imputation step, without requiring modifications to the model structure.

5.1 Limitations of the project

There are several limitations stemming from the inherent characteristics of the datasets used in this project.These limitations are not a result of analytic choices but rather reflect the practical boundaries of the available data.

- *Limited follow-up in Acupuncture dataset:* The Acupuncture study includes only two follow-up timepoints (3 and 12 months), which limits the ability to model time flexibly or detect subtle longitudinal trends. As a result, methods like linear mixed-effects models offer limited additional benefit in this context.
- *Weak treatment effect in VITAL:* The VITAL dataset shows relatively small therapeutic effects, making it difficult to detect meaningful differences between methods. This also reduces the practical impact of using more advanced imputation strategies.
- *No known data-generating mechanism:* Since we are working with one-off samples from an unknown data-generating process and a true, unobserved treatment effect—rather than simulated data—we cannot evaluate the performance of different analytical methods against a known ground truth. This limits our ability to draw general conclusions about their long-run properties.

5.2 Future work

In addition, there are methodological aspects that could be improved or explored further in future work. These include choices related to model specification, variable adjustment, and the diversity of imputation techniques employed. Addressing these areas could enhance the robustness and generalizability of findings, especially in more complex or data-rich clinical contexts.

- *Model specification:* Both linear regression and LME rely on assumptions such as linearity and normally distributed residuals. While these assumptions were not formally tested, the sample size in the VITAL dataset is likely large enough for the central limit theorem to provide valid inference. That said, future work could benefit from visual checks of residuals and distributions of pain scores at each timepoint to give a clearer sense of model fit and outcome behaviour.
- *Imputation strategy:* In the most of our analysis, we used predictive mean matching for imputing missing variables. While this is a valid and robust approach, many other imputation models as we introduced were not explored in detail. These could be considered in future work to assess sensitivity to imputation method.
- *Covariate adjustment in substantive models:* Our analysis models adjusted only for treatment group and baseline pain score in our substantive model to ensure comparability across methods. This low-level of adjustment in the substantive model is the reason we observed larger discrepancies between results using MI and those not using MI, as a fuller set of covariates was used during MI. Future work could explore how more complex substantive model specifications interacts with MI and influences treatment effect estimates.

Bodner, Todd E. 2008. "(Bodner2008)what Improves with Increased Missing Data Imputations?" *Structural Equation Modeling: A Multidisciplinary Journal* 15 (4): 651–75. https://doi.org/10.1080/10705510802339072.

Carpenter, James R., Jonathan W. Bartlett, Tim P. Morris, Angela M. Wood, Matteo Quartagno, and Michael G. Kenward. 2023. *Multiple Imputation and Its Application.* 2nd ed. Statistics in Practice. Wiley. https://onlinelibrary.wiley.com/doi/book/10.1002/9781119756118.

Graham, John W., Allison E. Olchowski, and Tamika D. Gilreath. 2007. "How Many Imputations Are Really Needed? Some Practical Clarifications of Multiple Imputation Theory." *Prevention Science: The Official Journal of the Society for Prevention Research* 8 (3): 206–13. https://doi.org/10.1007/s11121-007-0070-9.

Royston, Patrick. 2004. "(Royston2004)multiple Imputation of Missing Values." *The Stata Journal* 4 (3): 227–41. https://doi.org/10.1177/1536867X0400400301.

(Rubin1987) Multiple Imputation for Nonresponse in Surveys. 1987. John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470316696.fmatter.

White, Ian R., Patrick Royston, and Angela M. Wood. 2011. "Multiple Imputation Using Chained Equations: Issues and Guidance for Practice." *Statistics in Medicine* 30 (4): 377–99. https://doi.org/10.1002/sim.4067.

Characteristic
N = 342[1]
N = 353[1]
N = 371[1]
N = 332[1]
	Both
	Fish Oil on
	Neither
	Vitamin D o

Sex 1-male,2-female	
1	111(32%)
2	231(68%)
Body mass index at randomization, kg/m2	32 (7)
Missing	8
Current smoking 1-yes,0-no	26(7.8%)
Missing	7
Age at randomization to VITAL study,years	67 (7)
Baseline Aspirin use 1-yes,0-no	156(46%)
Missing	6
WOMAC Pain score at baseline	37 (19)
Missing	48
WOMAC Pain score at year 1	32 (19)
Missing	182
WOMAC Pain score at year 2	32 (20)
Missing	47
WOMAC Pain score at year 3	32 (21)
Missing	78
WOMAC Pain score at year 4	30 (20)
Missing	129
Have you had a knee replacement surgery? (1=Yes, 2=No)	
1	43(13%)
2	293(87%)
Missing	6
Unilateral knee pain 1=Yes 0=No	200(71%)
Missing	61
Bilateral knee pain 1=Yes 0=No	81(29%)
Missing	61
Frequency of knee pain 1=Never 2=<1 day/wk 3=1~2 days/wk 4=3~6 days/wk 5=daily	
1	1(0.3%)
2	16(5.4%)
3	44(15%)
4	58(20%)
5	169(57%)
9	7(2.4%)
Missing	47
Tylenol use at baseline 1=never 2=occational 3=daily 9=missing	
1	108(37%)
2	91(31%)
3	53(18%)
9	43(15%)
Missing	47
Nsaids use at baseline 1=never 2=occational 3=daily 9=missing	
1	62(21%)
2	98(33%)