Clinical Trials IV Assignment 1

Raul Unnithan
Durham University

March 3, 2025

Declaration

This piece of work is a result of my work, and I have complied with the Department's guidance on multiple submissions and the use of AI tools. Material from the work of others not involved in the project has been acknowledged, quotations and paraphrases suitably indicated, and all uses of AI tools have been declared.

Contents

1 Questions **2**

1.1 Sample Size 2

1.2 Allocation 2

1.2.1 Imputing Missing Data 2

1.2.2 Allocation 3

1.3 Analysis of Results 5

2 Trial Considerations **8**

2.1 Sample Size 8

2.2 Allocation 8

2.2.1 Imputing Missing Data 8

2.2.2 Allocation 10

2.3 Analysis of Results 11

Chapter 1 Questions

1.1 Sample Size

Section 2.1 means the sample size for a normally distributed primary outcome variable formula can be used to determine the number of participants for each arm:

$$N = \frac{2\sigma^2(z_\beta + z_{\frac{\alpha}{2}})^2}{\tau_M^2}.$$

Next, sub in the provided values of $\alpha = 0.05$, $\beta = 1 - 0.85 = 0.15$, $\sigma = 7$, and $\tau_m = 2$:

$$N = \frac{2(7^2)(z_{0.15} + z_{0.025})^2}{2^2} = 219.912392 < 220.$$

Therefore, we need at least 220 participants in each trial arm.

1.2 Allocation

1.2.1 Imputing Missing Data

The priority was finding out which columns contained the missing values. In this case, they were only in the `baseline` covariate.

So, the priority was working out the type of missing data, i.e., whether it was missing at random (MAR), missing not at random (MNAR), or missing completely at random (MCAR).

The first way we can test this is by using Hotelling's Multivariate T-Test. This test examines the differences between those observations with an observation of some partially observed variable - `baseline` in this case and those without. It comes with the following hypotheses:

- H_0: the participants' data are MCAR.
- H_1: the participants' data are MAR or MNAR.

The test on this dataset returns a p-value of < 0.001, meaning there is sufficient evidence to suggest the missing `baseline` covariate data is MAR or MNAR.

However, this test is accompanied by caution. The power of this test can be strongly influenced by sample size. So, it is combined with a more detailed approach, the Absolute Standardised Mean Difference (ASMD).

The general rule is that ASMD values over 0.1 are cause for concern, though again, this can be vulnerable to small sample sizes, which is not an issue here as $n = 220$ for each arm. Also, there is a plot showing each ASMD. So what we ultimately test is:

- H_0: ASMD $\leq 0.1 \Rightarrow$ the participants' data are MCAR.

- H_1: ASMD $> 0.1 \Rightarrow$ the participants' data are MAR or MNAR.

We set `includeNA=T` so we can see the missingness effect and the observed values of other variables.

On the participants' dataset, [`asmd_min`, `asmd_max`] = [0.340, 0.604] and here are all the `asmd` values for age, sex and BMI, respectively: [0.604, 0.340, 0.501]
Although there are very few of them, the ASMD values are quite large (much bigger than the advised 0.1). Hence, they are statistically significant, supporting that there is sufficient evidence to suggest the participants' data are MAR or MNAR.

However, the only way we could determine whether the mechanisms for `baseline` was MAR or MNAR would be to measure or otherwise procure some of the missing data. We do not have the necessary information to work out which is the case. We would now talk at length with the experts/clinicians who have a much better understanding of the probable missingness causes, but this is not possible here. We can use the MAR or MNAR determination, though, to decide how to impute the missing data.

Using this deduction, we can do imputation using a regression model. This is implemented using the `stan_glm` regression function from the package `rstanarm`. The `rstanarm` functions use MCMC to generate samples from the posterior distribution of the regression model. The `baseline` variable is continuous, so linear regression should work well. This approach works because all the other variables have no missingness, as it will not be able to impute a value for any case with missingness. We will now use this model to impute values for `baseline`.

We impute by sampling one value from the posterior distribution for each point, which is why we are using `rstanarm`. Here is a plot of the imputed vs. already present data:

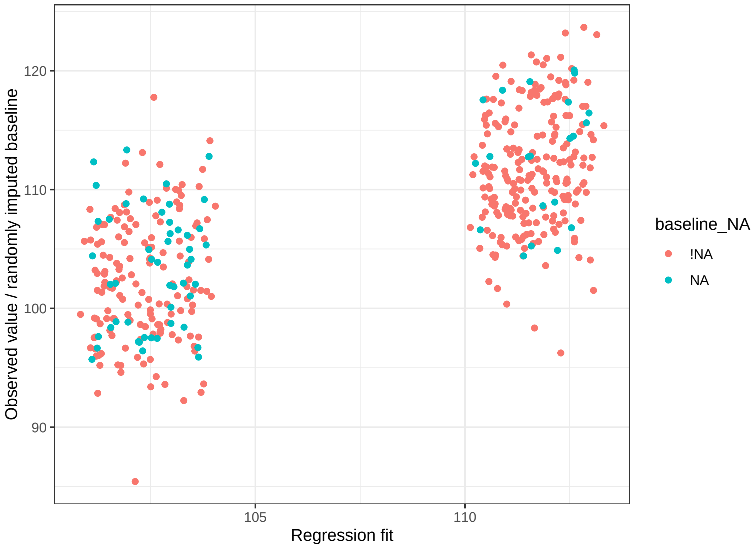


Figure 1.1: Complete Baseline Data

This imputed data is much more representative of the actual dataset than if we just used `predict` on the linear model, meaning we are ready for allocation.

1.2.2 Allocation

Before we can use any of our allocation methods, we're going to need to bin the 2 numeric variables: age and BMI. Since the age range is from 50-65, a natural bin is 50-54, 55-59, and 60-65. For BMI, we can bin using standard health metrics: a below-average BMI is anything less than 25. An average BMI is between 25 and 29, and an above-average BMI is anything more than 29. The numerical variables should spread well across each bin because this avoids bins with too few observations, leading to unreliable estimates. Here is the spread of `age, sex` and `BMI` across each bin:

Age Range:	50-54	55-59	60-65
Count:	166	135	139

Table 1.1: Age Group Distribution

BMI Category:	Below Average BMI	Average BMI	Above Average BMI
Count:	151	140	149

Table 1.2: BMI Category Counts

This is well-spread as desired. Now, we are ready for allocation.

Minimisation is used as an allocation method over the others. See 2.2.2 for justification. Minimisation aims to minimise differences between treatment groups. It was developed for use with strata as an alternative to RPBs. Minimisation balances individual prognostic factors rather than their interactions. It follows a set algorithm. The first patient is allocated using simple randomisation. Next, patients are recruited sequentially and need to be assigned to a trial arm. For each new patient, the level of each factor is listed.

Here, we have age, sex and BMI. Next, a sum is calculated to quantify the imbalance across all factors. For each factor level, the difference between the number of patients allocated to each treatment arm is computed. These differences are then summed across all factor levels as:

$$(n_A + x_{++} - n_B + x_{++}) + (n_A + +y_+ - n_B + +y_+) + (n_A + + + z - n_B + + + z),$$

where n_A and n_B represent the number of patients in treatment arms A and B, and x, y, z represent the levels of the factors.

Based on the imbalance sum, the new patient is allocated to a treatment arm with a certain probability. If the sum is negative, the patient is allocated to arm A with probability P, where $P > 0.5$. If the sum is positive, the patient is allocated to arm B with probability P. If the sum is zero, the patient is allocated to arm A with a probability of $\frac{1}{2}$. Some implementations use $P = 1$, making the allocation deterministic, while others use $\frac{1}{2} < P < 1$ to retain some randomness. Here, the latter is used.

The minimisation algorithm here accounts for the three covariates: `sex`, `age` and `BMI`. These are stored in the covariate matrix `covmat`. Each covariate is given equal weight, represented by `covwt`, ensuring no single covariate dominates the allocation process. There are two treatment arms in this trial, group A and group B, denoted as `trtseq` = (0,1). Finally, the allocation ratio is set at 1:1, meaning that, ideally, patients will be distributed equally across both arms.

For all subsequent patients, the function `Minirand` is used to determine treatment allocation. This function takes into account the covariate matrix, weighting scheme and treatment ratio to assign the new patient in a way that minimises imbalance across stratification factors. The final treatment allocations are then stored for trial results, which can be analysed to determine the treatment effect on the condition.

1.3 Analysis of Results

The trial has been run, and we have lots of data to analyse to try to assess what effect the treatment has had. τ will be used to denote the treatment effect on the condition.

Multiple methods exist for analysing results. 'Trawling' for the best possible model by trying lots of different things and inevitably settling on the one that leads to the most significant conclusion is poor practice and can increase the type I error rate (α).

This trial uses ANCOVA because it adjusts for treatment group baseline imbalances. See 2.3 for further justification. ANCOVA, in practice, involves fitting the following linear model to the observed outcomes x_i:

$$x_i = a_0 + \gamma b_i + \epsilon_i \quad \text{in group A}$$

$$x_i = a_0 + \tau + \gamma b_i + \epsilon_i \quad \text{in group B.}$$

Here, the ϵ_i are independent errors with distribution $N(0, \sigma_\epsilon^2)$, the b_i are the baseline measurements for $i = 1, \ldots, N_A + N_B$, for groups A and B with sizes N_A and N_B respectively, and a_0, γ are coefficients.

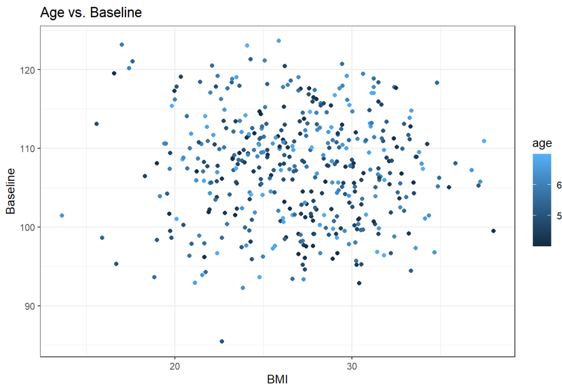
As with any linear model, we need to ensure that it is appropriate for our dataset. Two key things we need to check for are a linear effect across the range of the dataset and multicollinearity. The former is already dealt with because we have binned the necessary covariates into categories. Multicollinearity means making sure that none of the independent variables are highly correlated. This is not uncommon in clinical datasets since measurements are sometimes strongly related. It can mean choosing only one variable from a collection of two or more strongly related variables. We can check this by computing the variance inflation factor for each covariate and ensuring it is close to 1.

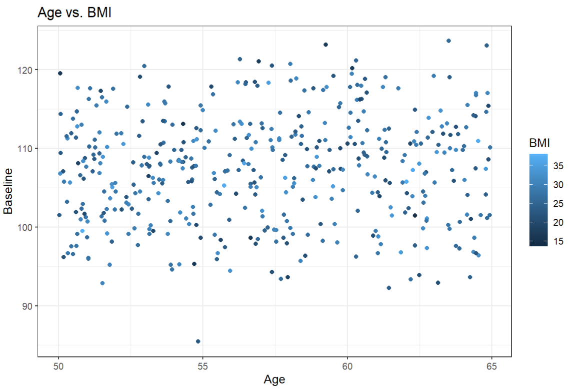
Now, we need to justify the covariates used in the ANCOVA model. This is done before the results are calculated, but it is clearer to do it in this section.

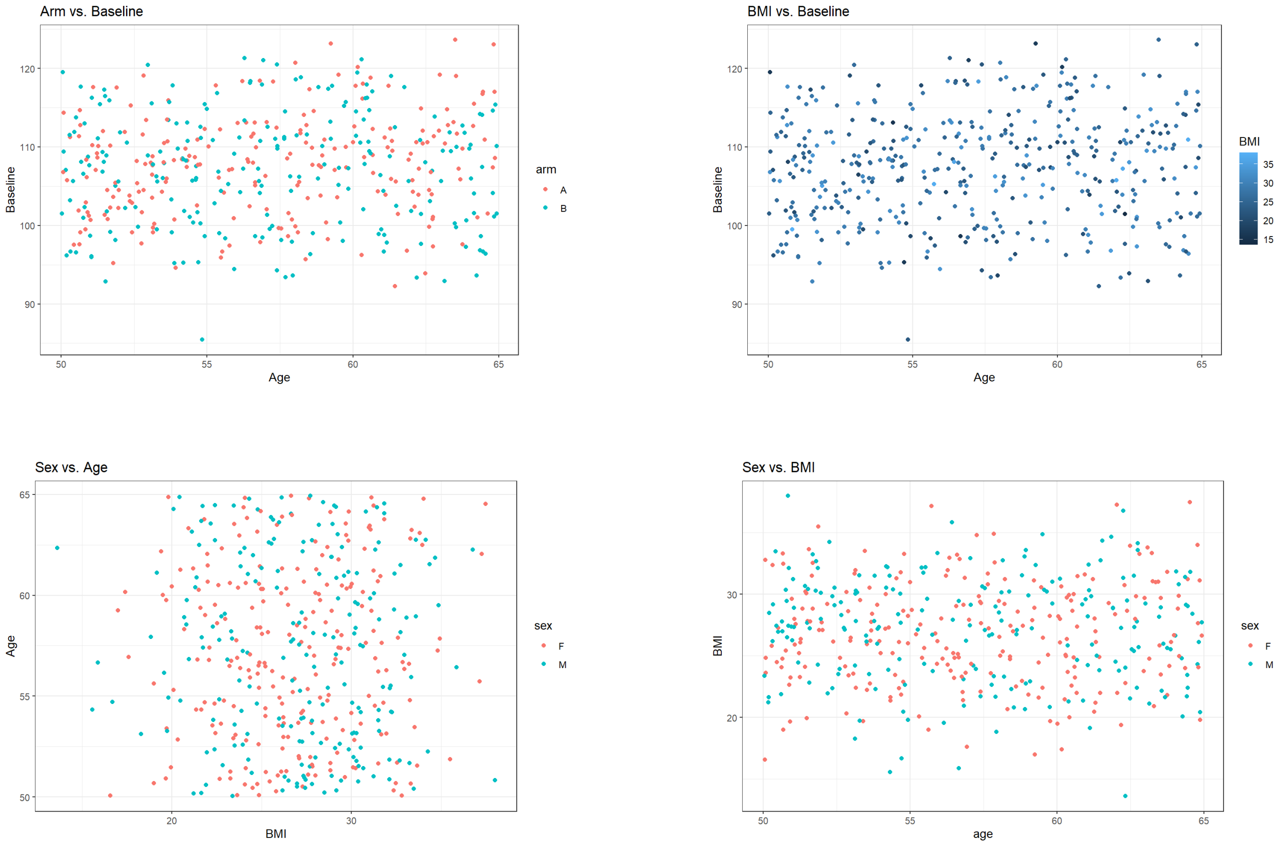
Adjusting for the baseline symptom score reduces residual variance and increases power, as it has a correlation of 0.7 with the outcome measurement. This ensures a more precise treatment effect estimate on the condition. Controlling for age, with a range of 50 to 65, and BMI, which has a mean of 27 and a standard deviation of 4, is also important as these factors may influence symptom severity and treatment response. Finally, the arm is essential to consider as it is key for hypothesis testing.

Including these covariates prevents confounding and ensures that any observed treatment effect is not biased by baseline differences between groups. Adjusting for these covariates also improves the interpretability of the results, ensuring that the treatment effect reflects a true change in symptoms rather than natural variation or baseline imbalances.

Now that we have justified including all the baseline covariates, we should consider the interaction terms that could be involved. Again, this is done pre-trial but explained here for clarity. The following plots were generated, and these indicate the model's significant covariates:







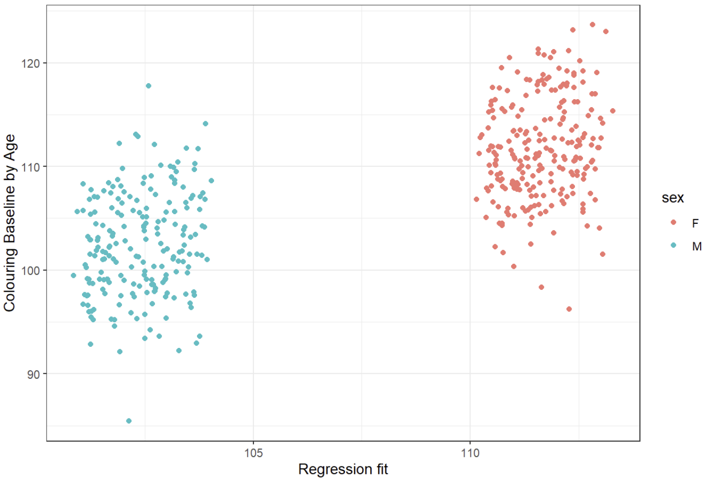


Figure 1.2: Sex vs. Baseline

From these plots, it is clear to see the only noteworthy interaction term to include is `sex:baseline`. This comes from the way sex clusters baseline in 1.2. Now, we can test a model including the sum of the baseline covariates and the `sex:baseline` interaction term. However, we still need to check for multicollinearity. This is done using the virtual inflation factor (`vif`) function from the `car` package. This gives the following range of VIF values:

Covariate	Baseline	Arm	Sex	Age	BMI	Baseline:Sex
VIF	3.5797	1.0050	446.3859	1.0301	1.0048	410.4810

Table 1.3: Variance Inflation Factors (VIF) for Each Covariate

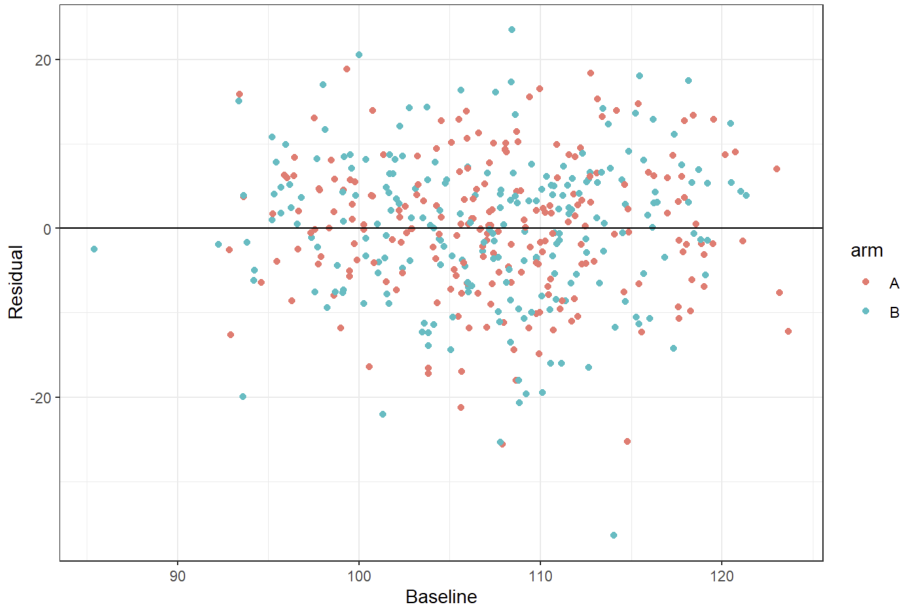
The VIF values for `sex` and `baseline:sex` are of particular concern, so one should be removed from the model. Let us remove `sex` as it has a higher VIF value. This gives us a new model and, hence, a new set of VIF values:

Covariate	Baseline	Arm	Age	BMI	Baseline:Sex
VIF	1.7084	1.0022	1.0301	1.0046	1.6772

Table 1.4: Variance Inflation Factors (VIF) for Each Covariate After Adjustment

These VIF values are roughly around one, so we have our final model to derive p-values and confidence intervals and hence determine if the treatment has had an effect on the condition.

We also know this model to be reliable looking at the residuals, which are spread apart nicely, too:



There are no clear patterns, and the distribution appears to be similar for each treatment group. The large sample allows us to assess the model's fit. Now, using the `armB` coefficient, we construct a 95& confidence interval for $\hat{\tau}$, we use (to 2 decimal places):

$$0.52781 \pm t_{0.975;435} \times 0.81179 = [-1.068, 2.123].$$

The model has $n - p = 435$ degrees of freedom because there are $n = 440$ data points, and we are estimating $p = 5$ parameters. This contains zero, and so our analysis has enabled us to conclude that **there is not** a significant reduction in the condition with the treatment. You can also see this in that $p = 0.65 > 0.05$.

Chapter 2 Trial Considerations

2.1 Sample Size

To calculate the required trial sample size, we assume that the outcome follows a normal distribution. The normal distribution of the baseline variable justifies this assumption because baseline and outcome have a strong correlation of 0.7, meaning the sample size formula is valid. We also want the smallest feasible sample size to make the trial as resource-efficient as possible.

2.2 Allocation

A huge amount of work in the planning and design of this trial goes into the random allocation: we want to eliminate all sources of bias (including those we cannot observe or are not even aware of) by randomly balancing the participants between the trial arms. If some of these participants' data are missing, we can no longer be confident that we still have this balance.

2.2.1 Imputing Missing Data

Dealing with missing data involves two main tasks: understanding the pattern(s) of missingness and processing the data to mitigate its effects.

The simplest thing we can do is discard the data for any participant who has some missing data. This is called a complete-case analysis because we only analyse data for participants whose data are complete. However, this approach has two main problems. First, if the missing data are not missing completely at random (MCAR), this can induce bias. Second, this approach can drastically reduce the amount of data.

When imputing, we could inadvertently introduce bias if we aren't careful (or if we are careful and unlucky) while imputing synthetic data. There are several methods to carry this out.

Hotelling's Multivariate T-Test

The first step was to examine the pattern of missingness: MCAR, missing at random (MAR), or missing not at random (MNAR), which was evaluated using Hotelling's multivariate t-test. This test examines the differences between those observations with an observation (of some partially observed variable) and those without. The test statistic is derived by assuming both groups are drawn from the same multivariate normal distribution. So, a high value of the test statistic (conversely, a low p-value) suggests that there are significant differences between the groups. Ultimately, this is a hypothesis test comparing: the null hypothesis that the participants' data are MCAR and the alternative hypothesis that the participants' data are MAR or MNAR.

There is a caution that comes with this test, however. The power of this test (and others like it) can be strongly influenced by sample size, so it is sensible to combine it with a more detailed approach, such as the Absolute Standardised Mean Difference (ASMD).

Absolute Standardised Mean Difference

The absolute standardised mean difference (ASMD) measures how different the values of the observed covariates are for missing versus observed values of each partially observed covariate. For every partially observed covariate, there is an ASMD for each other covariate.

Label the partially observed covariate X_M, and suppose $X_1, \ldots, X_K$ are the other covariates. The dataset is split into those cases with X_M observed and those with X_M missing. For each covariate $X_1, \ldots, X_K$, we find the absolute value of the difference in means and divide this by the standard deviation of that covariate. The ASMD is, therefore, always non-negative and should not be affected by sample size.

There is also a `Table 1`, so called because a summary table of this nature should always be included when summarising a dataset in terms of the difference between two groups. This is formatted a little strangely, as it is designed for use in printed works. This table includes the result of a statistical test (by default, a chi-squared test) showing whether the differences are statistically significant.

A general rule of thumb is that ASMD values over 0.1 are cause for concern, though again, this can be vulnerable to small sample sizes, which is not an issue here as $n = 220$ for each arm. Also, there is a plot showing each ASMD. So what we ultimately test is:

- H_0: ASMD $\leq 0.1 \Rightarrow$ the participants' data are MCAR.
- H_1: ASMD $> 0.1 \Rightarrow$ the participants' data are MAR or MNAR.

We set `includeNA=T` so that we can see the effect of missingness, as well as observed values, of other variables.

However, frustratingly, the only way we could determine whether the mechanisms for `baseline` was MAR or MNAR would be to measure (or otherwise procure) some of the missing data. We do not have the necessary information to work out which is the case. If this were a real trial, we would now talk at length with the experts/clinicians, who will have a much better understanding of the probable causes of missingness.

Next, we needed to use this conclusion to impute the missing data. Multiple methods exist to do this, and evaluating each of them and how well they apply is important. The first is mean imputation, which involves replacing missing values with the mean of the observed values for that variable. However, this method can introduce bias if the data are not MCAR, reduce the sample standard deviation, and distort relationships between variables. As we established initially, there is sufficient evidence to suggest that the missing data is MAR or MNAR.

Another method is Imputation Using Logic. This approach uses information about how a variable relates to other variables to fill in missing values. It is not applicable here as there are no clear relationships between `baseline` and the other covariates.

Imputation Using a Regression Model: This method uses regression models, such as those implemented via STAN functions from the `rstanarm` package, to impute missing values. This approach involves: using the missingness of other values as an input (by creating a nabular object), then removing variables from the model, and then working iteratively, generating temporary imputed values and cycling. Next, the variables are rounded with missingness. Finally, one value is sampled from the posterior distribution for each point.

The final imputation method is Multiple Imputation. This involves drawing multiple values from the posterior distribution rather than just one. This is a widely used approach to imputing missing data. However, since only 1 covariate has missing data, it was not the approach that was used as MICE is computationally expensive.

Computational efficiency is important for a clinical trial because it can lead to issues such as delays in analysis and decision-making, which can slow down interim analyses and trial adjustments. It can also increase cost and resource demand, requiring expensive hardware, software, and analyst time.

2.2.2 Allocation

Now that the data is complete and bins have been assigned, the arms need to be allocated. There are many ways to do this, but the most basic form is simple random allocation. In simple random allocation, each participant is assigned to one of the two trial arms with equal probability. This comes with lots of benefits: it generates a truly random sample. Each participant also has an equal probability of being assigned to each treatment group. The assignment of each participant is statistically independent of others. And finally, it does not require a centralised "master" randomisation.

This is a great idea, in theory. However, this can lead to a chance imbalance in group sizes, especially in smaller trials. Unequal group sizes are a problem because they can reduce the statistical power of the trial. It is also unreliable due to the `sample` function. This means there is a reliance on the seed for an allocation of low imbalance, which is not reliable for a clinical trial.

Random Permuted Blocks (RPBs) are another allocation method that helps avoid substantial imbalances in group sizes. However, if the block size is fixed and known, the allocation for some patients can become predictable, leading to potential selection bias. RPBs with Random Block Length can mitigate this by reducing the predictability of treatment allocation compared to fixed block size RPBs, thus ensuring patients are equally likely to receive each treatment. While using random block lengths makes it harder to predict, certain setups can still create occasional moments where the next allocation becomes more obvious.

Biased Coin Design is another allocation method. It adjusts the probability of allocation to maintain balance, making a participant less likely to be assigned to an over-represented group. However, if the imbalance is known, the probability of guessing the next allocation correctly is high, which invites bias. Urn models also adjust to balance, but they do this differently. They adjust allocation probabilities based on the balance of the design so far, favouring the under-represented group. The greater the imbalance, the higher the probability of reducing it. However, urn models have a glaring drawback. Near the start of the allocation, the probabilities are likely to change a lot to address the imbalance. Still, once a reasonable number of allocations have been made, it is likely to settle into simple random sampling (or very close).

The final two allocation methods I am considering are stratifying and minimisation. These are both more complex than the others and, hence, more suitable for allocation. Stratified sampling divides each factor into levels and assigns treatment separately for each combination of factor levels (strata), often using permuted block designs. Strata are determined by multiplying the number of levels in each factor:

$$\begin{aligned}\text{Total Strata} &= (\text{Levels of Sex}) \times (\text{Levels of Age}) \times (\text{Levels of BMI}) \\ &= 2 \times 3 \times 3 = 18 \text{ strata}\end{aligned}$$

Since there are 440 participants - 220 per arm, this means there are $\frac{440}{18} \approx 24$ participants per stratum, so 12 per arm. This is risky because some strata could have fewer than 10 participants per arm, leading to an imbalance. This is why minimisation was used.

2.3 Analysis of Results

There are several ways to analyse the results of a clinical trial. All of these ultimately lead to generating p-values and confidence intervals (CIs) for the treatment effect, τ.

Because the randomised allocation process should produce comparable groups, we can compare the primary outcome between the groups using CIs and p-values. We can do this simply using summary statistics on the outcome. However, this does not consider a balance of the design. This is where baseline values come in.

Baseline measurements are useful primarily for two reasons: they can assess the balance of the design and be used in the analysis. There are 2 ways to integrate baseline values. The classic way is to perform an unpaired t-test to determine the difference between the baseline and outcome variables. This is used to generate the required p-values and CIs. Using the baseline values in this way often reduces the p-value and shifts the CI slightly higher.

Another method is the “dodgy” approach. This involves examining each group separately to determine whether the outcome variable has significantly changed. One appeal of this analysis is that it can be used with a paired t test, which is generally more powerful, to derive the p-values and CIs.

However, these methods involve basing our analysis on the baseline values being statistically identical draws from the underlying distribution, and therefore, having the same expectation and variance. This is theoretically true but in real-life trials, there will be some imbalance in the baseline measurements for the different treatment arms.

This trial uses ANCOVA to generate CIs and p-values because it adjusts for baseline imbalances between treatment groups, leading to a less biased estimate of the treatment effect.