4-14-2023 2:00 PM

Multiple endpoints in randomized controlled trials: a review and an illustration of the global test

Lindsay Cameron,

Supervisor: Zou, Guangyong., *The University of Western Ontario*
Joint Supervisor: Jairath, Vipul., *The University of Western Ontario*
A thesis submitted in partial fulfillment of the requirements for the Master of Science degree in Epidemiology and Biostatistics

Abstract

A randomized controlled trial is often used to provide high quality evidence regarding treatment interventions. Due to the complex nature of many diseases, trials usually select multiple primary outcomes to capture the efficacy of the interventions. In this thesis, we conducted a literature search to determine the prevalence of the different types of multiple outcomes that have been used in randomized controlled trials. We also reviewed the corresponding statistical methods used to deal with such outcomes. In addition, we described the benefits of using global tests as a statistical method when there are multiple primary outcomes in order to answer the global question of whether the intervention is effective. As an illustration for the global test, we used data from a previously published trial in ulcerative colitis. The global tests included O'Brien's OLS, Lauter's test and O'Brien's rank-sum test, and all of the tests used produced statistically significant results with p-values less than 0.05. Global tests should be considered when using multiple outcomes as well as additional guidelines surrounding how multiple primary outcomes should be managed.

Keywords

Randomized controlled trials (RCTs), primary outcome, multiple outcomes, composite outcomes, co-primary outcomes, global test

Summary for Lay Audience

Medical research questions can be answered with many different types of study designs, each with its own strengths and weaknesses. A randomized controlled trial (RCT) is considered to provide the highest quality evidence. A RCT may be used to evaluate the effects of a new drug, treatment strategy, a different surgical protocol or a lifestyle change by comparing groups that receive the intervention of interest to a control group(s). When designing an RCT, it is important to have a clear research question. The primary outcome that is chosen for the study should attempt to answer the research question. Most research teams will choose a singular outcome that encompasses the research question, but sometimes the question may not be able to be answered by a single outcome. In this case, there would be multiple primary outcomes used in order to capture all of the information required. There are multiple ways to deal with multiple outcomes. First, one may choose one outcome as the primary and leave the rest as secondary outcomes. Second, one may use an established scoring system to create a composite outcome based on the multiple outcomes. Third, one may apply simultaneous tests to determine if at least one of the outcomes occurs or is statistically significant compared to a control group. Finally, one may treat multiple outcomes as co-primary outcomes. This thesis reports a literature search aimed at determining the prevalence and approaches used to handle multiple outcomes in randomized controlled trials in the medical literature.

There are many statistical methods that can be used to evaluate a dataset and there is not one specific method that is used when there are multiple primary outcomes present. In this thesis, we demonstrate one type of method that can be used which are global tests. Global tests provide certain advantages over other methods that can be used. To demonstrate how global tests can be

used, a dataset from a previously conducted clinical trial was used. The results produced from the global tests were then compared to the originally published results.

Acknowledgments

I would like to acknowledge both Dr. Guangyong Zou and Dr. Vipul Jairath for their help and guidance in developing and writing this thesis.

Table of Contents

List of Tables

List of Figures

List of Appendices

Chapter 1: Introduction

1.1: What Provides Quality Evidence?

When faced with the task of seeking the available evidence for a medical question, often the first step involves searching the relevant literature. For this search, the researcher wants to be confident that they are seeking the best information that is available to them. To do this, it is important to consider the type of study being used. Certain types of studies are structured to provide higher quality evidence than others. In the evidence pyramid, with the different types of studies, at the top of the pyramid is the evidence with the highest quality, which usually comes from randomized controlled trials (RCTs). A RCT is placed at the top of the pyramid because it is generally considered to provide the strongest inference for cause-effect arguments (Akobeng, 2005). This fact can be attributed to the randomization used in a RCT.

A RCT involves enrollment of participants and randomizing them into different intervention groups. The simplest RCT has a parallel two-group design in which participants are randomly assigned to either the new intervention group or the reference group. The new intervention may be a medication or other intervention, while the reference group can be a placebo or active comparator. The randomization process used in this type of study is designed to ensure that the two groups are comparable in terms of known and unknown confounders, and thus any difference observed between the outcomes at the end of the trial can be attributed to the new intervention. In other words, confounding can be reduced to a minimum in randomized trials.

1.1.1: How to Randomize

The simplest form of RCT design involves two comparator arms: one experimental group and one control group. Randomization is used to assign trial participants to their respective groups. This can be done in many ways, typically using statistical software to generate randomization sequence. When conducting the randomization process, it is important to keep the intended ratio between groups (arms). Most randomized trials use a 1:1 ratio, as such a design can have the best efficiency for a fixed totally sample size. If there is a large discrepancy in the group sizes, this can lower the precision when obtaining the results and may not represent the true treatment effect (Roberts & Torgerson, 1998).

Aside from a simple randomization strategy, certain methods can be used in order to avoid treatment arms with unequal group sizes. The first is the use of block randomization.

Block randomization ensures the allocation of participants into different arms according to the pre-specified ratio in each block. Common block size can range from 2 to 8. For instance, in a trial with two arms, and equal allocation to each arm, using a block size four, two participants would be assigned to the treatment arm and two would be assigned to the control arm (Sedgwick, 2014; Roberts & Torgerson, 1998). The pattern of allocation within each block of four can be of any combination, as long as two participants are assigned to each arm of the trial. With a block size of four, there would be six possible combinations for allocation (Sedgwick, 2014). Block randomization can also be used in trials that have more than two arms, as long as the chosen block size is a multiple of the number of trial arms (Sedgwick, 2014). Block randomization ensures that the number of trial participants in each arm follows the pre-specified ratio.

1.1.2: Advantage of Using a Randomized Controlled Trial

The only way to truly determine if an exposure or treatment has an effect on an individual or group of people, is if the study could be repeated with the intervention eliminated within the same time window. In this case, the question being asked is: “if we could go back in time and eliminate the exposure, would the outcome still occur?”. This situation is called the counterfactual (Bours, 2020). The counterfactual theory is based on the idea that a treatment/exposure is only causal if the outcome observed is different among those exposed compared to the same individuals when unexposed (Bours, 2020). Because the counterfactual scenario cannot actually be observed, an RCT is often used to provide an approximate answer to the counterfactual question. A RCT is a unique type of study because it is the only study that provides a method that is the closest to determining a causal effect. This attribute is due to the randomization process, making a RCT the optimal choice when it comes to evaluating treatments. Observational studies such as a cohort or case-control study, may provide a strong association between two variables, but cannot, on their own establish causality.

The key feature of randomization is to attempt to guarantee that participants’ baseline characteristics between comparator groups in a RCT are well balanced for all known and unknown confounders. In other words, the groups are comparable. Otherwise, the results may be confounded by other factors, which are referred to as confounders. Confounders are an independent variable that when present can result in an outcome that can be attributed some other exposure. Confounding is a type of bias that occurs when the true effect between an exposure and outcome becomes distorted because of the presence of a third variable (ie. a confounder). A confounder may falsely demonstrate an effect that is not truly there. A confounder may be confused with an effect modifier; however, an effect modifier is a variable

that results in a different magnitude of treatment effect across the variable. This may be that the treatment effect is larger in females compared to males. With a confounder the magnitude of the treatment effect remains similar across the variable. In order for a variable to be considered a confounder it needs to have the three following attributes. First, the variable must be associated with the exposure/treatment of interest. Second, the variable needs to be associated with the outcome and may or may not be causal. This could include being a risk factor for a disease. Third, the variable cannot be an intermediate on the causal pathway between the exposure and outcome of interest (Skelly et al., 2012). For example, when investigating the association between ice cream and drownings, it is found that there is a higher number of drownings when people are consuming ice cream. However, this does not mean that consuming ice cream causes people to drown. In this association the confounding variable would be warm temperatures. This variable is missing from the overall picture and helps provide an understanding of the association that is found. When temperatures outside rise, people are more likely to consume ice cream as a cold treat. People are also likely to go swimming to try and stay cool. Therefore, warm outside temperatures meet all three criteria in order to be considered a confounding variable. It is associated with both the exposure and the outcome, but it is not on the causal pathway because eating ice cream does not cause warm temperatures and in turn cause drownings. Therefore, it is important to consider any confounders that may influence the association between the exposure and outcome in order to accurately capture the true effect.

As mentioned, in order to determine the effect of an intervention, the treatment and control groups must be the same except for the intervention. The trial groups will be made up of different participants and it is important that the baseline characteristics between the trial arms are balanced. Baseline characteristics can consist of various risk factors such as age, sex or

weight, and past exposures or health conditions. When these baseline factors are not balanced between trial groups, this has the potential to introduce bias. Bias influences the results of a study and pulls the result further from the true effect (Bhide et al., 2018). This is due to the fact that the groups are no longer comparable. The control or placebo group is used to represent the counterfactual. However, when there is an imbalance of baseline factors, the control group no longer represents the counterfactual and is not an appropriate comparison to truly represent an intervention's effect. To avoid this imbalance between groups, randomization can be used for group allocation. Randomization tries to ensure that the baseline factors are distributed equally between the trial arms. This will also ensure that the groups are comparable and any difference in the treatment effect that is found between the groups, can be attributed to the treatment tested and not on different factors (Bhide et al., 2018).

1.2: Defining the Trial Outcome

Selection of primary outcome measures to answer the clinical question is crucial for a successful randomized trial. To avoid multiplicity problems, trials usually select a single primary outcome on which the conclusion of the trial can be drawn. A primary outcome can be defined as the outcome that the researchers deem the most important and provides the most compatible answer for the research question (Ferreira & Patino, 2017). A primary outcome should be well-defined, and always pre-specified as part of the study protocol and statistical analysis plan. This prevents any of the trial team from selectively picking specific results or only results that appear to be statistically significant (Andrade, 2015). The definition and structure of the chosen primary outcome will also influence other critical elements in the trial design process such as the sample

size calculation. Therefore, it is important to establish the nature and definition of the primary outcome at the study conceptualization stage.

Along with the primary outcome, researchers may choose to include a series of secondary or exploratory outcomes in the study. A secondary outcome may be related to the primary outcome and help provide more evidence, but usually cannot provide a decision alone, while exploratory outcomes are typically hypothesis generating (McLeod et al., 2019). A secondary outcome may also be used to capture a singular aspect of the primary outcome. Secondary and exploratory outcomes do not have as large of an impact on the study design of the trial as the primary outcome. The primary outcome informs the power and sample size calculations, whereas additional outcomes are typically not usually used for this purpose.

Because there are many different kinds of clinical trial designs and multiple areas of research study, the nature of a primary outcome can look different in each trial. The following section will discuss the different types and structures of primary outcomes that researchers can use when conducting a RCT.

1.2.1: Types of Primary Outcomes

The type of outcome used in a RCT differs between trials. There is no single outcome that is inherently better or more successful than others, the optimal outcome for a trial is based on the main research question of the trial.

The primary outcome is used to answer the overall research question of interest and determine if the intervention used is beneficial or not. Because each trial conducted is different, the results

that are provided can be more beneficial for or geared toward certain people. This could be the patients themselves, practitioners, or the general population.

Some researchers may choose a primary outcome that is relevant to patients, or the outcome may be directed more towards medical decision-makers. This decision can be used to separate primary outcomes into different types based on the targeted audience.

When it comes to patients of a particular medical illness or condition, the primary outcome that they would be concerned with is one that is clinically relevant to them. A clinically relevant outcome for patients is one that captures how they feel and/or function (McLeod et al., 2019). In this case, the outcome targets what would be deemed most important for patients. For some clinical scenarios, this could for example be to know if survival increased. Knowing if an intervention can prolong a person's life or quality of life is usually the important information for certain diseases. Aside from survival, patients would care about outcomes that can affect how they live their life (ie. quality of life, or other endpoints depending on the area of disease such as the incidence of heart attacks, stroke, hospitalization). These types of outcomes are serious health challenges and would be clinically relevant for patients as they would be something that patients would want to avoid. Providing patients with information from trials that use patient relevant outcomes, allows them to become part of the decision-making process because they are receiving the information that they care about most.

Despite it being the most important information for patients, not all RCTs will use a primary outcome that is clinically relevant. Many RCTs in the oncology field may use a biomarker to indicate cancer progression in patients. Researchers will use an outcome that is considered clinically important for doctors or medical decision makers. This type of outcome is called a surrogate or intermediate outcome. A surrogate outcome is a biological, physical or laboratory

value or measurement that is used in place of a more clinically relevant outcome (Fleming & DeMets, 1996). These outcomes are usually not geared towards patients, but people in the medical field or others who will use the information gained from the results. For example, in a trial that uses HIV patients as participants, the chosen outcome may be the CD4 counts of the participants. The CD4 counts provide more information to physicians than to the patients, but this measurement can be used to infer a more clinically relevant outcome for the patients. The patients would be interested in information regarding their survival time with HIV and not necessarily the associated lab results. The surrogate provides a "snapshot" that is considered to be associated with a concomitant clinically important outcome.

1.2.2: Use of Surrogate Outcomes

If patients themselves are more concerned with clinically relevant outcomes, why are surrogate outcomes even used in different clinical trials? Even if surrogate outcomes may not be the optimal outcome for some people, they do provide certain advantages when used as the primary outcome. For certain clinical trials, it may not be possible to include the most clinically relevant outcome as the primary outcome because it could be difficult to measure and increase the cost of the trial (Buyse et al., 2016). Certain endpoints may not be possible to be measured in all environments or doing so would be costly, therefore surrogate endpoints provide a replacement outcome that may be more manageable. Another advantage is the time length of the trial. The use of a surrogate outcome can reduce the length of the follow-up time for the trial (Hahn et al., 2021). This in turn can also benefit the cost of the trial. In addition, using a surrogate outcome can decrease the sample size needed for the trial (Buyse et al., 2016). Some of the more chronic

outcomes may occur at a low incidence which would require a large sample size in order to capture a large number of events. Using a surrogate outcome can allow for more events or measurements to occur without the need for increasing the number of people in order to capture an accurate picture of how the intervention works. A surrogate outcome may be useful when targeting disease progression instead of disease survival. By using a specific measurement or biomarker, a trial could conclude if the intervention appears to impact disease progression or does not make a difference. This outcome would allow for beneficial information without the needed extra resources for a larger trial investigating survival, or a clinical outcome that may take years to occur.

Surrogate outcomes may be more efficient in certain situations; however, they do have some disadvantages. A key disadvantage is that a surrogate outcome may not directly translate to a clinically important outcome. When using a clinical trial to evaluate a new medication, the surrogate outcome may be easier to observe, but the surrogate outcome may not capture other outcomes (D'Agostino Jr., 2000). Because medical conditions can be complex, a surrogate outcome may only explore one aspect of the condition and not the true impact of disease survival. Another problem can arise when interpreting the results of the outcome. Because a surrogate outcome provides a result that is less directly relevant, the interpretation of the result needs to focus on only what is there to report (D'Agostino Jr., 2000; Heneghan et al., 2017). The interpretation should not be used to infer a result with regard to a more clinically relevant outcome, as it may not be directly related.

The primary outcomes that have been discussed thus far have all been a singular outcome that was chosen by the research team. The only difference has been the area of focus and the target audience of the trial and outcome. There are other formats of primary outcomes that can be used

in a RCT. One alternate format is to combine multiple primary outcomes or endpoints. This could combine the different types of primary outcomes that have already been discussed. The remainder of this thesis will focus on the use of multiple outcomes or endpoints. The definition and use of this type of outcome will now be discussed.

1.3: Defining Multiple Outcomes/Endpoints

Most RCTs use a single primary outcome that encompasses the information that the research team hopes to gain. However, at times a singular outcome may not be able to capture the entire goal of the trial, or the medical problem of concern may be complex. In such cases, multiple endpoints are called for to address the clinical question. Multiple outcomes may also be used in new drug trials when it is unknown what area will demonstrate an effect from the drug or one when endpoint alone will not support approval (U.S. Food and Drug Administration, 2022).

There can be confusion surrounding the definition in the literature, some may consider multiple outcomes or endpoints as when a trial uses a primary outcome and one or more secondary outcomes (Snapinn, 2017). However, for this purpose, the term multiple outcome or endpoint will be used to describe more than one primary outcome used in a RCT. This will be the definition used moving forward.

The challenge when using multiple primary outcomes is that it can cause confusion surrounding the original research question (Vetter & Mascha, 2017). If too many outcomes are being tested, the overall goal of the trial may be lost and what the research team is actually trying to prove may be unclear. Confusion can also arise when trying to interpret the results when there are too many outcomes used as the primary outcome. One overall interpretation may not be possible

when there's multiple outcomes, therefore more than one interpretation may be needed. This may not aid in answering the research question.

1.3.1: Types of Multiple Outcomes

The term multiple outcome or multiple endpoints is an overall term that can have a different definition depending on the chosen structure of the trial. Ristl et al. (2019) conducted a literature search and identified three commonly used terms regarding the use of multiple endpoints. The first being a **combined endpoint**. This term is defined as an overall single measurement for each participant that is composed of several observations in a participant (Ristl et al., 2019). The second term used was **co-primary** endpoints. This term is defined by multiple outcomes that need to be affected by the treatment under investigation in order to establish that it causes an effect in either direction (Ristl et al., 2019). For example, in order for a new vaccine to be approved it may need to demonstrate both a decreased incidence of disease and elicit an immune response. This example requires a positive effect for both outcomes in order to be deemed successful. In other words, the test for each endpoint must reach significance for the trial to be declared positive.

The last term is **composite endpoint or outcome**. A composite outcome combines multiple individual events into one singular outcome using a scoring system (Ross, 2007). For example, one may define that only one of the individual events needs to be observed in a participant in order to classifying them as having the outcome of interest. This type of outcome is often used in cardiology based RCTs. A composite outcome could be the occurrence of a cardiac event including myocardial infarction, stroke or death. For the purpose of this thesis, we have decided

that a combined outcome is equivalent to a composite outcome and therefore, will just use the term composite outcome henceforth.

The definitions that have been discussed in this section are the definitions used moving forward in this thesis. In the next chapter, the types of multiple outcomes will be discussed more in-depth.

1.4: Objectives of This Thesis

The goal of this thesis is to conduct a literature search for the use of multiple outcomes in randomized controlled trials. Considering the information surrounding the use of multiple outcomes and the confusion that can arise in regard to the analysis and interpretation, this thesis also aims to provide an example analysis of a RCT that used multiple outcomes. Hence, there are two specific objectives for this thesis:

1) To determine the prevalence of the use of multiple primary outcomes in randomized controlled trials found in the major medical journals and how the outcomes were analyzed. This timeframe included trials published between July 2020 and July 2021.

2) To demonstrate how the results of a randomized controlled trial using multiple outcomes can be analyzed with the method of global tests using data from a previously conducted randomized controlled trial in the field of gastroenterology.

To achieve the first objective, a literature search will be conducted. The major general medical journals will be searched as well as the major gastroenterology journals because of the dataset that will be used. All of the RCTs will be identified and then it will be recorded whether a RCT used a singular primary outcome or multiple primary outcomes. If the trial used multiple primary outcomes, further details will be recorded including the method used to analyze the main outcomes.

To achieve the second objective, a dataset from a previously conducted gastroenterology RCT will be used, specifically a RCT in ulcerative colitis (UC; Feagan et al., 2005). The multiple primary outcomes that were used in the trial will be re-analyzed using different global test statistical methods for multiple outcomes.

Chapter 2: Literature Review

2.1: Introduction

This chapter will begin by providing more information regarding multiple outcomes. It will begin by discussing two different types of outcomes within the overall topic of multiple outcomes that will be of most interest for this thesis. The first will be **composite outcomes** which will include what composite outcomes are and how they are used, advantages and disadvantages, and their interpretation. The second type of outcome will be **co-primary outcomes**. This will also include what co-primary outcomes are and why they are used, and advantages and disadvantages.

The second part of this chapter will focus on methods for analysis. As mentioned previously, there is not a single standard method that is used exclusively for multiple outcomes. Therefore, there are multiple methods that can be used based on the type of outcome (ie. binary, time to event, etc.), and the chosen adjustment. Some of the more common methods for analyzing multiple outcomes, based on the format of the outcome, were chosen and summarized. This provides examples of methods that can be used when dealing with data that uses multiple outcomes as the primary endpoints.

The final part of this chapter will focus on a literature search. The literature search will be used to achieve the first objective of this thesis. It will explore multiple outcomes that are used in randomized controlled trials. The search will focus on how many trials are opting to use multiple outcomes for their primary outcomes, as well as the common methods used in the analysis process. This will include the two types of multiple outcomes that are of interest: composite and co-primary outcomes.

2.2: Different Types of Multiple Outcomes Used in Trials

The term "multiple outcomes" is an umbrella term that encompasses different types of outcomes. The two outcomes that are used most frequently within this category of outcomes are composite outcomes and co-primary outcomes. These two outcomes were chosen to be the focus for the literature review. Both of these terms will be explained further and why they are chosen to be a primary outcome in a randomized controlled trial.

2.2.1: Composite Outcomes

A composite outcome combines multiple individual events into one singular outcome using an established algorithm (Ross, 2007). In the context of binary events, a common scoring system is that only one of the individual events needs to be observed in a participant in order to classifying them as having the outcome of interest. This is the definition that will be used going forward when discussing a composite outcome.

2.2.2: Structuring a Composite Outcome

Many clinical trials have adopted the use of composite outcomes as the primary outcome in a trial in order to measure a treatment's effectiveness. A composite outcome could include combining multiple serious adverse events into an overall adverse event outcome. For example, in a cardiology RCT, the researchers may choose to create a composite outcome combining death, myocardial infarction and stroke. The outcomes of death, myocardial infarction and stroke

would be the individual components of the overall outcome. Another common scoring system is to use the coefficient from a linear regression to combine different items into a score. For example, the Crohn's disease activity index (CDAI) was developed using a multiple linear regression approach to combine multiple symptoms and clinical variables (Best et al., 1976). In particular, the calculation of CDAI is based on a symptom diary maintained by the patient for 7 days before evaluation. The weights for the eight determinants are as follows with weighting factors in brackets:

- Number of liquid stools in the past 7 days [×2],
- Severity of abdominal pain (scale from 0 to 3), [average in the past 7 days [×5]],
- General wellbeing (scale from 0 to 4, where 4 is 'terrible'), [average in the past 7 days [×7]],
- Presence of additional complications [number of complications [×20]]:
    - Arthritis or arthralgia;
    - Inflammation of the iris or uveitis;
    - Presence of skin disease: erythema nodosum, pyoderma gangrenosum, or aphthous ulcers;
    - Anal fissures, fistulae or abscesses;
    - Other fistulae; or
    - Fever in the past 7 days (body temperature higher than 100 °F100 °F / 37.8 °C37.8 °C)
- Use of antidiarrheal medication in the past 7 days (yes: 1 point, no: 0 points), [×30],
- Presence of an abdominal mass, where 0= none, 2=questionable, 5=definite [×10],
- Decreased hematocrit [×6]:

    - {47%−Hct (for men)42%−Hct (for women)
- Body weight standard weight [×1]

At the end we sum up the points to assess CDAI score.

How the outcome is recorded and evaluated during the statistical analysis can vary between trials. There is not one standard format for structuring a composite outcome. Some researchers may choose to record the outcome as time to event data. This format involves comparing the participants' time to one of the component events between the different trial arms. In this case, a participant only needs to experience one of the individual component outcomes. Alternatively, the outcome could be recorded in a binary format. This involves dichotomizing the overall outcome into participants who experienced the outcome and those who did not. As with the time to event format, participants only need to have experienced one of the individual component outcomes. This format will determine an event rate in each arm of the trial that can be compared. Other researchers may use a count format and count the number of events that occur in each arm of the trial to compare and evaluate. This process of analyzing the results of a trial using a composite outcome in the various formats will be discussed in a later section of this thesis.

The use of a composite primary outcome in a clinical trial to evaluate the effectiveness or efficacy of a treatment can have many advantages in the structure and design of the trial, as well as in the analysis process. However, there are disadvantages as well when choosing this structure for a primary outcome. These will now be discussed.

2.2.3: Advantages of Composite Outcomes

The choice of using a composite primary outcome does provide a number of advantages when it comes to the clinical trial design. The first one being that it eliminates the need for ranking a variety of outcomes and choosing just one for the primary outcome, avoiding the multiplicity issue (Freemantle et al., 2003). If all of the chosen outcomes for investigation are clinically relevant, it may be difficult to establish which is "the most important". By combining the chosen outcomes into a composite primary outcome, researchers no longer need to rely on a personal or random decision for choosing the primary outcome.

A composite outcome will increase the efficiency of the trial. Because of increasing medical care advancements, adverse and severe events for some medical conditions have become less frequent and survival time to event has increased (Ross, 2007). This means that the event rate in the trial will be low. In turn, the sample size would need to be large in order to accumulate enough events in each arm of the trial for comparison. By combining individual events into a single outcome, the event rate will increase, and this will make it easier to evaluate the differences between the arms of the trial (McCoy, 2018). Being able to detect more events, will increase the statistical precision and in turn, the efficiency of the trial (Freemantle et al., 2003).

The increased event rate now impacts other areas of the trial. The combined outcome decreases the sample size needed because the event rate is now large enough without the need for extra participants (Freemantle et al., 2003). This also results in a decreased follow-up period because separately the individual outcomes would require a longer follow-up period in order to accumulate enough events (Freemantle et al., 2003). Decreasing the size of a trial will increase efficiency of the trial, and it will also decrease the cost (Ross, 2007). Acquiring adequate funding

can be a challenge, therefore, it may be advantageous for researchers to conduct a clinical trial with a lower cost (Ross, 2007). The use of composite outcomes allows for this while still maintaining efficiency.

The decision to use a composite outcome creates a 'chain of events' as demonstrated above with a decrease in sample size and cost. They provide certain benefits and may make a trial that was not feasible, now possible (Ross, 2007). Therefore, composite outcomes are often a necessary choice for an outcome because they do not require as large of a trial compared to using single components of the composite and the trial can be conducted more quickly. This may be useful when the results are highly anticipated.

2.2.4: Disadvantages of Composite Outcome

Despite the numerous advantages that a composite outcome has, there are some disadvantages as well. When considering a composite outcome formatted as a time to event structure, complications can arise when it comes to recording and the analysis. Composite outcomes in this format typically group together any severe adverse events, including mortality or indicators of disease progression. Some trials may indicate that once one of the chosen events, that isn't fatal, is observed the participant does not have to remain in the trial (Chi, 2005). This decision does not affect the overall primary analysis but will affect the individual component analyses. When using a composite primary outcome, the individual components should always be investigated (Freemantle et al., 2003). The components can be recorded as secondary outcomes and their results should be reported along with the primary outcome results (Freemantle et al., 2003). If a participant's data is censored after experiencing one event, then that is the only data that is

available to analyze for the individual components. This can bias the individual component results, if a participant is more likely to leave the trial in one arm of the trial compared to the other (Chi, 2005). Participants should remain in the trial after one event for the entirety of the follow-up period in order to collect sufficient data on all of the individual components and minimize the risk of bias in the results (Chi, 2005).

Another disadvantage of using a composite primary outcome, arises when it comes to interpreting the results from the statistical analysis. The interpretation process will now be explained further.

2.2.5: Interpretation of the Outcome

Unlike with a single component primary outcome, interpretation of results from a composite outcome can lead to challenges. There is not one approach when it comes to interpretation (McCoy, 2018). Just using the overall measure of effect, as if it was a single outcome may not always be the best approach. When a trial is investigating a new drug and using a composite primary outcome, the final result may not be truly reflective of the drug's effect (Chi, 2005). This presents a challenge for clinicians and decision makers who are supposed to use these results in their clinical decision making.

Because a composite outcome is made up of individual components, there can be a discrepancy on the clinical importance of each component. In general, for a composite outcome, the results determine the overall effect when the individual components are of similar importance (U.S. Food and Drug Administration, 2022). However, the components may vary in severity, or some

components may be more relevant to patients than others. Therefore, an overall estimate of effect may not be truly representative. The result that is produced from the primary analysis, whether positive or negative, only applies to the components as a whole and not individually (Freemantle and Calvert, 2007). The results individually may vary between the components (Cordoba et al., 2010). Each component does not have to contribute equally to the overall result. One component may have a very large effect and pull the overall estimate in either a positive or negative direction. Cordoba et al. (2010), provide an example of a trial evaluating a drug and the composite outcome of death or chest pain. The result showed a large reduction in the outcome. They stated the result could be attributed to a decrease in death and chest pain, or the results could have been based strictly on a decrease in chest pain and no change in death or the rate of death could have even increased (Cordoba et al., 2010). Therefore, the results may not be representative of the whole picture when there are multiple components involved.

Not all trials use a composite primary outcome, but for those that do, should the results be ignored because of this confusion? Should a composite outcome be used when looking to make a medical decision? There is not one clear answer for these questions, but a composite outcome still provides valuable information as long as the reader interprets the results appropriately. McCoy (2018) provides some details on when a composite outcome should be used in the decision-making process for medical professionals. In a composite outcome, if the individual components are of approximately the same level of clinical importance to patients, then the result can be used between the components as long as the result is clinically and statistically significant (McCoy, 2018). In this situation the outcome could be used for recommendations. If the components change in the level of importance and are not clinically relevant to the patients, medical professionals should be cautious when deciding to use these results (McCoy, 2018). The

results may also be less informative if there is a gradient in the frequency of each component, with there being more events in the less important components and few or no events in the more important components (McCoy, 2018). In summary, the result from a composite outcome can provide valuable information and aid in making decisions when it comes to patient care. However, knowing how to interpret the results, observing the frequency of components and the relative importance of each component is critical to interpretation of the trial results.

2.2.6: Co-Primary Outcomes

The second type of outcome, within multiple outcomes, that will be discussed is co-primary outcomes. As with the use of a composite outcome, sometimes the overall goal or effect of interest cannot be accurately captured using a single outcome. By combining outcomes, it can provide more relevant information, which helps to determine the overall effect.

The definition of co-primary outcomes is a collection of outcomes that must all demonstrate an effect in order to classify the intervention of interest as having an effect. Therefore, in order to state that a treatment or intervention is beneficial, all outcomes must show a positive effect. A negative result occurs when the data acquired fails to provide a positive effect in all of the individual outcomes. If even one of the individual outcomes presents a negative result then the overall result becomes negative.

This definition of co-primary outcomes distinguishes the key difference between composite and co-primary outcomes. With a composite outcome only one of endpoints/components needs to

occur or have an effect, whereas with a co-primary outcome, all of the endpoints/components need to occur or show an effect (Hamasaki et al., 2018).

The advantages of using co-primary outcomes in place of a single primary outcome will now be discussed.

2.2.7: Advantages of Co-Primary Outcomes

The advantages of using co-primary outcomes as the chosen outcome of interest for a study, are similar in principle to that of a composite outcome. Both of these outcomes allow for a more in-depth view on how the intervention of interest is beneficial or not and encompasses more of the objective of interest. With co-primary outcomes, the individual outcomes may focus on different areas of interest. For instance, with a co-primary outcome that is composed of two separate individual outcomes, each outcome can target something different and be two completely separate outcomes, but both are important to measure for the condition of interest. The structure of this outcome allows the research team to gain information on two different areas while evaluating the effectiveness, leading to a more comprehensive understanding of the intervention of interest. For example, when studying infectious diseases, co-primary outcomes can be used to focus on both antimicrobial use and a clinically relevant outcome for the patient regarding safety or length of illness (Gillespie et al., 2018). Co-primary outcomes are often used in trials that are evaluating a new medication for approval because of the fact that multiple individual outcomes need to show effectiveness, to classify it overall as having an effect.

2.2.8: Disadvantage of Co-Primary Outcomes

As with the advantages discussed above, the disadvantages of using this format for a primary outcome are similar to that of a composite outcome. Because there may be more than one individual outcome being tested and each has to reach the significance level, maintaining the power of the study is an important consideration. Consideration of additional outcomes decreases the probability of detecting an effect (Chuang-Stein and Li, 2017). Losing power in a study because of this results in the same consequences that the sample size must increase to compensate for the loss of power, which in turn affects other areas of the study.

Another disadvantage to this type of outcome is that it can be more challenging to establish a treatment effect. Unlike with other outcomes that only need to test one hypothesis in order to find out if the result is significant, each outcome included in co-primary outcomes is tested with all outcomes needing to demonstrate a significant effect. This may deter research teams from using this type of outcome.

2.2.9: Comparison Between a Composite Outcome and Co-primary Outcome

Even though both use multiple endpoints, there are some differences between composite and co-primary outcomes. The analysis of a composite outcome looks at more of an overall view of the outcome. In most cases, the individual components of the outcome are analyzed separately during a secondary analysis and are not included within the primary analysis. In contrast, the individual outcomes in co-primary outcomes must be analyzed during the primary analysis because their results are needed in order to determine the result for the primary outcome.

Therefore, based on this idea, co-primary outcomes may provide more specific information initially, as a secondary analysis is not needed to evaluate the individual component outcomes as with a composite outcome.

Another issue that commonly arises when using multiple outcomes is the need to make an adjustment in order to control for the type I error. This issue is relevant in trials for which significance will be declared if at least one outcome reached significance, in this case the type I error needs to be controlled. A type I error occurs when a null hypothesis is incorrectly rejected when it is actually true for that dataset (Wason and Roberston, 2021). This conclusion results in a false positive. When more than one independent outcome is being tested, the type I error rate increases. When the type I error rate increases and no adjustments are made when conducting multiple tests, a multiplicity problem arises (U.S. Food and Drug Administration, 2022). This problem increases as the number of outcomes used goes up. However, the multiplicity problem can be managed by making the necessary adjustments during the statistical process.

An adjustment for the type I error is needed when using multiple primary outcomes, when detecting if there is an effect on at least one of the outcomes. However, an adjustment to control for the type I error is not needed when using co-primary outcomes (Hamaski et al., 2012). In this case, it is to determine if there is an effect on all of the outcomes. All of the individual outcomes should be tested at the same level of significance. The type I error does not need to be adjusted for, but the type II error does. A type II error occurs when we fail to reject the null hypothesis, however, there actually is a difference and the null hypothesis should be rejected (Sedgwick, 2014). This results in a false negative conclusion. When using co-primary outcomes, as the number of individual outcomes increases, so does the type II error (Hamaski et al., 2012). The type II error will also affect the power of the study, as a result, unless it is necessary to capture

the outcome of interest, using more than two co-primary outcomes is not recommended because of the loss of power (U.S. Food and Drug Administration, 2022). Therefore, it is important to remember any adjustments that may be needed when choosing the quantity of individual outcomes to include in a primary outcome.

2.3: Analyzing a Composite Outcome/Multiple Endpoint

Analyzing multiple outcomes, regardless of the structure, can be completed using various methods. Currently, there is not one standard method that is used widely. The most common methods include analyzing all outcomes separately as would be required with co-primary outcomes, a multiple testing approach such as global tests and, a single overall analysis, the latter two being used with composite outcomes (Table 1). The method that is chosen is based on what the original research question is asking (O'Brien and Geller, 1997). Is the research team looking for improvement in any one of outcomes or global improvement? Or is the team interested in knowing which individual outcome provides an effect? The research question itself should dictate what appropriate should be taken.

Table 1. Summarizing the main advantages and disadvantages when choosing between the standard methods that are used to evaluate data when there are multiple outcomes.

Analysis Method	Advantages	Disadvantages
Single overall analysis via composite outcome	Provides one overall result on the effectiveness of the intervention compared to the control. The statistical method can be more straightforward.	Does not provide any information on the individual components in order to determine which individual outcomes had an effect. Interpretation is also challenging when trying to apply a single result to more than one outcome. Lastly, there needs to be a valid scoring system.
Global test	Allows for a singular overall result, while considering the correlation between the outcomes. Creates a result by combining the results from individual tests. No scoring system needed.	The statistical methods may be more complex.
Separate individual analyses	Ability to determine separately if the individual outcomes have an effect. The results are easier to interpret as they are all separate.	There is not an overall statistical result. An overall result needs to be inferred from the individual tests performed. It is more difficult to establish a positive result because all individual results must demonstrate a positive result in order to state there is a positive effect overall.

Many researchers choose to combine the data and analyze the results as a single overall outcome, such as with a composite outcome. In this scenario, methods for one singular outcome can be applied. Using this idea, the individual components would need to be evaluated individually in a secondary analysis to gain individual information, as the primary analysis does not allow for individual interpretation. However, descriptive analysis of the individual components does not

provide confirmatory evidence (Schüler et al., 2014). The individual components within composite outcomes or multiple endpoints are correlated with each other. Each patient in a trial would have a set of data for each of the individual component outcomes. This means that these measurements are correlated with each other because they come from the same patient. When conducting the statistical analysis, if a method such as a simple linear or logistic regression is used, the correlation of these outcomes will not be accounted for. This can lead to inaccurate representation of the treatment effect and an inaccurate estimation of the variance (Tilley et al., 1999). In this situation there are multiple hypotheses being tested at the same alpha, which increases the probability of falsely rejecting a hypothesis. To prevent this, a multiple testing approach can be used. Multiple testing is used in order to seek information on the individual component outcomes found in composite or multiple outcomes, while controlling for the family-wise type I error rate (FWER; Ristl et al., 2019; Schüler et al., 2014).

The process of using a multiple testing method begins by creating an overall hypothesis, while taking into account the individual components. Ristl et al. (2019), define the term global test as an overall hypothesis with the null hypothesis being no treatment effect on any of the component outcomes and the alternative hypothesis of a treatment effect in at least one of the individual component outcomes.

When there are no well-established scoring systems for a composite outcome, global tests may be of value. By using a global statistical test (GST), a single overall test statistic is produced. A breakthrough for GST was the paper by O'Brien (1984). He proposed three very flexible methods to construct powerful tests. The first two are parametric methods, while the third one is nonparametric. In the parametric framework, O'Brien proposed to standardize the outcome variables such that different endpoints can be handled, followed by using ordinary least square or

weighted least square method to construct tests. O'Brien favored the non-parametric method using a rank-sum test and recommended the non-parametric method for general use. The test entails three steps. First, for each outcome, all patients are ranked according to their values on this outcome. Second, for each patient, the patient's ranks are summed to obtain a rank score. Finally, a t-test (or a one-way ANOVA if more than 2 groups are involved) is applied to compare the rank scores between the treatment and control group.

Using a GST allows for the correlation between outcomes when determining the treatment effect, which provides a more accurate representation of the treatment effect and variance estimate. A GST also has a higher power relative to single outcome tests or other multiple outcome testing procedures, when the intervention is shown to provide a benefit in each component outcome (Huang et al., 2009). Note that all GSTs focus on hypothesis testing rather than estimation of treatment effect.

2.3.1: Hierarchical Method of Analysis

The added complexity of multiple outcomes can present as a challenge. When the individual components vary in their relevancy and importance, the overall result cannot be directly applied individually. Due to the complications surrounding the interpretation of the results when multiple outcomes are present such that the determined overall result cannot be applied to the individual components, the strategy of multiple testing should be used. Using this strategy, the overall hypothesis for the outcome can be included as well as singular hypotheses for the individual components (Buyse 2010; Ristl et al., 2019).

Another method that can be used within multiple testing, is a hierarchical method. This choice of method is beneficial when is a large discrepancy in the clinical importance or estimated effect of the individual components, allowing for more individualized results. Using a hierarchical method requires an ordering of the components to be determined. The ordering is based on which component is expected to provide the largest effect. The hierarchical ordering method described by Schüler et al. (2014), begins with a hypothesis test of the overall outcome followed by each individual component beginning with the chosen top components, at a pre-stated significance level. This is a stepwise process in that the first hypothesis must be rejected before moving on to the next and the process is completed when the hypothesis cannot be rejected. This process is limited in that it strongly relies on the "correct" order of the components being chosen during the planning stage.

Similar to this hierarchical ordering method is the Bonferroni-Holm test. This test follows the same sequential process as the method described above, except a pre-determined order is not required and the significance level is not the same for each test. Using the Bonferroni-Holm test, the order that the components are evaluated in depends on their observed p-values and begins with the component with the smallest p-value (Schüler et al., 2014). As with the hierarchical order, analysis occurs in a stepwise process and components are evaluated only if the hypothesis is rejected, otherwise the procedure ends. Instead of evaluating the components at the same significance level, they are tested at adjusted significance levels that include α/n for the first hypothesis, $\alpha/(n\text{-}1)$ for the second hypothesis and continues to α, where n is equal to the number of components (Holm, S., 1979; Ristl et al., 2019). The adjustment of the significance levels is needed to control the global type I error rate (Schüler et al., 2014).

For the two methods described above, no assumptions are made about the underlying distribution of the outcome or the components. When more information is used in the analysis, the power of the test is increased, and it provides a less conservative approach. A different take on the Bonferroni approach is the Simes test. The Simes test is a semiparametric approach based on the Bonferroni method (Dmitrienko et al., 2012). The test is based off of the order of the p-values of the individual tests and tests the overall global hypothesis of the intersection of the individual hypotheses (Simes, 1986; Dmitrienko et al., 2012; Ristl et al., 2019). From the Simes test, it can be determined that at least one of the null hypotheses is false, but not which one (Dmitrienko et al., 2012). The Simes test provides greater power than the classical Bonferroni approach and is better suited when multiple there are multiple highly correlated test statistics (Simes, 1986).

Presented above are just some of the analysis methods that can be used when evaluating a composite outcome. In the literature there exists many variations and additions to previously established procedures. Each variation seeks to expand the previous method and add improvements to increase power, while controlling the FWER and providing a less conservative approach. As discussed, some methods utilize the underlying distributions in the procedure, while others are a non-parametric approach and make no assumptions on the distributions. Therefore, the choice of which method to use is at the discretion of the research team, and the overall goal of the result they hope to achieve. Select parametric and non-parametric methods will be discussed further in the next chapter.

2.3.2: A Time-to-event Outcome

One of the most common outcome methods used in RCTs is a time to event outcome. A time to event composite outcome may use death and other clinically relevant causes of death, but it may use non-fatal components such as hospitalization, disease progression or surgery. When using a time to event outcome, the most widely used method of analyzing is using the time to the first event experienced from the pre-determined individual events defined in the composite outcome (Ristl et al., 2019). This method is beneficial in RCTs with a small sample size because, as previously discussed, this method will increase the event rate for the outcome and in turn increase the power of the study.

Another method for a composite outcome involving time to event, is the method of a win ratio by Pocock et al (2012). This method begins by ordering the outcomes according to clinical importance, denoting as Y_1, Y_2, …, Y_m. The comparison between the treatment and control is done through the comparison of all pairs of patients. In each pair, one patient is taken from the treatment group while the other is from the control group. This way, every patient in the treatment is compared with every patient in the control, each time noting who 'won' in terms of Y_1, the most important outcome. If a winner cannot be declared, because of a tied score, or because of a censored event time (both or the smaller event time is censored), then compare the pair in terms of Y_2. If a tie occurs, compare them in terms of Y_3, etc., until a win or lose can be declared. If the pair is tied even when the comparison reaches Y_m, then the comparison is discarded. The win ratio is estimated by the number of wins by patients in the treatment group to the number of wins in the control group. See Redfors et al. (2020) for more guidance on the win ratio analysis.

2.3.3: A Binary Outcome

A large portion of RCTs use a binary composite outcome in contrast to the time to event outcome. A binary composite outcome combines multiple individual binary outcomes. Each individual outcome is dichotomized into yes/no to classify if a patient experienced the event or not. As with the time to event outcome previously described, the individual outcome variables can be combined into a new overall outcome variable. In this case, as with the time to event outcome, a patient only needs to experience one of the individual components to be given the "yes" status of the outcome; indicating that they experienced an outcome. If a patient were to experience multiple of the individual events, events that occurred after the first would be censored and would not be included in the primary analysis.

When a single overall variable is used, single outcome analysis methods can be used. In this case, logistic regression can be used to evaluate the difference between the treatment and control arms of the trial. This would provide an odds ratio summarizing the effect of the intervention on the outcome of choice between the two trial arms. A logistic regression model can also take into account the correlation factor between the outcomes. A logistic regression model can first be used assuming all of the observations for each patient in the trial are independent (Baraniuk et al., 2012). Using generalized estimating equations (GEEs), the logistic regression model can be refit, changing the coefficients and taking into account the correlation (Baraniuk et al., 2012). The odds ratio produced is now adjusted for the correlation and can provide information on whether to reject the null hypothesis or not, in the global test.

Another method that can be used for binary data is an exact method, such as Fisher's exact test. This method evaluates based on 2x2 tables. When using an outcome of multiple binary

outcomes, in order to control for the type I error rate exact tests can be based on the joint distribution from the multiple outcomes (Ristl et al., 2019).

Most trials that use a composite primary outcome will evaluate each of the individual outcomes in the secondary analysis to aid in interpretation. This allows the researchers and also the readers to see where the overall effect comes from, and which individual outcome may have been pulling the final result. The individual results also provide information for clinicians which outcome was affected the most and this may help aid in decision making for patient care.

Rauch and Kieser (2012), have proposed an alternative method that eliminates the need for the individual secondary analysis. This method combines the overall analysis with the analysis for the individual outcomes to all be completed in the primary analysis. Rauch and Kieser (2012), use a multiplicity adjustment method, to evaluate the composite outcome and all of the individual components at various predefined local alpha levels. The predefined alpha level for the composite outcome is selected to be close to that of the family wise alpha level and the individual components are evaluated at the remaining level (Rauch & Kieser, 2012). The local levels are determined and calculated using the correlation structure found between the composite outcome and the components.

2.3.4: Methods for Co-primary Outcomes

What makes the process for co-primary outcomes different than composite outcomes is that an overall effect measure is not needed. Co-primary outcomes need to evaluate all outcomes separately to determine for which outcomes the intervention was more successful or not compared to the control. Without the individual effects an overall answer to the research

question cannot be established. As mentioned previously, the definition of co-primary outcomes dictates that all included outcomes must demonstrate a positive effect to state that there is an overall positive effect. Based on this information, any statistical method could be used that compares groups to each other. For example, a simple t-test could be used to compare the intervention and control arms of the trial. The method chosen is at the research team's discretion and the data collected, as long as the included outcomes are evaluated separately.

2.4: A Literature Search of Recently Published Randomized Controlled Trials

The rest of this chapter will focus on the first objective of this thesis. The first objective of this thesis is to determine the prevalence of multiple outcomes used in a sample of randomized clinical trials. Multiple outcomes are often used when more information is needed to determine if a treatment is truly beneficial. Certain areas of study deal with diseases or other health problems that are more complex, and the use of a singular outcome may not cover the entirety of the health problem. The use of multiple outcomes can provide more detail in order to accurately capture all aspects of the disease or health related problem. In addition, multiple outcomes may be necessary to provide a complete and total understanding of the efficiency of the intervention.

To observe how often multiple outcomes are used, various medical journals will be searched to incorporate enough information. As discussed previously, because there are various ways of structuring multiple outcomes including co-primary outcomes and composite outcomes, therefore gathering information on the most common types that are used in randomized controlled trials will be of interest.

The second part of this objective is to determine how the multiple outcomes are being analyzed. There is not one standard method for analyzing multiple outcomes and some methods may not account for there being multiple outcomes. Observing the common statistical methods used and any adjustments used to account for multiple outcomes in the methods section, will be of interest.

2.4.1: Selecting the Journals for the Literature Search

In order to determine how many randomized clinical trials are using multiple outcomes, and also how they are being analyzed, a search of the major medical journals was conducted. In order to have a broad representation, a range of different medical journals were chosen to incorporate journals that focus on different medical areas and expertise. This decision allows us to determine if one area in particular tends to make use of multiple outcomes more than others, and also ensures that an adequate number of randomized controlled trials will be found.

First, the general medical journals were chosen. These journals included *The BMJ*, *The New England Journal of Medicine* (NEJM), *The Journal of the American Medical Association* (JAMA), and *The Lancet*. Several journals were also chosen from the field of gastroenterology, including *Gut*, *Gastroenterology*, *Annals of Internal Medicine*, and *The Lancet: Gastroenterology and Hepatology*. These journals were added to the search because the dataset that will be used later in this paper to demonstrate the analyzation of multiple outcomes, was taken from a gastroenterology RCT.

2.4.2: Search Criteria Used

The selected journals were searched for RCTs that were published, using the timeframe of one year. This timeframe included trials published between July 2020 and July 2021. This timeframe ensured that the information provided from the search captures the current use of multiple outcomes in trials and how they are analyzed.

Any trials from the search were determined by a combination of the article title and the abstract. Many trials were identified strictly by the title, however for any trial that did not include "randomized controlled trial" in the title, the abstract was read. Trials were excluded if they were a secondary analysis of a previously published trial, or a preliminary data trial published before the primary outcome data had been reached. All of the trials that met the criteria were reviewed, with a focus mainly on the methods section of each trial. A flow chart demonstrating the search process is included in an appendix.

2.4.3: Collecting Information from RCTs

After determining all of the clinical trials from the chosen journals, trials were categorized into 12 different categories based on the area of study. The categories included: neurology, cardiology, oncology, infectious disease, gastroenterology, nephrology, general surgical, obstetrics, mental health, optometric, lifestyle and medical practices. The category "medical practices" was used as a general category for any trial that did not fit into any of the other chosen categories and investigated a topic that would have an impact on the usual protocols/procedures in healthcare. For each clinical trial, it was recorded what type of primary outcome was used, if it was a singular outcome or a composite or co-primary outcomes. Information regarding the

format of how the primary outcome was recorded (ie. binary, time to event, categorical), was also documented. If a trial used a composite primary outcome, additional information was recorded. This information included how many individual components were included in the composite outcome, and how the sample size was determined and how the outcome was analyzed.

2.5: Results for Composite Outcomes

This section will focus on the results from the literature search. The literature search conducted allowed for the inclusion of different types of multiple outcomes used. The trials were separated into composite primary outcomes and co-primary outcomes. Each of the different types of outcomes found within the randomized controlled trial search will be discussed in their own section.

2.5.1: Composite Outcomes Used

First the results from the literature search regarding composite outcomes used in RCTs will be discussed, including how many were found and in what category, the number of individual components, and the structure of the outcome.

2.5.2: Prevalence of Composite Outcomes in the Major Medical Journals

From the literature search there were 291 trials found in total among all of the journals. Of these 291 trials, 53 (53/291, 18.2%) of them used a composite primary outcome. The number found within each selected journal is summarized below (Table 2).

Table 2. Summarizing the number of RCTs that were found in the literature search for each journal that used a composite primary outcome.

Journal Title	**Number of Trials that used a Composite Outcome**
The Lancet	21
JAMA	17
NEJM	6
The BMJ	6
Annals of Internal Medicine	2
Gastroenterology	1
Gut	0
The Lancet: Gastroenterology and Hepatology	0

2.5.3: Categories Used and the Number of Individual Components

The number of trials varied across the 12 chosen categories. The category that had the highest number of trials with composite primary outcomes was cardiology, with 15 trials. This result was expected as the majority of trials in the field of cardiology use a composite primary outcome

(Neaton et al., 2005; Ferreira-González, Busse et al., 2007). In the category of cardiology, the average number of individual component outcomes was approximately four, with the smallest number being two and largest being seven. Components in this category included major cardiac events such as stroke, myocardial infarction, angina, hospitalization, and death.

The second highest category was oncology with 11 trials. The average number of individual components for this category was approximately three, with the lowest being two and the highest being five. Components in this category included disease progression or reoccurrence and death.

Both the infectious disease category and medical practices category had seven trials in them. In the infectious disease category, the average number of individual components in an outcome was approximately three, with the lowest number being two and the highest being four. The components in this category included hospitalization, death and duration of hospitalization or oxygen use. In the medical practices category, the average number of individual components was three with the lowest number being two and the highest being nine. Because this was a general category, the individual components varied between the trials. Some of the components included hospitalization, death, transplant, and long-term oxygen therapy.

The category gastroenterology had four trials with composite primary outcomes. The average number of individual component outcomes was approximately three with the largest number being four and the lowest being two. The components included disease progression, infection, change in liver or kidney function and, death.

The category of nephrology and the category of general surgical, both had three clinical trials in them. The average number of individual components in the composite outcome in the nephrology category was approximately four, with the lowest number being three and the highest

being six. Some of the individual components included kidney injury progression, initiation of dialysis, rejection, and death.

In the general surgical category, the average number of individual components for the outcomes was five, with the lowest number being two and the highest being ten. Individual components in this category included the need for a blood transfusion, requirement for ventilation, post-operative complications or infections, and death.

The category of obstetrics had two clinical trials in it. One trial had a composite outcome made up of three individual components and the other trial had five. These individual components included fetal infection, fractures, brain injury, fetal or neonatal death including stillbirth.

Lastly, the category of neurology had one trial with a composite primary outcome. This trial had four components in the composite outcome and some these components were hospitalization and the need for additional treatments. The categories that did not have any trials that used a composite primary outcome were mental health, optometric and lifestyle.

2.5.4: Sample Size Calculation

As discussed previously, one of the main advantages of using a composite primary outcome is that it increases the power in a study to detect a difference between the groups. A study's power is directly correlated with the study sample size, in that, as the sample size increases, the power does as well. Unlike with studies that use a singular primary outcome, there is not one universal formula to use when calculating the sample size required for a trial that uses a composite primary outcome. The trials that were found with a composite outcome during the literature search, were analyzed to determine how the studies were powered and how the sample size was determined. It

was found that the majority of the trials used a power of 80%. There were 27 trials that used this power in their sample size determination. The next most common power used was 90% with 11 trials. The power with the third highest number of trials was 85% with five trials using this power in order to determine the sample size needed. The remainder of the trials had a power within the range of 75% to 99%.

A key component to determining the sample size needed for a trial, is the assumed or predicted incidence or event rate of the outcome of interest. The number of assumed events or event rate assumption usually comes from previous clinical trials that were conducted with the same or similar outcome(s) of interest. Having an assumed event rate for each arm of the trial, allows the researchers to determine the possible difference or measure of effect between arms. The proposed difference between arms is used to determine the required sample size. The use of a composite primary outcome in a trial may add a level of difficulty to this process. Using the chosen trials from the literature search, it was recorded how the authors recorded the assume event rate in the arms. Authors could have looked at the outcome as a whole and included an event from any of the chosen individual outcomes or look at each component outcome separately with its own event rate. Out of the 53 trials that used a composite primary outcome, only two trials used the individual outcome event rates in order to determine the sample size needed. The remainder of the trials used the assumed event rate for the outcome as a whole and did not include the event rate for the individual components.

2.5.5: Types of Composite Outcomes Found

Information regarding the type of outcome was also recorded. This included the format of how the composite outcome was recorded. For a composite primary outcome, the majority of trials chose to use a time to event outcome. When using a time to event outcome, time would be recorded, such as days, until one of the outcomes included in the composite outcome occurs. For this literature search, there were 26 (26/53, 49.1%) clinical trials that used a time to event outcome for recording any outcomes that occurred. The second highest format was binary. This format involves determining if a participant in the trial had one of the chosen outcomes or did not; a yes/no format. There were 24 (24/53, 45.3%) clinical trials that used a binary format for the outcome. There were two trials that used a count format for recording the outcome. Lastly, one trial used a categorical format for recording the outcome.

The results from the literature search regarding how the analysis was handled in the clinical trials will now be discussed.

2.5.6: Most Common Overall Method Found in the Literature Search

When analyzing a composite there are three overall methods that can be used. These include analyzing individual components separately, using a global test, or analyzing as a singular outcome. Within each, the methods discussed previously, or various other methods can be used. The literature search conducted provided information on which overall method is used most commonly when working with a composite outcome.

Among the trials that were found in this literature search, there was only 1 (1/53, 1.9%) trial that used a hierarchical method to evaluate the primary outcome. This trial used the win ratio to evaluate the composite outcome of death, duration of hospitalization, and the duration of supplemental oxygen use.

The most common method of analysis was to treat the composite outcome as a singular outcome. This combines all of data for each component and compiles it into a single overall outcome. This allows the primary outcome to be evaluated as a single outcome, while each individual component can be analyzed separately as a secondary outcome.

2.5.7: Analysis Methods Used Based on Outcome Type

The different types of outcomes used allowed for different analysis methods. For the time to event trials, the method of analysis was the Cox proportional hazards model or Cox proportional hazards regression. This method was used for the majority of the time to event trials. One trial used the win ratio (Pocock et al., 2012). In contrast to the Cox proportional hazards model that is used often with time to event or survival data, one trial that was found used the versatile test. This method combines the use of three log-rank tests which cover different time points in the trial (Menon et al., 2021). Outcomes that were binary, used multiple analysis methods. The most common method of analyzing a binary outcome was logistic regression. There were nine trials that used this method and one that used a Bayesian logistic regression model. The next most common method of analysis was a generalized linear model, which allows for a selection of the link function based on the data. There were three trials that used this method of analysis. The other methods found only included one or two trials and they were: Chi-squared test, Fisher's

exact test, Z-test or t-test, Global test, Mantel-Haenszel test, negative binomial regression, non-parametric analysis. Lastly, one trial that used a binary outcome was a non-inferiority trial. This trial used a confidence interval method in order to determine if the treatment is in the margin of interest. The two trials that used a count outcome for the composite outcome, used two different analysis methods. One trial used a Z-test to evaluate the outcome and the other used a binomial regression model. The last outcome format used was categorical. The trial with this format used Cochran-Mantel-Haenszel test to analyze the primary outcome data.

2.5.8: Summary of Composite Outcomes Found

In summary, there was a total of 53 trials found that used a primary composite outcome. These trials were found among RCTs pulled from different medical journals. The category that had the largest number of trials with a composite primary outcome was the cardiology category. This result was expected as RCTs that deal with a cardiac issue often use multiple outcomes in order to gather more information.

There were different methods used to analyze the outcomes. A time to event outcome was the most common one when the trials were reviewed. Therefore, a Cox proportional hazards model was a common method used in the data analysis process.

In the next section of this chapter, the information gathered from searching the trials found for those that used co-primary outcomes will be discussed.

2.6: Results for Co-Primary Outcomes

In addition to the composite primary outcomes discussed previously, the prevalence of co-primary outcomes used in RCTs that were found in the literature search will now be discussed. The search used the definition provided previous that in order to be classified as co-primary outcomes, all individual outcomes included must demonstrate a positive effect in order to classify the intervention as having a positive effect. In addition to this, the individual outcomes must be analyzed separately in order to determine the individual effects. Trials were not counted as co-primary outcomes if they did not meet these criteria, even if a trial states the use of co-primary outcomes. With this definition there were 10 (10/291, 3.4%) trials found that used co-primary outcomes.

Beginning with the medical practices category, this category had the highest number of trials that used co-primary outcomes. There were six trials found that met the definition of co-primary outcomes.

The category with the next highest number of trials with co-primary outcomes was gastroenterology. This category had two trials that used co-primary outcomes.

Lastly, both the categories of neurology and musculoskeletal each had one trial that met the definition of co-primary outcomes.

The remainder of the categories that were not mentioned did not have any trials that fulfilled the definition of co-primary outcomes. It is not surprising that there were so few trials that used co-primary outcomes. Because of the nature of co-primary outcomes, a level of difficulty is added compared to other forms of outcomes because of the fact that all individual outcomes have to achieve an effect. This could deter others from choosing this type of outcome. For some it may

be easier to choose an outcome that will demonstrate if there is an effect or not with one analysis only.

When searching through the RCTs found in the literature search, there were some trials that stated the use of co-primary outcomes, however upon further reading it was determined that the trials did not actually meet the requirements to be classified as co-primary outcomes. Two trials in the oncology category were found that did not appear to meet these requirements. The first trial used co-primary outcomes and did evaluate the individual outcomes separately and provided the individual results, however one of the outcomes was found to not be statistically significant and the intervention was still recommended (Sweeney et al., 2021). In this case, because one outcome was not statistically significant the intervention should not be recommended, regardless of the results of the other outcome. The second trial also stated the use of co-primary outcomes, however in the article there was no discussion on the need for both outcomes to demonstrate an effect as with other trials that used co-primary outcomes (Mittendorf et al., 2020).

In both of the categories of nephrology and gastroenterology, there was one trial that also appeared to be misclassified based on the definition used. The trial in the nephrology category described the use of co-primary outcomes and did analyze each of the individual outcomes separately, however in the article it discussed that there were two possible ways to achieve a positive result (Zarbock et al., 2020). According to the definition set previously, the only way to achieve a positive result is by obtaining a statistically significant result in all of the individual outcomes, therefore there should only be one way to obtain a positive result. The article found in the gastroenterology category was similar in the information that was found. This trial also appeared to use co-primary outcomes, but upon further reading in the trial they provided a

situation where one endpoint could be used to prove the intervention is beneficial (Mingrone et al., 2022).

2.6.1: Number of Individual Outcomes

When using co-primary outcomes in a RCT as the outcome of interest, the majority of research teams will limit the number of individual outcomes included to two. As mentioned, when more outcomes are added, it decreases the power and as a result the sample size must increase, which can affect other aspects of the trial like funds and resources. Therefore, two appears to be the most common number of individual outcomes that are included.

For the literature search conducted, the number of individual outcomes included was recorded. All of the trials found, across the different categories, that used co-primary outcomes, had two individual outcomes that were included. There were no trials found that used more than two individual outcomes in the multiple endpoint.

2.6.2: Methods Used to Analyze Within Each Category

The only requirement for analyzing co-primary outcomes is that the individual outcomes need to be evaluated separately. The method of analysis chosen is strictly dependent on the type of outcome that is used such as binary, continuous, time to event etc. The general methods include linear regression for continuous outcomes, logistic regression for binary outcomes and a Cox proportional hazards model for time to event data. Variations or adjustments may be made to these methods to better fit the dataset. There were six different methods used to evaluate the ten

co-primary outcomes that were found. In each trial, the same method was used for each of the outcomes included to determine their result.

There were three trials that used the Cochran-Mantel-Haenszel test to evaluate the outcomes. Each outcome was evaluated separately with both results provided as is required by the definition of co-primary outcomes used.

The methods of analysis of covariance (ANCOVA) and a mixed effects model, a form of linear regression, both had two trials each that used these methods. Lastly, the methods of a Cox proportional hazards model, Fisher's exact test and repeated measures least squares regression, all had one trial that used these methods.

Because the methods used for analysis were used separately for each outcome included, any adjustments needed for the inclusion of multiple outcomes is not needed. The outcomes are evaluated just as the process would be for a trial with a singular outcome. The trials stated prior to results that both outcomes would need to be significant for the intervention to be declared effective or in the results said that the co-primary outcomes were met.

2.6.3: Summary

After searching through the trials that were identified, all of the trials that used co-primary outcomes were identified. Co-primary outcomes were used less frequently compared to that of composite primary outcomes in RCTs. The category that had the largest number of co-primary outcomes was not the same as the category with the largest number of composite outcomes.

There were also a number of different methods that were used in the data analysis process for the trials found, with the most common method being the Cochran-Mantel-Haenszel test. Further discussion on the results found will continue in the final chapter of this thesis.

The next chapter will go more in-depth into some methods that may be used when analyzing multiple endpoints with the global test approach. To demonstrate the use of some of these methods, a dataset from a previously conducted trial will be used. The next chapter will focus on objective two of this thesis.

Chapter 3: Analyzing Randomized Trials with Multiple Outcomes of Different Types

Chapter 2 reviewed randomized controlled trials from the years 2020 to 2021. We found that 63 (63/291, 21.6%) of trials assessed co-primary and composite outcomes. It is well known that the key issue with trials of multiple outcomes is multiplicity problems, i.e., multiple tests without proper adjustment can increase type I error rates. There are three popular approaches to deal with multiplicity. First, pre-specify a single primary outcome, leaving the rest as secondary outcomes. Second, combine multiple outcomes into an aggregated outcome. Third, use a multivariate method to conduct a global test for all outcomes simultaneously. Each approach has its advantages and disadvantages. Using a singular outcome with multiple secondary outcomes, provides an easier approach when it comes to the analysis process because there is only one outcome to account for, however, a singular outcome may not capture all of the important aspects of the intended outcome. This could mean that the RCT does not provide enough detail regarding the outcome of interest. If multiple outcomes are combined into one, it allows for more information to be captured in the trial including the number of events that occur, but combining multiple outcomes that have varying levels of importance can lead to challenges when it comes to interpreting the overall result. Using a global test to simultaneously analyze all of the chosen outcomes is advantageous as it provides a singular method, but challenges can arise with the challenging statistical process. However, because the two defining features of trials with multiple outcomes are: i) outcomes are different types, ii) outcomes are usually correlated. Therefore, in balance, the global test has a lot to offer.

In our review, 207 (207/291 71.1%) of the RCTs found analyzed a singular primary outcome and kept any other outcomes of interest as secondary outcomes, 53 (53/291, 18.2%) of the RCTs

used a composite outcome created by combing multiple outcomes, and 10 (10/291, 3.4%) of the RCTs found used co-primary outcomes in the trial.

The purpose of this chapter is to popularize global tests by illustrating these methods using a previously published trial—MLN02. The organization of the chapter is as follows. We provide a brief description of this trial in Section 3.1. We then review global tests in Section 3.2, followed by the analyses of MLN02 in Section 3.3. We close the chapter with a discussion. The SAS code used for the analyses is presented in the appendix.

3.1: A Brief Description of the MLN02 Trial

Ulcerative colitis (UC) is a part of the group of inflammatory bowel diseases (IBD; Feagan et al., 2005). It causes inflammation of the colon and rectum and can cause abdominal cramping and bloody diarrhea. Treatments for illnesses included within IBD usually involve treating and managing the symptoms that occur with these illnesses and preventing complications such as surgery and hospitalization.

Treatment research for UC has investigated how to inhibit the recruitment of leukocytes that are brought to the inflamed area during the body's inflammatory response. The $\alpha_4\beta_7$ integrin is involved in the recruitment of leukocytes to the gut tract and is found on the surface of certain circulating T lymphocytes. There is a major ligand for the $\alpha_4\beta_7$ integrin found on the intestinal endothelium. It was considered that blocking the interaction between this integrin and the ligand may be an effective treatment for improving IBD conditions. MLN02 is an antibody that recognizes the $\alpha_4\beta_7$ integrin. Feagan et al. (2005), conducted a clinical trial to assess the use and effectiveness of treating UC patients with MLN02 therapy. This treatment was subsequently

shown to be effective in a phase 3 registration trial and is now a routine treatment for UC in the clinic.

3.1.1: Design and Outcomes in the MLN02 Trial

This was a randomized, double-blind, placebo-controlled trial, phase 2 trial. Eligibility for the study was adults with active UC. The trial arms consisted of two treatment groups and one placebo group. There were two different doses of MLN02: 0.5mg of MLN02 per kilogram and 2.0mg of MLN02 per kilogram. Using permutated blocks of three, 181 patients were randomized with 58 randomized to receive 0.5mg of MLN02, 60 randomized to receive 2.0mg of MLN02 and 63 randomized to receive the placebo.

To assess baseline disease conditions and final disease status, four different measurement scales were used. The first was the ulcerative colitis clinical score which is composed of four elements including rectal bleeding, stool frequency, patient assessment and overall physician assessment. Each individual component is scored from zero (normal) to three (severe disease). Combining all four of these individual components produces a score from zero (inactive disease) to twelve (severe disease activity). The next measurement scale used was the modified Baron score. This score attempts to classify the disease state endoscopically using a scale ranging from zero to four, with zero classifying normal mucosa, one granular mucosa with an abnormal vascular pattern, two friable mucosa, three microulceration with spontaneous bleeding, and four being gross ulceration. The third measurement scale was the Riley histopathological score, which has scores ranging from zero (no inflammation) to seven (severe acute inflammation). The last measurement score was based on a patient reported outcome using an inflammatory bowel

disease questionnaire. The scores from the questionnaire ranged from 32 to 224, with larger scores equating to a higher quality of life.

Patients in the trial were evaluated at weeks one, two, four and six, following randomization. The primary outcome of this trial was clinical remission status, a composite outcome, at six weeks after randomization. The status of clinical remission was defined by an ulcerative colitis clinical score of zero or one, in addition to a modified Baron score of zero or one. The secondary outcomes were the changes from baseline in each of the measurement scores. The primary outcome was analyzed using the Cochran-Mantel-Haenszel chi-square test.

3.1.2: Summary Statistics for the Outcomes

The results at the end of six weeks, showed that more patients in the two treatments groups achieved remission status compared to patients in the placebo group, with 33% of patients in the 0.5mg of MLN02 group achieving remission, 32% of patients in the 2.0mg of MLN02 group and only 14% of patients in the placebo group. When comparing each treatment group to the placebo group, the difference in the proportion of patients who achieved remission was statistically significant in each intervention group. The p-value was 0.020 for both the 0.5mg group and 2.0mg group when compared to the control group. These results indicated that use of MLN02 is an effective treatment in patients with UC.

Moving forward, only the 0.5mg intervention group will be used to demonstrate any effect that the intervention provides compared to the control group. The four individual outcomes will be indicated as $y1$ for the Riley histopathological score, $y2$ for the modified Baron score, $y3$ as the

ulcerative colitis clinical score, and lastly, *y4* will be a patient reported outcome pertaining to quality of life.

3.1.3: Summary of the Four Individual Outcomes

The three individual outcomes consisting of the Riley histopathological score, the modified Baron score and the ulcerative colitis clinical score, are a measure of disease of severity. These widely used scales allow for a universal classification on the current state of disease experienced by a patient and each measures a different, but important, facet of the disease. The way the scales are designed, a more severe or further progressed disease state receives a higher score. A low score indicates a more desirable state of disease or remission. Therefore, if a treatment is being tested, you would expect to see a decrease in these scores in order to declare it beneficial.

The last individual outcome was a patient reported outcome using an inflammatory bowel disease questionnaire. The other outcomes are focused on severity and state of disease, but this outcome explores a different aspect of the disease including how a patient is living with the disease including emotional and physical ability. The questionnaire is known as IBDQ-32 and is composed of 32 items and covers the four areas of: emotional and social function, and bowel and systemic symptoms (Yarlas et al., 2020). Each of the 32 items are scored on a 7-point scale with one representing the highest symptom occurrence or severity and seven representing the lowest symptom occurrence or severity (Yarlas et al., 2020). Based on this 7-point scale the scores from the overall questionnaire can range from 32 to 224. In contrast to the other three outcomes, improvement is indicated with a higher score on this scale. Therefore, to conclude that a

treatment or intervention is beneficial for a patient, we would expect to see an increase in the questionnaire scores (Yarlas et al., 2020).

There is a difference of direction when comparing the individual outcomes to each other. The aim of the outcomes focusing on disease severity is to achieve a low score overall, but the patient reported outcome aims to achieve a higher score. This makes it challenging to conduct any analyses on them as they are not being compared on the same scale. To compare them, the questionnaire scores were changed to negative values. For example, if a patient had a score of 32, the new score would be -32. Now all of the scores from the outcomes can be compared on the same scale. When looking at the negative data for the patient reported outcome, to determine if a participant in the intervention group had a beneficial score at the end of the trial, the negative score must be changed to its absolute value. Once the absolute value is used, the score follows the rules of the scale in that higher scores perform better.

Histograms were created for each individual outcome to visualize how the responses for each score differed between the two groups. Figure 1 demonstrates differences in the scores for the Riley histopathological scores between the intervention and control groups. Looking at just the control group in this graphic, the data appears to have more of a left skew with more participants having scores on the higher end of the scale. There did not appear to be many participants with a low Riley histopathological score in the control group at the end of the trial. Looking at the treatment group, there were more participants with scores on the lower end compared to the control group. The treatment group doesn't appear to follow perfectly a specific distribution, but there is a very slight right skew to the data. The larger peaks in the treatment group are closer the left side of the distribution. Because a lower Riley histopathological score indicates a more favourable state of disease, it appears that the participants in the treatment group performed

better than the control group. This result is consistent with the overall effect described previously.

Figure 2 compares the modified Baron scores between the treatment and control groups. Beginning with the control group, the distribution appears to be fairly symmetrical and normally distributed. However, the control group had more participants with higher scores compared to the treatment group. The treatment group has a very slight right skew in the data because the majority of participants in this group had scores on the lower end of the scale and very few participants on the right side of the distribution with higher scores. Because a lower modified Baron score indicates a better outcome, it appears that the treatment group performed better than the control group did. This histogram also appears to support the overall effect found.

The third histogram (Figure 3) compares the treatment and control groups' ulcerative colitis clinical scores. The data for the control group does not appear to follow a distinct distribution. It does not follow the bell shape of a normally distributed dataset. The distribution of the treatment group is right skewed. The majority of participants in this group had scores on the lower end with only a very few participants with higher scores. The control group had more participants with higher scores. There were less participants with the optimal small scores compared to the treatment group. Therefore, for the ulcerative colitis clinical scores it appears that the treatment group performed better than the control group.

The last histogram compared the scores from a patient reported outcome. The values for this outcome are negative to ensure it is being compared on the same scale as the other outcomes. The distributions for both groups are fairly similar, however the data for the control group is shifted to the right more than the treatment group. The treatment group has a higher frequency of participants with scores on the left side of the distribution. Because of the negative scale used,

when you take the absolute value of the scores, the treatment group has a higher frequency of larger scores which corresponds to a better outcome, suggesting that the treatment performs better than the control.

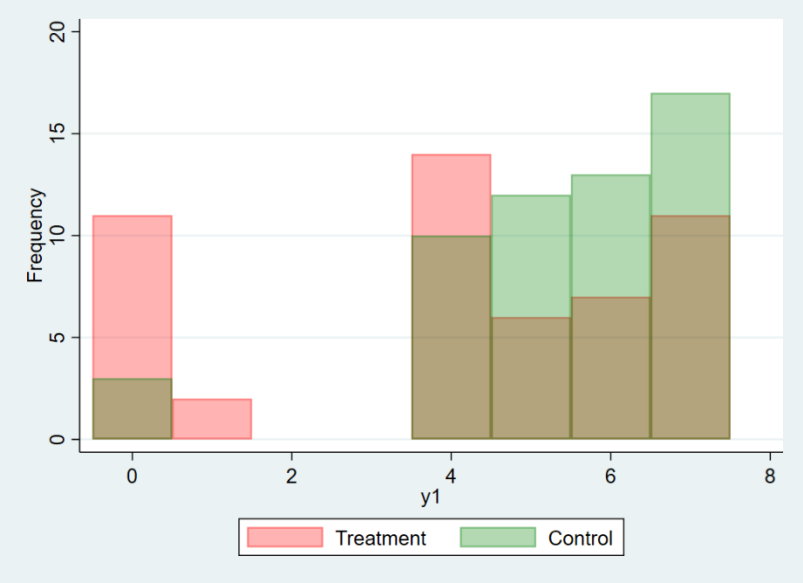


Figure 1. This histogram compares the frequency of the Riley histopathological scores of participants at the end of the trial in the placebo group and the treatment group receiving 0.5mg of MLN02.

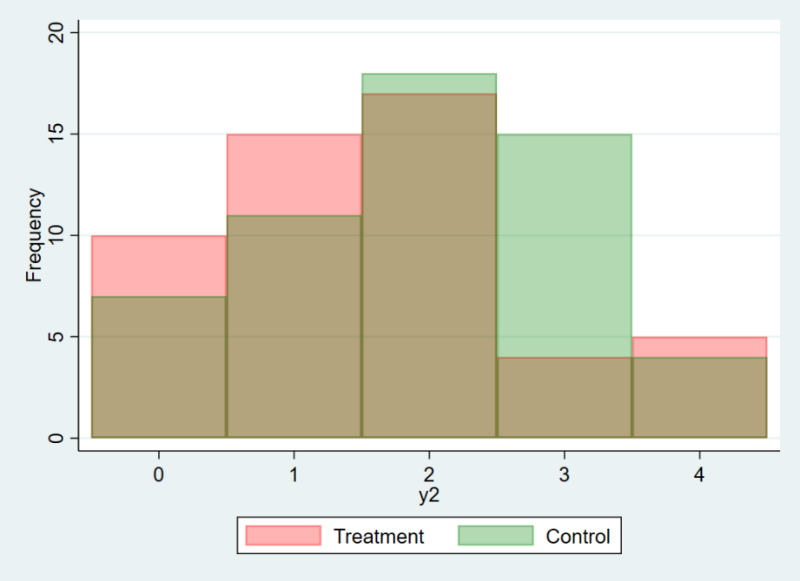


Figure 2. In this histogram the frequency of participants' modified Baron scores at the end of the trial are compared between the placebo group and the 0.5mg of MLN02 treatment group.

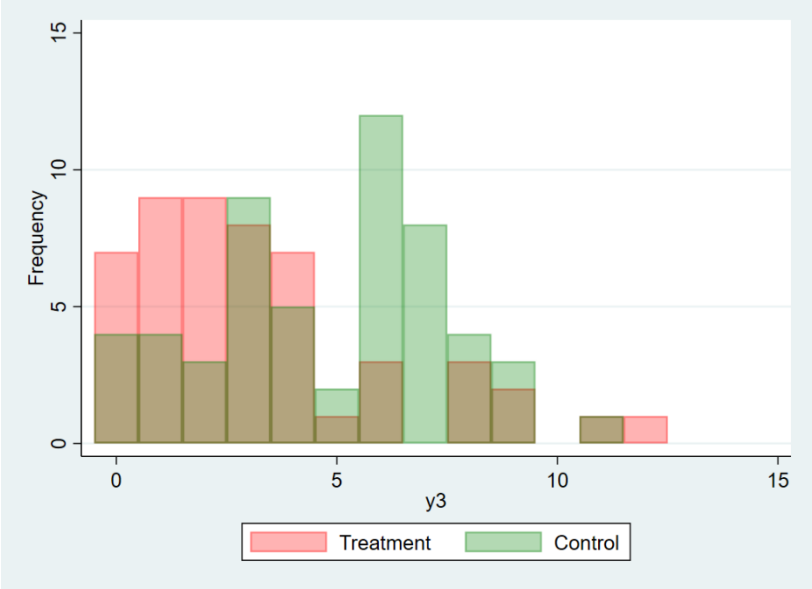


Figure 3. A histogram comparing the frequency of the ulcerative colitis clinical scores at the end of the trial between participants in the placebo and participants who were in the 0.5mg MLN02 treatment group.

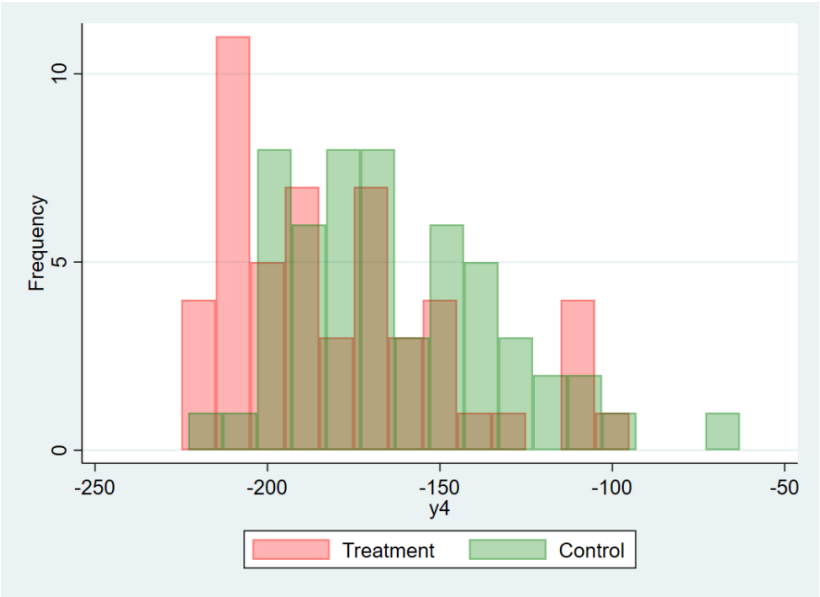


Figure 4. This histogram compares the frequency of patient reported outcome scores at the end of the trial between the participants in the placebo group and the participants who received 0.5mg of MLN02 in the treatment group.

3.1.4: Correlation Among the Outcomes

Unlike with a single outcome, when using multiple outcomes, correlation can play a factor. The individual outcomes can be correlated with each other. In the original MLN02 trial, the correlation among the outcomes was not investigated. Therefore, the correlation among the outcomes was calculated. Both the Pearson correlation (Pearson, 1920) and Spearman correlation (Spearman, 1904) were determined.

The Pearson correlation coefficient represents the linear relationship between two variables. The correlation coefficient is a scale-free and unitless measure. The coefficient measure can range from -1 to 1. The closer a coefficient is to either end of the range, indicates an ideal “perfect” linear relationship. A negative or positive sign dictates the direction of the linear relationship. A

positive correlation indicates that both x and y (both variables of interest) increase. A negative correlation occurs when y decreases as x increases. The correlation coefficient is calculated by dividing the covariance of the two variables of interest by the product of their standard deviations.

The Spearman correlation is a more robust way of determining the correlation, compared to the Pearson correlation. The Pearson correlation is sensitive to outliers, whereas the Spearman correlation uses a rank-based system that better tolerates any outliers. The Spearman correlation orders the values from least to greatest and assigns the values a rank based on this order. Assigning ranks makes any outliers only one unit higher than the previous value, preventing the outlier from influencing the correlation. Any ties that are found in the ranking process are assigned the average rank between the two tied values. As with the Pearson correlation, the Spearman correlation can take on any value between -1 and 1.

Using the MLN02 trial data, the Pearson and Spearman correlation were calculated to compare the four individual outcomes. As stated previously, the four individual outcomes include the Riley histopathological score (y1), the modified Baron score (y2), the ulcerative colitis clinical score (y3), and lastly, a patient reported outcome pertaining to quality of life (y4). The sample size for this data was 106 with 51 participants in the treatment group and 55 in the control group.

Beginning with the Pearson correlations, all of the correlations were above zero which indicates that there is a relationship between the individual outcomes (Table 3). The correlations were all positive, therefore as one outcome score increases so does the score of the other outcome. Based on these outcomes, this result can be expected as all of the individual outcomes pertain to an aspect of the symptoms experienced with UC with a higher score indicating worsening disease. Therefore, if one outcome is increasing then the other outcomes are also likely to increase with

worsening disease. None of the correlations found were very close to one, therefore there is only a moderate positive linear relationship among the individual outcomes. The two outcomes with the smallest correlation between them were the Riley histopathological score and the patient reported outcome score. The largest correlation was between the ulcerative colitis clinical score and the patient reported outcome score. All of the Pearson correlations were statistically significant with all p-values less than 0.001.

The Spearman correlations were similar to the Pearson correlations. All of the correlations were above zero, indicating a positive linear relationship so as one outcome score increases, the other outcome also increases (Table 4). In contrast to the Pearson correlations, the smallest correlation amongst the Spearman correlations was between the modified Baron score and the patient reported outcome score. The largest correlation was between the ulcerative colitis clinical score and the patient reported outcome score, which was the same for the Pearson correlations. The highest Spearman correlation is slightly higher than the largest Pearson correlation. As with the Pearson correlations, none of the Spearman correlations were very close to one, meaning there is only a moderate positive linear relationship among the individual outcomes. All of the Spearman correlations were statistically significant, with a p-value less than 0.001.

Table 3. Pearson correlations between the four individual outcomes of the Riley histopathological score, the modified Baron score, the ulcerative colitis clinical score (UCCS), and a patient reported outcome based on the IBDQ-32.

	Riley Score	**Modified Baron Score**	**UCCS**	**IBDQ-32**
Riley Score	1.000	0.471	0.511	0.421
Modified Baron Score	-	1.000	0.622	0.422
UCCS	-	-	1.000	0.685
IBDQ-32	-	-	-	1.000

Table 4. Spearman correlations between the four individual outcomes of the Riley histopathological score, the modified Baron score, the ulcerative colitis clinical score (UCCS), and a patient reported outcome based on the IBDQ-32.

	Riley Score	**Modified Baron Score**	**UCCS**	**IBDQ-32**
Riley Score	1.000	0.526	0.572	0.459
Modified Baron Score	-	1.000	0.641	0.416
UCCS	-	-	1.000	0.693
IBDQ-32	-	-	-	1.000

3.2: A Brief Review of Global Test for Multiple Outcomes

The MLN02 trial presents two key features that are common in practice. First, because the outcomes are of different types, i.e., one point has a different meaning depending on the endpoint, we cannot simply add scores within each subject and conduct analysis on the patient-specific sum of scores. The primary outcome is a composite outcome composed of different symptoms (i.e., bleeding, stool frequency), endoscopy and overall physician assessment. Each component of the outcome has a different clinical importance and uses a different scale, which prevents the measurements from being added together to create an overall score for each participant. Second, because the outcomes are correlated within subjects, conducting separate analyses for each outcome and then adjusting for multiple p-values using methods such as Bonferroni may lead to loss of power. Thus, we focus on methods that are appropriate for correlated outcomes of different types.

Both parametric and nonparametric methods have been proposed in the literature. Throughout this chapter, we use the following notations:

- Index *i* will be used for treatment group, *i*=0 for control group and *i=1* for treatment group
- n_i= the number of subjects in group *i*
- Index *j* will be used for subjects, $j = 1, 2, \dots, n_i$
- Index *k* will be used for the outcomes, $k = 1, 2, \dots, K$
- X_{ijk}= outcome *k* for the *j*th patient in the *i*th treatment group

3.2.1: Parametric Methods

By parametric methods, we refer to methods that rely on mean scores of the outcomes. These methods were developed commonly by assuming raw scores are normally distributed. In particular, O'Brien (1984) assumed that outcomes from the same subjects have a multivariate normal distribution. Outcomes among subjects are independent. Two versions of a test were developed by O'Brien to answer the question 'Is the treatment beneficial to the subjects?'. Because outcomes are made up of different types, both versions used the standardization of raw scores, but with different ways in the combination of tests constructed from the standardized scores. These two tests are now commonly referred to as O'Brien's OLS (ordinary least square test) and GLS (generalized least square test).

The null hypothesis for these tests is that there is no difference between the treatment and control. The alternative hypothesis is that at least one of the individual outcomes differs between the treatment and control. O'Brien's OLS is conducted with the following steps:

1) Convert outcome X_{ijk} to a standardized score Y_{ijk} by subtracting the outcome-specific mean and then divide by the pooled estimate of within-treatment standard deviation s_k, i.e.,

$$Y_{ijk} = \frac{X_{ijk} - \bar{X}_k}{s_k}$$

2) Conduct a two-sample t-test on Y_{ijk} for each outcome. Denote this by t_k.

3) Estimate correlation matrix R using all of the standardized scores Y_{ijk}.

4) Construct OLS by sum of t_k, and then divide the sum by the sum of all elements in R.

$$T_{ols} = \frac{\sum_{k=1}^{K} t_k}{\sum_{k,k'=1}^{K} R_{k,k'}}$$

Note that t_k may be expressed in terms of outcome raw scores X_{ijk} or the standardized scores Y_{ijk} as follows:

$$t_k = \frac{\frac{\bar{X}_{1.k} - \bar{X}_{0.k}}{s_k}}{\sqrt{\frac{1}{n_1} + \frac{1}{n_0}}} = \frac{\bar{Y}_{1.k} - \bar{Y}_{0.k}}{\sqrt{\frac{1}{n_1} + \frac{1}{n_0}}}$$

5) Refer to OLS to t-distribution with degrees of freedom of $n_0 + n_1 - 2$ to get p-values.

O'Brien's GLS is conducted by considering the weighted sum of the outcome-specific t-test statistics, with weights given by the column totals of the inverse of correlation matrix R, i.e.

$$T_{GLS} = \frac{\sum_{k=1}^{K} w_k t_k}{\sum_{k=1}^{K} w_k}$$

where w_k is the sum of elements in column k of the inverse correlation matrix R. Under the null hypothesis, T_{GLS} is distributed as a t-distribution with degrees of freedom of $n_0 + n_1 - 2K$ (O'Brien 1984).

Pocock et al. (1987) extended O'Brien GLS to handle combinations of binary, continuous, and survival outcomes, by replacing t_k with the appropriate tests for binary and survival outcomes.

One limitation of O'Brien's GLS and Pocock's test is that a weighted value may be negative. A negative treatment difference weighted by a negative weight may mislead an investigator to conclude efficacy in favor of a treatment, even though the treatment is worse than a control. In

addition, simulation results by Sankoh et al. (1999) have shown that OLS performed better than GLS in terms of controlling type I error rates. Therefore, we will focus on the OLS test.

Lauter (1996) proposed a class of tests that can be applied to trials with multiple outcomes. The methods also start by standardizing the outcomes. Instead of standardizing by the pooled within-group standard deviation as the O'Brien global tests, Lauter standardized each outcome by the standard deviation of both groups. This can be easily done using PROC STANDARD in SAS with options of mean=0 and STD=1. Since the standard deviation used by Lauter is no less than the within-group standard deviation used by O'Brien, Lauter's tests are usually smaller than O'Brien's test.

Simulation results by Logan and Tamhone (2004) suggest that Lauter's test performed well, especially when the effect sizes among outcomes do not vary dramatically, otherwise O'Brien's OLS is more powerful.

3.2.2: Nonparametric Methods

O'Brien (1984) also proposed a rank-sum test for multiple outcomes of different types. The test entails the following steps:

- Obtain outcome-specific ranks
- Sum up ranks across outcomes by subjects
- Conduct 2-sample t-test to subject-specific rank-sums

Simulation results by O'Brien (1984) and Sankoh et al (1999) suggest this test performed well without making normality assumption for the outcome scores.

3.2.3: Effect Sizes for Trials with Multiple Outcomes of Different Types

All of the above methods were developed to obtain p-values to answer the question of whether or not the treatment is more beneficial as compared with control. To quantify the extent or magnitude of treatment effect, we need estimates and confidence intervals for effect sizes, as recommended by the CONSORT statement.

As shown in Section 3.2.1, the OLS test statistic is a function of effect sizes, i.e., for outcome k, we can estimate effect size E_k, by the mean difference in standardized scores between two treatment groups. Exact confidence interval estimation under normality has been pointed out by Zou (2007). Since O'Brien OLS also has a t-distribution, we can apply the same method to this test. Specifically, this test may be re-written as:

$$T_{ols} = \frac{\sum_{k=1}^{K} E_k}{\sum_{k,k'=1}^{K} R_{k,k'}} \Big/ \sqrt{\frac{1}{n_1} + \frac{1}{n_0}}$$

Thus, we can refer to O'Brien global effect size as:

$$E_{ols} = \frac{\sum_{k=1}^{K} E_k}{\sum_{k,k'=1}^{K} R_{k,k'}}$$

for which the confidence interval may be used from the SAS code by Zou (2007).

Sample size and power implication with O'Brien OLS can be made clear with effect size E_{ols}. For example, suppose we wish to estimate a sample size for a 1:1 randomization trial with two

outcomes, which are correlated with a correlation of 0.50. The effect size used is also 0.50. Then, for a trial of 80% power at a two-sided 5% significance level, with a single outcome, the minimum total sample size is given by:

$$N = \frac{4\left(z_{\alpha/2} + z_{1-\beta}\right)^2}{\mathrm{ES}^2} = \frac{4 \times (1.96 + 0.84)^2}{0.5^2} = 126$$

In contrast to the trial that uses two outcomes, the $E_{ols} = \frac{2 \times 0.5}{\sqrt{2 \times 1.5}} = 0.577$, thus the minimum total sample size is:

$$N = \frac{4 \times (1.96 + 0.84)^2}{0.577^2} = 96$$

This is a reduction of sample size by 24%.

It is interesting to point out that the effect size is a 1-to-1 function of a more meaningful quantity, which may be defined as the win probability. The win probability is the probability that a randomly selected subject will do better than a randomly selected subject from the control. It is well-known that:

$$\mathrm{WinP} = \Phi^{-1}(E_k/\sqrt{2})$$

where Φ is the standard normal distribution function.

The win probability can be estimated without assuming the normality of the outcomes. The details are beyond the scope of this thesis. We refer the reader to Zou (2021) for theoretical development, simulation results, and SAS code for computation.

3.3: Analysis of MLN02 Trial

This section demonstrates the use of global tests as a way of analyzing trials that have multiple primary outcomes. These methods were not used in the original published MLN02 trial. The chosen tests for demonstration will compare the 0.5mg intervention group to the control group. The null hypothesis used for these tests is that there is no difference between the treatment and control. The alternative hypothesis is that at least one of the individual outcomes differs between the treatment and control. The sample size used was 106, with 51 participants in the treatment group and 55 participants in the control group.

Beginning with Lauter's test, the four individual outcomes were standardized by the standard deviation of both groups. Next, the subject-specific sum of standardized scores were determined, which were used to run Lauter's test. Lauter's test used a pooled method with the assumption of equal variances in the groups and 104 degrees of freedom.

The next method that was demonstrated was O'Brien's OLS test. This method begins by standardizing the endpoints. To accomplish this, the MLN02 data was converted to a long format instead of a wide format. This is useful with data that has repeated measures or multiple variables associated with each participant. Most datasets are usually formatted in wide format, which produces one row of data for each participant in the trial and a different column for each variable measured. In contrast, a dataset in a long format has multiple rows for each participant. In this format, each time point observation is a singular row resulting in one participant having many rows for one variable. Having the dataset in the long format makes the standardization process easier for this test. Next, the test was run in SAS using a mixed-effects model. A separate t-test for each individual outcome was created, as well as an overall estimate. For the individual

outcomes, the t-tests had 104 degrees of freedom and a standard error of 0.194. For the overall test, there was 104 degrees of freedom and a standard error of 0.153.

The last method to be demonstrated is O'Brien's rank sum test. The first step for this method is to get the subject-specific ranks. Then to run a t-test for each of the individual outcomes. The final step is O'Brien's rank test.

3.4: Discussion

The analyzation process started with Lauter's test and the use of standardizing the outcomes first before conducting a test. The t-value that was produced from this test was 3.30 with a p-value of 0.001. This result indicates that at the alpha level of 5%, this result is statistically significant, and the null hypothesis that there is no difference between the treatment and control, can be rejected. Therefore, based on this test, the intervention of 0.5mg appears to be superior to the control group in the effectiveness of treating UC based on the symptom scores.

The second method that was demonstrated was O'Brien's OLS test which had a t-value of 3.31 and the p-value was 0.001. This result indicates that the null hypothesis can be rejected at the 5% level because the p-value is less than 0.050. Therefore, overall, the 0.5mg intervention group had a better outcome compared to the control group. The t-value and p-value produced from this overall test is the same result that was found when conducting Lauter's test. Both tests are using the same data to test the overall effect of the intervention in comparison to the control group, the only difference being in the standardization process. Both of these tests are also parametric methods and therefore make assumptions about the distribution in the original population from which the sample for the trial was taken.

The overall result provides the information that at least one of the individual outcomes differs between the treatment and control group, but this result doesn't provide information regarding which individual outcome(s) are different. With O'Brien's OLS, each outcome was individually tested between the two interventions. The first outcome was the Riley histopathological score. The t-value was 3.200 with a p-value of 0.002. The second outcome was the modified Baron score. The t-value was 1.660 with a p-value of 0.099. The third outcome was the ulcerative colitis clinical score. The t-value was 2.810 and the p-value was 0.006. The last outcome was the patient reported outcome score. The t-value was 2.740 and the p-value was 0.007. For all of the individual outcomes except the second outcome, the modified Baron score, the null hypothesis can be rejected at the 5% level in favour of the alternate hypothesis. This result is statistically significant because the p-values are less than 0.050. Therefore, for three of the individual outcomes, the intervention group had a more favourable result than the control group. For the modified Baron score outcome, the p-value was greater than 0.050 and therefore, we fail to reject the null hypothesis at the 5% level. For this outcome the intervention did not show an improvement compared to the control group.

The last test that was used to demonstrate the use of global tests when there are multiple primary outcomes, is O'Brien's rank sum test. The overall O'Brien's rank sum test had a t-value of 3.450 and a p-value of 0.001. Therefore, the overall result for O'Brien's rank sum test provides information that the 0.5mg treatment group produced better scores for the four individual outcomes than the control group did. This result is statistically significant because the p-value produced is less than 0.050 and the null hypothesis can be rejected at the 5% level. This means that at least one of the individual outcomes differs between the intervention and control groups.

However, as with O'Brien's OLS test, knowing which individual outcome differs is unknown from this test alone.

Each of the individual outcomes were evaluated with a non-parametric t-test. As with the other individual t-tests conducted, there were three individual tests that indicated a positive and one that did not, which was the modified Baron score. The Riley histopathological score had a t-value of 2.820 and a p-value of 0.006. The ulcerative colitis clinical score had a t-value of 3.190 and a p-value of 0.002. The PRO score had a t-value of 3.180 and a p-value of 0.002. These tests all had a p-value less than 0.050 and in turn the null hypothesis can be rejected at the 5% level. For these outcomes, the intervention group provided a better score. The modified Baron score had a t-value of 1.860 and a p-value of 0.066. This p-value is greater than 0.050, therefore we fail to reject the null hypothesis. This result does not indicate that the intervention group performed better than the control group.

Including the individual tests in addition to the overall result test provides more information that may be beneficial to a reader. The overall result can provide a starting point by determining whether there actually is a difference between the investigating groups. However, as demonstrated with the previous analyses, no information about the individual outcomes is provided with this test.

The importance of an intervention should not be based strictly on the statistical significance of one test. Because the overall test is based on only one of the individual outcomes being different, the same result occurs whether it is just one outcome that is different or all of the outcomes. Aside from the statistical significance of a result, it is important to acknowledge and think about the clinical significance of a result as well. A result may indicate that there is difference between the treatment and control groups, but it could be determined that only one outcome out of the

four combined outcomes is actually different. If the effect only differs between the treatment and control group on one of the outcomes, can the treatment really be deemed successful when there was no difference between the two groups for the rest of the individual outcomes? Would one outcome alone be enough to influence decision makers on this treatment? The individual outcomes may also be of varying levels of clinical importance, and some may carry more weight in the minds of people who will use this data. This highlights the point that an outcome shouldn't rely on just the statistical significance alone. An intervention would appear to be more beneficial if all or almost all of the individual outcomes differed between the treatment and control group. This would provide clearer evidence that the intervention is superior to the control. However, this would only be determined by testing the individual outcomes in addition to the overall. This could be a step that is added to supplement the primary analysis and should be considered when using multiple outcomes, as demonstrated with this trial data.

Chapter 4: Discussion

The final chapter provides a discussion of this thesis. This chapter begins with a restatement of the purpose, followed by a summary of what was found and what this thesis can provide moving forward in research.

4.1: Purpose

A randomized controlled trial is considered the gold standard for providing quality evidence regarding a question of interest. The results that are garnered from RCTs, can aid in decision making, provide guidance for treatment plans and help to provide guidelines in the medical community. Therefore, a RCT should have a clear goal of what the research team hopes to achieve.

When a RCT is being designed, one of first key elements that needs to be decided is what will the primary outcome be, and does it capture what the goal of the trial is? There are several formats available to structure the primary outcome, each with their own strengths and weaknesses. When strictly one outcome will not capture what the researchers hope to gain from the trial, additional outcomes may be needed. Using a collection of multiple primary outcomes, can effectively capture all aspects of the intended goal.

Due to the benefits that using multiple outcomes provides, they can be found more readily in some medical specialties compared to others. For instance, multiple outcomes are often used in the cardiology field in order to capture a larger number of events that happen during the trial.

Aside from this area of study, multiple outcomes could be found in any RCT. This leads into the first objective of this thesis. The first objective was to determine the prevalence of multiple outcomes used in RCTs by conducting a literature search of the major medical journals over a timespan of one year. This literature search outlined what areas of research are using multiple outcomes more frequently, as well as the type of outcome being used and the method that is used to analyze.

This leads into the second objective of this thesis. There are various methods that can be used to analyze a primary outcome that is made up of multiple individual outcomes, depending on the approach that is taken. As mentioned, the methods used could involve looking at each individual outcome separately, analyzing as a singular outcome, or use a global test. A global test will provide a singular result, when using multiple correlated outcomes (Tilley et al., 1996). The second objective of this thesis was to demonstrate the use of global tests and how they are used with multiple primary outcomes. This was done by using a dataset from a previously conducted UC RCT.

These objectives were achieved in previous chapters of this thesis.

4.2: Summary

The common types of multiple outcomes that are found in the literature are a composite outcome and co-primary outcomes. A composite outcome combines two or more outcomes, individually called component outcomes, into a singular primary outcome (Dash et al., 2022). If a participant in the trial has an event described by one of the component outcomes included in the composite

outcome, then they would be considered as having experienced the composite outcome (Ferreira-González, Permanyer-Miralda et al., 2007). Participants in each of the treatment groups being tested could be counted as the number in each group who experienced the composite outcome or the time to the experienced event. The other type of outcome is co-primary outcomes, which defines a group of outcomes that need to demonstrate an effect in order to classify an intervention as having an effect (Ristl et al., 2019).

These were the definitions used to conduct the literature search for the use of multiple outcomes in the medical literature. The results from the search were sorted into RCTs that used a singular primary outcome and those that used multiple primary outcomes. The multiple outcomes were further differentiated into those that used a composite outcome and those that used co-primary outcomes. Overall, there were 291 trials found from the literature search. Of these trials found, 84 (84/291, 28.9%) were found to use multiple primary outcomes. Breaking the number of multiple outcome trials down into the separate types, there were 10 (10/291, 3.4%) RCTs found that used co-primary outcomes and 53 (53/291, 18.2%) RCTs used a composite outcome. When searching the results for trials, 21 (21/291, 7.2%) RCTs with multiple outcomes did not meet the definition of co-primary or composite outcomes and were not included in the results from the search as the literature search focused on co-primary and composite outcomes.

There were also numerous methods that were used in the trials that were found. Regarding the trials that used a composite outcome, the majority of the trials structured the composite outcome in a time to event format. Meaning that time was counted for each participant until one of the individual component outcomes occurs. A participant could have experienced one or more than one event, but only one event was necessary to be counted in the time to event. Using this format

of primary outcome, the outcome was analyzed, as a whole, most frequently using the Cox proportional hazards model.

With the trials that used co-primary outcomes, there was one method that was used slightly more frequently amongst the trials, which was the Cochran-Mantel-Haenszel test.

For the composite outcomes the most common overall method used to analyze data was to look at all of the data as a singular outcome. In this situation, in order to gain any information about each individual component, they must be included in a secondary analysis as the primary analysis is used for the outcome as a whole. There was one trial that used a composite primary outcome and utilized a hierarchical method. The method used was the win ratio or win probability. This method takes into account the clinical importance of each individual component and begins with the most clinically important outcome in a hierarchical order. For the co-primary outcomes, the outcomes were analyzed individually to determine the individual effects as is required for co-primary outcomes.

4.3: Results

In the previous chapter, this thesis demonstrated the use of global tests as a method of analyzation for multiple primary outcomes. These methods were not used in the original trial that the dataset was taken from. The original trial aimed to determine if more participants achieved clinical remission in the groups receiving either 0.5mg of MLN02 per kilogram or 2.0mg of MLN02 per kilogram, compared to participants in the placebo group at week 6 of the trial. Clinical remission was defined as an ulcerative colitis clinical score of 0 or 1, a modified Baron

score of 0 or 1 without the presence of rectal bleeding (Feagan et al., 2005). In the original trial, a binary composite outcome was defined by using two of the outcomes and the analysis was done using the Cochran-Mantel-Haenszel chi-square test. For illustration of the global test, the data was analyzed using Lauter's test, O'Brien's OLS and O'Brien's rank sum test. All of the methods were testing the overall null hypothesis that the number of participants achieving clinical remission does not differ between the groups. However, with the global tests demonstrated, only the 0.5mg MLN02 per kilogram treatment was used to compare to the placebo group.

It is important to compare the results that were produced using this different analysis method to the originally published results. This will help to determine if the results gained are similar to that of the previous trial or if the produced results provide any new information on how the intervention affected the participants.

In the original trial, the percentage of participants in remission was compared in each individual intervention group to the placebo and overall. Each comparison was statistically significant at the 5% level, with the overall comparison producing a p-value of 0.030. This p-value is different than the p-value produced by the global tests, but the result is still the same as the overall global tests that the null hypothesis can be rejected and conclude that there is a difference between the intervention and placebo groups.

Next, we can compare each scoring system used, between the intervention and placebo groups. Using the global tests, it was determined that each scoring tool differed between the treatment and placebo group except for the modified Baron score. For this score, the null hypothesis could not be rejected based on the p-value, indicating that there was no difference between the treatment and placebo group. The methods used in the original clinical trial indicate a similar

result to the global tests. The original trial results determined that the ulcerative colitis clinical scores, the Riley histopathological scores and the scores from the inflammatory bowel disease questionnaire were better in the treatment group compared to the placebo with p-values of 0.008, 0.030 and 0.003, respectively. However, there was no difference indicated between the groups with the modified Baron score. The p-value was 0.050, which is equal to alpha at the 5% level, therefore we fail to reject the null hypothesis. Therefore, despite different methods being used to analyze this dataset, the same overall results appear to be achieved.

4.4: Implications of this work

One of the main takeaways from this thesis is to observe how prominent the use of multiple outcomes is in RCTs. The multiple outcomes found varied across the different topic categories that were chosen. It was interesting to see which disease area was more likely to use multiple outcomes compared to the others. The literature search using a one-year timeframe (July 2020 to July 2021) only provided a snapshot of the RCTs that are published by choosing a few select journals. It would be interesting to conduct a similar search now and see if the number of RCTs using multiple primary outcomes changes.

Along with the prevalence of multiple outcomes, this thesis also provided the benefits that come with using multiple primary outcomes. The use of multiple outcomes can aid in other aspects of a clinical trial such as the trial size and cost of the trial. More information can also be gathered when more components are included in the primary outcome.

The literature search demonstrated that there were a variety of different methods that were used to analyze the outcomes that were found depending on how the outcome was structured. This includes the results from across the different types of outcomes that were included in the search. Because of this, it could be beneficial to have more guidelines surrounding the use of multiple outcomes in clinical trials. With many options available for methods to use, it may be helpful to have a set of guidelines that indicate what method should be used for each situation. The guideline could cover the different types of multiple outcomes (ie. composite, co-primary and more than one separate outcome) and what method would work best in each situation. This could also take into consideration how the outcome is structured such as in a binary, count, categorical or time to event format. The guideline would aid in providing instruction on where to start for the statistical method that should be used or what could be used based on that format.

The guidelines that a research team usually follows when conducting a randomized trial is the CONSORT statement. This statement provides standards for what should be included when writing the report for a RCT (Schulz et al., 2010). By following the checklist provided by this document, this ensures that authors of trials are transparent when writing about the design of the trial and the analysis of the data. However, the statement does not provide information regarding the use of multiple outcomes. The CONSORT statement cautions the use of multiple outcomes as it can lead to problems surrounding the interpretation of the result that is found (Tyler et al., 2011). Recently, there has been a 17-item recommendation added to extend the CONSORT statement. With this extension, the term ‘composite outcome’ is included with the recommendation to define all individual components (Butcher et al., 2022).

Based on this information, statistical methods surrounding multiple outcomes should be included in the recommendations. This should include global tests. Only three possible methods were

demonstrated in this thesis, however there are other global tests and adjustments that can be used. Some of these other methods were discussed in the second chapter. These methods would be important to include in any recommendations as they take into account the fact that there are multiple outcomes included and the correlations that are present.

When using multiple primary outcomes, there is more information that is collected compared to using a singular primary outcome. Because of this, there is more opportunity to omit certain information or results and only provide results that may be significant. Including a section in the guidelines that discusses what results should be included in all RCTs may be beneficial. For example, this could include standards that state that an overall result is needed as well as the results of the individual component outcomes, if the format of the outcome allows for this. The added guidelines would provide transparency between the authors of the trial and the readers. The addition of guidelines may also help with reader understanding. When someone is reading the report from a RCT in order to gain new information, it is helpful to have the full picture when trying to understand the results. This may involve including all of the information that was collected. The recent extension of the CONSORT statement added in the statistical methods that any methods used to account for multiplicity should be described (Butcher et al., 2022). This provides a step in the right direction for having transparency when using multiple primary outcomes. Therefore, adding information about multiple outcomes to a previously constructed guideline or creating a new one would be beneficial to those conducting the RCT and those that seek to gain information from it.

4.5: Guidance for Future Research

This thesis presented statistical methods that can be of use in methods of other randomized controlled trials. The global tests demonstrated provide a method of obtaining a singular overall result for all the outcomes included, while still considering the correlations between the individual outcomes. The methods demonstrated considered both parametric and non-parametric methods which can be used depending on the shape of the distribution of the dataset being used. Parametric methods are typically used with datasets that are approximately normally distributed, whereas non-parametric methods do not follow a specific distribution and can be used when a dataset does not meet the criteria for normally distributed data. Therefore, the two kinds of methods that were used provide a demonstration of how these tests can used when a research team is dealing with data from multiple primary outcomes.

From the MLN02 dataset, the scores from the individual outcomes could not be added together because of the different scales that were used. This made the use of global tests advantageous. There have been additions to the measurement tools used to score the state of disease for UC, therefore future research could evaluate the data using an established composite score and compare those results to the results from global tests. In addition, the results provided from the demonstrated global tests only provide the p-value to indicate statistical significance and if the treatment differed from the control. The CONSORT statement recommends the use of effect size to demonstrate the treatment effect in place of a p-value. Therefore, further steps could include estimating the effect size and associated 95% confidence interval.

In addition to the methods that were demonstrated in this thesis, a different way to analyze the dataset could be used. An overall method that can be used when there are multiple outcomes, is

to use a hierarchical method. This could be done using a method such as the win ratio. The win ratio takes into account the clinical importance of the individual component outcomes that make up the primary outcome. The method begins using the most clinically important outcome, as this is the outcome that is of most interest. This outcome is used to compare each participant in the treatment group with each participant in the control group and determines the "winners" in each comparison. If there is no difference, then the comparison continues to the next clinically important outcome until an overall result is achieved.

The individual outcomes in this dataset do have different levels of clinical importance. Using this method described, the hierarchical order that the individual outcomes would be used in begins with the ulcerative colitis clinical score as the most clinically important outcome. This outcome would be followed, in decreasing order of importance, by the modified Baron score, the Riley histopathological score and finally the patient reported outcome score would be of the lowest clinical importance for this group of outcomes.

The other methods demonstrated do not factor in the clinical importance of each individual outcome when determining an overall result. It would be interesting to use this dataset for re-analyzation again and use the win ratio. The win ratio could be used for both strengths of the treatment group in order to determine if the treatment provides more "wins" compared to the placebo group and is therefore a better option for participants. This could be a good starting point for understanding the treatment effect size.

This process may also inspire others who have not used these methods in an analysis process to try them with RCT data. This may include a previously conducted trial as was demonstrated here or a future trial that deals with multiple outcomes.

References

Akobeng, A.K. (2005). Understanding randomised controlled trials. *Archives of Disease in Childhood, 90*(8), 840-844.

Andrade, C. (2015). The primary outcome measure and its importance in clinical trials. *The Journal of Clinical Psychiatry, 76*(10), e1320-e1323.

Baraniuk, S., Seay, R., Sinha, A.K., & Piller, L.B. (2012). Comparison of the global statistical test and composite outcome for secondary analyses of multiple coronary heart disease outcomes. *Progress in Cardiovascular Diseases, 54*(4), 357-361.

Best, W.R., Becktel, J.M., Singleton, J.W., & Ker, F. Jr. (1976). Development of a Crohn's disease activity index. National cooperative Crohn's disease study. *Gastroenterology, 70*(3), 439-444.

Bhide, A., Shah, P.S., & Acharya, G. (2018). A simplified guide to randomized controlled trials. *Acta Obstetricia et Gynecologica Scandinavica, 97*(4), 380-387.

Bours, M.J.L. (2020). A nontechnical explanation of the counterfactual definition of confounding. *Journal of Clinical Epidemiology, 121*, 91-100.

Butcher, N.J., Monsour, A., Mew, E.J., Chan, A.W., Moher, D., Mayo-Wilson, E., Terwee, C.B., Chee-A-Tow, A., Baba, A., Gavin, F., Grimshaw, J.M., Kelly, L.E., Saeed, L., Thabane, L., Askie, L., Smith, M., Farid-Kapadia, M., Williamson, P.R., Szatmari, P., … Offringa, M. (2022). Guidelines for reporting outcomes in trial reports: the CONSORT-outcomes 2022 extension. *JAMA, 328*(22), 2252-2264.

Buyse, M. (2010). Generalized pairwise comparisons of prioritized outcomes in the two-sample problem. *Statistics in medicine, 29*(30), 3245-3257.

Buyse, M., Molenberghs, G., Paoletti, X., Oba, K., Alonso, A., Van der Elst, W., & Burzykowski, T. (2016). Statistical evaluation of surrogate endpoints with examples from cancer clinical trials. *Biometrical Journal, 58*(1), 104-132.

Chi, G.Y.H. (2005). Some issues with composite endpoints in clinical trials. *Fundamental & Clinical Pharmacology, 19*(6), 609-619.

Chuang-Stein, C., & Li, J. (2017). Changes are still needed on multiple co-primary endpoints. *Statistics in Medicine, 36*(28), 4427-4436.

Cordoba, G., Schwartz, L., Woloshin, S., Bae, H., & Gotzsche, P.C. (2010). Definition, reporting, and interpretation of composite outcomes in clinical trials: systemic review. *BMJ, 341*, c3920.

D'Agostino R. B., Jr (2000). Debate: The slippery slope of surrogate outcomes. *Current controlled trials in cardiovascular medicine, 1*(2), 76–78.

Dash, K., Goodacre, S., & Sutton, L. (2022). Composite outcomes in clinical prediction modeling: are we trying to predict apples and oranges? *Annals of Emergency Medicine*, *80*(1), 12-19.

Dmitrienko, A., D'Agostino, R.B. Sr., & Huque, M.F. (2012). Key multiplicity issues in clinical drug development. *Statistics in Medicine, 32*(7), 1079-1111.

Feagan, B.G., Greenberg, G.R., Wild, G., Fedorak, R.N., Paré, P., McDonald, J.W.D., Dubé, R., Cohen, A., Steinhart, A.H., Landau, S., Aguzzi, R.A., Fox, I.H., & Vandervoort M.K. (2005). Treatment of ulcerative colitis with a humanized antibody to the $\alpha_4\beta_7$ integrin. *The New England Journal of Medicine*, *352*(24), 2499-2507.

Ferreira, J.C., & Patino, C.M. (2017). Types of outcomes in clinical research. *Jornal brasileiro de pneumologia : publicacao oficial da Sociedade Brasileira de Pneumologia e Tisilogia*, *43*(1), 5.

Ferreira-González, I., Busse, J.W., Heels-Ansdell, D., Montori, V.M., Akl, E.A., Bryant, D.M., Alonso-Coello, P., Alonso, J., Worster, A., Upadhye, S., Jaeschke, R., Schünemann, H.J., Permanyer-Miralda, G., Pacheco-Huergo, V., Domingo-Salvany, A., Wu, P., Mills, E.J., & Guyatt, G.H. (2007). Problems with use of composite end points in cardiovascular trials: systematic review of randomised controlled trials. *BMJ*, *334*, 786.

Ferreira-González, I., Permanyer-Miralda, G., Busse, J.W., Bryant, D.M., Montori, V.M., Alonso-Coello, P., Walter, S.D., & Guyatt, G.H. (2007). Methodologic discussions for using and interpreting composite endpoints are limited, but still identify major concerns. *Journal of Clinical Epidemiology*, *60*(7), 651-657.

Fleming, T.R., & Demets, D.L. (1996). Surrogate endpoints in clinical trials: are we being misled. *Annals of Internal Medicine, 125*(7), 605-613.

Freemantle, N., & Calvert, M. (2007). Composite and surrogate outcomes in randomised controlled trials. *BMJ, 334*, 756.

Freemantle, N., Calvert, M., Wood, J., Eastaugh, J., & Griffin, C. (2003). Composite outcomes in randomized trials: greater precision but with greater uncertainty? *JAMA, 289*(19), 2554-2559.

Gillespie, D., Francis, N.A., Carrol, E.D., Thomas-Jones, E., Butler, C.C., & Hood, K. (2018). Use of co-primary outcomes for trials of antimicrobial stewardship interventions. *Infectious Diseases*, *18*(6), 585-597.

Hahn, A., Podbielski, A., Heimesaat, M.M., Frickmann, H., & Warnke, P. (2021). Binary surrogate endpoints in clinical trials from the perspective of case definitions. *European journal of microbiology & immunology*, *11*(1), 18–22.

Hamasaki, T., Evans, S.R., & Asakura, K. (2018). Design, data monitoring, and analysis of clinical trials with co-primary endpoints: a review. *Journal of Biopharmaceutical Statistics*, *28*(1), 28-51.

Hamasaki, T., Sugimoto, T., Evans, S., & Sozu, T. (2012). Sample size determination for clinical trials with co-primary outcomes: exponential event times. *Pharmaceutical Statistics*, *12*(1), 28-34.

Heneghan, C., Goladacre, B., & Mahtani, K.R. (2017). Why clinical trial outcomes fail to translate into benefits for patients. *Trials*, *18*, 122.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*(2), 65-70.

Huang, P., Goetz, C.G., Woolson, R.F., Tilley, B., Kerr, D., Palesch, Y., Elm, J., Ravina, B., Bergmann, K.J., Kieburtz, K., & Parkinson Study Group. (2009). Using global statistical tests in long-term Parkinson's disease clinical trials. *Movement Disorders: official journal of the Movement Disorder Society, 24*(12), 1732-1739.

Lauter, J. (1996). Exact t and F tests for analyzing studies with multiple endpoints. *Biometrics*, *52*(3), 964-970.

Logan, B.R., & Tamhane, A.C. (2004). On O'Brien's OLS and GLS tests for multiple endpoints. Recent developments in multiple comparison procedures (pp. 76-88). Institute of Mathematical Statistics.

McCoy, C.E. (2018). Understanding the use of composite endpoints in clinical trials. *The Western Journal of Emergency Medicine, 19*(4), 631-634.

McLeod, C., Norman, R., Litton, E., Saville, B.R., Webb, S., & Snelling, T.L. (2019). Choosing primary endpoints for clinical trials of health care interventions. *Contemporary Clinical Trials Communications*, *16*, 100486.

Menon, U., Gentry-Maharaj A., Burnell, M., Singh, N., Ryan, A., Karpinskyj, C., Carlino, G., Taylor, J., Massingham, S.K., Raikou, M., Kalsi, J.K., Woolas, R., Manchanda, R., Arora, R., Casey, L., Dawnay, A., Dobbs, S., Leeson, S., & Parmar, M. (2021). Ovarian cancer population screening and mortality after long-term follow-up in the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS): a randomised controlled trial, *Lancet*, *397*(10 290), 2182-2193.

Mittendorf, E.A., Zhang, H., Barrios, C.H., Saji, S., Jung, K.H., Hegg, R., Koehler, A., Sohn, J., Iwata, H., Telli, M.L., Ferrario, C., Punie, K., Penault-Llorca, F., Patel, S., Nguyen Duc, A., Liste-Hermoso, M., Maiya, V., Molinero, L., Chui, S.Y., & Harbeck, N. (2020). Neoadjuvant atezolizumab in combination with sequential nab-paclitaxel and anthracycline-based chemotherapy versus placebo and chemotherapy in patients with early-stage triple-negative breast cancer (Impassion031): a randomised, double-blind, phase 3 trial. *The Lancet*, *396*(10257), 1090-1100.

Neaton, J.D., Gray, G., Zuckerman, B.D., & Konstam, M.A. (2005). Key issues in end point selection for heart failure trials: composite end points. *Journal of Cardiac Failure, 11*(8), 567-575.

O'Brien, P.C. (1984). Procedures for comparing samples with multiple endpoints. *Biometrics, 40*(4), 1079-1087.

O'Brien, P.C., & Geller, N.L. (1997). Interpreting tests for efficacy in clinical trials with multiple endpoints. *Controlled clinical trials, 18*(3), 222-227.

Pearson, K. (1920). Notes on the history of correlation. *Biometrika*, *13*(1), 25-45.

Pocock, S.J., Ariti, C.A., Collier, T.J., & Wang, D. (2012). The win ratio: a new approach to the analysis of composite endpoints in clinical trials based on clinical priorities. *European Heart Journal*, *33*(2), 176-182.

Pocock, S.J., Geller, N.L., & Tsiatis, A.A. (1987). The analysis of multiple endpoints in clinical trials. *Biometrics, 43*(3) 487-498.

Rauch, G., & Kieser, M. (2012). Multiplicity adjustment for composite binary endpoints. *Methods of Information in Medicine, 51*(4), 309-317.

Redfors, B., Gregson, J., Crowley, A., McAndrew, T., Ben-Yehuda, O., Stone, G.W., & Pocock, S.J. (2020). The win ratio approach for composite endpoints: practical guidance based on previous experience. *European Heart Journal, 41*(46), 4391-4399.

Ristl, R., Urach, S., Rosenkranz, G., & Posch, M. (2019). Methods for the analysis of multiple endpoints in small populations: a review. *Journal of Biopharmaceutical Statistics, 29*(1), 1-29.

Roberts, C., & Torgerson, D. (1998). Randomisation methods in controlled trials. *BMJ*, *317*(7168), 1301.

Ross, S. (2007). Composite outcomes in randomized clinical trials: arguments for and against. *American Journal of Obstetrics and Gynecology, 196*(2), 119.e1-119.e6.

Sankoh, A.J., Huque, M.F., Russell, H.K., & D'Agostino, R.B. (1999). Global two-group multiple endpoint adjustment methods applied to clinical trials. *Drug information journal: DIJ/Drug Information Association*, *33*(1), 119-140.

Schüler, S., Mucha, A., Doherty, P., Kieser, M., & Rauch, G. (2014). Easily applicable multiple testing procedures to improve the interpretation of clinical trials with composite endpoints. *International Journal of Cardiology*, *175*, 126-132.

Schulz, K. F., Altman, D. G., Moher, D., & CONSORT Group. (2010). CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. *Journal of clinical epidemiology*, *63*(8), 834–840.

Sedgwick, P. (2014). Pitfalls of statistical hypothesis testing: type I and type II errors. *BMJ*, *349*, g4287.

Sedgwick, P. (2014). Treatment allocation in trials: block randomization. *BMJ*, *348*, g2409.

Simes, R.J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, *73*(3), 751-754.

Skelly, A.C., Dettori, J.R., & Brodt, E.D. (2012). Assessing bias: the importance of considering confounding. *Evidence-Based Spine-Care Journal*, *3*(1), 9-12.

Snapinn, S.M., & Jiang, Q. (2007). Responder analyses and the assessment of a clinically relevant treatment effect. *Trials*, *8*, 31.

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, *15*, 72-101.

Sweeney, C., Bracarda, S., Sternberg, C.N., Chi, K.N., Olmos, D., Shahneen, S., Massard, C., Matsubara N., Alekseev, B., Parnis, F., Atduev, V., Buchschacher, G.L., Gafanov, R., Corrales, L., Borre, M., Stroyakovskiy, D., Alves, G.V., Bournakis, E., Puente, J., … de Bono, J.S. (2021). Ipatasertib plus abiraterone and prednisolone in metastatic castration-resistant prostate cancer (IPATential150): a multicentre, randomised, double-blind, phase 3 trial. *The Lancet*, *398*(10295), 131-142.

Tilley, B.C., Illemer, S.R., Heyse, S.P., Li, S., Clegg, D.O., Alarcon, G.S. (1999). Global statistical tests for comparing multiple outcomes in rheumatoid arthritis trial. *Arthritis & Rheumatism*, *42*(9), 1879-1888.

Tilley, B.C., Marler, J., Geller, N.L., Lu, M., Legler, J., Brott, T., Lyden, P., & Grotta, J. (1996). Use of global test for multiple outcomes in stroke trials with application to the national institute of neurological disorders and stroke t-PA stroke trial. *Stroke*, *27*(11), 2136-2142.

Tyler, K.M., Normand, S.L.T, & Horton, N.J. (2011). The use and abuse of multiple outcomes in randomized controlled depression trials. *Contemporary Clinical Trials*, *32*(2), 299-304.

U.S. Food and Drug Administration. Center for Drug Evaluation and Research. (2022). *Multiple endpoints in clinical trials: guidance for industry*. Author.

Vetter, T. R., & Mascha, E. J. (2017). Defining the Primary Outcomes and Justifying Secondary Outcomes of a Study: Usually, the Fewer, the Better. *Anesthesia and analgesia*, *125*(2), 678–681.

Wason, J.M.S., & Robertson, D.S. (2021). Controlling type I error rates in multi-arm clinical trials: a case for the false discovery rate. *Pharmaceutical Statistics*, *20*(1), 109-116.

Yarlas, A., Maher, S., Bayliss, M., Lovely, A., Cappelleri, J.C., Bushmakin, A.G., DiBonaventura, M.D. (2020). The inflammatory bowel disease questionnaire in

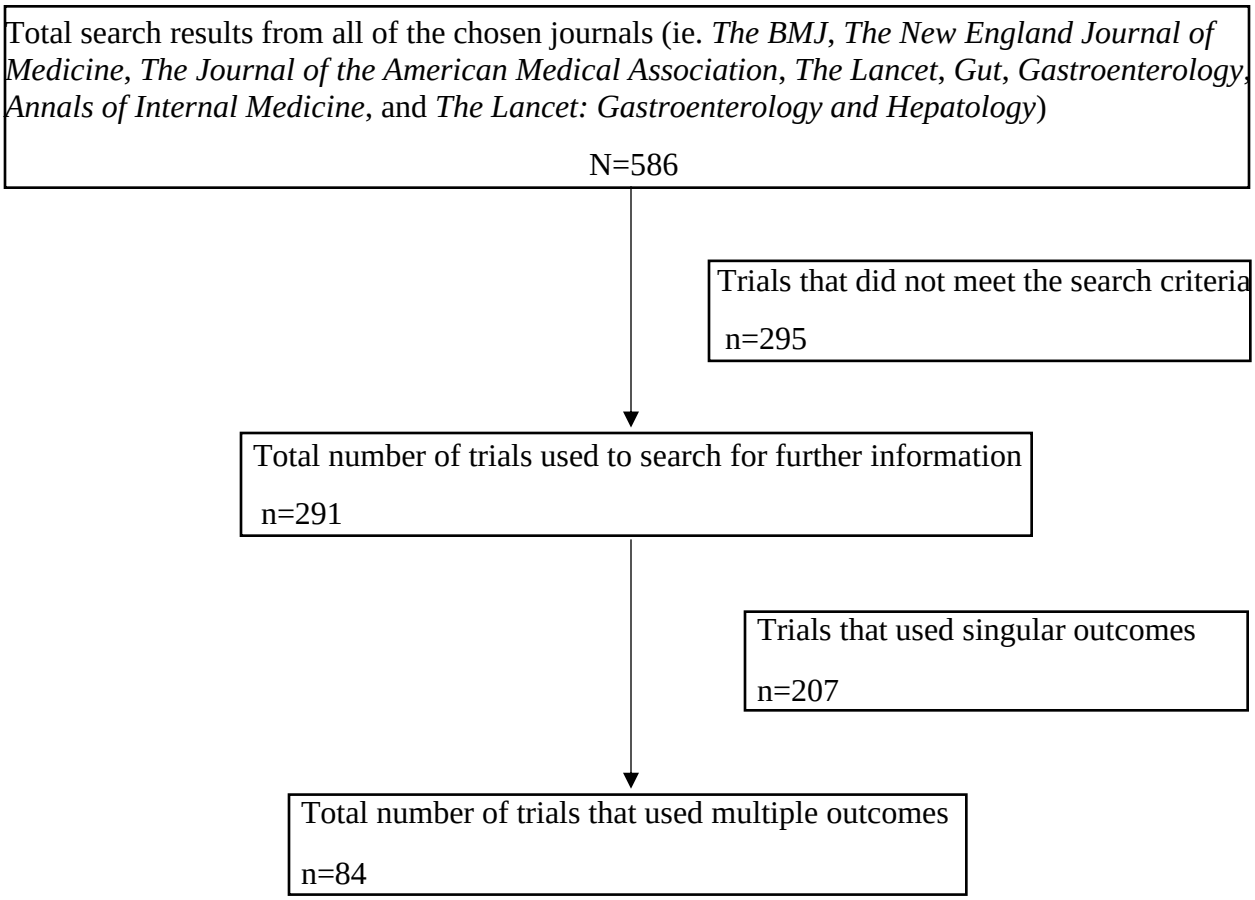
randomized controlled trials of treatment for ulcerative colitis: systematic review and meta-analysis. *Journal of Patient-Centered Research and Reviews, 7*(2), 189-205.

Zarbock, A., Küllmar, M., Kindgen-Milles, D., Wempe, C., Gerss, J., Brandenburger, T., Dimski, T., Tyczynski, B., Jahn, M., Mülling, N., Mehrländer, M., Rosenberger, P., Marx, G., Simon, T.P., Jaschinski, U., Deetjen, P., Putensen, C., Schewe, J.C., Kluge, S, … Meersch, M. (2020). Effect of regional citrate anticoagulation vs systemic heparin anticoagulation during continuous kidney replacement therapy on dialysis filter life span and mortality among critically ill patients with acute kidney injury: a randomized clinical trial. *JAMA*, *324*(16), 1629-1639.

Zou, G.Y. (2007). Exact confidence interval for Cohen's effect size is readily available. *Statistics in Medicine, 26*(15), 3054-3056.

Appendices

Appendix 1: Flow Chart to Demonstrate how the Literature Search was Conducted



Appendix 2: SAS Code used to Analyze the MLN02 Trial

```
options nocenter nofmterr ls=132;
ods graphics off;

*Feagan, B.G., Greenberg, G.R., Wild, G., Fedorak, R.N., Paré, P.,
McDonald, J.W., Dubé, R., Cohen, A.,
Steinhart, A.H., Landau, S. and Aguzzi, R.A., 2005. Treatment of
ulcerative colitis with a humanized
antibody to the a4ß7 integrin. New England Journal of Medicine,
352(24), pp.2499-2507.;


libname ll 'C:\Users\Lindsay\OneDrive\Documents\Thesis\Analysis';

%let za = probit(1-.05/2);
data mln02;
 set ll.mln02;   if tmtcode in (1,3);
 trt=(tmtcode in (1)); * only compare placebo=0 (trt=1) vs 0.5 mg
MLN02 (=0);
 y1=w4_mriley; * histology;
 y2=w4_mbaron; * endoscopic;
 y3=w4_uccs;   * clinical;
 y4=-w4_ibdqt;  * QoL;
 idn=id;
 subj=idn;
 group=trt;
 if y1>.; * get rid of missing values;
  keep  subj group   y1-y4;
run;


ods listing close;
ods output pearsonCorr=corr1(keep=y1 y2 y3 y4)
           SpearmanCorr=corr2(keep=y1 y2 y3 y4);
proc corr data=mln02 spearman pearson;
  var y1 - y4;
run;

ods listing;
proc print data=corr1;
title 'Pearson correlation';
run;
proc print data=corr2;
title 'Spearman correlation';
run;




** Lauter's SS;
* 1) standardization;
proc standard mean=0 std=1 data=mln02 out=stdra;
   var y1 y2 y3 y4;
run;
* 2) subject-specific sum of standardized scores;
```

```
data sumStdra;
   set stdra;
   array endp{*}  y1 y2 y3 y4;
   sumScore=0;
   do k=1 to dim(endp);
    sumScore = sumScore + endp{k};
     end;

ods listing close;
ods output ttests=tests(where=(Method='Pooled'));
proc ttest data=sumStdra;
   class group;
   var sumScore;
run;

ods listing;
proc print data=tests;
title 'Lauter test';
run;

*** OBrien's OLS test;
* 1) standardize endpoints;
** long-format works better;
proc sort data=mln02; by subj group;

proc transpose data=mln02 out=raLong(rename=(col1=score
_name_=outcome));
   by subj group;
run;


proc sort data=raLong; by outcome;
run;

ods listing close;
ods output summary=overMean (keep=outcome score_Mean);
proc means data=raLong; by outcome ;
    var score;
run;

proc sort data=ralong; by outcome group;

ods listing close;
ods output summary=stats(drop= score_Mean score_min score_max);
proc means data=raLong; by outcome group;
    var score;
run;


proc transpose data=stats out=TransN(rename=(col1=n1 col2=n2)
                                     drop=_name_ _label_) ; by
outcome;
  var score_n;
run;


proc transpose data=stats out=TransSD(rename=(col1=SD1 col2=SD2)
```

```
                                          drop=_name_ _label_); by
outcome;
  var score_StdDev;
run;


data NSD;
   merge TransN TransSD;
   poolSD = sqrt(((n1-1)*SD1**2+(n2-1)*Sd2**2)/(n1+n2-2));
   keep outcome poolSD;
run;



data stand4OLS;
  merge raLong overMean NSD; by outcome;
  Y =(score-score_Mean)/poolSD;
run;


* 2) run OLS using mixed-effect model;

ods listing;

proc mixed data=stand4OLS noclprint noitprint;
   class subj group outcome;
   model Y = outcome group outcome*group/ddfm=kr;
   repeated outcome/subject = subj type=un;

   estimate 'separate ttest: outcome1' group 1 -1  outcome*group  1 0
0 0  -1 0 0 0;
   estimate 'separate ttest: outcome2' group 1 -1  outcome*group  0 1
0 0  0 -1 0 0;
   estimate 'separate ttest: outcome3' group 1 -1  outcome*group  0 0
1 0  0 0 -1 0;
   estimate 'separate ttest: outcome4' group 1 -1  outcome*group  0 0
0 1  0 0 0 -1;

   estimate 'Overall' group 1 -1;
   title 'Test for each outcome and OBrien OLS test';
run;


*** O'Brien's rank-sum test;
* 1) get the subject-specific ranks,
  2) ran ttest for each outcome,
  3) rank O'Briens test;

proc rank data=mln02 out=ranks;
   var y1 y2 y3 y4;
run;


ods listing close;
ods output ttests=tests(where=(Method='Pooled'));
proc ttest data=ranks;
   class group;
```

```
   var y1 y2 y3 y4;

run;


ods listing;
proc print data=tests;
title 'Rank test for each outcome';
run;



data sumrank;
  set ranks;
  array endp{*} y1-y4;
  sumrank=0;
  do k=1 to dim(endp);
    sumrank=sumrank + endp{k};
  end;
run;

ods listing close;
ods output ttests=tests(where=(Method='Pooled'));
proc ttest data=sumrank ;
   class group;
   var sumrank;
run;

ods listing;
proc print data=tests;
title 'OBrien rank-sum test';
run;


title ;
```

Curriculum Vitae

Lindsay Cameron

EDUCATION

Master of Science : Epidemiology and Biostatistics — 2020 - present
Western University, London, Ontario, Canada

Bachelor of Science : Molecular Biology & Genetics — 2016 – 2020
University of Guelph, Guelph, Ontario, Canada

RESEARCH EXPERIENCE

Research Project in Molecular and Cellular Biology — 2019 – 2020

- Fourth year undergraduate research project with Dr. Ian Tetlow
- Examining starch levels in *Arabidopsis thaliana* with the addition of different maize branching enzymes
- Working in lab environment using various lab techniques such as PCR, gel electrophoresis, Western Immunoblotting and zymograms

CERTIFICATES

SAS Programming 1 : Essentials

- Introduction to basic SAS commands