11-7-2023 2:30 PM

Multi-Arm Randomized Control Trials in Inflammatory Bowel Disease: A Literature Review and an Illustration of Methods for Analysis

Sahiba Saini,

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

This thesis aimed to review the literature on multiple-arm randomized control trials in inflammatory bowel disease (IBD) and to illustrate how to analyze these trials, focusing on appropriately controlling the type 1 error rates. The literature review found 247 trials published from the inception of each database to April 2014, of which 122 (49%) trials were multiple-arm trials and of those, 59 (48%) trials were on ulcerative colitis and 63 (52%) on Crohn's disease. A published assessment tool was adopted to assess whether controlling of Type I error rates was needed. Despite the common use of this trial design and the need for multiple comparison procedures (MCPs) based on the assessment tool, only 20% of trials applied any MCPs. Failing to apply appropriate procedures may have inflated the Type I error rates, yielding false positive results. Data from a multiple-arm trial evaluating the efficacy of MLN02 (vedolizumab) was used as a case study to present how to implement common MCPs. Assessment tools regarding MCPs are helpful to have readily available to improve the scientific rigour of such research. Future research into the utility of online resources that guide multiple-arm trial design and analysis within IBD could promote such tools.

Keywords:

inflammatory bowel disease, ulcerative colitis, Crohn's disease, multi-arm trials, multiple comparison procedures, randomized control trials, type 1 error

Summary for Lay Audience

Inflammatory bowel disease (IBD) is a term that describes disorders involving long-term (chronic) inflammation of tissues in the digestive tract. Two common forms of IBD are ulcerative colitis (UC) and Crohn's disease (CD). Randomized controlled trials (RCTs) are the standard tool for evaluating the effect of treatments for IBD. This thesis reviewed the frequency of RCTs that used a design of more than two comparison groups. Such trials are termed multiple-arm trials. Since a critical statistical complication with such trials is controlling the false positive rate (Type I error) within the analysis, this thesis also enumerated and explored the procedures used to control the Type I error.

The literature review identified 217 randomized control trials about IBD, and among those, 59 UC and 63 CD disease trials had a multiple-arm design. An assessment tool evaluated these trials to determine if they possessed specific study characteristics, suggesting using a statistical procedure for multiple study groups to maintain statistical validity. Trials were also assessed on whether statistical procedures were initially implemented in their methods to determine the types of procedures already being used within IBD multiple-arm trials.

Data from an example study was analyzed and presented using some commonly used statistical procedures for trials with multiple study groups. This thesis established the importance of such statistical methods as they can help multiple group studies with specific characteristics control for statistical validity and maintain credible results and conclusions. Having tools and guidelines that assist with the decision-making around using such statistical methods could benefit the quality and efficacy of research work, especially within the field of IBD. Future research efforts could analyze online resources, which help set up the required statistical logistics for conducting multiple group IBD trials.

Acknowledgments

I would like to express my gratitude to Dr. Guangyong Zou and Dr. Vipul Jairath who have supervised the works of this thesis, and all those who have provided their gracious support and time with guiding, proofreading, and editing the content of this thesis.

Table of Contents

List of Tables

List of Figures

List of Appendices

Chapter 1: Introduction

1.1 Inflammatory Bowel Disease (IBD)

Inflammatory bowel disease (IBD) is a chronic inflammatory condition of the gastrointestinal (GI) tract, which arises from a modified immune response involving intermittent or aggressive inflammation of the intestines[1]. This condition has often been labelled as an idiopathic disease[2,3], which can be related to its complex etiology and unpredictable progression[4]. Since no permanent cure is available for IBD patients[5], those diagnosed with this chronic condition may only achieve remission of their symptoms using applicable induction and maintenance therapies[6]. Extensive research suggests that lifestyle factors, including diet, country of origin/ethnicity, immune health[7], genes and environmental conditions, predispose the development of IBD[8]. However, the consensus is that inconsistent behaviour of immune T-cells in conjunction with a patient's genetic history of IBD, degree of homeostatic health, or the type of gut flora present within the intestines[9] are significant determinants in the onset of IBD. These determinants can be informative in describing the overwhelming disease burden in North American and European countries and the rising incidence in certain Asian[10] and African countries[11] where IBD was once less common. This progression of disease burden from the countries of the West to the East can be associated with the increased income and modernization of emerging countries and, therefore, the advancement of the sedentary lifestyle, which is exacerbated by the Western diet[12].

A wide selection of IBD treatments are available and are individually tailored or modified over time to be most effective[13]. Treatments can also be phased out with patients due to reduced effectiveness over time or the development of inadvertent side effects[13]. Common pharmacological treatments used to treat symptoms of IBD include anti-inflammatory drugs, corticosteroids, immunomodulators (i.e., immunosuppressants), integrin antagonists, monoclonal antibodies, antibiotics, and probiotics[14]. Since the dosage administered for these treatments is a crucial part of attaining remission and possibly maintaining it[15], multiple research experiments aim to study the dose-response outcomes of these drugs utilizing a multiple-arm trial design. Due to multiple comparisons, this design may result in an inflated false positive rate, commonly called an inflated Type I error rate, if appropriate methods are not used [16,17]. This thesis explores

the assessment of multi-arm randomized control trial designs requiring multiple comparison adjustment procedures through an IBD trial case study.

1.1.1 The Past and Current Global State of IBD

Two common types of IBD are Crohn's Disease (CD) and ulcerative colitis (UC)[18]. A third type is called inflammatory bowel disease unclassified (IBDU), which describes cases that do not align with the specifications of either CD or UC[19]. Although distinctive in their manifestation of IBD, knowing these subtypes are advantageous in establishing a community of knowledge, resources, and relationships between physicians, clinicians, researchers, and patients[20]. Formulating this network of expertise is essential amidst the growing incidence of IBD globally[21]. It has been estimated that by the end of this decade, there will be almost 4 million patients living with this condition within the continent of North America[21]. Historically, IBD has been most pronounced in countries such as Canada, the United States[21], Australia, New Zealand[22], and Northern Europe regions, especially among the Caucasian heritage. With growing rates of IBD reported within Asia, Africa, Eastern Europe, and South America, the concern for IBD has progressed to affecting multiple races and ethnicities outside of the Northwestern world[22]. Figure 1 depicts a map published by Atlab and colleagues presenting the age-standardized disease burden of IBD worldwide in 2017[12].

Figure 1 presents that during 2017, the top six regions (in no specific order) with the most significant disease burden are the United States, United Kingdom, Norway, Poland, Croatia, and Slovenia[12], within the North American and North European region[21]. Following close with their high prevalence of IBD are other countries of Europe (i.e., Italy, Switzerland, Germany, Russia, Ukraine, Belarus), Asia (i.e., China), South America (i.e., Brazil, Paraguay, Argentina), North America (i.e., Mexico, Canada), and Australia[12]. Higher and growing prevalence proportions of IBD are observed to be complementary with a nation's increase in income which is an indicator of other factors such as a population's ability to access ready-made diets (that are less nutrient dense), tobacco, alcohol, healthcare tools which can diagnose the condition earlier, and education regarding IBD[12].

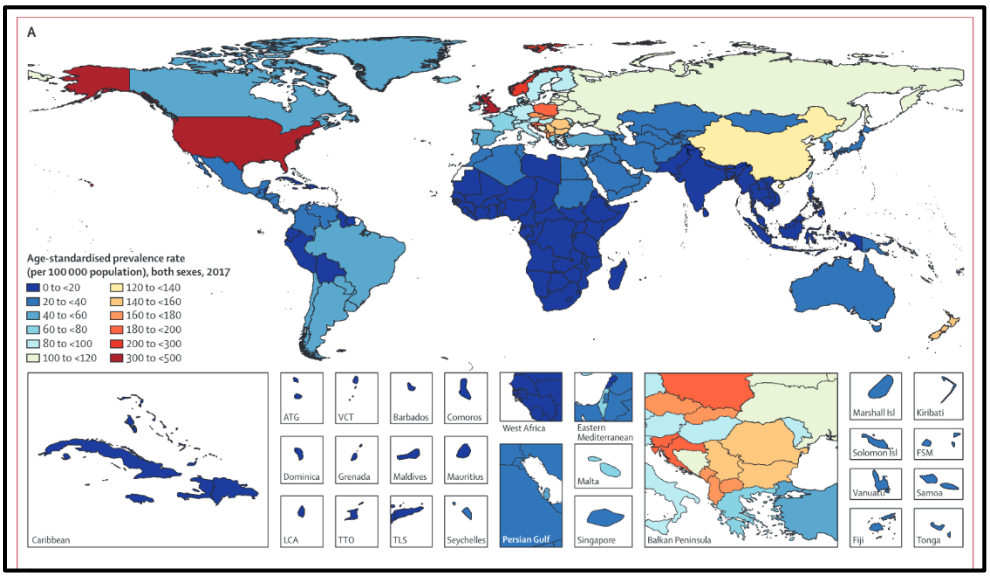

Figure 1: The global age-standardized prevalence distribution of IBD for every 100 000 persons in the year 2017[12]

The incidence of IBD is impacted by the migration of populations from countries with a reduced risk of IBD to countries where they are exposed to factors that drastically increase their risk[2]. Population data have frequently found that IBD risk is high for migrants who arrive in countries where a heightened risk for IBD is already present[23]. This can be a result of the complex interactions between the individual's genetic profile and an introduction into new environments where possible divergences in diet, lifestyle, and sanitation exist. A prominent example is migrants of South Asian descent who arrive in Canada and experience a higher likelihood of developing IBD than when they were in their country of origin, which has been attributed to their transition to adopting a Western lifestyle[23]. Furthermore, this gene and environment interaction can be carried forward to future generations as it has also been found that the risk for IBD is more pronounced in the offspring of immigrants than those native to the host country[22,24].

The diet of an individual also has a specific impact on their risk of IBD. The Western diet has been linked to many cases as a possible precursor for the emergence of IBD[6]. The Western diet is often high in sugars, carbohydrates, fats, meats, and calories[25] while low in plant fibres[18] formulating a greater likelihood of IBD[25]. Consumption of fast foods worldwide due to

globalization combined with sedentary lifestyles[26] result in certain noncommunicable diseases[27]. An example of such a case is China, whose accelerated path to westernization since the last century has been simultaneous with the development of IBD incidence among their population[21] such that their IBD disease burden trends are beginning to mimic those of leading Western countries[12]. Systematic reviews and meta-analyses have been capturing this effect on populations of Asian descent[28,29]and this was a significant topic of concern during a 2016 conference between the Chinese Society of Gastroenterology and the American Gastroenterological Association[30].

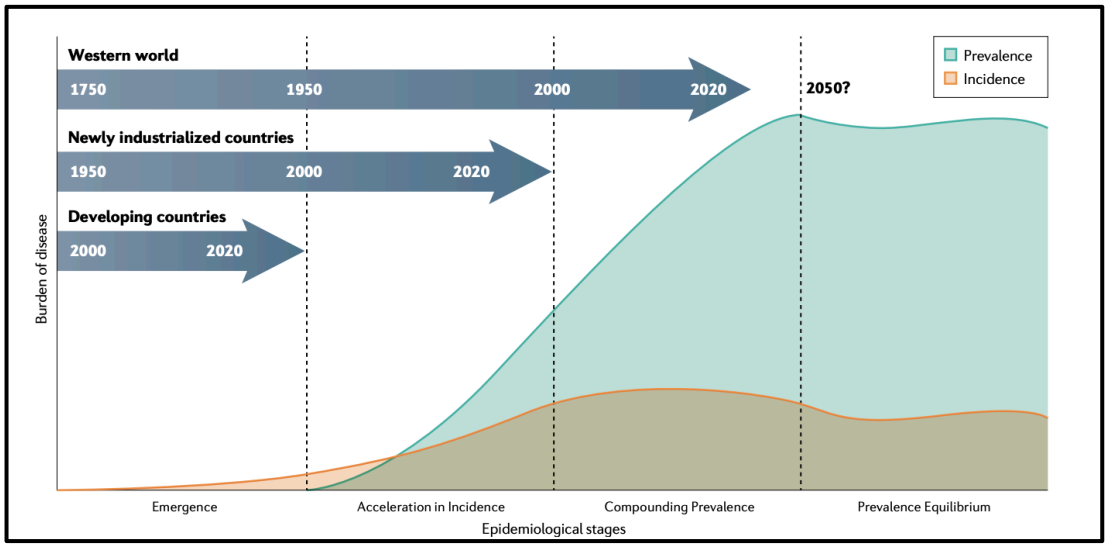

Figure 2: An area chart depicting the change in global prevalence and incidence of IBD over the four stages of the global expansion of the disease for developing, new industrialized, and Western countries[31]

The international burden of IBD can be best represented by Kaplan and Windsor's four stages of global expansion, as presented in Figure 2[31]. Their four-stage framework conceptualizes the global progression into the following stages: emergence, acceleration of incidence, compounding prevalence, and prevalence equilibrium [31]. According to this framework, the emergence of IBD is marked by its medical identification in primary publications of notable cases and the recognition of its various symptoms, which helped distinguish between CD and UC[31]. The 18th

and 19th centuries mark the emergence of IBD within North America, and the 1950s signify its end when CD was distinguished from UC[31]. However, other emerging nations like China began to experience their initial emergence of primary IBD cases during the beginning of the 1950s[31], as shown in Figure 3. The framework continues into the growing incidence phase which occurs in the latter half of the 20th century for North America[31]. In this stage Kaplan and Windsor describe how there is growth in the population at risk until the advent of the 21st century[31]. Emerging countries with increased incidences recognized later than Western countries followed the same trend of increased incidence followed by a plateau in incidence and an increase in prevalence due to the lengthy nature of the disease[31,32].

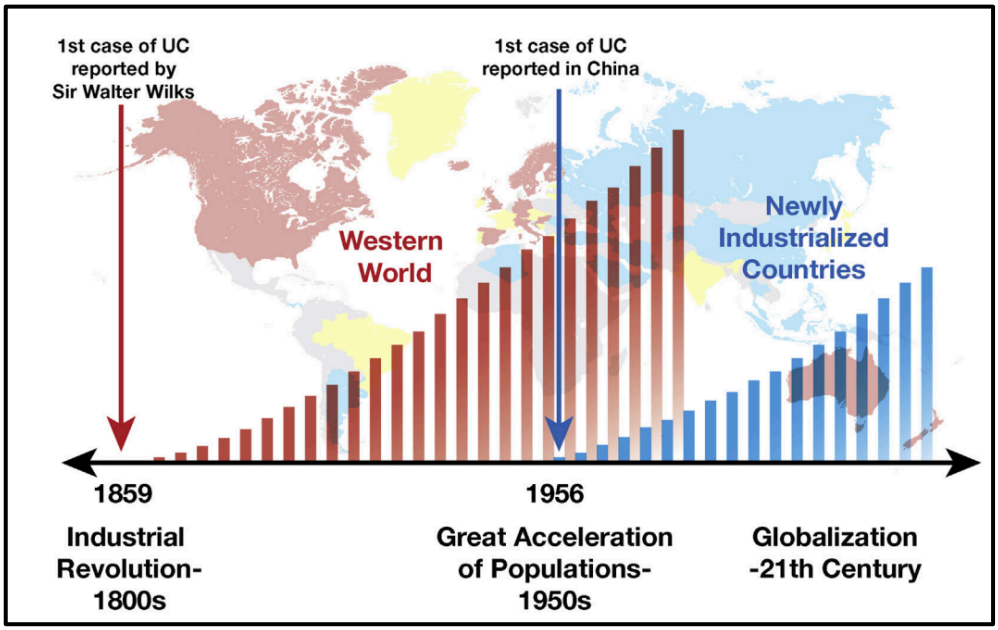

Figure 3: A timeline depicting the increase in IBD cases in the Western world since the 1800s, with a lag in the increase of IBD cases for newly industrialized countries which began later in the 1950s[33].

The next stage of magnified prevalence is due to the target population of IBD having developed a longer lifespan and the increased quality of detection tools within the Western world[31]. As the *baby boomer* cohort shifts into the geriatric demographic, there will begin a flow

of prevalent cases due to mortality resulting in a consistent flow of prevalence with decreased levels of growth compared to before[31]. Although this stage of prevalence represents a theoretical path that countries will begin to move towards soon, there is a possibility that significant global events, research discoveries within IBD or discoveries within related fields can result in an altered incidence and prevalence progression[31].

1.1.2 The Effects of IBD within Canada

Canada has one of the highest prevalence and incidence of IBD internationally [32]. Figure 4 presents IBD prevalence by Canadian provinces for 2008, with predicted values for 2018 and 2030 as collected by Kaplan and colleagues[32]. In 2008, Nova Scotia, Manitoba, and Saskatchewan had the highest burden for IBD within Canada, while Quebec, Ontario, and British Columbia had the lowest[32]. Nova Scotia is often highlighted for carrying a significant burden of IBD cases within Canada, and this has been mainly attributed to the gatekeeping of gastroenterology specialists[34]. Nova Scotian patients, on average, had to wait a quarter to half a year before seeing a specialist[34].

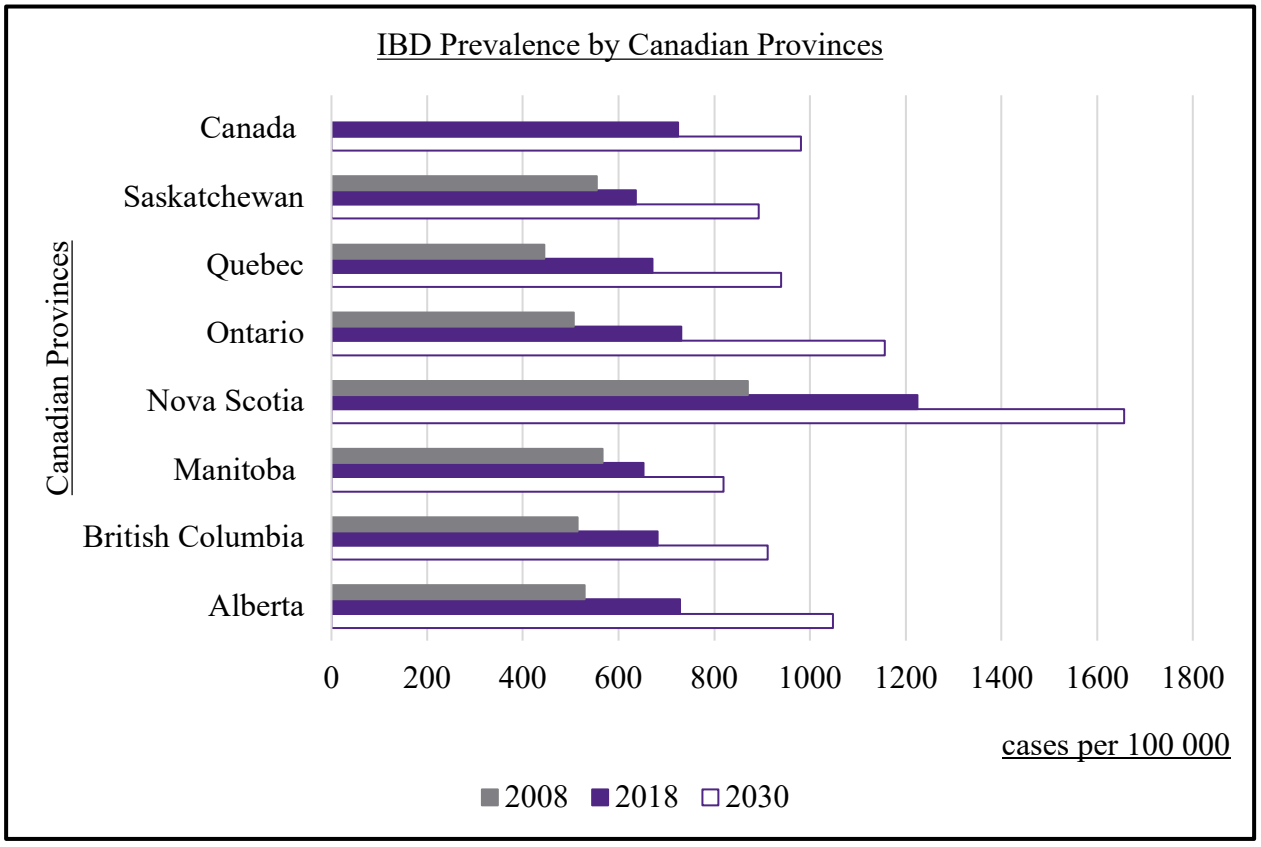

Figure 4: A grouped bar graph presenting the IBD prevalence for 2008 with predicted prevalence values for 2018 and 2030 by the provinces in Canada[32].

Reasons such as the inability to allow patients to access specialty care outside their assigned local specialists who already carry long waitlists and lengthy diagnostic procedures have been mentioned to delay their access[34].

For 2018 and 2030, Nova Scotia maintained the highest disease burden but is closely followed by Ontario[32]. Ontario is often highlighted for having high cases of inflammatory disease globally, possibly attributed to Ontario being the most populated province in Canada with a diverse demographic[35]. A longitudinal study conducted by Benchimol and colleagues measured changes in the prevalence of IBD in Ontario over almost a decade[34,36]**.** It determined a significant increase of 0.17% in disease burden for adolescent, adult, and geriatric populations[35].

The cost of IBD to the Canadian healthcare system was more than $2 billion in 2018[20], and these costs are expected to increase in the foreseeable future as the prevalence of IBD has been estimated to increase from 0.7% of the Canadian population to 1% by the end of this decade[33]. Data for 2018 showed approximately 270 000 Canadians would be living with IBD[32], costing $4000 in annual medical expenses per IBD patient[21]. In addition to the direct costs of having IBD, indirect and social costs exist[37]. Patients living with IBD experience high levels of mental distress due to heightened anxiety about their self-confidence, body image, and ability to maintain or form social relationships with others[38]. Such mental distress can lead to the deterioration of mental health (i.e., depression), social isolation, and unhealthy coping mechanisms, which generate further dysfunction in their daily lives[38]. Furthermore, family members and caretakers are emotionally distressed when supervising and supporting those with IBD[38].

A study led by Becker and colleagues analyzed survey data collected by *Crohn's and Colitis Canada* to present the impact of having IBD on the lives of its patients and those around them[39]. Graph A in Figure 5 presents data regarding participation in leisure activities and shows that most IBD patients find the condition significantly affects their lives[39]. In contrast most family members feel that the disease has some impact on their lives[39]. Graph B in Figure 5 presents their perspective on whether the condition has impacted relationships with others and provides insight into social limitations brought upon by this disease[39]. Most IBD patients and

their family members found the condition had some impact on their interpersonal relationships; however, there were still a handful of IBD patients who felt the disease had a significant impact[39]. Graph C in Figure 4 presents the survey response regarding mental wellness (2015). IBD patients found the condition had some impact on their mental health, and more than 50% of the family members also felt it had some effect on their mental wellness[39].

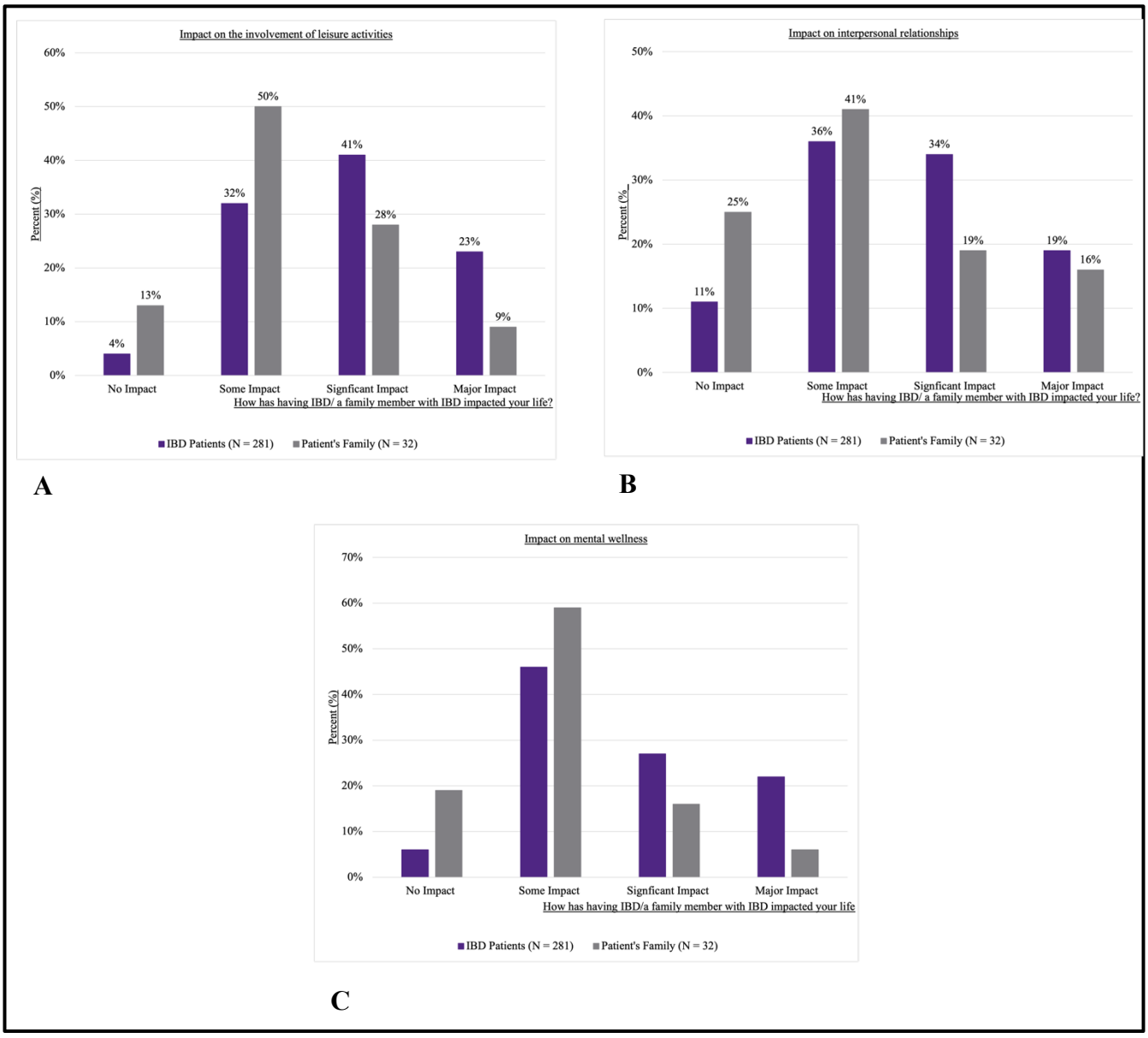

Figure 5: Series of three graphs of survey answer responses depicting IBD's social and mental wellness impact on the lives of patients and their family members. Graph A informs about leisure activities, Graph B informs about interpersonal relationships, and Graph C on mental wellness[39].

Such results emphasized the necessity of support tools and resources for the mental health of those with IBD and their caretakers or family members in Canada[39].

1.1.3 Differentiating between Crohn's Disease (CD) and Ulcerative Colitis (UC)

Although CD and UC are classified under the comprehensive group of IBD, they are generally distinctive in their pathology and symptomology[3]. The terminology used to describe both disease forms is typically based on the additional clinical, medical, or biological characteristics defining that unique case[2]. For example, CD can also be referred to as regional enteritis or regional/Crohn's ileitis[2] and can be assigned a further specified title dependent on whether the disease affects the colon, ileum, small intestine, or gastrointestinal (GI) tract[40]. Comparatively, UC cases can also be described with a more specific diagnosis of proctitis, proctosigmoiditis, pancolitis, or left-sided colitis, based upon the disease location[41]. The most significant difference between CD and UC is that CD can manifest anywhere in the gastrointestinal tract, which begins at the mouth and ends at the anus[40], compared to UC, which only occurs within the colon[42] and begins inflammation at the rectum[43]. Specific symptoms also help to differentiate a diagnosis of UC from CD. Blood in feces, generation then expulsion of mucus in the tract, and crypt branching are usually associated more often with UC[43]. In contrast, inflammation in elevated segments of the GI tract, blockage of the small intestines or colon, bulking of mass near the stomach, and fistulas are relevant to CD[42].

Diagnostic imaging tools are imperative in helping confirm Crohn's disease within patients[40]. When symptoms are subliminal, an ileocolonoscopic exam is conducted with a biopsy to ensure the patient can be diagnosed with Crohn's disease[40]. In more severe cases, patients undergo a computerized tomography (CT) scan instead to determine better the stage at which the condition

has progressed[40]. For ulcerative colitis, before a diagnosis is made, there must be standard UC symptoms and an endoscopic examination which shows inflammation beginning at the rectum and following continuously into the colon[44].

Table 1: Summary Table of Biological and Epidemiological Differences between CD and UC

	Crohn's Disease	**Ulcerative Colitis**
Where can the disease occur within the human body?	Occurs anywhere along the gastrointestinal tract[40]	Is restricted to occur within the colon[40]
Symptoms commonly associated with the disease?	Diarrhea[45] Reduction in body mass[45] Fever outbreaks[45] Blood in stool[40] Stomach pain[40]	Incontinence[46] Fatigue[46] Stomach cramps[46] Diarrhea with sporadic bleeding[41] Sensations of wanting to pass stool when not able to produce stool (tenesmus)[41]
Which is the most common surgery for managing this disease?	Surgical Procedure: Ileocecal resection[47]	Surgical Procedure: Colectomy[46]
The age group most likely to exhibit the disease?	From teenagers starting at 15 years to adults of 35 years of age[48]	From teenagers starting at 15 years to adults of 45 years of age[48]
The sex which is most likely to exhibit this disease?	Generally, trends of incidence have been higher in females than males[48, 49]	Generally, trends of incidence are equivalent in females and males[48,49]
Canadian province with the highest prevalence* of the disease?	Nova Scotia (as of 2008): = 412 per 100 000 persons[50]	Nova Scotia (as of 2008): = 350 per 100 000 persons[50]

Canadian province with the lowest prevalence* of the disease?	British Columbia (as of 2008) = 228 per 100 000 persons[50]	Quebec (As of 2008) = 168 per 100 000 persons[50]
Projection of disease burden* of this disease within Canada by the year 2030?	= 493 per 100 000 persons[50]	= 436 per 100 000 persons[50]

Note:

*Prevalence rates were standardized by age and sex for the population of Canada as for that pertaining year [50]

1.2 Navigating IBD Disease Management using RCTs

Since there is no cure for IBD at present, IBD patients must rely entirely on induction and maintenance therapies to achieve remission of symptoms[42]. Remission can be achieved through the prescribed use of anti-inflammatories, antibiotics, antibodies[51], corticosteroids, immunomodulators, biologics, or alternative medicine or therapies, individually or in combination[40]. For ulcerative colitis, drugs that can be used for management include olsalazine, balsalazide, mercaptopurine, infliximab, ciclosporin, or tacrolimus[42]. In contrast, drugs such as budesonide, thiopurines, methotrexate, and 6-mercaptopurine are more commonly used for CD[40]. Select drugs can treat UC or CD, including azathioprine, 5-aminosalicylates, and prednisone[40].

Since the goal is inducing remission and maintaining this state, many of these competing treatments are constantly being further studied and developed alongside others to improve the lives of IBD patients. Well-informed and robust RCTs can prove advantageous in studying how to achieve optimal efficacy of these drugs or alternative treatment approaches. Although RCTs are often conducted in a two-arm format, implementing multiple-arm trials can assist in advancing treatment strategies for patients with so many competing alternatives to study within the realm of IBD,

1.2.1 From Traditional Two-Arm RCT Design to Multi-Arm Trials

The classic set-up for a randomized control trial consists of a single treatment or experimental group and a control or comparison group[52]. This direct comparison of a single treatment group with that of a placebo group allows an investigator to determine whether the occurrence of a phenomenon can be attributed to the treatment or mere chance[52]. In an RCT, each group of participants can also be referred to as a study arm as these groups are an extension of the study's main sample, which in turn is desired to represent their population of interest[52].

Two-arm trials are the most used design for RCTs as they offer the simplest comparison method. A trial with more than two arms is called a multi(ple)-arm trial. Although multi-arm trials can be set up not to require a standard control group (i.e., non-inferiority trials)[53], this thesis will focus only on randomized trials, which are placebo controlled. Multi-arm trials can be categorized under adaptive trial designs as they don't conform to the fixed trial design, which includes the classic two-arm studies since they address the required study objectives with modifications that do not alter the overall integrity of the statistical validity of the trial[54].

However, available adjustments still maintain the trial's effectiveness and include multiple outcome assessment periods, increased transparency on the choice of multi-arm design, and statistical adjustment procedures if required for multiple testing[54]. A famous trial which has utilized the multi-arm design to test different interventions simultaneously to accelerate the evaluation process effectively is the STAMPEDE trial[54,55]. This multi-arm trial demonstrates the qualities of an adaptive design as the investigators are willing to drop any arms which were not successful as the trial progressed[54]. However, since multi-arm trials often introduce specific complexities and considerations, they are less commonly observed in the literature than two-arm trials[56].

Trials with multiple arms can be advantageous in cost-effectiveness, reduced turnover time, and practical resource allocation[54]. Phase three trials which require new treatments or drugs to be tested against a standard, could benefit most from multi-arm trials as they can be expensive and resource intensive to conduct[57]. Using a multiple-arm set-up in a phase three study, it would

be possible to collectively test different treatments and dose quantities or schedules against a standard in one trial, reducing the need for an otherwise elongated assessment process[57]. However, it is imperative that multi-arm trials clearly state their methods and analysis steps. Hence, it is coherent to the audience how certain complexities were responded to and to inform further how future multi-arm trials in the field can be conducted[58]. Juszczak and colleagues have written an extension of the Consort 2010 Statement, a guideline focusing on the conduct of randomized control trials[58]. Since the original Consort document generally focuses on two-arm trials, Juszczak et al. have formulated a list to inform the types of information multi-arm trials should address in their publications to increase the reproducibility and clarity of such research[58]. The guideline touches upon some of the twenty-five items outlined in the original consort and provides an extension of which things should give more detail regarding multi-arm designs[58]. Such extensions include explicit detail on segments of the study, such as an outline of the hypotheses, trial set-up description, sample size estimation, effect size, and statistical analysis pursued[58]. In the following chapters, this thesis will explore the use of this trial design for RCTs within IBD literature and expand on their utility around multiple comparison procedures, which aim to control the type 1 error rate during the analysis of multiple arms.

1.3 Objectives of this Thesis

This thesis aims to review the basic features of the multiple-arm design and its frequency in evaluating treatment effects for inflammatory bowel disease. The thesis contains a literature review to assess which multi-arm trials within this research area would be suggested to use multiple comparison procedures and, from those recommended trials, which have used these procedures to retain the statistical validity of their results. Upon assessing this, the thesis provides an example of integrating commonly used multiple comparison procedures into data analysis of an IBD randomized control trial. Recommendations are presented on the steps that may be taken when choosing the appropriate statistical analysis method for specific data when completing a multiple-arm study analysis. This thesis aims to use its content to remind readers that multiple arm trials are valuable, especially in IBD. Still, it is essential to continually evaluate its requirement for appropriate multiple adjustment methods to improve the robustness of their research methods and the replicability of their results when suggested.

Chapter 2: A Brief Review of Methods for Multiple Comparisons Procedures

2.1 A Brief Introduction to Multi-arm Trials

Whether to test multiple interventions with one another, to test them against their paired combinations, or to a common control group, multi-arm trials provide an effective solution for testing two experimental groups at once[58]. Compared to a two-arm randomized control trial, using a multi-arm randomized control trial design includes several advantages, such as the ability to evaluate more than one research question at a time, an increased pace at which interventions can be assessed, and the ability to give patients quicker access to interventions which can optimize their health[59]. Additional to its research and patient-related benefits, multi-arm trials provide logistical advantages such as a reduction in the budget that would otherwise be required to conduct separate investigations of the research questions[53] and the collective use of resources, personnel, and knowledge to address both projects[60]. However, the application of multi-arm trials is more complex, and statistical challenges remain in the analysis stages of the trial[53]. The critical statistical challenge is maintaining a nominal Type I error rate, often set at the 5% level[60]. This chapter will introduce some standard multiple comparison procedures (MCPs) used to preserve the type 1 error rate in multi-arm trials when necessary, and the utility of these techniques will also be further explored in the proceeding chapters.

2.2 Type I Error in Two Arm Trials

The Type 1 error rate can be defined as the probability that a test statistic rejects the null hypothesis of no effect correctly when no effect exists[60]. This is also referred to as a false positive rate[61] and is denoted by the alpha (α) symbol [61].The α is a critical value that must be defined during the statistical analysis stages or prior and is just as crucial as obtaining a p-value using appropriate test statistics[62]. The test statistic is used to determine a respective p-value, which is defined as the probability of obtaining a test statistic that is at least as large as the one

obtained from the trial, assuming the null hypothesis is true[62]. The alpha cut-off, which was primarily determined, is then used to assess the p-value where a p-value less than the alpha level suggests rejection of the null hypothesis[63]. For the conventional alpha level of 5%, there is a 5% chance of reaching an incorrect conclusion and a 95% chance of arriving at the correct conclusion[64]. In the case of a traditional two-arm trial that is placebo-controlled or otherwise, only a single comparison is made. Therefore, only a single null hypothesis is being assessed against the alpha cut-off of, for example, 5%[60]. With only a single null hypothesis, the desired alpha cut-off level is preserved, and that is why there is no requirement for additional correction steps needed when dealing with two-arm trials and their type 1 error rate[60].

2.3 Type I Error Rate in Multiple Arm Trials

Multi-arm, which features multiple comparisons, will have more than one null hypothesis and this can lead to the inflation of the type 1 error when statistical adjustments are not made, especially when recommended[60]. To provide further context on why simply setting a 5% cut-off cannot translate to a case of multiple comparisons, the following Example 1 should be considered.

Example 1:

Traditional Two-Arm Trials: Case 1

Comparison Group 1 vs. Comparison Group 2	H_0 with α =5%

Alternative Multi-Arm Trials: Case 2

Comparison Group 1 vs. Control Group	$H_{0\text{-}1}$ with α =5%
Comparison Group 2 vs. Control Group	$H_{0\text{-}2}$ with α =5%

In Case 1, when a single comparison is being conducted for a traditional two-arm trial, there is only one corresponding null hypothesis (H_0), and so the probability of a false positive occurring is considered for one event[60]. However, in Case 2, there are three overall comparison

groups and two comparison combinations, as shown. Each comparison has its respective null hypothesis and its own cut-off alpha. If Case 2 still adopts a 5% alpha level for each comparison, the overall type I error of incorrectly rejecting at least one null hypothesis will be larger than the intended 5% level[64]. This is because there are two events with their individual probability. Since each has a 95% chance of being correct (based on the 5% cut-off level), the probabilities, when multiplied for each comparison, are 1- (0.95 x 0.95) = 0.0975. In general, in a multi-arm trial with k comparisons, the error rate of mistakenly rejecting at least one null hypothesis is given by:

$$1 - 0.95^k$$

Thus, the error of erroneously rejecting at least one null hypothesis, regardless of which and how many null hypotheses are true, is termed the familywise error rate (FWER)[60]. There must exist some form of management when it comes to error rates because inflated results from RCTs can result in the approval of ineffective drugs or interventions, and that is why many regulatory organizations expect strict control of these rates when performing multi-arm trials[60] . Such include The International Council for Harmonization of Technical Requirements for Pharmaceuticals for Human Use (ICH) guidelines on statistical principles for clinical trials[65], which states that: "*in confirmatory analyses, any aspect of multiplicity…should be identified in the protocol*; *adjustment should always be considered and the details of any adjustment procedure… should set out in the analysis plan*[66]." Examples of procedures which can improve the statistical analysis of multiple-arm trials are discussed.

2.4 Background on four types of MCPs

The *Bonferroni Correction* is a conservative approach named after Carlo Bonferroni, a statistician who developed the initial concept[67]. It is a commonly used approach for controlling the type 1 error rate by correcting the α cut-off for each comparison to ensure that they sum to the overall desired α[67]. As described later, the method works by dividing the desired α by the number of required statistical significance tests it plans to perform concurrently[68] resulting in a

smaller quotient than the overall α. Doing this ensures that, as a collective, each test does not surpass the cut-off level and therefore reduces the chance of type I error occurring[69]. The Bonferroni method does have a few drawbacks such that it is comparatively a less powerful procedure, and this has an impact on the type II error rate[69]. But, in clinical settings, a type I error is considered more problematic than a type II error[69]. Additionally, the Bonferroni test establishes an overarching universal null hypothesis which, once rejected, shows that the comparisons are statistically different [68]. This approach makes it problematic to verify which comparisons from the collective are responsible for this difference[67]. Furthermore, the Bonferroni test determines the significance level for each comparison based on the total number of comparisons. Formulating the significance level in this manner can result in a range of significance values which might be drastically low if this procedure happens to be used in a trial with numerous comparisons[70]. Therefore, this test, although commonly utilized, would not be ideal for multiple comparison trials with many arms[71].

An extension of Bonferroni's procedure is the *Holm* or *Bonferroni-Holm procedure*, which was introduced by Sture Holm in 1979[72]. Inspired by the method of Carlo Bonferroni, the Holm method was aimed at using the basic concepts of the Bonferroni method but targeted increasing the power of the correction while still retaining the ability to control the type 1 error rate[73]. Holm's method is viewed as a more reasonable procedure compared to the Bonferroni method because it generates its adjusted p-value not by just dividing the significance level by the number of comparisons but by dividing it by the number of comparisons minus the rank of the comparison plus one, which allows the adjusted p-value to take into consideration the comparisons being evaluated[73].

An alternative method that ensures better stability for both the type I and type II error rates is the *Hochberg Sequential Approach*[74], as it retains more statistical power than the Holm and Bonferroni method[75]. Like Holm, this method depends on ranking the comparisons; however, Holm is a step-down procedure, and Hochberg is a step-up procedure, which will be explained further in the chapter[76]. While Bonferroni controls the family-wise error rate, Hochberg tackles the false discovery rate through its sequential order of testing the calculated statistics[77].

A fourth method, named *Dunnett's test,* compares each of the many experimental groups are analyzed against a common control group using the *Studentized t-statistic* distribution[78].

Dunnett's test aims to compare the experimental groups' mean value with the control group's mean value and considers their absolute difference against the test statistic[78]. The critical value for Dunnett's procedure considers any correlation in the test statistics of each comparison[79]. Examples of additional tests which can also be used to control for type 1 error in multiple comparison analyses but are not included in this thesis are: Tukey's HSD (honestly significant differences) test, Sidak's test, Fisher's LSD (least significant difference) test, and Scheffe's test[80]. These tests are not included as they were not evident within the literature review studies in Chapter 3.

2.5 Single-step Procedures for Multiple Comparison

Single-step MCPs are unique because they consider a single hypothesis at the time and do not assess them collectively, allowing these hypotheses to be evaluated in no predefined order[81]. The following two sections will review two examples of single-step MCPs: the *Bonferroni* method and the *Dunnett* method, with further information about these techniques being provided.

2.5.1 Bonferroni Method

The most known Bonferroni adjustment for multiple comparisons is a single-step procedure based on the Bonferroni inequality [82]. To understand the underlying context of the Bonferroni correction, refer to Example 2.

Example 2:
Let H_1 and H_2 be two null hypotheses. The Bonferroni inequality states[82],

$$\Pr(\{\text{reject } H_1\} \text{ or } \{\text{reject } H_2\}) \leq \Pr(\{\text{reject } H_1\}) + \Pr(\{\text{reject } H_2\}) \quad (1)$$

i.e.,

$$1 - \Pr(\{\text{reject } H_1\} \text{ or } \{\text{reject } H_2\}) \geq 1 - [\Pr(\{\text{reject } H_1\}) + \Pr(\{\text{reject } H_2\})] \quad (2)$$

or

$$\Pr(\{\text{not reject } H_1\} \text{ AND } \{\text{not reject } H_2\}) \geq 1 - [\Pr(\{\text{reject } H_1\}) + \Pr(\{\text{reject } H_2\}) \quad (3)]$$

From the above example, two null hypotheses are recognized, and the first line states that the probability of rejecting null hypothesis one (H_1) or rejecting null hypothesis two (H_2) must be less or equal to the sum of the probability of rejecting H_1 and the probability of rejecting H_2[82]. This probability principle from line 1 can then be written in its complement form, and this is shown as two options in lines 2 and 3. Line 2 states that one minus the probability of rejecting H_1 or rejecting H_2 is greater or equal to one minus the probability of rejecting H_1 and rejecting H_2. Otherwise, this can be written as shown in line 3, where the probability of not rejecting H_1 and H_2 is greater or equal to 1 minus the probability of rejecting H_1 and rejecting H_2.

Therefore, in the case of the two null hypotheses, if one uses 0.05/2 for each hypothesis, then the probability of making the correct decision is at least 95%, and the FWER is at most 5%. This suggests that for a multi-arm trial with k comparisons, one can use α/k for each null hypothesis to control the FWER [83].

This procedure can also be applied using the Bonferroni adjusted p-values, denoted here by $\tilde{p}_j$[84]:

$$\tilde{p}_j = \min\{kp_j, 1\}$$

where, pj refers to the raw p-values for the jth hypothesis, and $\tilde{p_j}$ refers to the adjusted p-values[71]. This value can be interpreted as evidence against the corresponding null hypothesis when all comparisons are considered collectively[84]. To evaluate the statistical significance of each adjusted p-value and its corresponding null hypothesis, one must compare each of those adjusted p-value with the desired nominal alpha level for the trial[84]. However, a limitation of the Bonferroni procedure remains that it lacks the potential to detect a correlation between test statistics and, therefore, can be a very conservative procedure, leading to low statistical power[67].

2.5.2 Dunnett Method

Several alternative MCPs, including the Dunnett procedure, use the correlation between the generated test statistics to perform multiple comparisons[85]. Dunnett's procedure proves advantageous when comparing multiple active interventions with a single shared control group [78]. It generates test statistics for the comparison between the experimental arm and the control arm using the following equation[78]:

$$\text{corr}(\bar{y}_i - \bar{y}_0, \bar{y}_{i'} - \bar{y}_0) = \frac{\text{var}(\bar{y}_0)}{\sqrt{\text{var}(\bar{y}_i - \bar{y}_0)\text{var}(\bar{y}_i - \bar{y}_0)}} = \sqrt{\frac{n_i n_{i'}}{(n_i + n_0)(n_{i'} + n_0)}}$$

Within the equation, $\bar{y}_i$ and $\bar{y}_0$ denote means of *i*th experimental arm and control arm, respectively, with their corresponding sample sizes denoted by n_i and n_0, where $i = 1, 2, .. k$. This equation will generate critical values from a bivariate normal distribution, and the test statistic will require an assumption of normal data with common variance[78]. The absolute differences between each experimental arm and the control groups are compared to the test statistic value to determine whether the difference can be deemed statistically significant per this procedure[78].

A disadvantage of single-step procedures is that they may have low power[86], while the advantage is that they can be used to obtain confidence intervals through the simple inverting of tests[87]. Although Dunnett's test requires the assumption of a normal distribution[88] and does not allow comparisons between experimental arms that are not the placebo, the procedure can perform one-tailed or two-tailed testing[76].

2.6 Stepwise Procedure for Multiple Comparison: Step-up and Step-down

MCPs are set up to examine whether an individual comparison is statistically significant based on whether another comparison is also statistically significant[89]. The comparison, which is used to determine the statistical outcome of the other, is set by its place in the predefined order and its statistical outcome[76]. Stepwise procedures can be categorized as step-down procedures or step-up procedures[76]. Step-down procedures begin with the first comparison within the predefined order and continue through the order until it discovers a statistically different comparison[90]. This procedure then deems all preceding comparisons in the list not statistically significant and all those following it to be statistically significant[90]. Step-up procedures instead begin at the last comparison in the predefined order and continue through the order until it discovers a comparison which is not statistically significant[90]. The following two sections will

review two examples of single-step MCPs: the *Holm* method and the *Hochberg* method, with further information about these techniques being provided in subsequent sections.

2.6.1 Holm Method: Step-down Procedure

The Holm method is a step-down procedure based on the Bonferroni method and is sometimes referred to as the Bonferroni-Holm method[76]. This method is praised for being a simplistic approach to account for multiple comparisons[88], and it is adaptive to use in most data analysis scenarios[91]. The Holm method performs analysis of multiple comparisons by placing the p-values of each comparison in ascending order and following the proceeding outlined in steps[90].

Let $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(k)}$ be the ordered p-values and $H_{(1)}, H_{(2)}, \ldots, H_{(k)}$ be the corresponding hypotheses.

Step 1: if $p_{(1)} > \alpha/k$, then stop and retain all hypotheses. Otherwise, reject $H_{(1)}$ and proceed to step 2

Step 2: if $p_{(2)} > \alpha/(k-1)$, then stop and retain $H_{(2)}, \ldots, H_{(k)}$. Otherwise, reject $H_{(2)}$ and proceed to step 3 …

Step k-1: if $p_{(k-1)} > \alpha/2$, then stop and retain $H_{(k)}$. Otherwise, proceed to step *k.*

Step k*: if* $p_{(k)} > \alpha$*,* then retain $H_{(k)}$. Otherwise, reject $H_{(k)}$.

Alternatively, the procedure can be carried out by first obtaining the adjusted p-values and then comparing the adjusted p-values, each with an overall alpha value of 5%. The adjusted p-values (letter p with ~ on top) for each comparison are obtained through the following process[92] .

For $H_{(1)}$, $\tilde{p}_{(1)} = kp_{(1)}$

For $H_{(2),}$ $\tilde{p}_{(2)} = \max\left[\tilde{p}_{(1)}, (k-1)p_{(2)}\right]$ …

For $H_{(k-1)}$, $\tilde{p}_{(k-1)} = \max\left[\tilde{p}_{(k-2)}, 2p_{(k-1)}\right]$

For $H_{(k)}$, $\tilde{p}_{(k)} = \max\left[\tilde{p}_{(k-1)}, p_{(k)}\right]$

The adjusted p-value for each comparison must be compared with the desired overall alpha level (commonly 5%) to determine whether a hypothesis being tested is statistically significant[92].

2.6.2 Hochberg Method: Step-up Procedure

Hochberg's method proposed a procedure very similar to Holm's, except that it starts from the largest p-value (i.e., the most insignificant comparison) found at the bottom of the predefined list and follows up this list[74,90]. The steps used to conduct this procedure are shown below[74].

> Again, let $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(k)}$ be the ordered p-values and $H_{(1)}, H_{(2)}, \ldots, H_{(k)}$ be the corresponding hypotheses.
> Step 1, if $p_{(k)} \leq \alpha$, stop and reject all $H_{(1)}, H_{(2)}, \ldots, H_{(k)}$. Otherwise, retain $H_{(k)}$ and proceed to step 2
> Step 2, if $p_{(k-1)} \leq \alpha/2$, step and reject $H_{(2)}, \ldots, H_{(k)}$. Otherwise, retain $H_{(k-1)}$ and proceed
> …
> Step *k-1*, if $p_{(2)} \leq \alpha/(k-1)$, stop and reject $H_{(k-1)}, H_{(k)}$. Otherwise, retain $H_{(2)}$ and proceed
> Step *k*, if $p_{(1)} \leq \alpha/k$, reject $H_{(k)}$. Otherwise, retain $H_{(k)}$.

The generated adjusted p-values for each comparison are obtained through the following process[74].

> For $H_{(k)}$, $\tilde{p}_k = p_{(k)}$
> For $H_{(k-1)}, \tilde{p}_{(k-1)} = \min\left[2p_{(k-1)}, \tilde{p}_{(k)}\right]$ …
> For $H_{(2)}$, $\tilde{p}_{(2)} = \min\left[(k-1)p_{(2),}\, \tilde{p}_{(1)}\right]$
> For $H_{(1)}$, $\tilde{p}_{(1)} = \min\left[kp_{(1)}, \tilde{p}_{(2)}\right]$

Like Holm's method, to determine whether a p-value is statistically significant, the adjusted value for each comparison must be compared with the desired overall alpha level, usually 5%[90].

2.7 Summary and Additional MCPs

This chapter has summarized four procedures for multiple comparisons commonly seen in multi-arm trials and will be referenced later in this thesis. Except for the Dunnett which uses GLM function, all procedures are based on raw p-values and have been implemented in the SAS procedure PROC MULTTES. This analysis will be featured in Chapter 4 of this thesis.

Other procedures mentioned in the literature are *O'Brian-Flemming* and *Lan-DeMets,* which are used for interim analyses and not necessarily specific to multi-arm trials. The use of a hierarchical testing procedure as a method to control for inflation of error is common as well, however since this procedure is often established prior to data analysis[93], it is referenced as a *varied procedure* further along on the thesis. There is also mention of the procedure MCP-Mod[94] which is an integrative methodology using MCPs and modeling to determine dose relationships. Due to its two-step nature, this method is recorded for its frequency of use in the next chapter but could not be compared with traditional MCPs. The Benjamini-Hochberg procedure is not included in further sections of this thesis as it is more relevant to genomic studies where the number of hypotheses being tested is numerous and usually in the order of thousands or millions[95,96].

Chapter 3: Literature Review

3.1 Introduction: Multiple Comparison Adjustments

Multi-arm trials have become a pragmatic option within drug studies as they provide an efficient alternative to traditional two-arm trials[97]. These studies expedite information inexpensively regarding prevalent topics such as dose-finding, competing treatment drugs, or efficacy and safety[97]. The field of IBD is active in examining and testing many available treatment options and determining optimal dose ranges of conventional drugs; therefore, multi-arm trial design can be helpful in collectively expanding information while attaining lower costs, reduced resources/time, increased opportunity for participation, and convenience[97]. Although multi-arm trials can be an attractive option, it is essential to acknowledge that certain design elements must undergo consideration to reduce possible analysis errors and inappropriate extrapolation of results[57,60]. Even with the focus of this thesis being on statistical procedures that can increase the internal validity of a trial[57,60], logistical considerations are also crucial in trial design. These include implementing a shared control group, periodic analysis for adaptive trial designs[57], planned *stop* rules, recording/reporting recruitment strategies, and blinding of participants[58] .These approaches are helpful to ensure the trials generate robust results.

Maintaining internal validity in trials is critical to ensure the trial design can generate results representative of the intervention's true effect size[98]. The random error caused by the inflation of type 1 error can lead to decreased internal validity, and therefore it must be closely monitored in multi-arm trials[98]. The type I error rate is the odds of having a false positive or accidentally rejecting a null hypothesis that should have failed to be rejected[79]. As introduced in Chapter 2, each study's type I error rate is constrained by the significance level, defined as the probability that the null hypothesis will be rejected if this study was performed an inexhaustible number of times[69]. This is referred to as alpha (α) and is commonly predetermined to be 0.05[69]. With multiple arms in a trial, the goal is to be vigilant of the family-wise error rate (FWER), which is the chance of at least one of the comparisons resulting in a type I error[60]. To ensure that statistical errors are accounted for in specific cases of the multi-arm trial design implementation, various forms of adjustments can be used, ranging from more to less conservative[60].

3.1.1 Are MCPs Necessary for Consideration?

However, there is still some disagreement on whether using MCPs in the cases of multi-arm trials is necessary[79]. Those who argue that MCPs are not mandatory often say this because the different experimental arms are concerned with various lines of inquiry[60]. Therefore, these lines of inquiry are viewed as separate hypotheses that would have been otherwise tested in different trials if not for this multi-arm set-up[60]. Additional arguments include that when traditional two-arm trials conduct analysis, they are not required to implement comparison procedures for all the other trials in the literature investigating similar topics, so why should multi-arm trials be required to do so instead[79]? It has been observed through literature that MCPs are not relevant in the case of exploratory trials as these trials aim solely to investigate the topic and are not forming conclusions or implementing change as is done in confirmatory trials[60]. Yet, in the cases of confirmatory trials, MCPs should be considered as they provide a solution for reducing the type 1 error and ensuring that the correct conclusions are being translated from the analysis to the application[79]. Encouraging the use of MCPs can reduce the number of studies that limit the recommendation of treatment strategies which otherwise could result in better outcomes for the patients of concern[79].

3.1.2 Determining whether to implement an MCP

Figure 6 presents a decision tool from an article written by Odutayo and colleagues[17] that helps navigate whether an MCP is necessary for a specific trial design. This assessment tool will be referenced consistently in the next few sections.

This tool presents three main decision points for the reader to follow to determine whether the trial in question would fall under the recommendation of using an MCP or would not necessarily benefit from one.

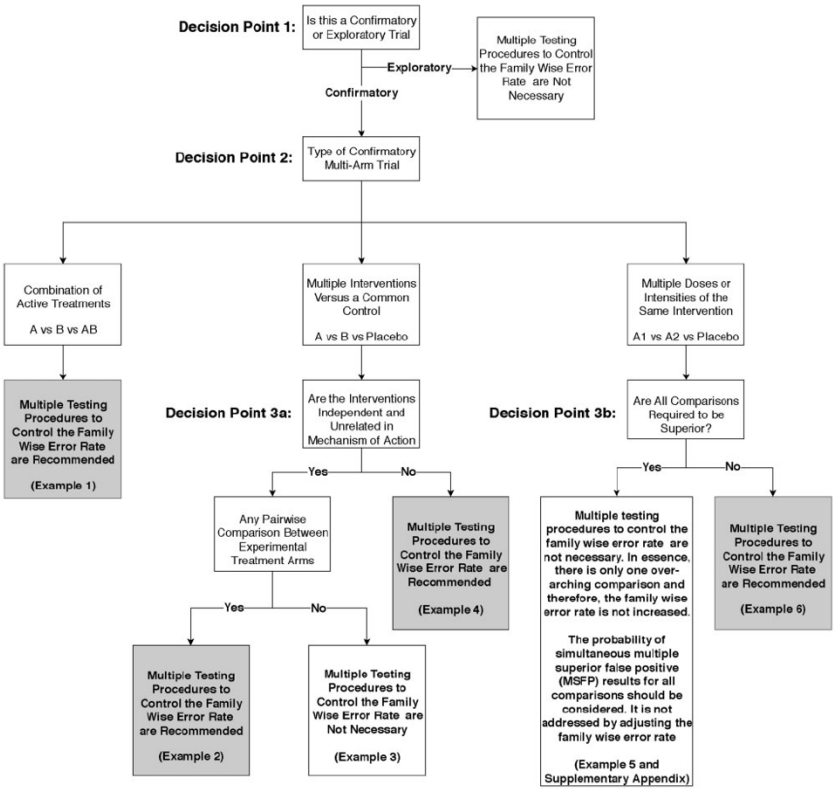

Figure 6: This is a published assessment tool for MCP that can help determine if recommending an MCP is required to maintain the type 1 error for a research trial[17]. It begins with Decision 1 which determines the type of trial being assessed. Decision 2 will classify the multi-arm trial by the designation of the arms in the trial, and Decision 3 will vary by the choice in Decision 2. Decision 3a is relevant to multiple intervention arms compared with a placebo, and it determines whether the interventions are related or not in their approach to attaining results. Decision 3b is relevant to multiple dose arms compared with a placebo and it determines whether the comparisons require superiority in their testing.

Decision 1 in the chart opens the reader to whether the trial is a confirmatory or an exploratory trial[17]. An exploratory trial is pursued to survey and begin a preliminary investigation into a topic of interest without attempting to answer specific questions or draw any conclusions[60]. Alternatively, a confirmatory trial has a clear objective or hypothesis that it plans to address within the trial and results in conclusive ideas of knowledge that could be translated into applicable settings[60]. According to the tool[17], trials considered exploratory are optional to implement an MCP and are generally excluded from pursuing further consideration because their intent for research is preparatory. Confirmatory trials follow through to Decision 2, which lets the reader determine the type of confirmatory trial associated with the research. The three main trial designs are those which compare individual experimental arms with combinations of those experimental arms, those which compare intervention arms with a placebo, and those which compare dose arms regarding doses with a placebo. Since this thesis concerns placebo-controlled trials, the first type of confirmatory trial listed will not be referenced further, and the latter two types will be mentioned further. Based on this decision, the framework for the thesis will be limited to Decision 3a or Decision 3b.

3.1.3 Purpose of this Literature Review

Within the field of IBD, it is unclear how often and why multi-arm trials can be used to pursue research about CD and UC. Additionally, it is not entirely understood how much consideration RCTs within this topic area give to MCPS when they have study characteristics which align with MCP recommendation tools such as the one put forth by Odutayo and colleagues[17], and the types of MCPs these trials use in cases requiring such statistical adjustments. This literature review aims to fill such gaps in knowledge by informing on the proportion of IBD trials using the multi-arm design, the current utilization of MCP in multiple-arm trials, and the types of MCPs being used in IBD research.

3.2 Methods

A general sample of placebo-controlled IBD RCTs was obtained from two published peer-reviewed systematic reviews and meta-analyses to analyze multi-arm trials in recent IBD literature extensively. The RCTs were retrieved from a database of IBD trials used for these systematic reviews and meta-analyses. Since these papers examined induction and maintenance trials for CD[99] and UC[100], the trials were organized by CD (induction and maintenance) and UC (induction and maintenance) within the database. These synonymous research papers aimed to assess placebo groups' general response and remission rate in placebo-controlled induction and maintenance trials concerning CD[99] and UC[100]. The trials included in these systematic reviews were aggregated from searching the databases of the Cochrane Library, EMBASE, Medline, and or the Cochrane Inflammatory Bowel Disease Review Group's Specialized Trials Register from the years the databases were established up till April of 2014[99,100]. The searches were not limited to specific languages and included abstracts from external sources, which were manually searched to retrieve additional studies[99,100]. However, the search strategy for these papers did require that the included studies were placebo controlled RCTs and had adult participants who were patients of active CD/UC and undergoing induction (2 weeks plus) or maintenance (4 months plus) treatments with corticosteroids, immunosuppressants, or aminosalicylate[99,100]. Trials were also required to have recruited and examined the involved patients using outlined assessment tools, which included the Disease Activity Index, Mayo Clinic Score, and Harvey-Bradshaw Index[99,100]. Trials which investigated fistulizing Crohn's disease, hospital-based cases with extreme ulcerative colitis, and utilized antibiotics, probiotics, or alternative complimenting treatments or technologies were excluded from these papers[99,100].

A repository of RCTs was formed for this thesis from the trials of the original database based on whether each trial met the requirements for this review. Since the papers had already established that all included trials were RCTs about IBD, the only inclusion criteria for this review were already met. Trials were excluded from the repository if they only had an abstract, as this review intended to extract data from the trial's methodology and result analysis. All accepted trials for the literature review were organized in the repository in four groups: CD – induction, CD – maintenance, UC – induction, and UC – maintenance. For each group, the trials were numbers starting from 1 in a non-random order to assign an identifier.

The decision tool from the publication by Odutayo and colleagues[17] was used to formulate the extraction table for this review. The extraction table recorded the title of the publication, the authors, and the number of arms in the trial. Trials with two arms were excluded from further data extraction, while trials with more than two arms were considered for further analysis. *Decision 1* was used to classify multi-arm trials as confirmatory, exploratory[17], or not specified based on their characteristics. Trials classified as exploratory were excluded from further analysis as researchers pursued these trials to examine the topic area without the onus of analysis[17]. Confirmatory or not specified trials were advanced to the next decision point.

Decision 2 focused on organizing the multi-arm trial by whether the experimental arms were intervention-based or dose-based. As mentioned, a *combination of active treatments* has been excluded from further consideration as the trials in this review must be placebo controlled. This means that trials were grouped into *multiple interventions vs. control*, *multiple doses vs. control*[17], and a third additional group of *multiple interventions and doses vs. control* which was amended to the original framework [17].

Trials which belonged to the multiple interventions vs. control group had to be analyzed by *Decision 3a,* which assessed if the multiple experimental arms studied were unassociated in their manner of effect or not[17]. If these experimental arms were related to the method of effect, then it was assessed if the trial had implemented an MCP, and this information was recorded along with the type of procedure used[17]. Trials which used a specified multiple comparison procedure (such as those described in Chapter 2), or statistical software was recognized as 'adjusted using MCP', while those which used hierarchical or alternate adjustments were recognized as 'adjusted using varied procedure'. However, if the experimental arms were not related, then the trial was assessed to determine if the multiple experimental groups were compared against each other in a pairwise fashion, in addition to being compared with the control group[17]. If the unrelated experimental arms were compared against one another, then these trials were assessed for whether they implemented an MCP, and this information was recorded along with the type of procedure used[17]. If the unrelated trials were not compared against one another, then they were excluded from assessing if they required an MCP.

Trials that belonged to the multiple doses vs. control group had to be analyzed by *Decision 3b*, which assessed if the comparisons required superiority. If the trial did require superiority, then trial was excluded from requiring an MCP[17]. Otherwise, if there was no

superiority requirement then the trial was assessed for implementation of an MCP. This information was recorded along with the type of procedure used[17]. As for trials which used a statistical software, they were recorded as *'adjusted using MCP'* while those which used hierarchical or alternate adjustments were recognized as *'adjusted using the varied procedure'*.

Trials belonging to the multiple interventions and multiple doses vs. control group were required to be analyzed through both *Decision 3a* and *Decision 3b* and their respective lines of assessment questions. It was also recorded when a trial was excluded from using a multiple comparison procedure but used one anyway.

After analyzing these trials using the questions from the decision flowchart[17], additional information regarding these trials was extracted using the following prompts:

- Whether the trial had an adaptive or non-adaptive design.
- Whether the trial was industry sponsored or not.
- Whether a justification was provided for the use of multiple experimental arm trial design.
- Whether the statistical power was calculated to reflect the multiple experimental arms.
- Whether the sample size was calculated to reflect the multiple experimental arms.
- Total Sample Size and Randomization Ratio.
- Study Phase – Phase I, II, III, IV, or any combination.
- Type of journal the trial was published in –Speciality, Medical, or General Sciences.

These prompts were used to only extract supplementary information about the trials and were not the leading focus of the review.

3.3 Results

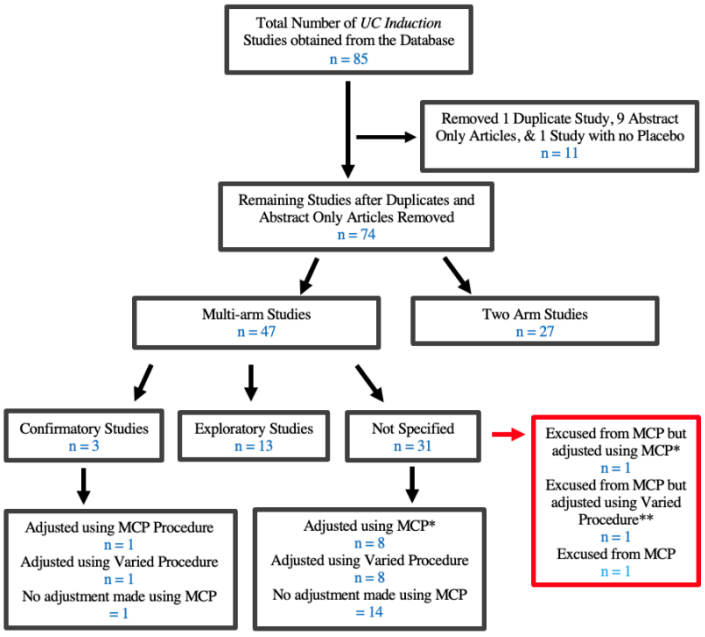

Figure 7: Flow chart depicting the breakdown of the trials included in the UC induction disease classification.

Note:

*Study compared multiple interventions and multiple doses, where the multiple interventions the approach utilized an MCP procedure, but the multiple doses approach was excused from requiring an MCP procedure

** Study compared multiple interventions and multiple doses, where the multiple interventions approach utilized a varied procedure, but the multiple doses approach was excused from requiring an MCP procedure.

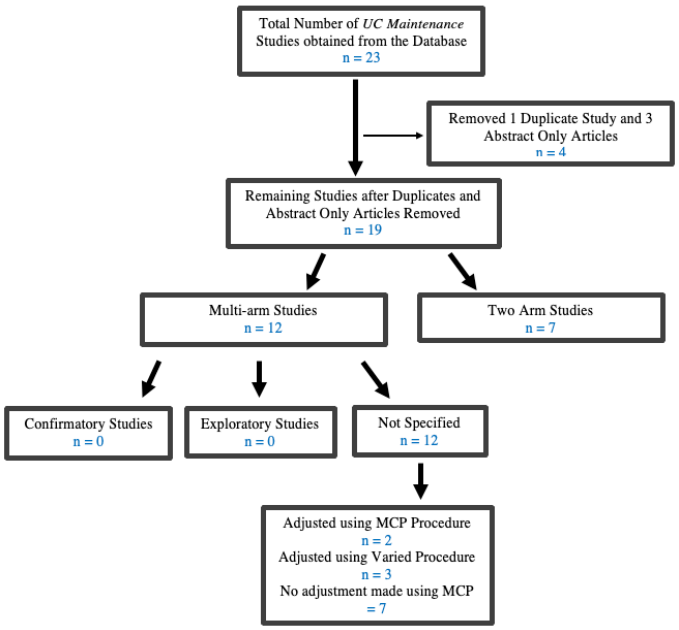

Figure 8: Flow chart depicting the breakdown of the trials included in the UC maintenance disease classification.

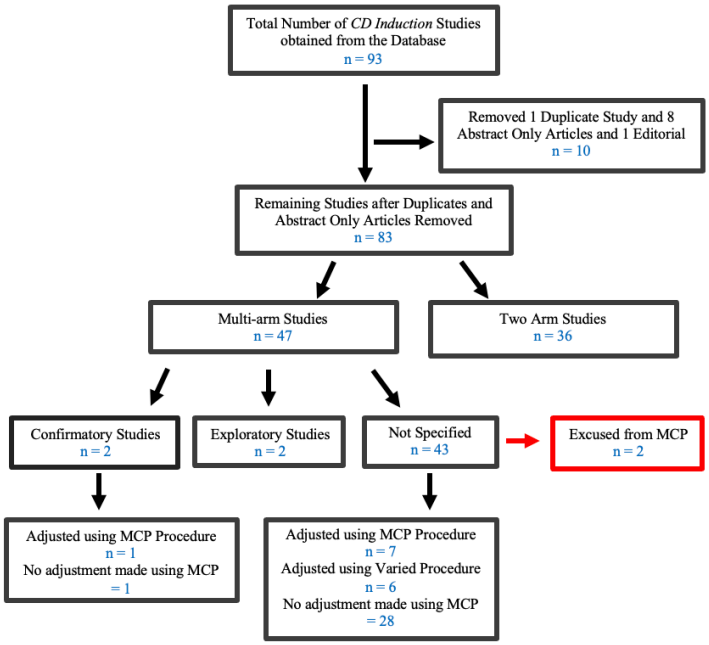

Figure 9: Flow chart depicting the breakdown of the trials included in the CD induction disease classification.

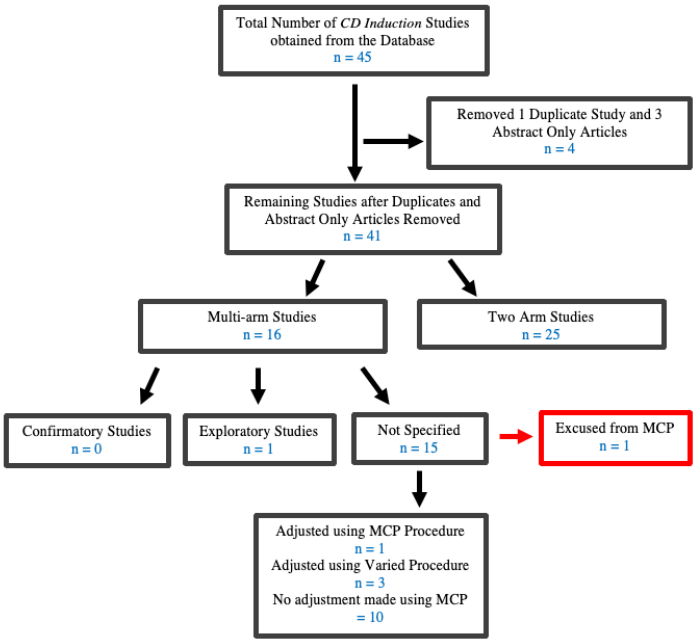

Figure 10: Flow chart depicting the breakdown of the trials included in the CD maintenance disease classification.

Table 2: Specific Details of the Trials categorized by their disease classification

	Ulcerative Colitis Induction Trials n (%); **n= 74**; **n= 47**; **n= 34** **n = 33**	**Ulcerative Colitis Maintenance** Trials n (%); **n= 19**; **n =12**; **n= 12** **n = 12**	**Crohn's Disease Induction** Trials n (%); **n= 83**; **n= 47**; **n= 45** **n = 43**	**Crohn's Disease Maintenance** Trials n (%); **n= 41**; **n =16**; **n= 15** **n = 14**
Number of arms included in study				
Two	27 (36)	7 (37)	36 (43)	25 (61)
Three	28 (38)	10 (53)	16 (19)	9 (22)
Four	9 (12)	1 (5)	22 (27)	7 (17)
Five	8 (11)	1 (5)	7 (9)	
Six	2 (3)		1 (1)	
Seven			1 (1)	
Trial Design				
Two-arms	27 (36)	7 (37)	36 (43)	25 (61)
Multi-arms	47 (64)	12 (63)	47 (57)	16 (39)
Type of Trial				
Dose	44 (94)	12 (100)	43 (92)	13 (81)
Intervention	1 (2)		3 (6)	3 (19)
Dose/Intervention	2 (4)		1 (2)	
Phase of Trial				
Phase 1			1 (2)	
Phase 2	21 (62)	4 (33)	30 (67)	7 (47)
Phase 3	11 (32)	8 (67)	14 (31)	8 (53)
Phase 2/3	1 (6)			

Adaptive or Non-adaptive				
Adaptive	5 (15)	4 (33)	15 (33)	7 (47)
Non-adaptive	29 (85)	8 (67)	30 (67)	8 (53)
Type of journal the trial published in?				
Specialty	23 (68)	8 (67)	37 (82)	12 (86)
Medical	10 (29)	4 (33)	6 (13)	2 (14)
General/Science	1 (3)		2 (5)	1 (7)
Was the trial industry sponsored/funded?				
Yes	45 (94)	10 (83)	39 (87)	13 (93)
No	3 (6)	2 (17)	6 (13)	1 (7)
Was justification provided for conducting a multi-arm trial?				
Yes	6 (18)	1 (8)	13 (30)	1 (7)
No	27 (82)	11 (92)	30 (70)	13 (93)
Does the calculated power reflect multiplicity?				
Yes	6 (18)	1 (8)	5 (15)	1 (7)
No	27 (82)	11 (92)	38 (88)	13 (93)
Does the calculated sample size reflect multiplicity?				
Yes	4 (12)	1 (8)	1 (2)	0 (0)
No	29 (88)	11 (92)	42 (98)	14 (100)

Note First ***n*** value depicts the total number of trials after exclusions.

The second ***n*** value depicts the total number of multi-arm trials for each disease classification.

The third ***n*** value depicts the total number of multi-arm trials after the exclusion of exploratory trials.

The fourth ***n*** value depicts the total number of multi-arm trials which were expected to use an MCP.

Ulcerative Colitis ***Induction*** n = 33	Multiple Comparisons Procedure	9 (27)	Bonferroni	2 (22)
			Dunnett	1 (11)
			Hochberg	3 (34)
			Holm	2 (22)
			MCP-Mod	1 (11)
	Varied Procedure	9 (27)		
	No Adjustment Made	15 (46)		
Ulcerative Colitis ***Maintenance*** n = 12	Multiple Comparisons Procedure	2 (17)	Bonferroni	1 (50)
			Dunnett	-
			Hochberg	-
			Holm	-
			MCP-Mod	1 (50)
	Varied Procedure	3 (25)		
	No Adjustment Made	7 (58)		

Figure 11: Flow-chart providing the breakdown of the Ulcerative Colitis trials and their use of an MCP and if so the type of MCP.

Crohn's Disease ***Induction*** n = 43	Multiple Comparisons Procedure	8 (19)	Bonferroni	4 (50)
			Dunnett	2 (25)
			Hochberg	2 (25)
			Holm	-
			MCP-Mod	-
	Varied Procedure	6 (14)		
	No Adjustment Made	29 (67)		
Crohn's Disease ***Maintenance*** n = 14	Multiple Comparisons Procedure	1 (7)	Bonferroni	1 (100)
			Dunnett	-
			Hochberg	-
			Holm	-
			MCP-Mod	-
	Varied Procedure	3 (21)		
	No Adjustment Made	10 (72)		

Figure 12: Flow-chart providing the breakdown of the Crohn's Disease trials and their use of an MCP and if so the type of MCP.

Figure 7 presents the breakdown of the trials included in the UC induction classification. Eighty-five induction trials were obtained from the UC induction trial repository, from which 11 trials were excluded for being either a duplicate study from the list or for only having an abstract. After exclusions, 74 trials remained, from which 47 were classified as multi-arm trials and 27 as two-arm trials. The two-arm trials were excluded from further analysis. From the multi-arm trials, three were confirmatory trials, 13 were exploratory, and 31 were not specified. Any exploratory trials were also excluded from the analysis. From the confirmatory trial, one trial was excused from requiring an MCP, while two other trials were partially excused but still required an MCP. The partially excluded trials were both multiple-interventions and multiple-dose trials. Even though the trials were excused from requiring an adjustment for their multiple doses analysis, they were recommended an adjustment for the multiple-interventions analysis. These trials are asterisked in Figure 7 to highlight that they should not be double counted as they are included in both the excluded and required MCP groups. From the trial not specified, eight were adjusted using an MCP, eight trials were adjusted using a varied procedure, and 14 trials did not use any adjustment procedure. From the confirmatory trials, one trial was adjusted using an MCP, one trial was adjusted using a varied procedure, and one trial did not use an adjustment procedure.

Figure 8 presents a similar breakdown of the trials included in the UC maintenance classification. Twenty-three maintenance trials were obtained from the UC maintenance repository, from which four were excluded for being either a duplicate study from the list or for only having an abstract. After exclusions, 19 trials remained, from which 12 were classified as multi-arm trials and seven were two-arm trials. These two-arm trials were excluded from further analysis. From the multi-arm trials, there were no confirmatory or exploratory studies. The 12 trials which were not specified had two trials which were adjusted using an MCP and three which used a varied procedure, and seven which did not use an adjustment procedure.

Figure 9 presents the breakdown of the trials included in the CD induction classification. Ninety-three induction trials were obtained from the CD induction repository, from which ten were excluded for being either a duplicate study from the list or for only having an abstract. From the 83 remaining trials, 47 were classified as multi-arm trials and 36 as two-arm trials. The two-arm trials were excluded from further analysis. From the multi-arm trial, two were confirmatory trials, two were exploratory, and 43 were not specified. Exploratory trials were

excluded from the analysis. Two trials were excused from requiring an MCP adjustment from the not specified trials. Seven trials were adjusted using an MCP, six trials were adjusted using a varied procedure, and 28 trials did not use an adjustment procedure. Among the confirmatory trials, one trial was adjusted using an MCP, and one did not use an adjustment procedure.

Figure 10 presents the breakdown of the trials included in the CD maintenance classification. Forty-five Crohn's disease maintenance trials were obtained from the CD maintenance repository. From the trials, four were excluded for being either a duplicate study from the list or for only having an abstract. From the 41 trials, 16 were classified as multi-arm trials and 25 as two-arm trials. The two-arm trials were excluded from further analysis. Of the multi-arm trials, one was an exploratory trial, and 15 were not specified. The exploratory trials were also excluded from the analysis. Among the not specified trials, one trial was excluded from requiring an MCP, one was adjusted using an MCP, three were adjusted using a varied procedure, and ten trials did not use an adjustment procedure.

Table 2 provides a summarized breakdown of how many studies which were obtained from the designated trial repositories were multi-arm trials and their respective features. This table consists of four columns to reflect the four disease classifications used to organize the trials in the database. Found underneath each disease classification are four *n* (total number of trials) values which descend in value to reflect how trials were segregated during analysis to obtain relevant data. The first (blue) ***n*** value depicts the total number of trials which remained after removing trials which were duplicates, abstract only, or were not placebo controlled. This ***n*** value is relevant to the total for the first two rows, which analyze the number of arms included in the trials and whether the trials were multi-arm or two-arm trials. For all condition classifications except CD maintenance trials, there were more multi-arm trials than two-arm trials. Among the multi-arm trials, three and four-arm trials were the most common.

The (green) ***n*** value in each column represents the total number of trials which remain after excluding all two-arm trials. This ***n*** value is reflected in the totals for the third row, which provides the breakdown for the types of multi-arm trials. Multiple doses compared with placebo were the most common multi-arm trial type for each classification. Each column's (yellow) ***n*** value represents all multi-arm trials, excluding exploratory ones. The rows in this section were intended to collect data regarding study features. They were inclusive to all multi-arm trials regardless of whether they required a multiplicity adjustment. This ***n*** value is representative of

rows four through seven. Row four provides a breakdown of the trial phase, showing that UC and CD maintenance trials were more often phase 3 trials, while UC and CD induction trials were more often phase 2. The next row informs whether trials were adaptive or non-adaptive in design. Trials from all condition classifications were most often non-adaptive than adaptive. The following two analyze the journal type in which the trials were published and whether the trial was sponsored or funded through an industry-related entity. Each condition classification had trials published in specialty IBD or gastrointestinal journals, while the remaining were published in medical journals, and fewer trials were in general science journals. Over 80% of trials were industry sponsored or funded among all the classifications.

The final (purple) ***n*** value is representative of all the multi-arm trials which were suggested to use an MCP and is relevant to the final three rows of the table. The proceeding row displays whether trials provided a statement of justification for why a multi-arm trial design was chosen. More than 70% percent of the trials did not provide a justification statement. The following two rows show whether trials provided a power and sample size calculation with consideration for multiplicity. Only up to 18% of trials across all condition classifications provided a power value adjusted for multiplicity. At the same time, up to 12% of trials across all categories provided a sample size calculation adjusted for multiplicity, apart from CD maintenance studies which had no trials which adjusted their sample size calculation.

Figure 11 and Figure 12 present a summary of the utilization of MCPs for each classification, along with the breakdown of which MCP was used to control the type 1 error rate. Figure 11 is specific to UC trials which were recommended to use an MCP. It displays the number and percentage (in brackets) of trials that utilized an MCP, a varied procedure or failed to implement an adjustment strategy. Within the UC induction group, most trials attempted to implement an adjustment strategy, whether an MCP or a varied procedure. For the 27% of trials in this group that implemented an MCP, the most used procedures in descending order were: Hochberg, Holm, and Bonferroni, Dunnett and MCP-Mod. While in the UC maintenance group, most trials did not attempt to implement an adjustment strategy. Among the 17% of trials in this group that implemented an MCP, the only used procedures were Bonferroni and MCP-Mod.

Figure 12 is specific to CD trials which were recommended to use an MCP. It displays the number and percentage (in brackets) of trials that utilized an MCP, varied procedure or failed to implement an adjustment strategy. Most trials did not attempt to implement an adjustment

strategy within the CD induction group. Of the 19% of trials in this group that implemented an MCP, the most common procedure in descending order was Bonferroni, then Dunnett and Hochberg. Similarly, most trials did not attempt to implement an adjustment strategy in the CD maintenance group. Only one trial in this group implemented an MCP and it was the Bonferroni method.

Additional tables and figures regarding the collected data can be found in the Appendix. Tables A through D presents the summary tables for the data during the extraction stage for only the multi-arm trials. The master table in the Appendix will also contain all extracted data and presents the complete list of all trials within each condition classification. The sequence of the columns for these tables reflects the layout of the decision tool table formulated by Odutayo and colleagues[17]. The primary column in these tables provides the identifier for the trials, which was used when organizing trials during the extraction phase of the review. This value was assigned to studies based on the order in which these studies were imported from the storage database and has been assumed as random. The following column lists how many arms were found in these multiple-arm studies. This value includes all experimental arms and the placebo arm. The subsequent column dealing with decision 1 (D1) records whether the trial is exploratory or confirmatory in its investigation. Exploratory trials were excused from multiplicity adjustments and did not undergo further analysis. The remaining trials were assessed using decision 2, which distinguished them by the type of multiple-arm trial. Knowing this determined whether further information would be extracted using decision 3a or 3b. Trials which examined multiple interventions compared to placebo were assessed using decision 3a. In contrast, trials which examined multiple doses compared to placebo were assessed using decision 3b. However, the few studies that analyzed a trial with multiple interventions and doses of the same intervention required that decisions 3a and 3b were used for assessment. The goal of Decision 3a was to investigate whether the interventions studied in the trial were unrelated or related in their manner of effect. If they were not, the trial was evaluated for whether it used an adjustment for multiplicity. This question was answered in a *yes* or *no* manner. If the answer was *no*, then it was commented on whether an MCP was used, and if it was a *yes*, it was then assessed if the comparisons were pairwise or not. If a pairwise comparison was conducted, it was commented on whether an MCP was used. However, if there were no pairwise comparisons, the analysis was excused from using a multiplicity adjustment.

Alternatively, decision 3b assessed whether the trial's objective was the superiority of all experimental doses compared to the placebo. If a trial required superiority, then it was excused from a multiplicity adjustment; however, if there was no requirement for the superiority of all experimental doses, then it was commented on whether a multiplicity adjustment was used. The following two columns were used to gather information on whether the power and sample size calculations for these trials also included an adjustment for multiplicity, and these were assessed in a *yes*-and-*no*-answer format.

Figures A, B, C, and D in the Appendix section present the distribution of the study designs included in each condition classification. Figure A shows that after excluding 12 percent of the studies, the UC induction trials included 56 percent of multi-arm studies and 32 percent of two-arm studies. Figure B shows a similar distribution of study designs for the UC maintenance trials, with 17 percent of studies excluded, 52 percent as multi-arm trials, and 31 percent as two-arm studies. As for CD, Figure C shows that from the induction studies, 11 percent were excluded, 53 percent were multi-arm, and 36 percent were two-arm trials. In Figure D, CD maintenance studies had 9 percent excluded, 36 percent multi-arm trials, and 55 percent two-arm studies. From all condition classifications, except the CD induction trials, multi-arm studies comprised the most significant proportion of trial designs compared to two-arm trials.

Figures E, F, G, and H in the Appendix section present the distribution of the study phases of multi-arm trials in each IBD condition classification. Figure E shows that UC induction trials had 6% phase 1, 60% phase 2, 32% phase 3, and 2% phase 2/3 trials. Figure F shows that the UC maintenance trials had 33% phase 2 and 67% phase 3 trials. Figure G shows that from the CD induction trials, 2% were phase 1, 65% were phase 2, and 33% were phase 3 trials. The CD maintenance trials comprised 44% of phase 2 and 56% of phase 3 trials. From all condition classifications, phase 2 and phase 3 trials consisted of the most significant proportion of the study phase.

Figures I, J, K, and L in the Appendix section present the distribution of multi-arm trial types among the multi-arm trials of each condition classification. For the UC induction trials in Figure I, 86% of trials were multiple doses versus placebo, while 10% were multiple interventions versus placebo, and 4% were multiple dose and intervention versus placebo. The UC maintenance trials in Figure J had 100% of the trials as multiple doses versus placebo. The CD induction trials in Figure K had 84% of trials as multiple doses versus placebo and 16% as

multiple interventions versus placebo. The CD maintenance trials in Figure L had 69% of trials as multiple doses versus placebo and 31% as multiple interventions versus placebo. Multiple doses versus placebo were the most common multi-arm trial type from all condition classifications.

3.4 Discussion:

The total sample of IBD trials included in this review showed that multiple-arm RCTs were commonly conducted. Within the four distinctive condition classifications, UC induction, UC maintenance, and CD induction had more than 50% of their trials as multi-arm in design. Among these three groups, three arms were prominent in the UC induction and UC maintenance groups, while four arms were most common in the CD induction group. Although multi-arm trials were less common in the CD maintenance group, three-arm trials were most common within those multi-arm trials. In scientific research, three-arm trials have been praised for ensuring assay sensitivity, the concept of separating interventions with effectual outcomes from those not producing effectual outcomes[101]. Multiple doses versus placebo were the most popular multi-arm trial type, with over 80% of trials from each condition classification using this design. Multi-arm trials provide researchers with the convenience of comparing competing treatment dosages, and it increases productivity as ineffective dosages can become easier to identify and exclude[97]. Since there are numerous treatment strategies available to induce and maintain treatment for patients based on their type of IBD and specific to their physiological requirements[102] there is less focus in IBD research for intervention studies and more on finding the optimal benefit from these treatment strategies through dose-finding trials. This can possibly explain the low number of multiple interventions versus placebo trials across the four condition classifications. This does not mean there is a lack of value in researching the effectiveness of IBD treatments against one another. Instead, the current concentration of research in this field is development on novel therapies to improve the therapeutic window.

The trial phases can provide important information regarding the research objectives that researchers aim to achieve using the multiple-arm design. Most trials in the UC induction and CD induction classification were phase 2, while those in the UC maintenance and CD maintenance group were mainly phase 3. The use of phase 2 clinical trials in induction studies in

both UC and CD induction studies can indicate that IBD induction treatments have established that their treatments are effective (at specific doses). However, the most efficacious treatment (doses) still needs to be determined to achieve the desired outcome of inducing remission[103]. Conducting phase 2 trials on diverse patient profiles using the multi-arm design is valuable because IBD symptoms can range from moderate to severe. Since induction treatments combat the disease into remission, they may benefit from the phase 2 trial design so that competing treatments are tested to evaluate which meet the specific needs of different IBD patient profiles[104]. UC and CD Maintenance trials were most often phase 3 trials, and this trial design focuses on assessing the safety of the proposed treatment and comparing it to the most accepted alternative treatment[105]. The increased use of the phase 3 trial design in IBD maintenance trials indicates that there is still testing around improving available or new treatment regimens while ensuring their effectiveness against the current standard treatment. Although phase 2 trials are commonly dose-finding studies, phase 3 trials within this literature review were also found to compare treatment dosages, which could be attributed to dose optimization. Further exploration of testing the treatment dosage against the standard dose within these phase 3 trials could suggest that researchers are trying to achieve more extended periods of remission for IBD patients using this multi-arm trial design. Whether these trials aim to research comparing multiple novel treatments in phase 2 studies or successful treatments with the current standards in phase 3 studies, multi-arm trials can work to introduce treatment approaches that can better accomplish a state of remission for patients of various profiles.

The utilization of the terms *confirmatory* and *exploratory*, as suggested by the referenced decision tool of Odutayo and colleagues[17], was challenging to implement due to the lack of use within studies to classify their objectives and methods. Since this was an issue which was noticed early during the data extraction step, the category of *not specified* was added to account for this, and it was then assumed that studies which fell under this category had no intention of being exploratory and so could not be excused from the analysis. To put this shortcoming into perspective, approximately 66% of studies fell under not specified from the UC induction group and 100% of studies from the UC maintenance group. Approximately 91% of studies were *not specified* within the CD induction group and 94% in the CD maintenance group. Such high percentages are indicative of this approach for organizing studies to be challenging, and an alternative strategy may prove more appropriate. It is possible, though, that instead of searching

for the explicit use of the words *confirmatory* and *exploratory* throughout the objectives and methods section, their context can be understood and assessed for whether the authors are using their study to examine a research area or if they also intend to collect and use data towards objectives that have implications in the field. Although phase 1 trials are usually pursued for exploratory research[106], this review had a single trial in the CD induction group, which had a phase 2 trial's properties of being a dose-escalating study[106]. Due to this, it was not simply assumed in this literature review that any phase 1 trials were exploratory studies or that other phase trials were confirmatory; therefore, the third *unspecified* category was implemented as a pertinent option.

Excluding the studies exempted from requiring a multiple comparison procedure, the number of studies that had implemented one within each classification was less than 30%. Less than half the trials within each condition classification implemented a statistical adjustment strategy which could aid in improving their likelihood of statistical error and ensure that their analysis of results could yield a robust interpretation and conclusion. However, some trials implemented varied procedures to aid in the control their error rate. Hochberg was the most used multiple comparison procedure within UC induction trials, with approximately 34% of the MCP-adjusted studies using this procedure. In this group, there were also nine trials which used a hierarchical testing approach to maintain the error rate and 15 trials which did not attempt any adjustment strategy out of the total of 33 studies which were recommended to use an MCP. Among the UC maintenance group, two studies implemented a multiple comparison procedure; one trial utilized Bonferroni while the other adjusted using the MCP-Mod program. Within the same group, three trials used a varied procedure, and seven attempted no MCP out of the total of 12 studies which were recommended to use an MCP. Considering all the UC trials, the prominent procedures were Hochberg and Bonferroni. The increased use of the Hochberg procedure within these trials is informative as this procedure has an increased control on the family-wise error rate (FWER)[95]. As mentioned in Chapter 2, the Hochberg method employs an individual assessment of each considered p-value. It is used for less conservative multiple comparison testing and is optimal for researchers who wish to employ a method with greater statistical power than Holm's and Bonferroni's procedures[74]. Furthermore, this multiple comparison technique has been praised for being a straightforward procedure which can be

conveniently used in common spreadsheet programs[77] and not just statistics-based programs, making it a feasible option for researchers of all backgrounds.

The most used multiple comparison procedure within the CD induction trials was Bonferroni, with 50% of the MCP-adjusted studies having used this adjustment procedure. Six trials used a varied procedure in this group and 29 that had not attempted an adjustment strategy out of a total of 43 studies that were recommended to use an MCP. For the CD maintenance trials, there was a single study which used Bonferroni as its MCP of choice out of a total of 14 studies which were recommended to use an adjustment procedure. Additionally, three studies used a varied adjustment procedure, and ten did not attempt any adjustment. The most used adjustment procedure was Bonferroni among all the UC and CD trials in this review. This choice of MCP could be popular due to its straightforward procedure[107], making it an accessible choice for first-time users of an MCP or for those who are not entirely familiar with the range of various alternative MCP options. Bonferroni's procedure is ranked to be more conservative than Holm and Hochberg's procedures which may make it appealing to those who wish to implement a more rigorous standard for their trials. However, the latter is more effective at retaining power in a trial[73] and controlling for the family-wise error rate[95,108]. Multi-arm RCTs could benefit from providing more descriptions of the MCP they choose to implement and background on why this MCP is a better fit for their trial set-up. Providing justifications for the type of MCP could demonstrate that the researchers have ensured that it can retain the proper balance between the type 1 error rate and statistical power for their results based on the trial design. It can also provide an opportunity for the trials that do not choose to implement an MCP to present their case about why their research could fare better without one.

Trials in this review also did not justify their pursuit of a multi-arm design majority of the time. Less than 30% of trials in each classification provided a justification statement. However, induction trials were relatively more likely to provide a justification statement than maintenance trials. It is essential to consider that research teams will only sometimes justify choosing a multi-arm trial when conducting dose-finding studies as it is the appropriate design for their research purposes.

However, as mentioned by Juszczak and colleagues[58], trials must work towards showcasing transparency, which can include disclosing their choice for a multi-arm trial design. Increased transparency regarding specific researcher choices (i.e., choosing to do a multiple-arm design, implementing an MCP strategy or deciding which MCP to use) can provide further insight for others who plan to utilize the multiple-arm trial design for their research.

Trials also scarcely included why they did not choose to implement an MCP when recommended in their methods or statistical analysis sections. It is difficult to determine if a trial is set up for a rigorous analysis or rather statistical error without context for why an MCP was not used for a multi-arm trial. This puts the responsibility on the research audience to analyze the trials' characteristics to determine whether it would benefit from the use of an MCP, and this is only if the audience is well informed about multiplicity in RCTs and the role of MCPs. By including some description concerning the MCP thought process within a multi-arm trial, there can be increased reliability of the analysis and results presented in a trial, and this can also allow for replicability of results in other trials. Furthermore, trials that did not implement an MCP but were recommended to do so are at risk of s misinterpretation of their results. This can lead to alternate conclusions being drawn, resulting in unseen effects that may exist but need to be recorded by the analysis of the trial. Since most trials were phase 2 or 3, not considering an MCP when recommended has implications, resulting in interventions and dose regimes being administered to IBD patients that can interfere with their remission.

Moreover, reporting sample size and power and considering these calculations for multiplicity can help increase transparency in multiple-arm trials[58]. Less than 20% of trials had their statistical power value reflect multiplicity and less than 15% had their sample size value reflect multiplicity across each condition classification. Other statistical measures must reflect multiplicity adjustments, primarily when a trial has already implemented an MCP for its analysis. This allows for consistency in statistical calculations but also is essential for the transparency and replicability of studies[109].

3.5 Concluding Summary:

Multi-arm trials are being increasingly used within the field of IBD and more often to design trials which study multiple doses simultaneously compared to a placebo. Among the trials recommended to use an MCP to improve their control on the type 1 error rate, almost half these trials would not implement an adjustment procedure to aid with this. This reduced initiative to consider an MCP can be attributed to the fact that other trials already within the field do not do the same or that the researchers may have referenced other sources which may have recommended against it. Additional reasons can also be that the researcher needs to be made aware that an adjustment could increase the internal validity of their trial, or they instead believe such an adjustment adds no further benefit to their results. Further research not captured in this review could evaluate the perspective of researchers regarding the use of MCPs for multi-arm trials in IBD to understand better how decisions around MCP usage and the type of MCP used are made.

The future research which can be pursued based on this review is a review of what tools and references researchers commonly reference to determine whether to implement an MCP within their research. Such an evaluation can clarify how researchers base their decision on using MCPs and the type they choose to implement.

Chapter 4: Analysis of MLN02

4.1 Introduction: The MLN02 Multi-arm Trial

To better comprehend the application of MCPs used for statistical analysis in inflammatory bowel disease, data from a multi-arm trial conducted by Feagan and colleagues[110] was analyzed. The trial focuses on treating ulcerative colitis in patients using a humanized antibody named *MLN02* to target the integrin $\alpha_4\beta_7$ within the gut[110]. This chapter will analyze the components of this trial using the MCP methods from Chapter 2 along with the MCP decision tool presented by Odutayo and colleagues[17] in Chapter 3 of this thesis to demonstrate how such resources can help determine whether a multi-arm trial would benefit from implementing an MCP. This section will compare whether any differences between the study arms were statistically significant using a standard multi-arm comparison method and then using the MCPs outlined in Chapter 2. This section will also compare the interpretations made through each line of results.

4.1.1 Integrin $\alpha_4\beta_7$ and MLN02

Integrin cells such as $\alpha_4\beta_7$ have been essential in formulating antibody inflammation treatments because they interact with immune cells to provide a targeted response during inflammatory episodes[111]. Integrin cells fall under the collective classification of adhesion molecules and are intermembrane proteins which regulate communication from within the cell to the content of the extracellular matrix or vice versa[112]. Integrins consisting of the α and β subunits are known as heterodimers as they consist of two different glycoproteins and can exist in 24 variations within the human body[113]. Among the α_4 family, integrins $\alpha_4\beta_1$ and $\alpha_4\beta_7$[114] exist, and these integrins aid in cell communication and transport during immune responses, making them crucial players in inflammation responses[115]. For example, the role $\alpha_4\beta_1$ has in inflammation is to allow for leukocyte extravasation by providing attachment of leukocytes through *tethering* and then allowing for these molecules to prepare for transport across the endothelial border. This border monitors movement between tissues and the bloodstream using the *rolling* motion[116]. Contrarily, $\alpha_4\beta_7$ possesses the function of *homing* particular leukocytes to their assigned organ tissue[114], through the gut-associated lymphoid tissue (GALT), and then also

performs tethering and rolling actions[117] to allow the leukocytes to cross the endothelial border and affirm their inflammation-related immune response[111]. Specifics as to how these movements occur are presented in Figure 11.

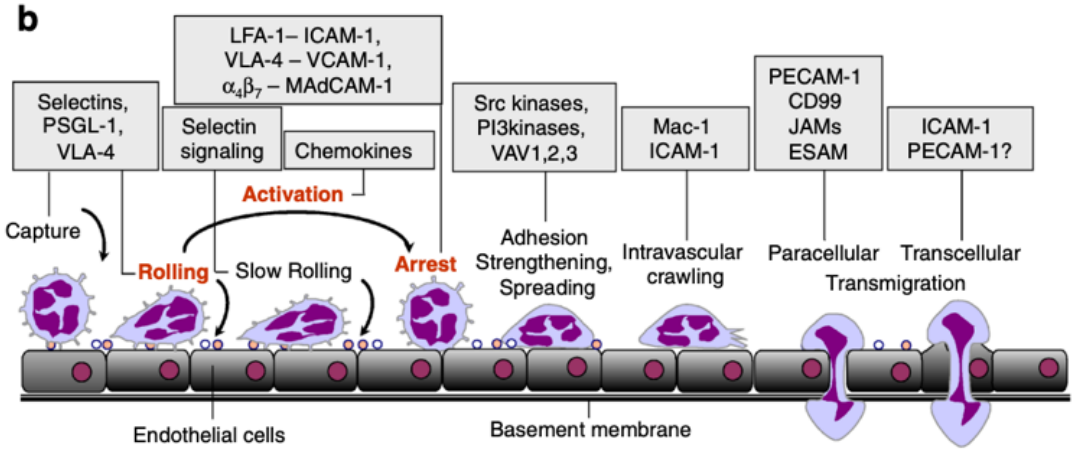

Figure 13: The stages in the movement of leukocytes which is known as TEM[118].

This movement of leukocytes is referred to as TEM or *transendothelial migration*[119]. Significant leukocytes which participate in the inflammation response are the effector T immune cells[118] which have a short lifespan in this role because their purpose is to migrate the immune signal towards the inflicted tissue and then possibly transform into memory T cells to prevent any future instances of attack on the tissue[120]. The subgroups of effector T cells which most commonly interact with inflammation are the T helper 1, T helper 2, or T helper 17 cells. It has been observed that T helper 1 cells are concentrated in cases of Crohn's disease and T helper 2 cells for ulcerative colitis[118]. Ensuring the movement of leukocytes to afflicted or pathologically compromised tissue is an automatic response by the immune system and therefore, this course of action is carefully controlled to prevent unrequired and extended episodes of inflammation[119].

Specific targeting of integrins to prevent the movement of lymphocytes and their engagement in the occurrences of inflammation is referred to as anti-adhesion therapy and is useful in instances of managing inflammatory bowel disease, multiple sclerosis, or even psoriasis[112]. Integrin antagonists or humanized forms of monoclonal antibodies have been of great interest in producing this effect and, especially since the later 2000s, have become an accepted form of therapy for IBD conditions[121]. Since it is known that the dysregulation and excess inflammation which occur throughout inflammatory bowel disease can be attributed to an

imbalanced immune system, the method of repression of leukocyte migration during these inflammation episodes can help alleviate a main symptom of IBD and restore balance to the body's immune system[118]. Various drugs which work to bring about such an effect include vedolizumab (previously known as MLN02), etrolizumab, natalizumab, etrolizumab, and alicaforsen[118].

4.1.2 Description of the Trial Design and Eligibility

This trial investigates the effectiveness of antibody therapy in managing ongoing ulcerative colitis in selective patients using a randomized trial design with two experimental dosage groups and a placebo for comparison[110]. As summarized at the beginning of this chapter, integrin $\alpha_4\beta_7$ found in the intestinal tissue is important in the transportation of leukocytes to areas of inflammation, and therefore the trial focuses on MLN02, a monoclonal antibody which is specific to $\alpha_4\beta_7$ and has been humanized to block biological transportation without activating other reactions to reduce inflammatory responses[110]. Since the main feature of ulcerative colitis is inflammation which is limited to the intestines and rectum[43], MLN02 is presented as a viable option for providing relief for active UC patients with knowledge that a dosage of 2.0 milligrams per kilogram of a patient's weight has been shown to be safe and reliable in attaining an effect of relief from inflammation in prior investigative trials[110].

Participants for this trial were recruited during the years 2000 to 2003 from twenty of the included university medical facilities and had to have active UC to qualify[110]. Patients were considered as having active UC if their inflammation presented itself being 25 centimetres or more from the anal verge, if they had matched a score of between five to nine on the ulcerative colitis clinical score tool in addition to either a score of two or more on the Baron score, and one or more on the stool frequency/rectal bleeding tool[110]. Patients could also only be included if they had no prior therapeutic program for ulcerative colitis and/or if they had been taking mesalamine for a period of four weeks plus before the screening stage of the trial with consistent dosages for this drug in the past 14 days[110]. Patients were excluded from the trial if they had taken a corticosteroid treatment four weeks before the screening stage or, specifically, parenteral corticosteroids within six weeks prior[110]. Patients were also excluded if they had received immunosuppressive therapeutic treatment in the prior three months to the screening stage or applied topical creams which contained mesalamine or corticosteroids in the prior seven days[110].

The goal of this chapter is to only formulate an analysis based on the degree of differences between the different arms of this trial with acknowledgment of the potential use of multiple comparison procedures.

4.1.3 The MLN02 Multi-arm Trial: Is an MCP necessary?

Before considering whether the trial conducted by Feagan and colleagues[110] possesses study characteristics which could benefit from an MCP implementation (as assessed using the tool presented by Odutayo and colleagues[17]), it is imperative to review the statistical analysis methodology conducted by the researchers of the trial to compare the primary outcome in the study arms. The primary outcome in this trial was defined as achieving clinical remission from the symptoms of UC in the sixth week of the trial, and this was determined using the measures of ulcerative colitis clinical score and modified Baron score and an absence of a physical symptom[110]. To compare the degree of differences between the clinical remission point estimates of each of the study groups, the researchers conducted a Cochran-Mantel-Haenszel chi-square test with statistical adjustments for patients who were mesalamine users prior to the trial commencing and labelling study participants who did not complete the course of the trial as required as not having achieved the primary endpoint[110]. The trial did not implement an MCP for its statistical analysis and did not provide any details directly in the text whether there was any consideration or justification for why an MCP was not implemented.

In the following table, each decision point in the tool by Odutayo and colleagues[17] will be used to assess whether this trial would be recommended to use an MCP to help control type-1 error in its multi-arm design.

Table 3: Assessing the MLN02 multi-arm trial for an MCP recommendation

Decision Point	MLN02 Trial
Decision Point 1: Is this trial exploratory or confirmatory in nature?	Although the text does not explicitly make mention of whether the trial is a confirmatory or exploratory trial, since the goal of the trial is assessing the efficacy of the drug, we can assume it is not an exploratory study. So, although not specified this trial will follow further assessment through this tool and will not be excluded.
Decision Point 2: Type of confirmatory multi-arm trial?	This trial is aimed at assessing the efficacy of two dosage levels compared to a control arm. The two dosages were 0.5 mg/kg body weight and 2.0 mg/kg body weight of MLN02. That makes this study a "multiple doses of same intervention vs. control" multi-arm trial.
Due to the response for Decision 2, this will follow through to **Decision Point 3b**.	
Decision Point 3b: Are the comparisons required to be superior?	The trial does not mention that the comparisons rank in any order of superiority so therefore the comparisons are not required to be superior.
Recommendation: Multiple comparison procedure is recommended for this trial.	

4.2 Methods for Analysis

Since the outcome of interest in this trial was clinical remission in the sixth week, this assessment was interested in comparing the three treatment groups and their level of health in the sixth week.

The collected data for this trial was provided in the form of a *SAS* dataset (titled *MLN02*), where the variables of interest were treatment assignment (*tmtcode*) and week 6 ulcerative colitis clinical scores (*w6_uccs*). The three treatment groups were: *tmtcode1* (0.5 mg/kg of MLN02) referred to as treatment 1, *tmtcode2* (2.0 mg/kg of MLN02) which is referred to as treatment 2, and *tmtcode3* (placebo) referred to as treatment 3.

4.2.1 Descriptive Statistics

Descriptive statistics and graphs were generated using SAS to summarize the spread of the week six ulcerative colitis clinical scores data points within each treatment group. This included the number of observations, means, standard deviations, minimum and maximum values, boxplots with quartiles, clustered bar graphs, and histograms. The SAS code used to generate the descriptive statistics and graphs can be found in Appendix 3.

4.2.2 Multiple Comparison Procedures

The four multiple comparison procedures presented in Chapter 2 were used to generate p-values adjusted for multiplicity using SAS. The results for the Bonferroni, Holm, and Hochberg methods were generated using the *MULTTEST* procedure on SAS. The results for the Dunnett method were generated using the *GLM* procedure on SAS. The multiple comparison procedures were used to assess whether there was a statistically significant difference in the health outcome of the three treatment groups.

4.3 Results

4.3.1 Descriptive Statistics

In the initial assessment of the sixth-week health outcome data by the treatment assignment group, the descriptive statistics in Table 4 were generated using *SAS*. In the first column of the table, there are the treatment allocations for treatments 1 through 3 which have been defined in the previous section. The proceeding column provides information regarding the total observations of week 6 ulcerative colitis clinical scores for each treatment group. Among the groups in descending order: treatment 3 had 63 total observations, treatment 2 had 60 and treatment 1 had 58. The proceeding column provided the total non-missing observations for each treatment group which were 61 for treatment 3 and 55 for both treatments 1 and 2. The descriptive statistics values presented for each treatment in the table were the mean, standard deviation, minimum and maximum values. Treatment 1 had an approximate mean value of 2.71 for the UC clinical score, with standard deviations of approximately 2.50. The minimum value which treatment 1 could take on was 0 on the UC clinical score, with the highest value being 8. Treatment 2 had an approximate mean value of 3.53 for the UC clinical score, with standard deviations of approximately 3.21. The minimum value which treatment 2 could take on was 0 on the UC clinical score, with the highest value being 12. Treatment 3 had an approximate mean value of 4.23 for the UC clinical score, with standard deviations of approximately 2.82. The minimum value which treatment 3 could take on was 0 on the UC clinical score, with the highest value also being 12.

Table 4: Descriptive Statistic Values for the Three Treatment Groups

Descriptive Statistics for Week 6 UC Clinical Scores (UCCS) by Treatment Assignment						
Treatment Assignment	**Total Observations**	**Total (non-missing) Observations**	**Mean**	**Standard Deviations**	**Min. Value**	**Max. Value**
Treatment 1 (0.5 mg/kg)	58	55	2.71	2.50	0	8
Treatment 2 (2.0 mg/kg)	60	55	3.53	3.21	0	12
Treatment 3 (placebo)	63	61	4.23	2.82	0	12

Note: Values presented in this table were obtained using SAS

4.3.2 Boxplots and Quartiles

Boxplots and quartile values shown in Figure 14 and Table 5 were generated using *SAS* to provide a visual presentation of the degree of difference between the mean values of treatments and provides summaries of the range of values that existed for the sixth week UC clinical scores for each group. The boxplot for treatment 1 shows that the minimum value for the UC clinical score was 0, with the highest being 8, and the mean is approximated at 2.7. The median value for treatment 1 is shown to be 2 which is less than its mean value, and the first quartile is found to be at 1 and the third quartile at 4. The boxplot for treatment 2 shows that the minimum value for the UC clinical score was 0 with the highest being 12 and the mean is approximated at 3.5 The median value for treatment 2 is shown to be 3 which is less than its mean value and the first quartile is found to be at 1 and third quartile at 6. The boxplot for treatment 3 shows that the minimum value for the UC clinical score was 0 with the highest being 12 and the mean is approximated at 4.2 The median value for treatment 3 is shown to be 4 which is also less than its mean value and the first quartile is found to be at 2 and third quartile at 6.

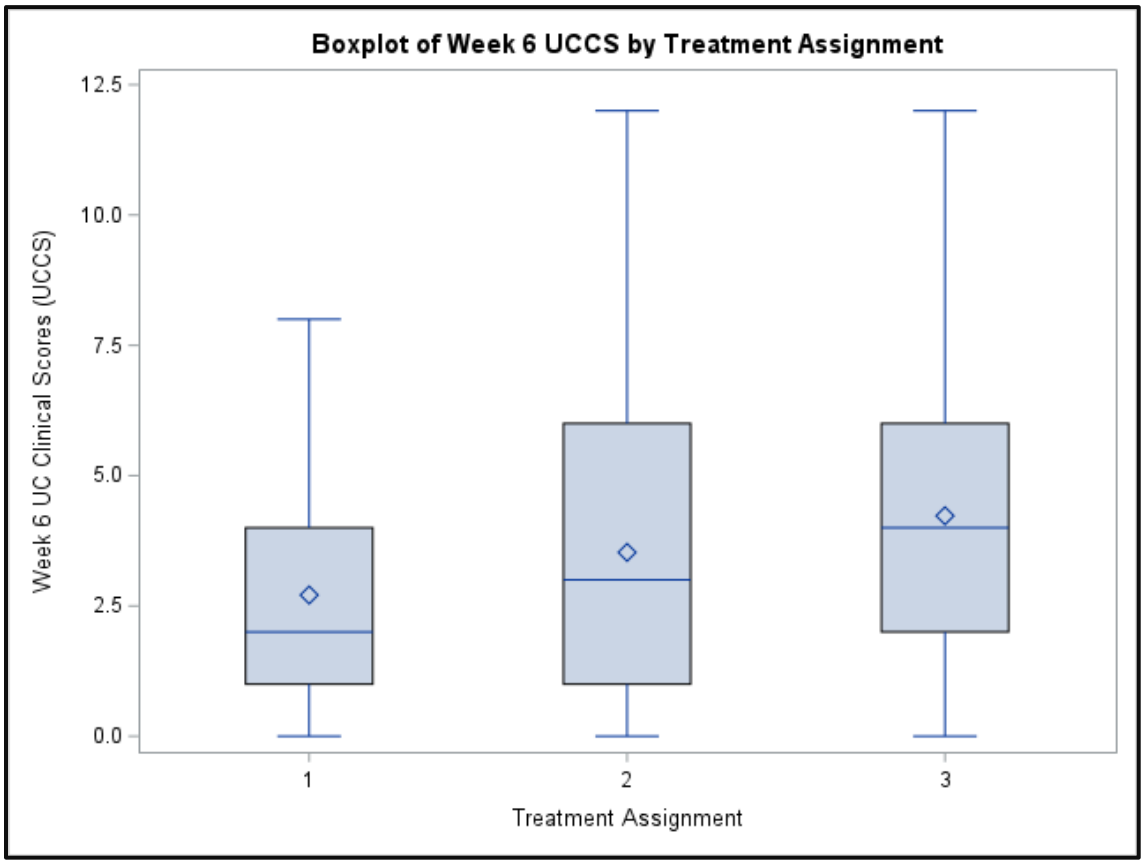

Figure 14: Boxplots display the means and range of the values of week 6 UC clinical scores found in the three treatment groups and their quartile placements, obtained using SAS.

Table 5: Quartile values for the three treatment groups

Quartile	**Treatment 1 (0.5 mg/kg)**	**Treatment 2 (2.0 mg/kg)**	**Treatment 3 (placebo)**
Q_1 (25%)	1	1	2
Median	2	3	4
Q_3 (75%)	4	6	6

Note: Values presented in this table were obtained using SAS

4.3.3 Frequency Bar Graph

The cluster bar graph shown in Figure 15 was used to display the most commonly occurring UC clinical scores for each treatment group and to help understand the distribution and rank of scores within the groups. For treatment 1, the top three most commonly occurring UC clinical scores were 0, 1, and 2, with the bottom three occurring scores being 5, 3, and 6/8. Treatment group 1 is the only group that does not have scores of certain values, including scores 7 and 9 through 10. For treatment 2, the top three most commonly occurring UC clinical scores were 0, 1, and 2, with 9, 10, and 12 being the bottom three occurring scores. For treatment 3, the top five commonly occurring UC clinical scores were 3 and 1/2/6/7, with the bottom three occurring scores being 9/ 10/12. The lower UC clinical scores were most frequent in treatment group 1, while the higher clinical scores were most frequent in treatment group 2.

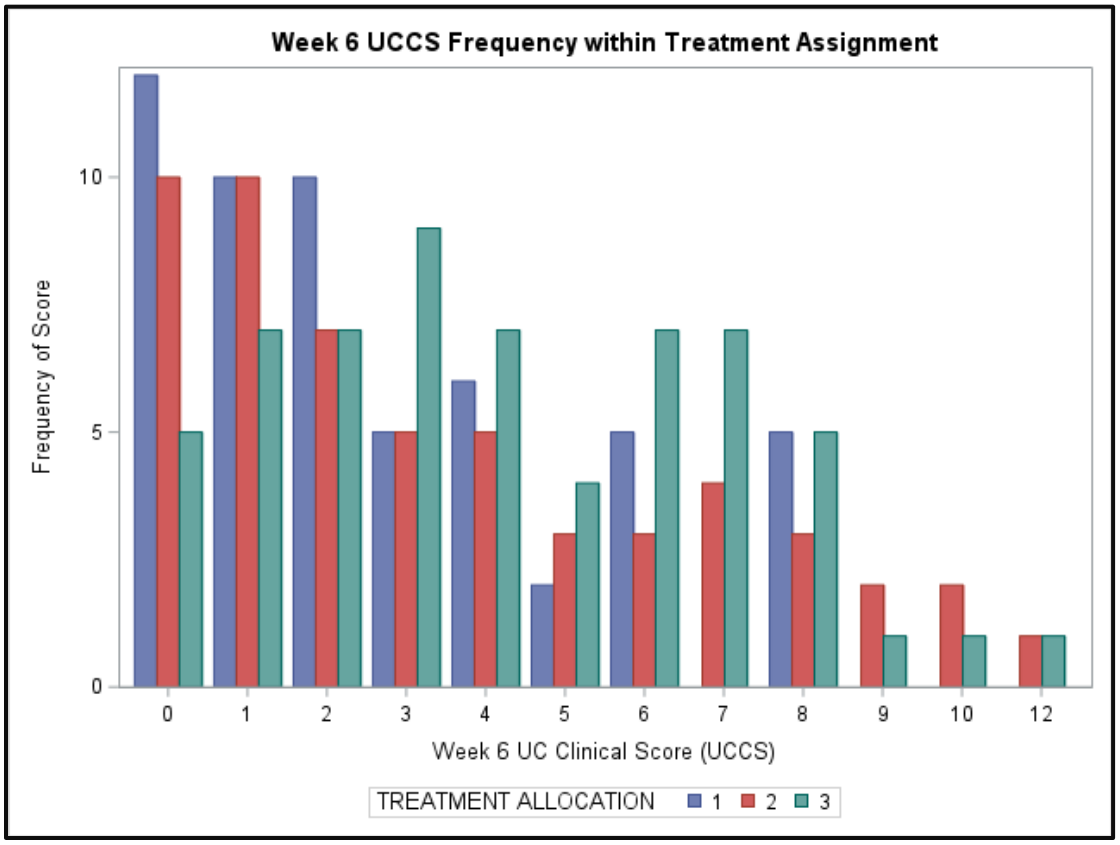

Figure 15: Cluster bar graph which displays the frequency of week 6 UC clinical scores as they occurred within each treatment group, obtained using SAS.

4.3.4 Histograms

The histogram shown in *Figure 17* presents a similar range of data to the previous 2 figures but shows the distribution of each treatment group superimposed onto another for better comparison. When comparing the distribution of the three treatment groups, it can be noted that there is greater variability in treatment 2 compared to treatment 1 and 3. Furthermore, the data from treatment 1 and 2 are skewed to the right, with more frequency of their values being of the lower range from the UC clinical scores. Treatment 3 appears to be less skewed compared to the other two groups, but it still maintains to have more values occurring in the lower and middle range on the UC clinical score scale. Treatment 2 has a distribution curve which is of lower height than treatment 1 and 3. This indicates this group has a higher standard deviation value and greater variability among the UC clinical scores reported at week 6.

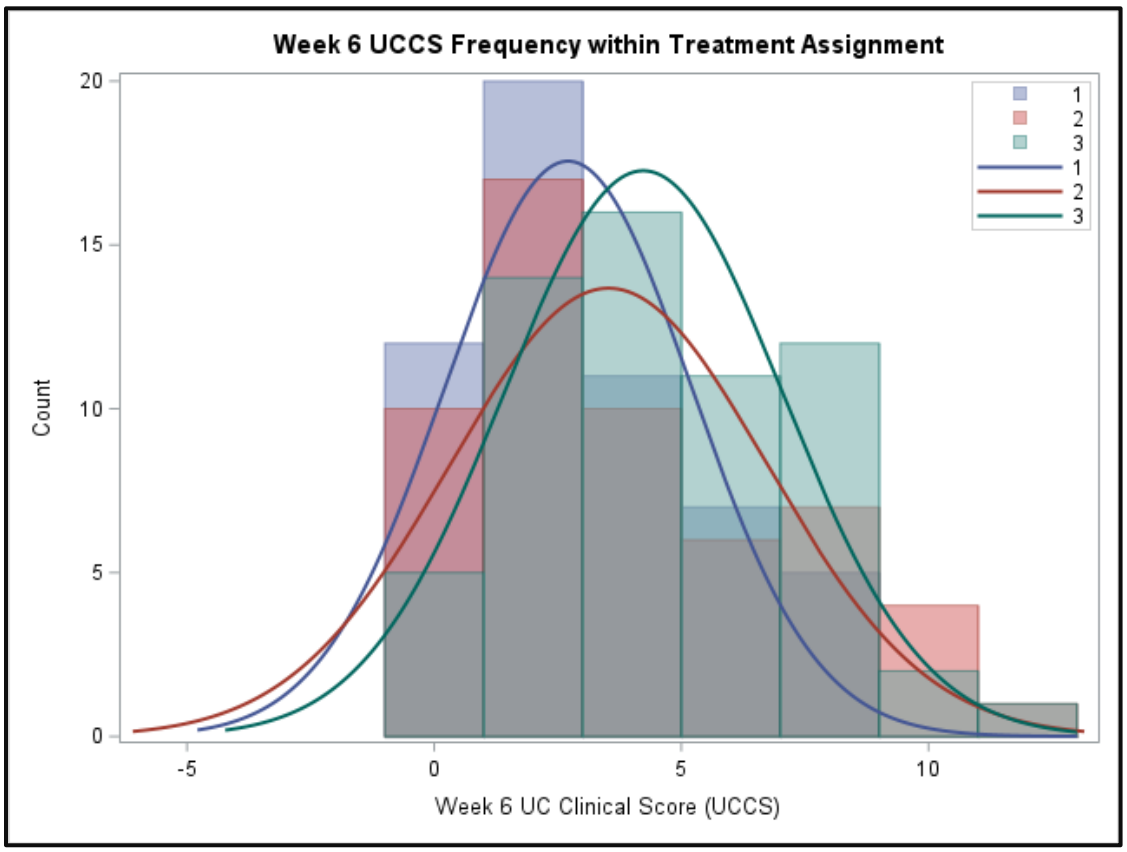

Figure 16: Histograms which depict the distribution of the week 6 UC clinical scores for each treatment group, obtained using SAS.

4.3.5 Descriptive Statistics Summary

To give preliminary insight into the difference which may exist between the three trial arms, descriptive statistics were produced using *SAS* in Chapter 4. In Table 4, the means for the primary outcome of each treatment group were presented, and from all groups, treatment 1 had the least mean value, treatment 2 had the second to least mean value, and treatment 3 (the control group) had the largest. This comparison of means can also be seen in the boxplot presented in Figure 14. The boxplot also shows that the range of values is contained between 0 and 8 for treatment 1; however, treatment 2 and treatment 3 have a larger range of 0 through 12. This can be visually observed in Figure 15, which provides the frequency of each outcome for each of the three groups. Treatment 2 also has the largest interquartile range (IQR) of 5, while treatment 1 has an IQR of 3, and treatment 3 has an IQR of 4. This indicates that the data has a greater spread in treatment 2, and this is also evident as this group has the largest standard deviation among the three with a value of 3.21, as shown in Table 4. This can be visually observed in the distribution curves found in Figure 16. Treatment 1, and treatment 3 have narrower curves than treatment 2, which support the initial thought that the data points in treatment 2 have more spread. Understanding which group(s) is statistically different among the three is difficult with just descriptive statistics, but these items provide insightful information regarding how the data for the primary outcome is situated in each trial arm.

4.3.6 Multiple comparisons procedures

Table 6 presents the adjusted p-values generated using the ORTHOREG procedure and the Shaffer-Simulated Method in SAS. Using this step-down method, these p-values were used to determine whether each treatment group was statistically different from the other arms in the trial. The generated adjusted p-values for each treatment group were p=0.1809 for treatment 1 (0.5mg/kg), p=0.0040 for treatment 2 (2.0 mg/kg), and p=0.0649 for treatment 3 (control). The alpha p-value of 0.05 was used to set the statistical level and is also presented below.

Table 6: The Shaffer-Simulated Method generated p-values for the three treatment groups

Treatment Assignment	Adjusted* (initial) p value	Alpha value
Treatment 1 (0.5 mg/kg)	0.1809	0.05
Treatment 2 (2.0 mg/kg)	0.0040	
Treatment 3 (placebo)	0.0649	

Note:

*The p-values are listed as adjusted as they are generated using the Shaffer-Simulated Method as the initial p-values, obtained using SAS

Table 7 presents the p-values for each treatment group using the multiple comparison adjustment procedures of interest: Bonferroni, Step-down Bonferroni (Holm), and Hochberg. The results of the MCP adjusted p-values are listed as test 1 for p= 0.1809, which is respective to the 0.5 mg/kg experimental group, test 2 for p=0.0040, which is respective to the 2.0 mg/kg group, and test 3 for p=0.0649 which is respective to the control group. For treatment 1, the p-value does not witness a change when using the stepwise procedures of Holm and Hochberg but does when using the Bonferroni method resulting in a p-value of 0.5427. For treatment 2, the p-value adopts a different value when using the three types of multiple comparison procedures; however, the p-value of 0.0042 is consistent across all three tests. For treatment 3, the p-value adopts a different value when using the three types of multiple comparison procedures and is given the p-value of 0.1298 for both the Holm and Hochberg and a p-value of 0.1947 for the Bonferroni method.

Table 7: Bonferroni, Holm, and Hochberg adjusted p-values for the three treatment groups

	Treatment 1 (0.5mg/kg)	**Treatment 2 (2.0 mg/kg)**	**Treatment 3 (placebo)**
Adjusted* (initial) p-value	0.1809	0.0014	0.0649
Bonferroni	0.5427	0.0042	0.1947
Holm	0.1809	0.0042	0.1298
Hochberg	0.1809	0.0042	0.1298

Note: Values are obtained using the MULTEST procedure on SAS.

*The p-values are listed as adjusted as they are generated using the Shaffer-Simulated Method as the initial p-values.

Table 8 presents the p-values for each treatment group using an additional multiple comparison adjustment procedures of interest: the Dunnett method. The Dunnett method was assessed separately using the GLM procedure on SAS, while the other methods were assessed using the MULTTEST procedure on SAS. The results for the Dunnett method only present adjusted p-values for treatment 1 (0.5 mg/kg) and treatment 2 (2.0 mg/kg) as this method requires comparison to treatment 3, the control group. Compared to the initial generated p-value of 0.1809, the adjusted p-value for treatment 1 had changed to 0.0028 using Dunnett's method. For treatment 2 the initial p-value of 0.0014 has changed to 0.1171 using the Dunnett's method.

Table 8: Dunnett adjusted p-value for the two experimental arm groups

	Treatment 1 (0.5 mg/kg)	**Treatment 2 (2.0 mg/kg)**
Treatment 3 (placebo)	0.0028	0.1171

Note: Values are obtained using the GLM procedure on SAS.

4.3.7 Multiple comparisons procedures summary

Table 9 presents the p-values by their respective MCP methods generated using SAS. Underneath each p-value, it is listed whether the p-value was found to be significant based on the chosen alpha level of 0.05.

Table 9: Summary Table of P-Values

	0.5 mg/kg	2.0 mg/kg	Control
Initial p-value Generated using: ***Shaffer-Simulated Method***	0.1809	0.0014*	0.0649
Significance	p>0.05	p<0.05	p>0.05
Bonferroni Generated using: ***Multtest Procedure***	0.5427	0.0042*	0.1947
Significance	p>0.05	p<0.05	p>0.05
Holm Generated using: ***Multtest Procedure***	0.1809	0.0042*	0.1298
Significance	p>0.05	p<0.05	p>0.05
Hochberg Generated using: ***Multtest Procedure***	0.1809	0.0042*	0.1298
Significance	p>0.05	p<0.05	p>0.05
Dunnett Generated using: ***GLM Procedure***		0.1171	Control Group
Significance	p<0.05	p>0.05	

Using the *Multtest Procedure* on SAS, results were generated using the Bonferroni, Holm, and Hochberg correction methods. Using the Bonferroni method, the adjusted p-value was 0.5427 for treatment 1, 0.0042 for treatment 2, and 0.1947 for treatment 3. For the Holm method, the p-values were 0.1809 for treatment 1, 0.0042 for treatment 2, and 0.1298 for treatment 3. The p-values generated by Hochberg's method were identical to those generated by Holm's. Using the *GLM Procedure* on SAS, results were generated using the Dunnett method, and since this method compares each experimental arm with the common control arm, there were no results for treatment 3, the placebo. The generated p-values using this method were 0.0028 for treatment 1 and 0.1171 for treatment 2, with treatment 1 having statistically significant results when compared to the nominal alpha value of 0.05.

Chapter 5: Discussion and Concluding Remarks

5.1 Summary of Chapter

In Chapter 4, this thesis used an example IBD RCT[110], investigating the efficacy of a drug used to manage ulcerative colitis named *MLN02* using multiple arm trial design to provide an example of how to ensure investigators can get benefit from their analysis of their results from consulting a multiple comparison procedure recommendation tool. Although the trial did not implement an MCP originally, although recommended by the decision tool[17], this chapter will explore the results generated from the MCPs analysis conducted in Chapter 4 using SAS.

5.2 Discussion

From the original results of the trial, Feagan and colleagues[110], used the Cochran–Mantel–Haenszel test and determined that for their primary outcome of achieving clinical remission by the sixth week of the trial as determined by an improvement in the ulcerative colitis clinical score and Baron score, all three arms were statistically significant. When compared collectively, they produced a p-value of 0.03, and each experimental arm was statistically significant when compared with the control arm. Pairwise comparison generated a p-value of 0.02[110]. According to the analysis conducted by Faegan and colleagues[110], the primary outcome was successfully achieved using doses of 0.5 mg/kg and 2.0 mg/kg.

For the MCP analysis in Chapter 4, the initial p-values were generated using the *ORTHOREG* procedure, which computed these results for the multi-arm trial using the Shaffer-Simulated Method. The p-values were 0.1809 for treatment 1, 0.0040 for treatment 2 and 0.0649 for treatment 3. Based on these initial p-values, it can already be noted that using a nominal alpha of 0.05, treatment 2 is the only treatment arm with statistically significant results. When using the *MULTTEST* procedure to produce results adjusted using the Bonferroni technique and with an alpha of 0.05, treatment 2 was the only treatment arm with statistically significant results. These results were identical for Holm's and Hochberg's method. Although one uses a step-down approach and the other uses a step-up approach, the factor multiplied by the p-value was the same for Holm and Hochberg's method. The conclusion for both methods was consistent

with Bonferroni's with 2.0 mg/kg being the only statistically significant arm. However, when Dunnett's method generated p-values using the *GLM* procedure, it was treatment 1 or 0.5 mg/kg which had statistically significant results when compared to treatment 2 or 2.0 mg/kg.

Bonferroni's method is at the forefront of MCPs used in multi-arm trials mainly due to its ease of application[122]. However, this method needs more statistical power, which can become problematic in multiple comparisons when there exists more than one null hypothesis that should not be accepted within the list of comparisons, which can impact the type 2 error rate within a trial[123]. Additionally, the Bonferroni test is often criticized for being conservative, and this is because this test generated p-values for each comparison based on the number of comparison groups involved[122]. The Bonferroni p-value in Table 7, when compared to the p-values of the other three procedures, are either equal or larger in value. Since this trial has three arms, each p-value is multiplied by the factor of 3. This results in higher p-values for all groups except treatment 2 (this is due to the rank of this comparison) when compared to the values generated by Holm or Hochberg. This demonstrates the conservative nature of this approach as the p-values generated using this procedure are less likely to be lower in value than the desired alpha and, therefore, less likely to be statistically significant when compared to the p-values generated by the alternative methods. This is why the Bonferroni method becomes problematic when there are many comparisons[124]. If this trial had many study arms, such as 10, the factor by which these p-values would be multiplied would further reduce the chance of statistically significant results. Therefore, although this approach is simple, those using it must consider these additional considerations.

Offered commonly as alternatives to Bonferroni's method are the stepwise methods of Holm and Hochberg's methods. The Holm and Hochberg method generates p-values based on the rank of the comparison arm, among others [122]. In the case of Holm's method, this is conducted in ascending order, so treatment 2 is multiplied by 3, treatment 1 is multiplied by 2, and treatment 3 is multiplied by 1. For Hochberg's method, this is conducted in descending order, so the p-value of treatment 1 remains the same, while treatment 3 is multiplied by a factor of 2 and treatment 2 is multiplied by 3. Based on the adjusted p-values for Holm's and Hochberg's methods, treatment 2 was the only group statistically significant for the primary outcome. Compared to Bonferroni's approach, these two methods are praised for retaining statistical power. Between these both, Hochberg's method has also been noted as being more

powerful than Holm's method. However, it is difficult to observe this distinction in this case because this trial only has three comparison groups. The conclusions are different from the original results of the trial and the results attained using these three MCPs. Although Faegan and his colleagues [110] conclude that both treatment doses are statistically significant, the conclusion using these three procedures is that only the 2.0mg/kg dose (treatment 2) is statistically significant for the primary outcome.

The Dunnett method utilizes a different methodology (as opposed to the other three approaches) for generating a test statistic, as presented in Chapter 2. Therefore, the p-values output was produced using the *GLM* Procedure on SAS. According to Dunnett's method, treatment 1 had a statistically significant difference in the primary outcome compared to the control group, with a p-value of 0.0028. At the same time, treatment 2 was not statistically significant, with a p-value of 0.1171. The outcome using Dunnett's is opposite to that generated using the earlier three approaches, where treatment 2 instead of 1 was statistically significant. This difference in results can be due to the Dunnett method not being able to compare all trial arms against one another collectively; instead, it only allows for comparison across experimental and shared control groups[76]. Due to the approach in comparison, the Dunnett method may highlight statistically significant results between each group and the control, which may not be present when all groups are compared collectively[122]. This is likely why Dunnett's method finds that treatment 1 is statistically significant, as it could detect statistical differences between this group and the placebo group, which could not be detected when using the other three MCPs. Conversely, Dunnett's method couldn't detect enough statistically significant differences between treatment 2 and the placebo group, so its results differed. However, this pairwise format of Dunnett's method does allow researchers to conduct one-tailed or two-tailed analyses[76]. This makes the Dunnett method more applicable to cases where researchers plan to conduct multi-arm non-inferiority trials or determine how proposed multiple interventions compare to the control. In the case of this trial [110],, although the results for the Dunnett method could not provide insight as to whether there were any statistically significant differences in the primary outcome between the three groups overall, it was able to detect a substantial difference between treatment 1 and the control group which was not seen using the other tests. The results from Dunnett's method are still different from those of the original authors, who deemed both comparison groups statistically significant [110] for the primary outcome.

Although the original trial methods do not include the use of the MCP analyses presented in Chapter 4, the interpretation of results can still vary by the type of multiple comparison procedure implemented. Therefore, researchers must evaluate whether to implement an MCP for their analysis and determine if the trial would statistically benefit from implementing one. Furthermore, they must also assess which MCP would best fit their scope of research as each MCP retains advantageous qualities when compared with one another.

5.2 Conclusion

5.2.1 Concluding Remarks

As the prevalence of IBD within Canada is set to rise to about 1% within the next couple of years[125], there is a need for a variety of interventions which can provide effective long-term management of the symptoms associated with IBD for the Canadian population who must live with these conditions. Nonetheless, the research within the field of IBD is constantly evolving to assist patients with managing their symptoms. Multi-arm trials are a powerful research tool which provides researchers with the leverage to be cost-effective and time-efficient[97] at testing multiple competing drugs or interventions to establish those which can lead to optimal benefits. In some cases, IBD patients can also benefit from the cooperative effects of using more than one intervention[59] or can build tolerances to their initial treatment and must adopt new therapies to manage symptoms[14] in which multi-arm trials have the capabilities of testing interventions for these unique circumstances while allowing the comparison with alternative interventions and a control arm at the same time.

Multiple-arm trials are becoming increasingly popular within IBD literature, but their implementation can have statistical consequences on the type 1 error rate. Since the type 1 error rate determines the possibility of retaining a false positive result, it can impact the results and their interpretation extracted from the research. Researchers must contemplate the implementation of MCPs, using decision tools such as those proposed by Odutayo and colleagues[17], to determine whether their investigation can benefit from its integration. For these reasons, this thesis aimed to investigate the utilization of the multi-arm trial design and MCPs within multi-arm trials in IBD.

The main takeaways of this thesis were not to enforce the use of MCPs within trials which use the multi-arm design but rather first to introduce the already commonly utilized multiple comparison procedures within the field of IBD so the readers can be familiarized with them. Second, to present an MCP implementation tool and highlight that such tools can help researchers determine whether their research can statistically benefit from an MCP. Third, to promote transparency around the researcher's acknowledgment of the statistical impact which can occur when using a multi-arm design and their justification on whether they choose to implement an MCP.

From the results obtained from the literature review in Chapter 3, it was observed that multiple-arm trials have become increasingly common within IBD literature. However, there needs to be more acknowledgment of whether these trials require the implementation of an MCP. The need for consideration of these statistical tools is concerning since most of the trials within the literature review were phase 2 or 3 and confirmatory trials. This means the results from the research can lead to the development of management options for patients with IBD in the clinical setting. Through the increase in transparency regarding the consideration of MCPs, those conducting the research or those who wish to replicate or advance prior research can know the statistical approaches used to ensure that the data analyzed has increased reliability. Knowing such information can help clinical and healthcare stakeholders who evaluate research stemming from multi-arm trials assess the feasibility of implementing IBD management options. They ensure confidence in their patients and institutions that these therapies can provide optimal relief for their IBD symptoms.

Based on the findings of this research, this thesis encourages the continued use of multi-arm trials as they are a helpful tool, especially in competitive and resource-sensitive fields. Along with their advantageous qualities, multiple-arm trials provide researchers with an alternative method of studying their topic of interest. Therefore, they are a valuable methodology that others should take the time to familiarize themselves with. In addition to understanding the basis of multiple-arm trials, it is also essential to advance the common knowledge of the multiple comparison procedure decision-making tools, such as that of Odutayo and colleagues[17]. Such resources can help researchers assess the rigour and reliability of the trials they read in literature and those they pursue. Knowing these tools can increase the transparency around the statistical

analysis decisions that researchers may make when conducting multiple comparison trials and allow them to justify whether their research can benefit from an MCP implementation and whether they decide to pursue their use or not. Awareness of what MCPs are, the types of MCPs available, and how to consider their use within the field of IBD can help improve their utility in examining optimal management solutions for patients of Crohn's disease and ulcerative colitis.

5.2.2 Limitations and Future Directions: Where to go from here?

Specific multiple comparison procedures can easily produce confidence intervals for their estimated values. Some procedures, like Bonferroni's, can provide adjusted confidence intervals, while alternative procedures like Tukey's, Scheffé's, and Dunnett's produce simultaneous confidence intervals[126]. Juszczak and colleagues' extension of the Consort 2010 Statement has listed the reporting of confidence intervals for all individual comparisons in multi-arm trials as essential. Although this thesis did not record the frequency of trials within IBD, which reported adjusted confidence intervals when using an MCP, information in this topic area can be crucial in gauging how critical the dependability of an estimate is in IBD multi-arm trials. Further research into this topic area can also inform primarily on how often adjusted confidence intervals are reported within IBD multi-arm RCTs, whether a researcher's choice of MCP is dependent on their ability to generate confidence intervals, the important of confidence interval reporting in muti-arm trials for purposed of trial replicability and reliability.

Additional research efforts can also be focused on the utility of online resources which can help formulate multi-arm trial designs and provide guidance on statistical analysis techniques for multi-am trials in IBD. For example, a team of researchers has cultivated a software tool[127] which aids in multi-arm trial design, MCP implementation, and sample size calculation. Advanced understanding of the current knowledge and utilization of such internet-based tools among researchers who perform multi-arm IBD trials can determine which resources are commonly referenced when planning multi-arm trial and their analysis.

Citations

1. Laass MW, Roggenbuck D, Conrad K. Diagnosis and classification of Crohn's disease. *Autoimmun Rev*. 2014;13(4-5):467-471. doi:10.1016/j.autrev.2014.01.029
2. Podolsky D. Inflammatory Bowel Disease. *New England Journal of Medicine*. 1991;325(13):928-937. doi:doi:10.1056/nejm199109263251306
3. Sartor RB. Mechanisms of disease: Pathogenesis of Crohn's disease and ulcerative colitis. *Nat Clin Pract Gastroenterol Hepatol*. 2006;3(7):390-407. doi:10.1038/ncpgasthep0528
4. Weimers P, Munkholm P. The Natural History of IBD: Lessons Learned. *Curr Treat Options Gastroenterol*. 2018;16(1):101-111. doi:10.1007/s11938-018-0173-3
5. Kaplan GG. The global burden of IBD: From 2015 to 2025. *Nat Rev Gastroenterol Hepatol*. 2015;12(12):720-727. doi:10.1038/nrgastro.2015.150
6. Jeong DY, Kim S, Son MJ, et al. Induction and maintenance treatment of inflammatory bowel disease: A comprehensive review. *Autoimmun Rev*. 2019;18(5):439-454. doi:10.1016/j.autrev.2019.03.002
7. Baumgart DC, Carding SR. Inflammatory bowel disease: cause and immunobiology. *The Lancet*. 2007;369(9573):1627-1640. doi:10.1016/s0140-6736(07)60750-8
8. Nikolaus S, Schreiber S. Diagnostics of Inflammatory Bowel Disease. *Gastroenterology*. 2007;133(5):1670-1689. doi:10.1053/j.gastro.2007.09.001
9. Boirivant M, Cossu A. Inflammatory bowel disease. *Oral Dis*. 2012;18(1):1-15. doi:10.1111/j.1601-0825.2011.01811.x
10. Burisch J, Munkholm P. Inflammatory bowel disease epidemiology. *Curr Opin Gastroenterol*. 2013;29(4):357-362. doi:10.1097/MOG.0b013e32836229fb
11. Hodges P, Kelly P. Inflammatory bowel disease in Africa: What is the current state of knowledge? *Int Health*. 2020;12(3):222-230. doi:10.1093/inthealth/ihaa005
12. Alatab S, Sepanlou SG, Ikuta K, et al. The global, regional, and national burden of inflammatory bowel disease in 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet Gastroenterol Hepatol*. 2020;5(1):17-30. doi:10.1016/S2468-1253(19)30333-4
13. Hazel K, O'Connor A. Emerging treatments for inflammatory bowel disease. *Ther Adv Chronic Dis*. 2020;11. doi:10.1177/2040622319899297
14. Cai Z, Wang S, Li J. Treatment of Inflammatory Bowel Disease: A Comprehensive Review. *Front Med (Lausanne)*. 2021;8. doi:10.3389/fmed.2021.765474
15. Robinson M. Medical Therapy of Inflammatory Bowel Disease for the 21st Century. *European Journal of Surgery*. 1998;(582):90-98. doi:10.1080/11024159850191517
16. De SK, Baron M. Step-up and step-down methods for testing multiple hypotheses in sequential experiments. *J Stat Plan Inference*. 2012;142(7):2059-2070. doi:10.1016/j.jspi.2012.02.005
17. Odutayo A, Gryaznov D, Copsey B, et al. Design, analysis and reporting of multi-arm trials and strategies to address multiple testing. *Int J Epidemiol*. 2020;49(3):968-978. doi:10.1093/IJE/DYAA026
18. Nikolaus S, Schreiber S. Diagnostics of Inflammatory Bowel Disease. *Gastroenterology*. 2007;133(5):1670-1689. doi:10.1053/j.gastro.2007.09.001
19. Zhou N, Chen WX, Chen SH, Xu CF, Li YM. Inflammatory bowel disease unclassified. *J Zhejiang Univ Sci B*. 2011;12(4):280-286. doi:10.1631/jzus.B1000172

20. Mulder DJ, Noble AJ, Justinich CJ, Duffin JM. A tale of two diseases: The history of inflammatory bowel disease. *J Crohns Colitis*. 2014;8(5):341-348. doi:10.1016/j.crohns.2013.09.009
21. Windsor JW, Kaplan GG. Evolving Epidemiology of IBD. *Curr Gastroenterol Rep*. 2019;21(8). doi:10.1007/s11894-019-0705-6
22. Ng SC, Shi HY, Hamidi N, et al. Worldwide incidence and prevalence of inflammatory bowel disease in the 21st century: a systematic review of population-based studies. *The Lancet*. 2017;390(10114):2769-2778. doi:10.1016/S0140-6736(17)32448-0
23. Agrawal M, Burisch J, Colombel JF, Shah CS. Viewpoint: Inflammatory Bowel Diseases among Immigrants from Low- to High-Incidence Countries: Opportunities and Considerations. *J Crohns Colitis*. 2020;14(2):267-273. doi:10.1093/ecco-jcc/jjz139
24. Benchimol EI, Mack DR, Guttmann A, et al. Inflammatory bowel disease in immigrants to Canada and their children: A population-based cohort study. *American Journal of Gastroenterology*. 2015;110(4):553-563. doi:10.1038/ajg.2015.52
25. Lewis JD, Abreu MT. Diet as a Trigger or Therapy for Inflammatory Bowel Diseases. *Gastroenterology*. 2017;152(2):398-414.e6. doi:10.1053/j.gastro.2016.10.019
26. Hu FB. Globalization of diabetes: The role of diet, lifestyle, and genes. In: *Diabetes Care*. Vol 34. ; 2011:1249-1257. doi:10.2337/dc11-0442
27. Fedacko J, Takahashi T, Singh RB, et al. Western diets and risk of non-communicable diseases. In: *Functional Foods and Nutraceuticals in Metabolic and Non-Communicable Diseases*. Elsevier; 2022:3-21. doi:10.1016/b978-0-12-819815-5.00042-2
28. Li T, Qiu Y, Yang HS, et al. Systematic review and meta-analysis: Association of a pre-illness Western dietary pattern with the risk of developing inflammatory bowel disease. *J Dig Dis*. 2020;21(7):362-371. doi:10.1111/1751-2980.12910
29. Cui G, Yuan A. A systematic review of epidemiology and risk factors associated with Chinese inflammatory bowel disease. *Front Med (Lausanne)*. 2018;5(183). doi:10.3389/fmed.2018.00183
30. Yang Y, Owyang C, Wu GD. East Meets West: The Increasing Incidence of Inflammatory Bowel Disease in Asia as a Paradigm for Environmental Effects on the Pathogenesis of Immune-Mediated Disease. *Gastroenterology*. 2016;151(6):e1-e5. doi:10.1053/j.gastro.2016.10.034
31. Kaplan GG, Windsor JW. The four epidemiological stages in the global evolution of inflammatory bowel disease. *Nat Rev Gastroenterol Hepatol*. 2021;18(1):56-66. doi:10.1038/s41575-020-00360-x
32. Kaplan GG, Bernstein CN, Coward S, et al. The Impact of Inflammatory Bowel Disease in Canada 2018: Epidemiology. *J Can Assoc Gastroenterol*. 2019;2(Supplement_1):S6-S16. doi:10.1093/jcag/gwy054
33. Kaplan GG, Ng SC. Understanding and Preventing the Global Increase of Inflammatory Bowel Disease. *Gastroenterology*. 2017;152(2):313-321.e2. doi:10.1053/j.gastro.2016.10.020
34. Heisler C, Mathias H, Morrison JB, Kits O, Veldhuyzen Van Zanten S, Jones J. Understanding access to IBD Speciality Care in Nova Scotia through the Patient Lense. *J Can Assoc Gastroenterol*. 2018;1(2):335. https://academic.oup.com/jcag/article/1/suppl_2/335/4916736
35. Benchimol EI, Manuel DG, Guttmann A, et al. Changing age demographics of inflammatory bowel disease in Ontario, Canada: A population-based cohort study of

epidemiology trends. *Inflamm Bowel Dis*. 2014;20(10):1761-1769. doi:10.1097/MIB.0000000000000103
36. Burns EE, Mathias HM, Heisler C, et al. Access to inflammatory bowel disease speciality care: The primary healthcare physician perspective. *Fam Pract*. 2021;38(4):416-424. doi:10.1093/fampra/cmab006
37. Kuenzig ME, Benchimol EI, Lee L, et al. The Impact of Inflammatory Bowel Disease in Canada 2018: Direct Costs and Health Services Utilization. *J Can Assoc Gastroenterol*. 2019;2(Supplement_1):S17-S33. doi:10.1093/jcag/gwy055
38. Jones JL, Nguyen GC, Benchimol EI, et al. The Impact of Inflammatory Bowel Disease in Canada 2018: Quality of Life. *J Can Assoc Gastroenterol*. 2019;2(Supplement_1):S42-S48. doi:10.1093/jcag/gwy048
39. Becker HM, Grigat DM, Ghosh FRCPC FRCCPE S, et al. living with inflammatory bowel disease: a crohn's and colitis canada survey. *Can J Gastroenterol Hepatol*. 2015;29(2):77.
40. Veauthier B, Hornecker JR. Crohn's Disease: Diagnosis and Management. *Am Fam Physician*. 2018;98(11):661-669.
41. Langan RC, Gotsch PB, Krafczyk MA, Skillinge DD. Ulcerative Colitis: Diagnosis and Treatment. *American Family Physician*. 2007;76(9):1323-1331. www.aafp.org/afp.
42. Baumgart DC, Sandborn WJ. Gastroenterology 2 Inflammatory bowel disease: clinical aspects and established and evolving therapies. *www.thelancet.com*. 2007;369(9573):1641-1657. doi:10.1016/S0140-6736(07)60751-X
43. Conrad K, Roggenbuck D, Laass MW. Diagnosis and classification of ulcerative colitis. *Autoimmun Rev*. 2014;13(4-5):463-466. doi:10.1016/j.autrev.2014.01.028
44. Feuerstein JD, Moss AC, Farraye FA. Ulcerative Colitis. *Mayo Clin Proc*. 2019;94(7):1357-1373. doi:10.1016/j.mayocp.2019.01.018
45. Torres J, Mehandru S, Colombel JF, Peyrin-Biroulet L. Crohn's disease. *The Lancet*. 2017;389(10080):1741-1755. doi:10.1016/S0140-6736(16)31711-1
46. Ungaro R, Mehandru S, Allen PB, Peyrin-Biroulet L, Colombel JF. Ulcerative colitis. *The Lancet*. 2017;389(10080):1756-1770. doi:10.1016/S0140-6736(16)32126-2
47. Fumery M, Seksik P, Auzolle C, et al. Postoperative Complications after Ileocecal Resection in Crohn's Disease: A Prospective Study from the REMIND Group. *American Journal of Gastroenterology*. 2017;112(2):337-345. doi:10.1038/ajg.2016.541
48. Kaplan G, Benchimol E, Bernstein C, et al. *Crohn's and Colitis Canada: 2018 Impact of Inflammatory Bowel Disease in Canada*.; 2018.
49. Bernstein CN, Blanchard JF, Rawsthorne P, Wajda A. Epidemiology of Crohn's Disease and Ulcerative Colitis in a Central Canadian Province: A Population-based Study. *Am J Epidemiol*. 1999;149(10). https://academic.oup.com/aje/article/149/10/916/172954
50. Coward S, Clement F, Benchimol EI, et al. Past and Future Burden of Inflammatory Bowel Diseases Based on Modeling of Population-Based Data. *Gastroenterology*. 2019;156(5):1345-1353.e4. doi:10.1053/j.gastro.2019.01.002
51. Mulder DJ, Noble AJ, Justinich CJ, Duffin JM. A tale of two diseases: The history of inflammatory bowel disease. *J Crohns Colitis*. 2014;8(5):341-348. doi:10.1016/j.crohns.2013.09.009
52. Altman DG, Schulz KF, Moher D, et al. The Revised CONSORT Statement for Reporting Randomized Trials: Explanation and Elaboration. *Annals of Intern Medicine*. 2001;134(8):663-694. www.annals.org

53. Emmerson J, Todd S, Brown JM. Recommendations for designing and analysing multi-arm non-inferiority trials: a review of methodology and current practice. *Curr Control Trials Cardiovasc Med*. 2021;22(1). doi:10.1186/s13063-021-05364-9
54. Millen GC, Yap C. Adaptive trial designs: What are multiarm, multistage trials? *Arch Dis Child Educ Pract Ed*. 2020;105(6):376-378. doi:10.1136/archdischild-2019-317826
55. James ND, Sydes MR, Clarke NW, et al. Systemic therapy for advancing or metastatic prostate cancer (STAMPEDE): A multi-arm, multistage randomized controlled trial. *BJU Int*. 2009;103(4):464-469. doi:10.1111/j.1464-410X.2008.08034.x
56. Baron G, Perrodeau E, Boutron I, Ravaud P. Reporting of analyses from randomized controlled trials with multiple arms: A systematic review. *BMC Med*. 2013;11(1). doi:10.1186/1741-7015-11-84
57. Freidlin B, Korn EL, Gray R, Martin A. Multi-arm clinical trials of new agents: Some design considerations. *Clinical Cancer Research*. 2008;14(14):4368-4371. doi:10.1158/1078-0432.CCR-08-0325
58. Juszczak E, Altman DG, Hopewell S, Schulz K. Reporting of Multi-Arm Parallel-Group Randomized Trials: Extension of the CONSORT 2010 Statement. *JAMA - Journal of the American Medical Association*. 2019;321(16):1610-1620. doi:10.1001/jama.2019.3087
59. Wason J, Magirr D, Law M, Jaki T. Some recommendations for multi-arm multi-stage trials. *Stat Methods Med Res*. 2016;25(2):716-727. doi:10.1177/0962280212465498
60. Wason JMS, Stecher L, Mander AP. Correcting for multiple-testing in multi-arm trials: Is it necessary and is it done? *Trials*. 2014;15(1). doi:10.1186/1745-6215-15-364
61. Banerjee A, Chitnis U, Jadhav S, Bhawalkar J, Chaudhury S. Hypothesis testing, type I and type II errors. *Ind Psychiatry J*. 2009;18(2).
62. Mudholkar GS, Chaubey YP. On defining P-values. *Stat Probab Lett*. 2009;79(18):1963-1971. doi:10.1016/j.spl.2009.06.006
63. Moyé LA. P-Value Interpretation and Alpha Allocation in Clinical Trials. *Ann Epidemiol*. 1998;8:351-357.
64. Cowles M, Davis C. On the Origins of the .05 Level of Statistical Significance. *Am Psychol*. 1982;37(5):553-558. doi:10.1037/0003-066X.37.5.553
65. Baber N. International conference on harmonisation of technical requirements for registration of pharmaceuticals for human use (ICH). *Br J Clin Pharmacol*. 1994;37(5):401-404. doi:10.1111/j.1365-2125.1994.tb05705.x
66. ICH Expert Working Group. *International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use ICH Harmomised Tripartite Guideline Statistical Principles for Clinical Trials E9*.; 1998.
67. Armstrong RA. When to use the Bonferroni correction. *Ophthalmic Physiol Opt*. 2014;34(5):502-508. doi:10.1111/opo.12131
68. Andrade C. The P value and statistical significance: Misunderstandings, explanations, challenges, and alternatives. *Indian J Psychol Med*. 2019;41(3):210-215. doi:10.4103/IJPSYM.IJPSYM_193_19
69. Sedgwick P. Pearson's correlation coefficient. *BMJ* . 2012;345(7864). doi:10.1136/bmj.e4483
70. Green J, Britten N. Education and debate Qualitative research and evidence based medicine. *BMJ*. 1998;316(7139):1230-1232. doi:10.1136/bmj.316.7139.1230
71. Chen SY, Feng Z, Yi X. A general introduction to adjustment for multiple comparisons. *J Thorac Dis*. 2017;9(6):1725-1729. doi:10.21037/jtd.2017.05.34

72. Holm S. Board of the Foundation of the Scandinavian Journal of Statistics A Simple Sequentially Rejective Multiple Test Procedure A Simple Sequentially Rejective Multiple Test Procedure. *Source: Scandinavian Journal of Statistics*. 1979;6(2):65-70.
73. Aickin M, Gensler H. Adjusting for Multiple Testing When Reporting Research Results: The Bonferroni vs Holm Methods. *Am J Public Health*. 1996;86(5):726-728. doi:10.2105/ajph.86.5.726
74. Hochberg Y. A sharper Bonferroni procedure or multiple tests of significance. *Biometrika*. 1988;75(4):800-802. doi:https://doi.org/10.1093/biomet/75.4.800
75. Chen SY, Feng Z, Yi X. A general introduction to adjustment for multiple comparisons. *J Thorac Dis*. 2017;9(6):1725-1729. doi:10.21037/jtd.2017.05.34
76. Lee S, Lee DK. What is the proper way to apply the multiple comparison test? *Korean J Anesthesiol*. 2018;71(5):353-360. doi:10.4097/kja.d.18.00242
77. Thissen D, Steinberg L, Kuang D. Quick and Easy Implementation of the Benjamini-Hochberg Procedure for Controlling the False Positive Rate in Multiple Comparisons. *Journal of Educational and Behavioral Statistics*. 2002;27(1):77-83.
78. Dunnett CW. A Multiple Comparison Procedure for Comparing Several Treatments with a Control. *J Am Stat Assoc*. 1955;50(272):1096-1121.
79. Wason JMS, Robertson DS. Controlling type I error rates in multi-arm clinical trials: A case for the false discovery rate. *Pharm Stat*. 2021;20(1):109-116. doi:10.1002/pst.2059
80. Pizarro JN, Guerrero E, Galindo PL. Multiple comparison procedures applied to model selection. *Neurocomputing*. 2002;48:155-173. www.elsevier.com/locate/neucom
81. Dudoit S, Van Der Laan MJ, Pollard KS. Multiple testing. Part I. Single-step procedures for control of general type I error rates. *Stat Appl Genet Mol Biol*. 2004;3(1). doi:10.2202/1544-6115.1040
82. Galambos J. Bonferroni Inequalities. *The Annals of Probability*. 1977;5(4):577-581. doi:doi:10.1214/aop/1176995765
83. Holland BS, Diponzio Copenhaver M. An Improved Sequentially Rejective Bonferroni Test Procedure. *Biometrics*. 1987;43(2):417-423. doi:10.2307/2531823
84. Wright PS. Adjusted P-Values for Simultaneous Inference. *International Biometric Society*. 1992;48(4):1005-1013. https://about.jstor.org/terms
85. Dunnett CW, Tamhane AC. Step-Up Multiple Testing of Parameters with Unequally Correlated Estimates. *Biometrics*. 1995;51(1):217-227. https://www.jstor.org/stable/2533327
86. Romano JP, Shaikh AM, Wolf M. Multiple Testing. In: *The New Palgrave Dictionary of Economics*. Palgrave Macmillan UK; 2010:1-5. doi:10.1057/978-1-349-95121-5_2914-1
87. Agresti A, Bini M, Bertaccini B, Ryu E. Simultaneous confidence intervals for comparing binomial parameters. *Biometrics*. 2008;64(4):1270-1275. doi:10.1111/j.1541-0420.2008.00990.x
88. Ludbrook J. Multiple comparison procedures updated. *Clin Exp Pharmacol Physiol*. 1998;25(12):1032-1037. doi:10.1111/j.1440-1681.1998.tb02179.x
89. Welsch RE. Stepwise Multiple Comparison Procedures. *Source: Journal of the American Statistical Association*. 1977;72(359):566-575.
90. Dunnett CW, Tamhane AC. A Step-Up Multiple Test Procedure. *Source: Journal of the American Statistical Association*. 1992;87(417):162-170.

91. Giacalone M, Agata Z, Cozzucoli PC, Alibrandi A. Bonferroni-Holm and permutation tests to compare health data: Methodological and applicative issues. *BMC Med Res Methodol*. 2018;18(1). doi:10.1186/s12874-018-0540-8
92. Aickin M, Gensler H. Adjusting for Multiple Testing When Reporting Research Results: The Bonferroni vs Holm Methods. *Am J Public Health*. 1996;86(5):726-728. doi:https://doi.org/10.2105/AJPH.86.5.726
93. Koch GG, Gansky SA. Statistical Considerations for Multiplicity in Confirmatory Protocols. *Ther Innov Regul Sci*. 1996;30(2):523-534. doi:10.1177/009286159603000228
94. Thomas N. Understanding MCP-MOD dose finding as a method based on linear regression. *Stat Med*. 2017;36(27):4401-4413. doi:10.1002/sim.7424
95. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B (Methodological)*. 1995;57(1):289-300.
96. Ferreira JA. The Benjamini-Hochberg method in the case of discrete test statistics. *International Journal of Biostatistics*. 2007;3(1). doi:10.2202/1557-4679.1065
97. Jaki T, Wason JMS. Multi-arm multi-stage trials can improve the efficiency of finding effective treatments for stroke: A case study. *BMC Cardiovasc Disord*. 2018;18(1). doi:10.1186/s12872-018-0956-4
98. Akobeng AK. Assessing the Validity of Clinical Trials. *J Pediatr Gastroenterol Nutr*. 2008;47(3):277-282. doi:0.1097/MPG.0b013e31816c749f
99. Jairath V, Zou G, Parker CE, et al. Systematic review with meta-analysis: placebo rates in induction and maintenance trials of Crohn's disease. *Aliment Pharmacol Ther*. 2017;45(8):1021-1042. doi:10.1111/apt.13973
100. Jairath V, Zou G, Parker CE, et al. Systematic review and meta-analysis: Placebo rates in induction and maintenance trials of ulcerative colitis. *J Crohns Colitis*. 2016;10(5):607-618. doi:10.1093/ecco-jcc/jjw004
101. Chuang-Stein C. Assay Sensitivity. *Encyclopedia of Statistical Sciences*. Published online 2004:1-5.
102. Fakhoury M, Negrulj R, Mooranian A, Al-Salami H. Inflammatory bowel disease: Clinical aspects and treatments. *J Inflamm Res*. 2014;7(1):113-120. doi:10.2147/JIR.S65979
103. Van Norman GA. Phase II Trials in Drug Development and Adaptive Trial Design. *JACC Basic Transl Sci*. 2019;4(3):428-437. doi:10.1016/j.jacbts.2019.02.005
104. Daher S, Khoury T, Benson A, et al. Inflammatory bowel disease patient profiles are related to specific information needs: A nationwide survey. *World J Gastroenterol*. 2019;25(30):4246-4260. doi:10.3748/wjg.v25.i30.4246
105. Stanley K. Design of randomized controlled trials. *Circulation*. 2007;115(9):1164-1169. doi:10.1161/CIRCULATIONAHA.105.594945
106. Francillon A, Pickering G, Belorgey C, et al. Exploratory Clinical Trials: Implementation Modes & Guidelines, Scope and Regulatory Framework. *Therapies*. 2009;64(3):155-159. doi:10.2515/therapie/2009022
107. Bender R, Lange S. Adjusting for multiple testing-when and how? *J Clin Epidemiol*. 2001;54:343-349.
108. Van der Weele TJ, Mathur MB. Some Desirable Properties of the Bonferroni Correction: Is the Bonferroni Correction Really So Bad? *Am J Epidemiol*. 2019;188(3):617-618. doi:10.1093/aje/kwy250

109. Prager EM, Chambers KE, Plotkin JL, et al. Improving transparency and scientific rigor in academic publishing. *J Neurosci Res*. 2019;97(4):377-390. doi:10.1002/jnr.24340
110. Feagan BG, Greenberg GR, Wild G, et al. Treatment of Ulcerative Colitis with a Humanized Antibody to the a4b7 Integrin. *N Engl J Med*. 2005;352(24):2499-2507. doi:10.1056/NEJMoa042982
111. Elices MJ. The Integrin α4β1 (VLA-4) as a therapeutic Target. In: *Ciba Foundation Symposium*. ; 2007:79-90. doi:10.1002/9780470514719.ch7
112. Ren G, Roberts AI, Shi Y. Adhesion molecules. *Cell Adh Migr*. 2016;5(1):20-22. doi:10.4161/cam.5.1.13491
113. Takada Y, Ye X, Simon S. The integrins. *Genome Biol*. 2007;8(5):215.1-215.9. doi:10.1186/gb-2007-8-5-215
114. Yu Y, Zhu J, Mi LZ, et al. Structural specializations of α 4β 7, an integrin that mediates rolling adhesion. *Journal of Cell Biology*. 2012;196(1):131-146. doi:10.1083/jcb.201110023
115. Hesterberg PE, Winsor -Hines D, Briskin MJ, et al. Rapid Resolution of Chronic Colitis in the Cotton-top Tamarin With an Antibody to a Gut-Homing Integrin a4b7. *Gastroenterology*. 1996;111:1373-1380.
116. Baiula M, Spampinato S, Gentilucci L, Tolomelli A. Novel Ligands Targeting α4β1 Integrin: Therapeutic Applications and Perspectives. *Front Chem*. 2019;7:489. doi:10.3389/fchem.2019.00489
117. Hoshino H, Kobayashi M, Mitoma J, Sato Y, Fukuda M, Nakayama J. An integrin α4β7•IgG heterodimeric chimera binds to MAdCAM-1 on high endothelial venules in gut-associated lymphoid tissue. *Journal of Histochemistry and Cytochemistry*. 2011;59(6):572-583. doi:10.1369/0022155411404416
118. Mosli MH, Rivera-Nieves J, Feagan BG. T-cell trafficking and anti-adhesion strategies in inflammatory bowel disease: Current and future prospects. *Drugs*. 2014;74(3):297-311. doi:10.1007/s40265-013-0176-2
119. Mueller SN, Gebhardt T, Carbone FR, Heath WR. Memory T cell subsets, migration patterns, and tissue residence. *Annu Rev Immunol*. 2013;31:137-161. doi:10.1146/annurev-immunol-032712-095954
120. Kumar B V., Connors TJ, Farber DL. Human T Cell Development, Localization, and Function throughout Life. *Immunity*. 2018;48(2):202-213. doi:10.1016/j.immuni.2018.01.007
121. Lau MS, Tsai HH. Review of vedolizumab for the treatment of ulcerative colitis. *World J Gastrointest Pharmacol Ther*. 2016;7(1):107. doi:10.4292/wjgpt.v7.i1.107
122. Barnett MJ, Doroudgar S, Khosraviani V, Ip EJ. Multiple comparisons: To compare or not to compare, that is the question. *Research in Social and Administrative Pharmacy*. 2022;18(2):2331-2334. doi:10.1016/j.sapharm.2021.07.006
123. Perneger T V. What's wrong with Bonferroni adjustments. *BMJ*. 1998;316(7139):1236-1238. doi:10.1136/bmj.316.7139.1236
124. Bland MJ, Altman DG. Multiple Significance Tests: the Bonferroni method. *BMJ*. 1995;310(6973):170. doi:10.1136/bmj.310.6973
125. Coward S, Clement F, Benchimol EI, et al. Past and Future Burden of Inflammatory Bowel Diseases Based on Modeling of Population-Based Data. *Gastroenterology*. 2019;156(5):1345-1353.e4. doi:10.1053/j.gastro.2019.01.002

126. Bender R, Lange S. Adjusting for multiple testing-when and how? *J Clin Epidemiol*. 2001;54:343-349.
127. Grayling MJ, Wason JMS. A web application for the design of multi-arm clinical trials. *BMC Cancer*. 2020;20(1). doi:10.1186/s12885-020-6525-0

Appendices

Appendix 1

Table A: Results for the multiple-arm trials only in the Ulcerative Colitis (Induction) group

Trial #	# arms in the trial	**D1:** Confirmatory/ Exploratory	**D2:** Type of Multi-arm Trial	**D3A**: Are interventions unrelated?	**D3B:** Comparisons Superior? MCP Adjustment made?	Does power reflect multiple arms?	Does the sample size calculation reflect multiplicity?	Sample Size	Justification for multi-arm trial?	Adaptive or Nonadaptive trial?	Phase of Trial
1	5	Explora-tory	Multiple Doses v. Placebo	-	-	-	-	-	-	-	-
5	3	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But **varied** procedure used	No	No	170	No	Non-adapt	2
6	3	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. MCP Procedure used: **Hochberg**	No	No	746	No	Adapt	3
8	3	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But no MCP procedure used	No	No	181	No	Non-adapt	2
9	3	Explora-tory	Multiple Doses v. Placebo	-	-	-	-	-	-	-	-
10	4	Explora-tory	Multiple Doses v. Placebo	-	-	-	-	-	-	-	-
11	4	Explora-tory	Multiple Doses v. Placebo	-	-	-	-	-	-	-	-
12	3	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But no MCP procedure used	No	No	123	No	Non-adapt	2

13	3	Not Specified	Multiple Doses + Interventions v. Placebo	No – Adjusted using **Holm**	Superiority for all required. **Excused** from MCP.	-	-	343	No	Non-adapt	3
14	4	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. MCP adjustment **Hochberg**	No	No	168	No	Non-adapt	2
18	3	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. MCP adjustment **Holm**	No	No	280	No	Non-adapt	3
22	3	Explora-tory	Multiple Doses v. Placebo	-	-	-	-	-	-	-	-
26	3	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. MCP adjustment **Dunnett**	No	No	65	No	Non-adapt	2
28	3	Explora-tory	Multiple Doses v. Placebo	-	-	-	-	-	-	-	-
30	3	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But no MCP procedure used	No	No	390	Yes	Adapt	3
31	4	Not Specified	Multiple Doses v. Placebo	-	Superiority for all required. **Excused** from MCP.	-	-	84	No	Non-adapt	2
33	6	Explorat-ory	Multiple Doses v. Placebo	-	-	-	-	-	-	-	-
34	4	Explora-tory	Multiple Doses v. Placebo	-	-	-	-	-	-	-	-
35	3	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But no MCP procedure used	No	No	364	No	Non-adapt	2

36	3	Confirma-tory	Multiple Doses v. Placebo	-	Superiority not required for all. But **varied** procedure used	Yes	Yes	774	Yes	Non-adapt	2/3
38	4	Explora-tory	Multiple Doses v. Placebo	-	-	-	-	-	-	-	-
39	3	Confirma-tory	Multiple Doses v. Placebo	-	Superiority not required for all. MCP adjustment **Bonferroni**	Yes	Yes	252	No	Non-adapt	2
40	3	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But **varied** procedure used	Yes	No	197	No	Non-adapt	2
42	3	Explora-tory	Multiple Interventions v. Placebo	-	-	-	-	-	-	-	-
43	4	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But **varied** procedure used	No	No	490	No	Adapt	3
44	5	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But **varied** procedure used	No	No	194	No	Non-adapt	2
45	4	Not Specified	Multiple Interventions v. No Placebo	-	-	-	-	-	-	-	-
46	6	Explora-tory	Multiple Doses v. Placebo	-	-	-	-	-	-	-	-
48	3	Confirma-tory	Multiple Doses v. Placebo	-	Superiority not required for all. But no MCP procedure used	No	No	165	No	Non-adapt	2/3

49	5	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. MCP adjustment **Bonferroni**	Yes	No	354	Yes	Non-adapt	3
50	4	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But **varied** procedure used	No	No	106	No	Adapt.	2
51	3	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But **varied** procedure used	No	Yes	156	No	Non-adapt	2
52	5	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. MCP adjustment **MCP-Mod**	Yes	No	250	No	Non-adapt	2
53	3	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But no MCP procedure used	No	No	149	Yes	Non-adapt	2
54	5	Explora-tory	Multiple Doses v. Placebo	-	-	-	-	-	-	-	-
55	3	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But **varied** procedure used	Yes	Yes	961	No	Non-adapt	3
57	3	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But no MCP procedure used	No	No	186	No	Non-adapt	2
58	3	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But no MCP procedure used	No	No	21	No	Non-adapt	2
59	3	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But no MCP procedure used	No	No	158	No	Non-adapt	2

70	3	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But no MCP procedure used	No	No	136	Yes	Non-adapt	3
72	3	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But no MCP procedure used	No	No	39	No	Adapt	3
73	4	Not Specified	Multiple Doses and Interventions	No – But **varied** procedure used	Superiority for all required. **Excused** from MCP.	No	No	410	No	Non-adapt	3
75	3	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But no MCP procedure used	No	No	159	No	Non-adapt	2
76	5	Explora-tory	Multiple Doses v. Placebo	-	-	-	-	-	-	-	-
77	5	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But no MCP procedure used	No	Yes	112	No	Non-adapt	2
78	3	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But no MCP procedure used	No	No	80	Yes	Non-adapt	3
79	3	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But no MCP procedure used	No	No	124	No	Non-adapt	2
81	5	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. MCP adjustment **Hochberg**	No	No	587	No	Non-adapt	2

Table B: Results for the multiple-arm trials only in the Ulcerative Colitis (Maintenance) group

Trial #	# arms in the trial	**D1:** Confirmatory/ Exploratory	**D2:** Type of Multi-arm Trial	**D3A**: Are interventions unrelated?	**D3B:** Comparisons Superior? MCP Adjustment made?	Does power reflect multiple arms?	Does the sample size calculation reflect multiplicity?	Sample Size	Justification for multi-arm trial?	Adaptive or Nonadaptive trial?	Phase of Trial
1	3	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But **varied** procedure used	No	No	170	No	Adapt	2
8	4	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But no MCP procedure used	No	No	45	Yes	Non-adapt	2
9	3	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But no MCP procedure used	No	No	123	No	Non-adapt	3
13	3	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But no MCP procedure used	No	No	534 364	No	Non-adapt	3
14	3	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But no MCP procedure used	No	No	464	No	Adapt	3
15	3	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But **varied** procedure used	No	No	197	No	Non-adapt	2
16	3	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. MCP adjustment **Bonferroni**	No	No	541	No	Non-adapt	3
18	3	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But **varied** procedure used	No	No	216	No	Non-adapt	3
19	5	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. MCP adjustment **MCP-Mod**	No	No	219	No	Non-adapt	2

20	3	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But no MCP procedure used	Yes	Yes	961	No	Adapt	3
21	3	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But no MCP procedure used	No	No	274	No	Non-adapt	3
23	3	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But no MCP procedure used	No	No	39	No	Adapt	3

Table C: Results for the multiple-arm trials only in the Crohn's Disease (Induction) group

Trial #	# arms in trial	**D1:** Confirmatory/ Exploratory	**D2:** Type of Multi-arm Trial	**D3A**: Are interventions unrelated?	**D3B:** Comparisons Superior? MCP Adjustment made?	Does power reflect multiple arms?	Does the sample size calculation reflect multiplicity?	Sample Size	Justification for multi-arm trial?	Adaptive or Nonadaptive trial?	Phase of Trial
2	4	Confirm.	Multiple Doses v. Placebo	-	Superiority not required for all. MCP adjustment **Hochberg**	No	No	265	No	Non-adapt	2
7	4	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But no MCP procedure used	No	No	30	No	Non-adapt	2
10	4	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But no MCP procedure used	No	No	249	No	Adapt	2
16	3	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But no MCP procedure used	No	No	237	No	Non-adapt	2
17	3	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. MCP adjustment **Hochberg**	No	No	608	No	Non-adapt	3
19	3	Not Specified	Multiple Doses v. Placebo	-	Superiority for all required. **Excused** from MCP.	No	No	121	Yes	Non-adapt	2
20	5	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But no MCP procedure used	No	No	95	No	Non-adapt	3
21	4	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. MCP adjustment **Bonferroni**	No	No	248	No	Non-adapt	2
23	4	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But no MCP procedure used	No	No	258	No	Non-adapt	3

24	4	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But no MCP procedure used	Yes	No	299	Yes	Non-adapt	2
25	3	Confirm.	Multiple Doses v. Placebo	-	Superiority not required for all. But no MCP procedure used	No	No	133	No	Non-adapt	2
27	3	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But no MCP procedure used	No	No	36	No	Non-adapt	2
29	4	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. MCP adjustment **Bonferroni**	No	No	436	Yes	Adapt	3
32	4	Not Specified	Multiple Interventions v. Placebo	No— But no MCP procedure used	-	No	No	455	No	Non-adapt	2
33	3	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But no MCP procedure used	No	No	79	No	Adapt	2
34	3	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But no MCP procedure used	No	No	84	No	Non-adapt	2
35	3	Not Specified	Multiple Interventions v. Placebo	No – But no MCP procedure used	-	No	No	84	No	Non-adapt	2
36	4	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But no MCP procedure used	No	No	246	No	Adapt	2
37	3	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But no MCP procedure used	No	No	280	No	Adapt	2
38	4	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But no MCP procedure used	No	No	45	Yes	Adapt	3
39	3	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But no MCP procedure used	No	No	201	Yes	Non-adapt	2

40	5	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But no MCP procedure used	No	No	207	Yes	Non-adapt	2
43	3	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But no MCP procedure used	No	No	169	Yes	Adapt	2
50	4	Not Specified	Multiple Doses and Interventions	No – But no MCP procedure used	Superiority not required for all. But no MCP procedure used	No	No	164	No	Adapt	2
53	4	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But **varied** procedure used	No	No	451	No	Adapt	3
54	5	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But **varied** procedure used	No	No	526	No	Adapt	3
56	4	Explora-tory	Multiple Doses v. Placebo	-	-	-	-	-	-	-	-
58	3	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But **varied** procedure used	No	No	121	Yes	Non-adapt	2
61	4	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. MCP adjustment **Bonferroni**	No	No	262	No	Non-adapt	2
63	6	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But no MCP procedure used	No	Yes	220	Yes	Adapt	3
64	4	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. MCP adjustment **Dunnett**	No	No	76	Yes	Non-adapt	2
65	3	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But **varied** procedure used	Yes	No	220	No	Adapt	2

69	4	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But no MCP procedure used	No	No	527	Yes	Non-adapt	3
70	5	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. MCP adjustment **Bonferroni**	Yes	No	329	Yes	Non-adapt	3
71	4	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But **varied** procedure used	Yes	No	291	No	Non-adapt	2
72	5	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But **varied** procedure used	No	No	284	No	Non-adapt	3
73	4	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But no MCP procedure used	No	No	187	No	Non-adapt	2
74	4	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But no MCP procedure used	Yes	No	310	No	Non-adapt	2
77	4	Not Specified	Multiple Interventions v. Placebo	Yes – no pairwise – **Excused** from MCP	-	No	No	295	No	Adapt	2
79	3	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But no MCP procedure used	No	No	77	No	Non-adapt	2
81	4	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But no MCP procedure used	No	No	108	No	Adapt	2
83	4	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But no MCP procedure used	No	No	130	No	Non-adapt	2
85	3	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But no MCP procedure used	No	No	200	No	Non-adapt	3

87	7	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. MCP adjustment **Dunnett**	No	No	40	Yes	Non-adapt	1
88	5	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But no MCP procedure used	No	No	45	Yes	Non-adapt	3
90	3	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But no MCP procedure used	No	No	90	No	Adapt	3
92	5	Explora-tory	Multiple Doses v. Placebo	-	-	-	-	-	-	-	-

Table D: Results for the multiple-arm trials only in Crohn's Disease (Maintenance) group

Trial #	# arms in trial	**D1:** Confirmatory/ Exploratory	**D2:** Type of Multi-arm Trial	**D3A**: Are interventions unrelated?	**D3B:** Comparisons Superior? MCP Adjustment made?	Does power reflect multiple arms?	Does the sample size calculation reflect multiplicity?	Sample Size	Justification for multi-arm trial?	Adaptive or Nonadaptive trial?	Phase of Trial
2	3	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But no MCP procedure used	No	No	499	No	Non-adapt	3
11	3	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But no MCP procedure used	No	No	75	No	Non-adapt	3
15	3	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But no MCP procedure used	No	No	573	No	Adapt	3
17	4	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. MCP adjustment **Bonferroni**	Yes	No	435	No	Adapt	3
20	3	Not Specified	Multiple Interventions v. Placebo	No – But no MCP procedure used	-	No	No	84	No	Non-adapt	2
22	4	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But no MCP procedure used	No	No	246	No	Adapt	2
24	3	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But no MCP procedure used	No	No	280	Yes	Adapt	2
25	3	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But no MCP procedure used	No	No	194	No	Non-adapt	2

26	3	Not Specified	Multiple Interventions v. Placebo	-	Superiority not required for all. But no MCP procedure used	No	No	143	No	Non-adapt	3
30	4	Explora-tory	Multiple Doses v. Placebo	-	-	-	-	-	-	-	-
32	4	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But **varied** procedure used	No	No	451	No	Non-adapt	3
33	4	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But **varied** procedure used	No	No	526	No	Adapt	2
35	3	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But **varied** procedure used	No	No	220	No	Non-adapt	2
36	4	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But no MCP procedure used	No	No	527	No	Non-adapt	3
39	4	Not Specified	Multiple Interventions v. Placebo	Yes – no pairwise – **Excused** from MCP	-	No	No	274	No	Adapt	2
43	3	Not Specified	Multiple Doses v. Placebo	-	Superiority not required for all. But no MCP procedure used	No	No	82	No	Adapt	3

Appendix 2

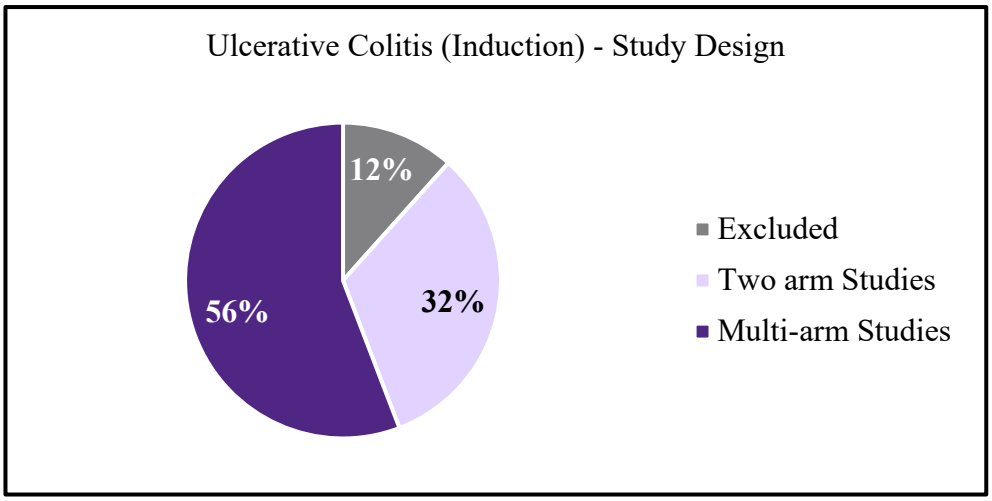

Figure A: Pie chart representing the breakdown of the study designs among the induction trials included in the ulcerative colitis disease classification.

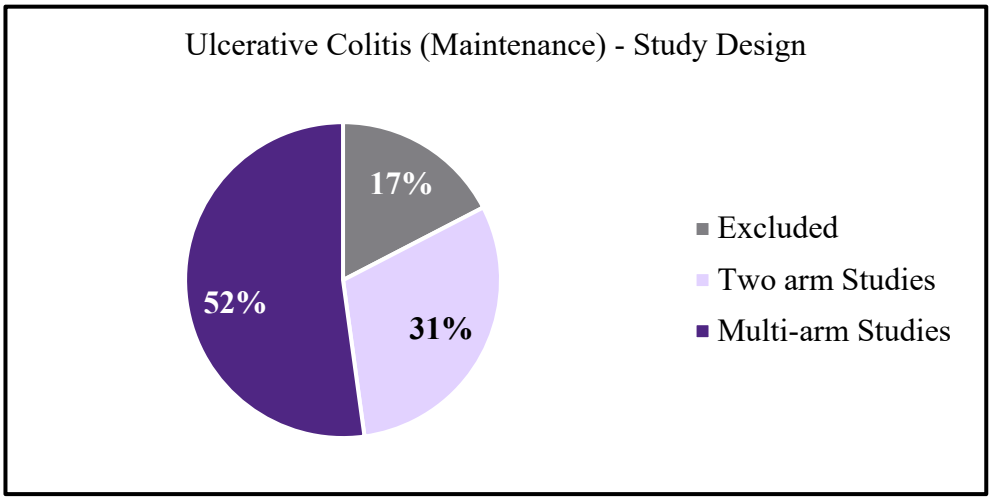

Figure B: Pie chart representing the breakdown of the study designs among the maintenance trials included in the ulcerative colitis disease classification.

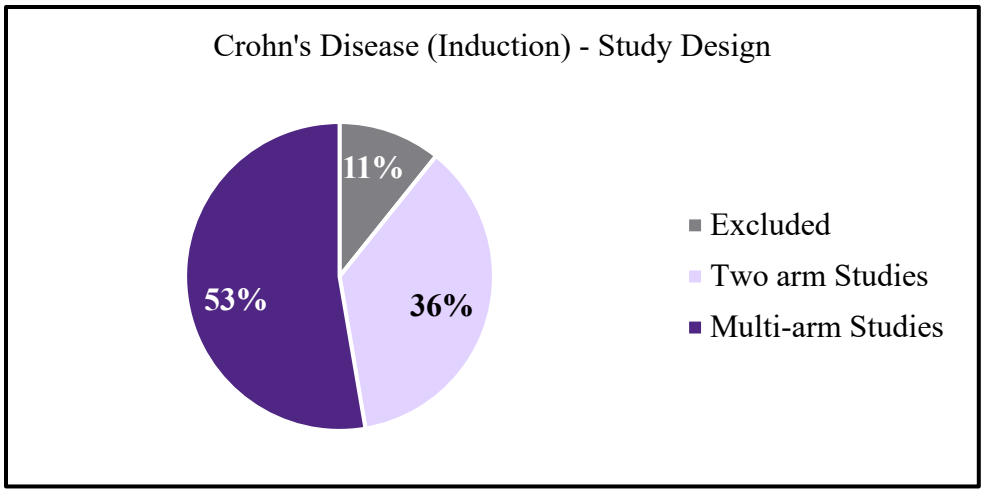

Figure C: Pie chart representing the breakdown of the study designs among the induction trials included in the Crohn's disease classification.

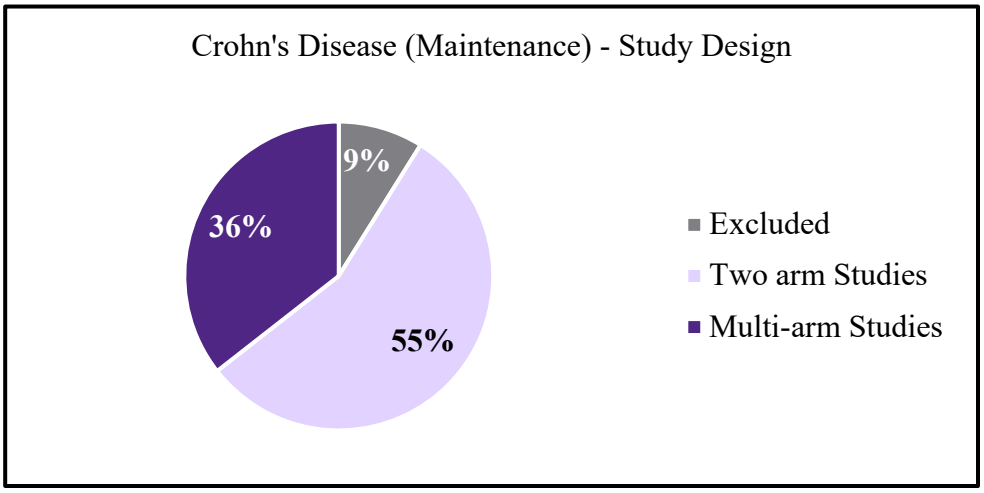

Figure D: Pie chart representing the breakdown of the study designs among the maintenance trials included in the Crohn's disease classification.

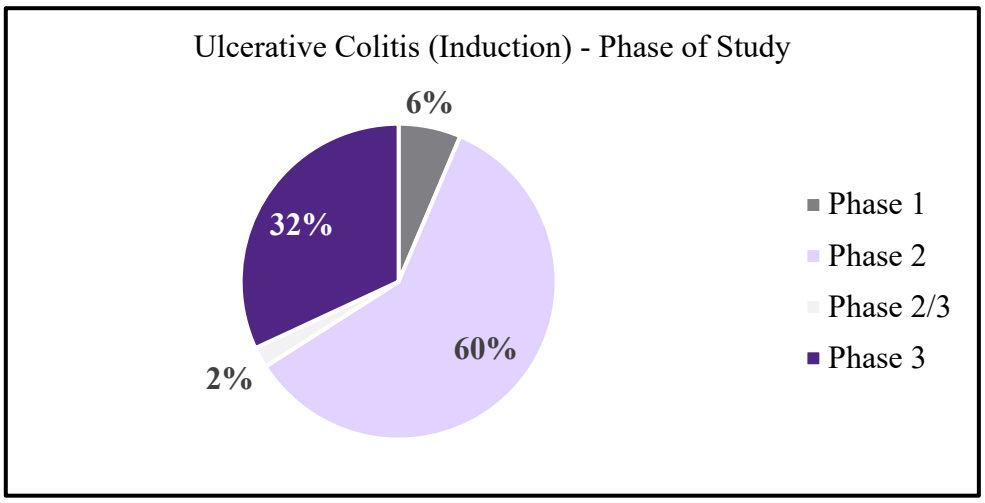

Figure E: Pie chart representing the breakdown of which study phase the induction trials included in the ulcerative colitis disease classification were in.

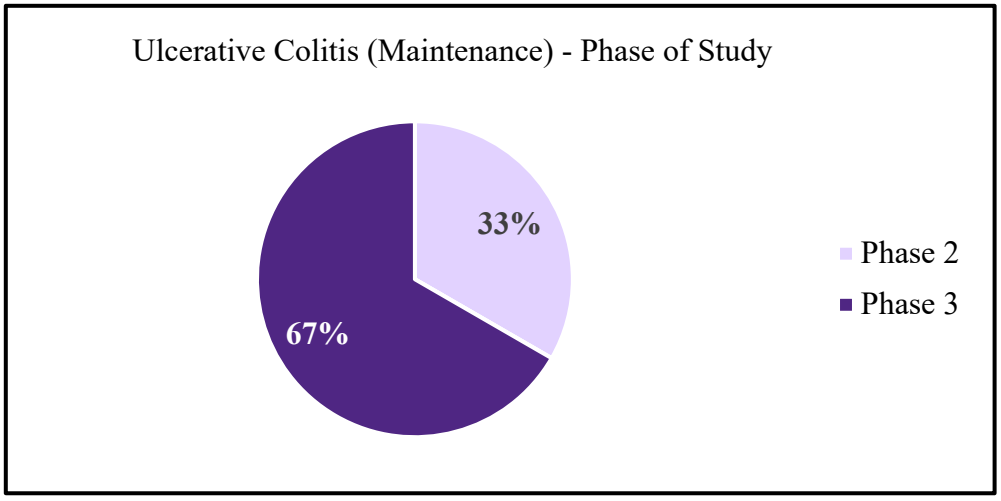

Figure F: Pie chart representing the breakdown of which study phase the maintenance trials included in the ulcerative colitis disease classification were in.

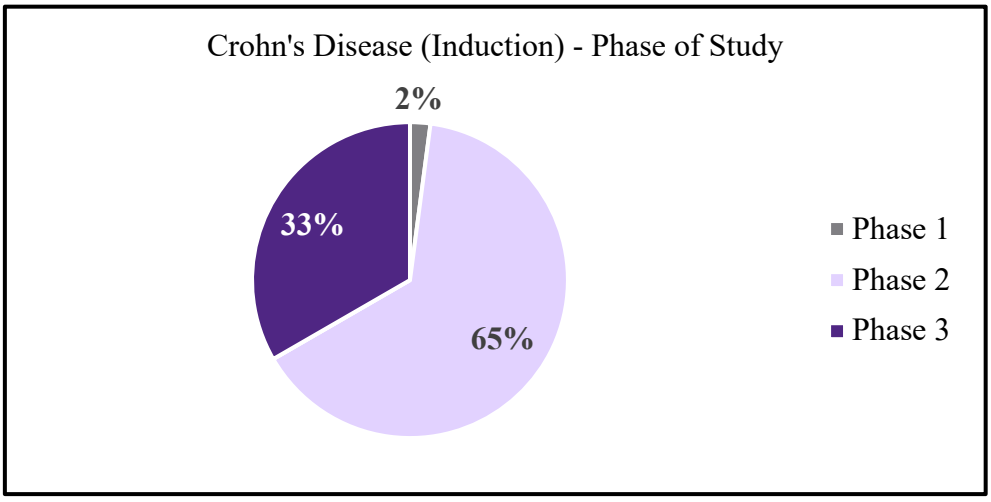

Figure G: Pie chart representing the breakdown of which study phase the induction trials included in the Crohn's disease classification were in.

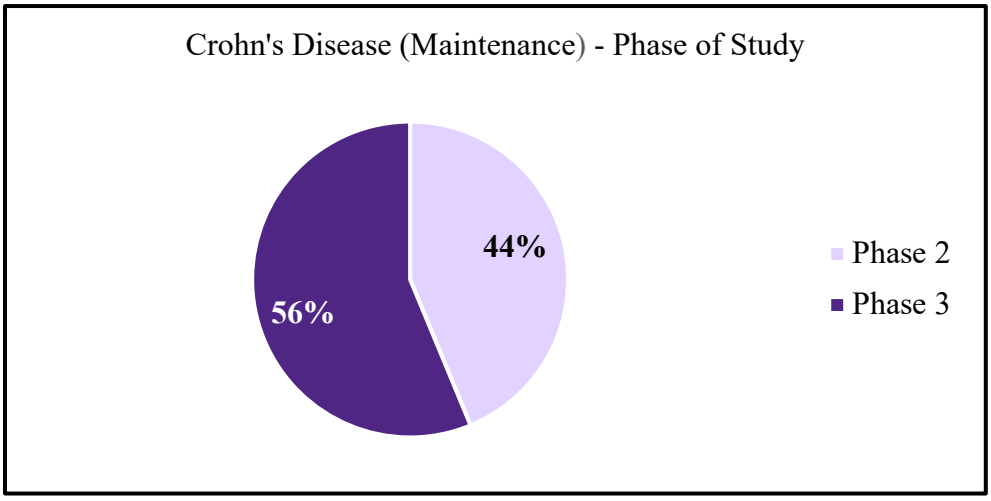

Figure H: Pie chart representing the breakdown of which study phase the maintenance trials included in the Crohn's disease classification were in.

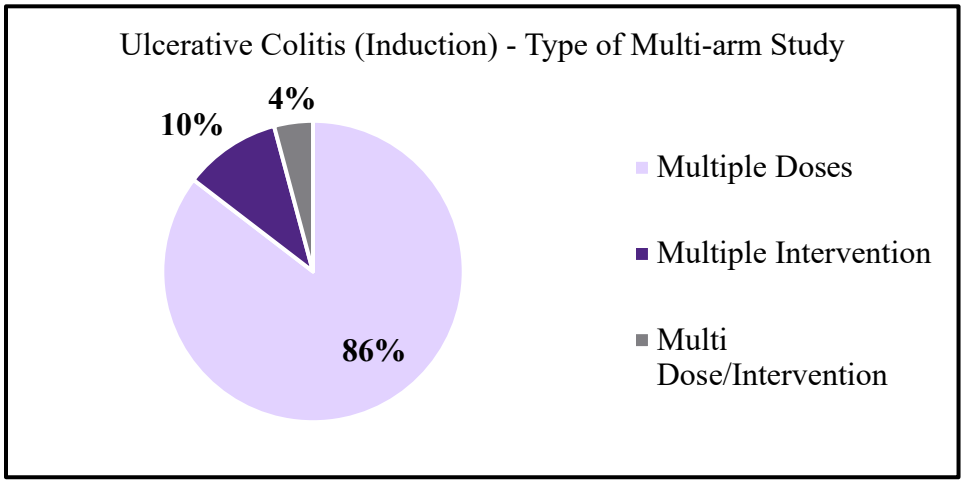

Figure I: Pie chart representing the breakdown of which types of multi-arm trials induction trials were included in the Ulcerative Colitis disease classification.

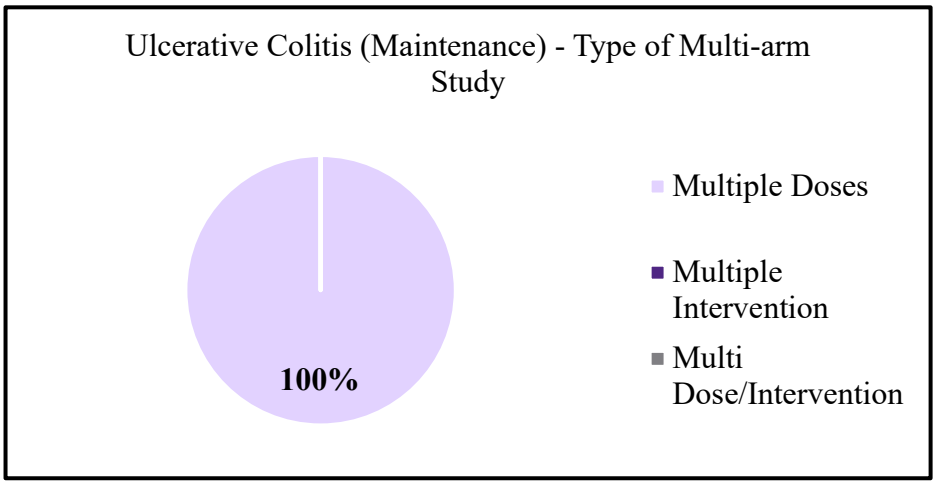

Figure J: Pie chart representing the breakdown of which types of multi-arm trials maintenance trials were included in the Ulcerative Colitis disease classification.

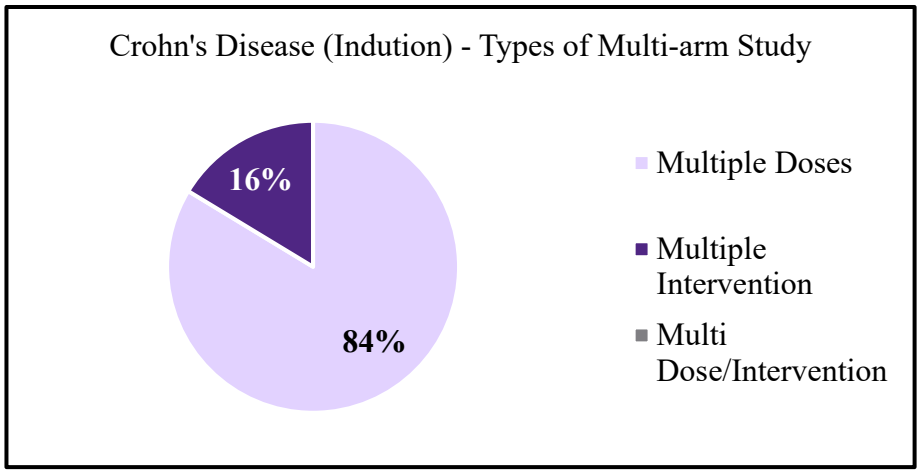

Figure K: Pie chart representing the breakdown of which types of multi-arm trials induction trials were included in the Crohn's Disease classification.

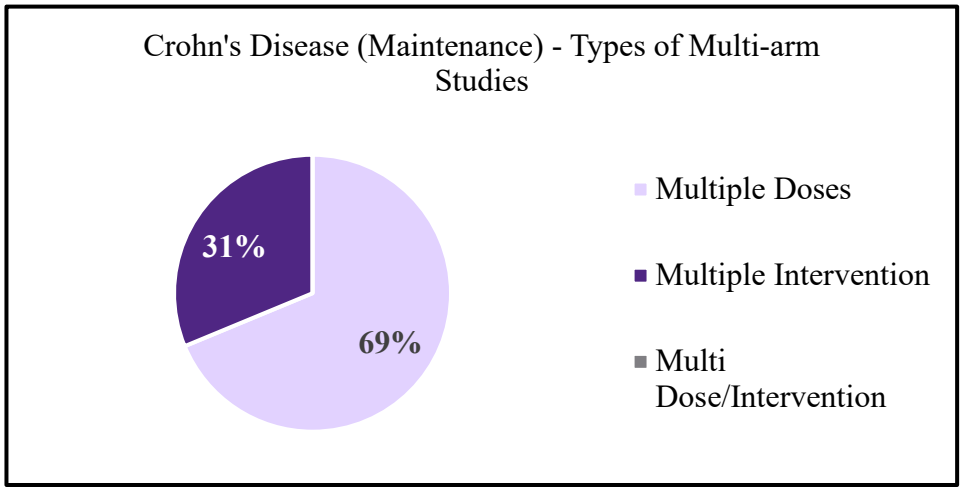

Figure L: Pie chart representing the breakdown of which types of multi-arm trials maintenance trials were included in the Crohn's Disease classification.

Appendix 3

The collected data for this trial was provided as a *SAS* dataset (*mln02*) and was set up with the following code to allow for further designated analysis in Chapter 4.

```
*Feagan, B.G., Greenberg, G.R., Wild, G., Fedorak, R.N.,
ParÈ, P., McDonald, J.W., DubÈ, R., Cohen, A.,
Steinhart, A.H., Landau, S. and Aguzzi, R.A., 2005.
Treatment of ulcerative colitis with a humanized
antibody to the a4fl7 integrin. New England Journal of
Medicine, 352(24), pp.2499-2507.;

options nocenter nofmterr ls=132;
*ods graphics off;
libname tmp1 'C:\Users\[*Insert HD name]\Desktop';
data mln02;
set tmp1.mln02;
run;
```

To assess the week 6 UC clinical scores, the following SAS code was used to generate the following descriptive statistic values and figures to assess the health outcomes between the three groups.

1) Number of Observations, Means, Standard Deviations, Minimum/Maximum values - Table

```
proc MEANS data=tmp1.mln02;
class tmtcode;
var w6_uccs;
run;
```

2) Boxplots and Quartiles

```
title "Boxplot of Week 6 UCCS by Treatment Assignment";
proc SGPLOT data=tmp1.mln02;
   vbox w6_uccs / category=tmtcode;
   yaxis label="Week 6 UC Clinical Scores (UCCS)";
   xaxis label="Treatment Assignment";
   run;
```

3) Boxplots and Quartiles – Table

```
proc univariate data=tmp1.mln02;
class tmtcode;
var w6_uccs;
output out=quartile_data
pctlpts = 25 50 75
pctlpre = Q_;
run;
```

4) Frequency Bar Graphs

```
proc SGPLOT data =tmp1.mln02;
vbar w6_uccs / group = tmtcode GROUPDISPLAY = CLUSTER;
title 'Week 6 UCCS Frequency within Treatment Assignment';
yaxis label="Frequency of Score";
xaxis label=" Week 6 UC Clinical Score (UCCS)";
run;
```

5) Histograms

```
title "Distrbution of Week 6 UCCS by Treatment Assignment";
proc SGPLOT data=tmp1.mln02;
histogram w6_uccs / group=tmtcode transparency=0.5
scale=count;
density w6_uccs / type=normal group=tmtcode;
keylegend / location=inside position=topright across=1;
xaxis label="Week 6 UC Clinical Score (UCCS)";
run;
```

To assess the week 6 UC clinical scores, the following SAS code was used to generate p-values according to the four multiple comparison procedures presented in Chapter 2.

```
*Establish the probability model for analysis;
%let za = probit(1-.05/2);
proc contents data=thesis.mln02;
run;

*Set up a separate dataset for analysis. Recall that
tmtcode=1, .5 mg, 2=2.0mg, 3=placebo;
data mln02;
set thesis.mln02;
run;

*Run an Ancova for w6_uccs;
```

```
proc orthoreg data=thesis.mln02;
class tmtcode ;
model w6_uccs = b_uccs  tmtcode;
lsmeans tmtcode/adjust=simulate(acc=0.0002 seed=1234 ) cl
stepdown(type=logical);
ods select diffs;
run;

proc print data=diffs;
run;

*Insert the p-value based analysis;
data pp;
input p@@;
cards;
.1809 .0014 .0649;

*Use the Multtest function to adjust the p-values using
Bonferroni, Holm, and Hochberg;
proc multtest inpvalues(p)=pp bon holm hochberg ;
run;

*Use the GLM function to produce p-values using the Dunnet
method;
proc glm data=mln02;
class tmtcode;
model w6_uccs = b_uccs tmtcode;
lsmeans tmtcode /adjust=dunnett pdiff=control('3');
run;
```

Appendix 4

Permission for Re-use Figure 1:

Dear Miss Sahiba Saini,

Thank you for your email.

We hereby grant you permission to reprint the material below at no charge in your thesis subject to the following conditions:

1. If any part of the material to be used (for example, figures) has appeared in our publication with credit or acknowledgement to another source, permission must also be sought from that source. If such permission is not obtained then that material may not be included in your publication/copies.

2. Suitable acknowledgment to the source must be made, either as a footnote or in a reference list at the end of your publication, as follows:
"This article was published in Publication title, Vol number, Author(s), Title of article, Page Nos, Copyright Elsevier (or appropriate Society name) (Year)."

3. Your thesis may be submitted to your institution in either print or electronic form.

4. Reproduction of this material is confined to the purpose for which permission is hereby given

5. This permission is granted for non-exclusive world English rights only. For other languages please reapply separately for each one required. Permission excludes use in an electronic form other than submission. Should you have a

specific electronic project in mind please reapply for permission.

6. Should your thesis be published commercially, please reapply for permission.

This includes permission for the Library and Archives of Canada to supply single copies, on demand, of the complete thesis. Should your thesis be published commercially, please reapply for permission- Canada
This includes permission for UMI to supply single copies, on demand, of the complete thesis. Should your thesis be published commercially, please reapply for permission- ROW

7. Posting of the full article online is not permitted. You may post an abstract with a link to the Elsevier websitewww.elsevier.com, or to the article on ScienceDirect if it is available on that platform.

8. Article can used be in the University library if it is embedded in the thesis and not used commercially.

Kind regards,

Thomas Rexson Yesudoss
Copyrights Specialist
ELSEVIER | HCM - Health Content Management

Visit Elsevier Permissions

From:
Date: Wednesday, August 30, 2023 07:04 PM GMT

Submission ID: 1369492
Date: 30 Aug 2023 8:04pm

Name: Miss Sahiba Saini
Institute/company: University of Western Ontario

Type of Publication: Journal

Title: The Lancet Gastroenterology & Hepatology
Auhtors: Sudabeh Alatab,Sadaf G Sepanlou,Kevin Ikuta,Homayoon Vahedi,Catherine Bisignano,Saeid Safiri,Anahita Sadeghi,Molly R Nixon,Amir Abdoli,Hassan Abolhassani,Vahid Alipour,Majid A H Almadi,Amir Almasi-Hashiani,Amir Anushiravani,Jalal Arabloo et al.
Year: 2020
From page: 17
To page: 30
ISSN: 2468-1253
Volume: 5
Issue: 1
Article title: The global, regional, and national burden of inflammatory bowel disease in 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017

I would like to use: Figure(s)

Quantity of material: 1
I am the author of the Elsevier material: No
Elsevier author is involved in my project: No

In what format will you use the material: Print and Electronic
Translation: No

Proposed use: Reuse in a thesis/dissertation

Material can be extracted: No

Curriculum Vitae

Name:	Sahiba Saini
Post-secondary Education and Degrees:	University of Toronto Toronto, Ontario, Canada 2015-2020 H.BSc The University of Western Ontario London, Ontario, Canada 2020-2023 M.Sc
Related Work Experience	Research Intern Ontario Health – Population Health and Prevention Team 2023