EMPIRICAL APPLICATION OF STATISTICS FOR CONTINUOUS OUTCOMES

EMPIRICAL APPLICATION OF DIFFERENT STATISTICAL METHODS FOR ANALYZING CONTINUOUS OUTCOMES IN RANDOMIZED CONTROLLED TRIALS

BY: Shiyuan Zhang, B.SC.

A Thesis Submitted to the School of Graduate Studies in Paritial Fulfillment of the Requirements for the Degree Master of Science

Masters of Science in Health Research Methodology (2013)

Department of Clinical Epidemiology and Biostatistics, McMaster University

1280 Main St W, Hamilton, ON L8S 4L8

TITLE: Empirical Application of Different Statistical Methods For Analyzing Continuous Outcomes In RCTs

Author: Shiyuan Zhang (BSc)

Supervisor: Dr. Lehana Thabane

Number of pages: 48

Abstract:

Background: Post-operative pain management in total joint replacement surgery remains to be ineffective in up to 50% of patients and remains to have overwhelming impacts in terms of patient well-being and healthcare burden. The MOBILE trial was designed to assess whether the addition of gabapentin to a multimodal perioperative analgesia regimen can reduce morphine consumption or improve analgesia of patients following total joint arthroplasty. We present here empirical application of these various statistical methods to the MOBILE trial.

Methods: Part 1: Analysis of covariance (ANCOVA) was used to adjust for baseline measures and to provide an unbiased estimate of the mean group difference of the one year post-operative knee flexion scores in knee arthroplasty patients. Robustness test were done by comparing ANCOVA to three comparative methods: i) the post-treatment scores, ii) change in scores, iii) percentage change from baseline.

Part 2: Morphine consumption, taken at 4 time periods, of both the total hip and total knee arthroplasty patients was analyzed using linear mixed-effects model (LMEM) to provide a longitudinal estimate of the group difference. Repeated measures ANOVA and generalized estimating equations were used in a sensitivity analysis to compare robustness of the methods. Additionally, robustness of different covariance matrix structures in the LMEM were tested, namely first order auto-regressive compared to compound symmetry and unstructured.

Results: Part 1: All four methods showed similar direction of effect, however ANCOVA (-3.9, 95% CI -9.5, 1.6, p=0.15) and post-treatment score (-4.3, 95% CI -9.8, 1.2, p=0.12) method provided the highest precision of estimate compared to change score (-3.0, 95% CI -9.9, 3.8, p=0.38) and percent change (-0.019, 95% CI -0.087, 0.050, p=0.58).
Part 2: There was no statistically significant difference between the morphine consumption in the treatment group and the control group (1.0, 95% CI -4.7, 6.7, p=0.73). The results remained robust across different longitudinal methods and different covariance matrix structures.

Conclusion: ANCOVA, through both simulation and empirical studies, provides the best statistical estimation for analyzing continuous outcomes requiring covariate adjustment. More wide-spread of the use of ANCOVA should be recommended amongst not only biostatisticians but also clinicians and trialists. The re-analysis of the morphine consumption aligns with the results of the MOBILE trial that gabapentin did not significantly reduce morphine consumption in patients undergoing major replacement surgeries. More work in area of post-operative pain is required to provide sufficient management for this patient population.

Special thanks to Dr. Lehana Thabane for continuous support, guidance, and mentorship during not only the conduct of this thesis project but also throughout the two-year master's education. A great deal of work and dedication was put forth by Dr. Thabane to ensure the successful completion of this project and the Master's degree.

Special thanks to Dr. James Paul for providing support in the conduct of this study and assistance with the Master's education. The clinical expertise in the topic area was of great importance in the proper conduct of the project.

Special thanks to Dr. Gary Foster for guidance and assistance in the study and specifically help with statistical analysis of the project. The help is greatly appreciated and has definitely played an integral role in the completion of the project.

Table of Contents

List of Figures and Tables

List of all Abbreviations and Symbols

ANCOVA	Analysis of Covariance
AR (1)	First Order Auto-Regressive
CCA	Complete-Case Analysis
CJRR	Canadian Joint Replacement Registry
CONSORT	Consolidated Standards of Reporting Trials
CS	Compound Symmetry
GEE	General Estimating Equations
LMEM	Linear Mixed Effects Model
MAR	Missing at Random
MCAR	Missing Completely at Random
MCMC	Markov Chain Monte Carlo
MI	Multiple Imputation
MOBILE	The Morphine Consumption In Joint Replacement Patients, With And Without Gabapentin Treatment, A Randomized Controlled Study
NSAID	Non-Steroidal Anti-Inflammatory Drug
RCT	Randomized Controlled Trials
RM-ANOVA	Repeated Measures Analysis of Variance
THA	Total Hip Arthroplasty
TKA	Total Knee Arthroplasty
UNS	Unstructured

Declaration of Academic Achievement

Student contribution to work:

The two sections of the thesis will be submitted for publication and below outlines the contribution to the paper.

Under the supervisor of Dr. Thabane, with various discussions with Dr. Foster and Dr. Paul, I conceptualized and design the study, including writing the initial proposal. As the project progressed, I obtained the MOBILE data from Toni Tidy and I conducted the statistical analysis. I also drafted the manuscript of the two studies.

1.0 Overarching Introduction

1.1 Clinical problem

Randomized controlled trials (RCTs) are the foundation of evidence-based decision making in healthcare and epidemiology [1]. RCTs are also the source of primary research [2, 3]. Efficacy of intervention, such as a new drug, surgery, or a novel way to provide better treatment to patients, can be compared through a RCT by providing a systematic method with low risk of bias. Proper conclusions from a RCT can only been drawn with the most appropriate statistical analyses. Indeed the proper design and reporting of results is strongly advocated by guidelines such as the Consolidated Standards of Reporting Trials (CONSORT) [4, 5].

Generally, the statistical analyses of an RCT are aimed at addressing three questions. These include; Does a population effect exists between the groups examined; what is the magnitude and precision of the effect if it does exist; and lastly does the statistical effect also have clinical importance [6]? It can be suspected that a proper design and methodology will ensure these statistics are addressed in a clear, concise and unbiased way. However, the proper conduct of an RCT is often more difficult than scientists and clinicians have envisioned. Some of the difficulties include having well defined outcomes that are measured in an appropriate way to reduce error [6]. Additionally, the statistical method of analysis must be appropriate for the type of data that provide the best possible parameter estimate. Potential sources of error may also arise from choosing incorrect time periods for measuring the outcome. More specifically, the strength of the evidence can differ greatly based on the specifics of the analysis plan. For instance, Whiting-O'Keefe et. al. examined the statistical methodology of health research studies published in some of the highest impact journals, such as Lancet, the New England Journal of Medicine, and Medical care [7]. In 20 of 28 health care experiments (71%), there was an inappropriate use of patient-related observations as the unit of analysis. More importantly, the conclusion formed was directly influence by this error as there was an erroneous increase in the power of the experiment to detect differences between the two intervention groups.

Statistical power is the probability that a test of the null hypothesis will correctly yield statistical significance when the null hypothesis is false. It is the chance of finding a difference between the treatment groups when such a difference exists. A number of factors influence the study power, including the sample size, the variability of treatment effects, and the statistical methods adopted [8]. Although it is generally accepted that an increase in sample size will increase the study power, the subsequent increase in the recruitment length and cost often makes this choice undesirable. Out of the three variables mentioned, the one often neglected is the differences in statistical power that can be achieved by adopting the most appropriate statistical method. A well developed statistical plan that incorporates the best statistical method will be the more cost-effective method to obtain the maximal study power.

More recently, a sample of 50 clinical trials reported in four major medical journals were examined to determine the extent of proper use of subgroup analysis and adjustment for baseline covariates [9]. Although two-thirds of the studies presented subgroup analyses, they were mostly inappropriate due to the lack of statistical tests for interaction. The author also concluded that there was a need for better defined statistical analysis plan for uses of baseline data, especially covariate-adjusted analyses and subgroup analyses. Moreover, the adoption of standards for statistical reporting must be improved. Investigators and journals need to adopt improved standards of statistical reporting, and exercise caution when drawing conclusions from subgroup findings.

Another source of error in the statistical method of trials comes from the improper modeling to detect and summarizing data patterns [10]. However, there is a higher risk of bias using the more efficient method of modeling. Incorrect assumptions within the analysis can potentially compromise estimates and tests derived from the model.

As these problems continue to persist in clinical trials, more attention must be paid to having a predefined statistical plan and methodology to ensure the publication of studies with the lowest risk of bias. The following studies attempt to examine two common areas of statistical analysis that may be at risk of bias when choosing the most appropriate statistical method. The first section critically appraises different statistical methods for adjusting trial results with the baseline covariates followed by the use of empirical data from the trial conducted at McMaster University to link theoretical framework with empirical data. The trial from which the empirical data was collected from is described in the following section. The second part of this paper examines different longitudinal methods of analysis and provides a sensitivity analysis of a published trial to provide inference of the trial results in a longitudinal manner. Similarly, the theoretical foundations of longitudinal analysis are summarized briefly followed by the empirical application of these statistical methods with the fore-mentioned trial.

1.2 The morphine consumption in joint replacement patients, with and without gabapentin treatment, a randomized controlled study (MOBILE) trial

Total hip arthroplasty (THA) and total knee arthroplasty (TKA) are major joint replacement surgeries after which patients can experience intense acute and chronic pain [11]. From 2008 to 2009, the Canadian Joint Replacement Registry (CJRR) registered 24,253 hospitalizations for hip replacements in Canada, with a majority of patients (63%) being 65 years of age or older [12]. Similar numbers are seen for knee replacement surgeries. There has been a 10-year increase of 101%, and a corresponding annual 6% increase in the number of joint replacement surgeries in Canada. As the prevalence of patients requiring joint arthroplasty increases, more attention must be paid to the surgical techniques and analgesia provided in the postoperative setting for this population.

Major surgery such as THA and TKA often leads to persistent acute and chronic pain in 10-50% of the patients following surgery [13]. There have been different strategies for providing adequate analgesic effects in this target population, including wound infiltration with local anaesthetic, peripheral nerve blockade with local anaesthetic, epidural local anaesthetic, oral or injectable non-steroidal anti-inflammatory drugs (NSAIDs), and systemic opioid (intravenous, intermittent, or patient-controlled analgesia) [14, 15]. Although each of the above mentioned strategies has advantages and short-comings, there has been a shift in anaesthesiology towards the use of a combination of strategies, which is often termed multimodal analgesia. Multimodal analgesia is defined as the use of a combination of opioid and non-opioid to manage postoperative pain, with the rationale behind such intervention being achieving sufficient analgesia due to additive effects, while minimizing the dose of individual drug [14, 16]. This also has the advantages of quickened recovery, shortened hospitalization time, and improved patient functionality [14, 15, 16]. Multimodal analgesia will not only allow for better patient pain management while reducing side effects, but can also significantly reduce healthcare costs.

1.3 Opioid for Pain Management and its Disadvantages

There has been a long history of using opioid as the main source of analgesia, but a shift from this practice of perioperative pain management is slowly recognized in different fields of medicine [15, 17, 18]. This is primarily due to some of the safety issues related to the use of opioids. For instance, high doses of opioids can raise safety concerns, such as respiratory depression. In a meta-analysis published by Walder et al. on the efficacy and safety of PCA for acute postoperative pain management, 30 of 180 patients (17%) had hypoxia (SaO2, 90%) [19]. Other side effects include high incidence of nausea, vomiting, confusion and delirium, constipation, and pruritus [17, 20]. Opioids can also hinder fast recovery and discharge due to its long duration of action in some patients. These unwanted side-effects of opioids are all related to prolonged rehabilitation and decrease in functional outcomes after procedures such as total knee arthroplasty and total hip arthroplasty.

1.4 Alternative Strategies to Opioids

Tissue trauma from surgery oftentimes sensitizes the peripheral nociceptors leading to central neuronal sensitization[13, 21]. Due to this biological rationale, adjuvants such as ketamine and gabapentinoids (gabapentin and pregablin) have shown to play a role in perioperative and postoperative pain management by providing preventive analgesia benefit, reducing postoperative opioid use, and decreased pain level[17, 22].

1.5 Gabapentin as an Adjuvant for Pain Management and the MOBILE TRIAL

"The **m**orphine c**o**nsumption in joint replacement patients, with and without ga**b**apentin treatment, a random**i**zed control**le**d study" (MOBILE) trials [2, 3] are designed to assess whether the addition of gabapentin, an anti-convulsive drug traditionally used for chronic pain management, to a multimodal perioperative analgesia regimen can reduce postoperative morphine consumption or improve analgesia following total hip or knee arthroplasty. Secondary outcomes such as pain score, range of motion, and side effects were also compared. Previous randomized controlled trials (RCTs) and a meta-analysis [23, 24] of 8 placebo-controlled, randomized trials showed gabapentin was shown to reduce pain scores, opioid consumption and other side effects. However, in the MOBILE trial, the primary outcome of 72 hour cumulative morphine consumption did not show a statistically significant difference between the gabapentin group and the control group [2, 3] .

1.7 Objectives

The specific objectives of the studies are defined in more details in the two sections. The overall objective of the studies is to empirically demonstrate the impact of statistical methods on the results of the study. The analysis can also help add more confidence to the results published in the MOBILE trials [2, 3].

1.8 Implications

The implications of the study are to communicate the importance of choosing the appropriate statistical method when designing clinical trials. Continuous outcomes are common in trials and appropriate analysis plans can have implications for sample size, study power, and the validity of study results. A meaningful reduction in the number of participants can be achieved with a proper analysis plan, which translates into decreasing the number of patients exposed to potentially harmful adverse events of the intervention. Furthermore, a simplified clarification on the statistical methods can benefit statisticians and trialists alike in creating harmonized language and improve the efficiency of the conduct of clinical trials.

PART I: EMPIRICAL COMPARISON OF FOUR BASELINE COVARIATE ADJUSTMENT METHODS IN ANALYSIS OF CONTINUOUS OUTCOMES IN RANDOMIZED CONTROLLED TRIALS

Master’s Thesis, Health Research Methodology,
Department of Clinical Epidemiology and Biostatics,
McMaster University

Shiyuan Zhang, James Paul, Manyat Nantha-Aree, Norman Buckley, Uswa Shahzad, Ji Cheng, Antonella Tidy, Justin DeBeer, Mitchell Winemaker, David Wismer, Dinshaw Punthakee, Victoria Avram, Lehana Thabane

2.0 **Introduction:**

Continuous outcomes are one of the most commonly used types of outcomes in clinical trials. They are easy to interpret for statistician and clinicians alike. For instance blood pressure, glucose sugar, or FEV1 are continuous in nature and understandable without requiring much manipulation to the data.

In a number of research fields, such as psychology and education, pain, and quality of life research, a common RCT design involves the measure of the primary outcome of the two groups at two time points, before, also known as baseline or covariate, and after the treatment [25]. This type of baseline-controlled design can be a very statistically powerful design to evaluate casual factors since adjustment of unbalanced covariates can be properly done in order to isolate the factors at work [26-29]. This design is often of great use to evaluators because it can control for all of the major threats to internal validity, such as maturation, selection, and instrumentation [30].

2.0.1 Clinical Problem: Inconsistency in choosing the method for baseline adjustment

Although seemingly straight forward, the statistical comparison of a continuous variable in an RCT that has both a pre and post-treatment score present an interesting challenge for clinicians and statisticians. The statistical properties of baseline adjustment methods are indeed very complex and often poorly understood, resulting in confusion when choosing the most appropriate statistical strategy [31]. Assman et. al. analyzed a sample of 50 trials from four top medical journals *BMJ*, *JAMA*, *The Lancet*, and *New England Journal of Medicine* [9] and reported the use of a number of different covariate adjustment methods.

The lack of consistency in the literature when the pre-post design is used further contributes to the difficulty of having a standard statistical method. These inconsistencies are often over whether to use covariate adjustment and the criteria used for selecting baseline factors for which to adjust. Most trials emphasised the simple unadjusted results and covariate adjustment usually made negligible differences.

2.0.2 Critical Appraisal of Four Adjustment Methods

There are a number of different baseline adjustment methods that are commonly used in clinical trials. Their popularity arises from a number of different factors, including ease of interpretation, ease of analysis, convenience, and historical reasons. Statisticians have also evaluated these methods based on whether the method is able to achieve the most appropriate size of estimate, precision, and p-value for the treatment difference [32, 33]. The four methods examined are post-treatment, analysis of covariance, change score, and percent change score. Specifically, for each method, a brief description of the method, and the advantages and disadvantages are described.

i) Post-Treatment:

In this method, analysis is done on the outcome of interest with no covariate adjustment. There are also a number of advantages of comparing strictly the unmodified outcome, including minimal influence by a secondary outcome, interpretation of the result is straight forward, and least time consuming. Moreover, for most clinical trials, analyses which adjust for baseline covariates are in close agreement with the simpler unadjusted treatment comparisons [9].

Other rationales of using the post-treatment score method with a simple ANOVA or T-test is that in practice, randomization allows for a balanced baseline measure in both treatment groups and thus any covariate adjustment is deemed unnecessary. This assumption of balanced baseline variable can be violated even in the perfectly designed RCT, and the effect is especially magnified in trials with a small sample size [34]. Indeed, through many simulation studies, the analysis of the post-treatment score when the baseline is adjusted can lead to different results [35].

ii) Analysis of Covariance (ANCOVA):

In recently published literature, the use of analysis of covariance as the statistical method of choice for analysis of intervention effect and adjustment for baseline variable has been advocated repeatedly [8, 36-40]. In order to adjust for the baseline measurement of the same variable, ANCOVA allows the use of the baseline result as a covariate in the analysis. In addition to the grouping factor and the outcome of post-treatment score, an additional covariate term is introduced that allows for a statistical adjustment based on the baseline score.

The reason for using ANCOVA in the analysis of continuous outcomes in studies is many fold. In RCTs, when treatment and placebo group have the same expected mean baseline values, both the post-treatment score method and the ANCOVA model will provide an unbiased estimate of the true treatment effect [34]. However, one advantage ANCOVA provides is a higher efficiency even under unbiased conditions and with controlled α-levels. Furthermore, even in the presence of measurement error or other within-patient variations, the ANCOVA approach based on the observed data still provides an unbiased estimate with better precision and a more powerful test than the ANOVA approach.

iii) Difference score between post-treatment score and baseline:

Many clinicians and clinical trials that deal with quality of life, such as studies in oncology, often have their primary outcome as a change in outcome calculated by subtracting the baseline value from the follow-up or post-treatment value. This is often referred to as a change score or sometimes the gain score, and it represents directly the change that is experienced or measured before and after the treatment. However, an obvious issue with the change score method is that the regression-to-the-mean effect can be substantially different between the two treatment groups.

Using change score may provide an advantage over using ANCOVA because it is not necessary to assume that baseline variables are measured without error [36]. Even in ANCOVA models that attempts to adjust for measurement error, the estimate of treatment effect remains biased [41].

iv) Percent difference score between post-treatment score and baseline:

An extension to the change score that preserves some of the information of the baseline result is by computing the percent change score. Percent change score is calculated by normalizing the change score by the baseline data. It can be considered an improvement to represent the change of pre and post-treatment since it is normalized, which increases comparability across subjects. However, one major concern in terms of statistical analysis is whether or not normalizing the data would alter the distribution of the data and may introduce additional complexity to the analysis.

2.0.3 Study Objective

In this study, four baseline adjustment methods were used to demonstrate empirically if the ANCOVA is statistically efficient compared to statistical analysis by post-treatment, change score, or percent change score. To compare methods, an analysis of the flexion range of motion is conducted using ANCOVA and the other three methods outlined above. The robustness of the latter three methods was evaluated by qualitatively comparing the direction and magnitude of effect, the precision of the estimate, and the ease of interpretation of the results. The study also seeks to identify the best approach to handle missing data in clinical trials. This is done through the use of multiple imputation (MI), where a sensitivity analysis is done by conducting a complete-case analysis (CCA).

2.1 Methods:

2.1.1 Current Literature Review and Summary

A search of published literature on baseline adjustment methods was conducted to summarize the available information on the use of these methods and their impact on the results of studies. Three different designs of studies were included in the search, descriptive, empirical, and simulation studies. Descriptive studies, in this study, are systematic reviews or compilation studies where a cohort of studies is summarized to address the current knowledge on the topic area. Empirical studies are studies where data from another previously published study is used to re-analyse with other statistical methods to compare and contrast these methods. Simulation studies are studies where statistical methods are compared, in terms of statistical power and other parameters using statistical and mathematical simulation. Studies of these designs were compiled and results were summarized in terms of the impact of the method on the results and the author's comments on the choice of the most appropriate statistical method.

We also sought to conduct an empirical study to examine the four baseline covariate adjustment methods. Empirical data from the MOBILE trial [1]was used to demonstrate the performance of different baseline adjustment methods and their effect on the statistical comparison. The statistical properties of covariate adjustments were examined in terms of several aims. These include: (i) direction of treatment difference; (ii) magnitude of treatment difference; (iii) precision of treatment difference [32, 33].

2.1.2 Description of MOBILE trial: Total Knee Arthroplasty [1]

In a single-centre, blinded, randomized, placebo-controlled study, 102 patients were randomized to receive either gabapentin or placebo, in addition to standard of care, 2 hours before undergoing total knee arthroplasty. Morphine consumption at 72 hours was the primary endpoint of the trial. Secondary outcomes included knee flexion score, pain score on a visual analogue scale, and side effects. In the trial, the statistical analysis was done without the adjustment of baseline covariates and a post-treatment score comparison was done for the knee flexion variable.

For patients in the total knee arthroplasty treatment, the variables used for analysis were flexion range of motion pre-operatively and one year after the procedure. Other than patient ID, group assignment, and the two flexion range of motion values, a few baseline and postoperative variables were also included in the data collection process. These variables were: gender, weight, height, ASA, blood pressure systolic, blood pressure diastolic, pain at rest at the four differenct time points, pain with passive movement at the time points, pain with weight at different time points. These variables were included in the MI strategy to have the most complete imputation method without any bias. In the MI process, all of these variables were imputed in a coherent process, using the Markov chain Monte Carlo (MCMC) method. Five iteration of the MI process was used for the three analyses of the longitudinal morphine consumption data and MIAnalyze was used to combine the estimates from each of the five iterations. A detailed analysis flow-diagram is included in the appendix.

2.1.3 Analysis of Post-Operative Knee Flexion Range of Motion, adjusting for baseline

Four baseline adjustment methods were used to obtain a statistical comparison of the control against the active group. ANCOVA, with the pre-treatment ROM as a covariate was used to determine treatment effect after the data has been imputed with MI. The treatment effect estimate, 95% confidence interval, and p value of the estimates of the mean difference were summarized. All statistical tests were performed using SAS (SAS 9.2 (32) English).

2.1.4 Sensitivity Analysis

Two robustness tests were conducted for this section.

i) Robustness of Post, Change, and Percent-Change Analysis:

Three other methods were used as comparators to the ANCOVA analysis to determine the sensitivity of the results to the choice of baseline adjustment method. The other three baseline adjustment methods were: comparing post-treatment score only with no baseline adjustment, subtracting the baseline from the post-treatment score and comparing the change score, and comparing the percent change score obtained from normalizing the change score described in the previous method by dividing the change score by the pre-treatment score **(Figure 2.1)**. For the purposes of this robustness test, only the MI results of the different methods were used. The robustness of each of the three methods was determined based on a number of factors, including direction and magnitude of effect, and the precision of the estimate. To further illustrate the robustness and potential difference between ANCOVA and the other three methods, a forest plot of the 95% CI of each of the 4 analyses was used to provide a visual comparison.

ii) Robustness of Complete-Case Analysis:

To compare whether the results were affected qualitatively by the implementation of MI, the results obtained by using complete-case analysis for each of the baseline adjustment method is used as the comparator. The robustness test is done by comparing the two types of data handling method based on a number of factors, including direction and magnitude of effect, precision of estimate. To further illustrate the difference between the two missing data handling methods, a forest plot of the 95% CI of each of the 8 analysis was used to provide a visual comparison.

2.2 Results:

2.2.1 Highlight of Literature on Use of Baseline Adjustment Methods

The use of these four different methods has been documented in the literature numerous times, through both theoretical simulation studies and studies using empirical data. For instance, Tu et al. 2005 used empirical results from periodontal research to demonstrate that different statistical methods have a substantial impact on study power [8]. In this study, it was demonstrated that with substantial variability in the correlation between the baseline and post-treatment score, ANCOVA should be used in preference to change score or percentage change score, as the appropriate method that is able to adequately reduce the Type II error rates. Other examples of studies are summarized in **Table 2.1**. In short, most empirical studies reported that ANCOVA is often the most appropriate statistical method to adjust for baseline covariates when analyzing randomized studies of continuous outcome. Similarly, theoretical and simulation studies show ANCOVA has the highest statistical power and is the method of choice. However, in nonrandomized studies, ANCOVA may yield biased results and the method of change score should be used as the primary method of analysis. Lastly, a number of descriptive studies that summarizes analysis from a large number of individual studies have also shown that the method of baseline adjustment still causes confusion for researchers. Specifically, there seems to be no single method that is consistently used and often times no justification is provided for the choice of the method that was used.

2.2.2 Analysis of Treatment Effect

The mean group difference, with the gabapentin treatment as the reference group was -5.5. This means the knee flexion range of motion of the patients in the active group was 5.5 less, on average, than those patients in the control group. The 95% CI is -11, 0.25, p value 0.068, and the difference between the two treatment arms is not statistically significant, although there is a trend towards less range of motion in the gabapentin group.

2.2.3 Sensitivity Analysis

i) Sensitivity to Method of Analysis/Baseline Adjustment (Figure 2.2, Table 2.2,2.3)

The primary analysis for the range of motion data was conducted using ANCOVA with MI for handling missing data. The first set of sensitivity analyses compared ANCOVA with the other three methods of adjusting for baseline data and conducting statistical comparison. The three methods all had similar direction of effect, where the group mean of flexion range of motion in the control group was higher than that of the active group. Moreover, the magnitude of the result of each of the three methods was similar to that obtained from the ANCOVA method, ranging from -3.9 to 4.3. The precision of the results was lower with change score and percent change score methods, which had a larger CI and p-value.

Comparing the group effect using post-treatment scores and without using a baseline adjustment had the most similar results compared to ANCOVA method. Along with having the same direction and almost identical magnitude of effect, the 95% CI was narrower, and had a correspondingly smaller p-value. Overall, the results of the post-treatment scores remained robust and the findings were consistent. The results obtained from the change score and percent change score methods had larger deviations compared to the primary analysis. Although they had the same direction of effect, the magnitude of effect was less and a wider 95% CI was obtained. Of of the three comparator methods, post-treatment offered the most robust method of analysis compared to ANCOVA, followed by change score and percent change score methods offered the least favourable method in terms of precision and magnitude of effect (Figure 2.2).

ii) Sensitivity to Missing Data (Figure 2.2, Table 2.2, 2.3)

The second set sensitivity analyses were done for different methods of handling missing data for flexion range of motion. For the knee range of motion score, 4% and 25% of the data was missing at baseline and post-treatment, follow-up time, respectively. The primary analysis for the range of motion data used ANCOVA with MI for handling missing data. When complete-case analysis was used, the mean group difference was -2.5 (95% CI -7.0, 2.3, p=0.27). Without any method of handling the missing data, the analysis of ANCOVA remained robust. The results maintained the direction of effect, where both estimates suggested trends favouring the control group. Moreover, the magnitude of the effect was also similar, where the point estimate was different by only 1.4. The precision was similar between MI and complete-case analysis. The p-value of the complete-case analysis was larger, due to the decrease in the precision of the analysis caused by lower number of cases available for analysis.

Comparing results between the two methods of handling missing data amongst the other three methods of analysis (post treatment, change score, and percent change), the exact same trends were observed. All results had the same direction of effect, with a slightly smaller magnitude of effect, smaller confidence interval, and larger p-values. The details of these results are summarized in Table 2.2 and 2.3.

2.3 Discussion:

2.3.1 Key Findings

The present study uses empirical data from TKA patients from the MOBILE trial to determine the treatment effect on the post-operative flexion range of motion score. With the method of ANCOVA while using MI to handle missing data, the results suggest the difference between the gabapentin and control group was not statistically significant.

Through sensitivity analysis on the method of analysis and method of handling missing data, it was found that all methods remained robust and the overall findings were consistent. It should be noted, however, that while the result were similar, there are distinguishable features. For instance, it was found that using the post-treatment score alone as the variable for statistical comparison yielded the most similar results to those from the ANCOVA method. In a simulation study published by Vickers 2001 [39], the statistical power of these four baseline adjustment methods were compared at different baseline to outcome score correlations. The results of this empirical study corroborate the simulation study, where at low correlation levels, ANCOVA and post-treatment score maintain statistical power at around 70%. Moreover, for change score and percent change (fraction) score, the statistical power decreases dramatically with the decrease of the baseline to outcome score correlation, where at a correlation of 0.2, the statistical power falls to around 50%.

In this study there was no significant correlation between the baseline and post-operative range of motion, it would interpreted as a near-zero correlation. Although the lowest correlation used in the simulation study was 0.2, it would be confidently extrapolated that for the analysis of data with no correlation, ANCOVA and post-treatment methods will have the highest statistical power. Furthermore, change score and percent change score will have even lower statistical power than those at the 0.2 correlation level. However, these conclusions were made on a qualitative level and no power calculation was done to determine the exact power of the various methods.

There are a number of methods commonly used for handling missing data. For a long time the preferred method of handling missing data was single imputation, by either imputing with the grand mean, or with last observation carried forward. There is increasing evidence and support for MI to be used as the primary method of handling missing data [42]. In the present study, the sensitivity analysis for the methods of handling missing data suggests that the conclusion drawn was not affected whether it was complete-case analysis or MI. Moreover, in the MOBILE trial, and in this study, we sought to follow an intention-to-treat principle, where all patients randomized were analyzed. The methods conducted for statistical comparison were quite robust even when missing data was ignored.

2.3.2 Key limitations

The brief literature review presented in **Table 2.1** was not conducted as a thorough systematic literature search. The summary provided was meant to highlight recent literature on the topic of covariate baseline adjustment methods, through simulation or empirical studies. It provides information that highlights the advantages and disadvantages of these various methodologies. Moreover, the results and conclusions drawn from these studies were not used to make inferences on the superiority of one method over the others. Nonetheless, the information gathered from this exercise helps to distill and assimilate large amounts of information in order to provide a quick overview of some of the research conducted in this topic area.

One of the main limitations of the empirical analysis portion of the study is the nature of the variable used in this study, which is the flexion range of motion. In the MOBILE trial, sample size estimation was based on the primary outcome, which was the cumulative morphine consumption at 72 hours post-operative. Therefore, with a sample size of 101, there may have been a lack of power to detect a difference between the two groups even if a true difference exists. As such, interpretation of the results must be made with caution due to the potential lack of sufficient power.

In this study, missing data was handled by using MI with five iterations. Although MI has been recognized as the most appropriate method for imputing missing data, it assumes that the missingness is either missing completely at random (MCAR) or at least missing at random (MAR) [42]. Currently, methods are not available for analysis data missing not at random. However, testing for the type of missing data mechanism is difficult, especially when there is a lack of auxiliary information, such as demographic, social characteristics of the participants. For the purposes of this study, the missingness was assumed to be MAR since only a small portion of the data is missing. Moreover, the robustness of the complete-case analysis further suggests that there was not a substantial amount of missing.

The comparison of this study, using empirical data, to previous simulation studies suggests similarity in the finding. However, interpretation of the results should be that obtained from an empirical study, where the characteristics of the data used may influence the results generated. This is to say that although change score and percent change have been suggested as the less statistically efficient method of adjusting for baseline, if the baseline data is highly correlated with the post-treatment score, change and percent change scores may be a valid and easily interpreted method to be used. Regardless, since ANCOVA has been shown, in a variety of studies [8,9,31,36,40,49,50,51,53], as the most statistically efficient method to analyze continuous outcomes with a baseline variable, it is suggested that the adoption of other methods of handling baseline data be used with caution.

2.4 Conclusion:

In this study, a comparison of the most commonly used methods of adjusting baseline data of a continuous outcome in RCT was done using an empirical dataset. The study results suggest that ANCOVA is a statistically efficient method of analyzing data of this nature and especially the use of change and percent change scores should be employed with caution since the statistical power of these methods is highly dependent on the correlation between the baseline and the outcome. A number of future studies should be conducted to strength the interpretation of this study. Simulation studies with a correlation of baseline and outcome lower than 0.2 can help strengthen the conclusion of this study. Moreover, empirical data that was the primary outcome of the study should be used in order to ensure the validity of the study.

Future studies may look at logistic regression and how the method of covariate adjustment effects the results. [43]. For instance, it is known that when the covariate included in the trial is that of a binary or survival nature, the adjustment methodology and implications are completely different. The omission of a balanced covariate has dramatic effects on the estimate of treatment effect and this effect is magnified when a highly prognostic covariate is included in the analysis. Investigating some of these scenarios and developing a complete empirical study based on those set out in this study would be of great interest [32].

PART II: RE-ANALYSIS OF MORPHINE CONSUMPTION FROM THE MOBILE TRIAL USING LONGITUDINAL STATISTICAL METHODS

Master's Thesis, Health Research Methodology, Department of Clinical Epidemiology and Biostatics, McMaster University

Shiyuan Zhang, James Paul, Manyat Nantha-Aree, Norman Buckley, Uswa Shahzad, Ji Cheng, Antonella Tidy, Justin DeBeer, Mitchell Winemaker, David Wismer, Dinshaw Punthakee, Victoria Avram, Lehana Thabane

9/23/2013

3.0 Introduction:

Major surgery such as total hip arthroplasty (THA) and TKA often leads to persistent acute and chronic pain in 10-50% of the patients following surgery [13]. Unrelieved pain after surgery increases heart rate, systemic vascular resistance, and circulating catecholamines, placing patients at risk of myocardial ischemia, stroke, bleeding, and other complications. There have been different strategies for providing adequate analgesic effects in this target population, including wound infiltration with local anaesthetic, peripheral nerve blockade with local anaesthetic, epidural local anaesthetic, oral or injectable non-steroidal anti-inflammatory drugs (NSAIDs), and systemic opioid (intravenous, intermittent, or patient-controlled analgesia) [14, 15]. Although each of the above mentioned strategies has advantages and short-comings, there has been a shift in anaesthesiology towards the use of a combination of these strategies, which is often termed multimodal analgesia. Multimodal analgesia is defined as the use of a combination of opioid and non-opioid to manage postoperative pain, with the rationale behind such intervention being achieving sufficient analgesia due to additive effects, while minimizing the dose of individual drug [14, 16]. This also has the advantages of quickened recovery, shortened hospitalization time, and improved patient functionality. Multimodal analgesia will not only allow for better patient pain management while reducing side effects, but its use can also significantly reduce healthcare costs.

The "The **m**orphine c**o**nsumption in joint replacement patients, with and without ga**b**apentin treatment, a random**i**zed control**le**d study" (MOBILE) trials [2, 3] is designed to assess whether the addition of gabapentin, an anti-convulsive drug traditionally used for chronic pain management, to a multimodal perioperative analgesia regimen can reduce postoperative morphine consumption or improve analgesia following total hip or knee arthroplasty. Secondary outcomes such as pain score, range of motion, and side effects were also compared. Previous randomized controlled trials (RCTs) and a meta-analysis [24] of 8 placebo-controlled randomized trials showed that gabapentin reduced pain scores, opioid consumption and other side effects. However, in the MOBILE trial, the primary outcome of 72 hour cumulative morphine consumption did not show a statistically significant difference between the gabapentin and the control groups [2, 3].

One of the potential sources of variability in trial results is the method of analysis. The choice of statistical analysis method can have a substantial influence on the statistical power and sample size of the trial [8, 44]. The primary objective of this study is to conduct an empirical re-analysis of the MOBILE trials by analyzing the primary outcome, morphine consumption, in a longitudinal manner, rather than to treat the outcome as one cumulative score. More specifically, if the primary outcome of 72 hour cumulative morphine consumption was analyzed longitudinally, instead of cross-sectionally, would the result of no treatment difference remain robust? Morphine consumption, which was measured at four time points, will be analyzed using linear mixed-effects model (LMEM), using a first order auto-regressive covariance matrix structure (AR(1)). Secondary outcomes of the study include a number of sensitivity analysis to determine the robustness of the results based on longitudinal method of choice, method of handling missing data, and choice of covariance matrix structure. Specifically, to determine the sensitivity to method of analysis, sensitivity analyses with two other longitudinal methods, ie repeated measures ANOVA and generalized estimating equations (GEE) - assuming AR(1) covariance structure - will be tested. The robustness of method of handling missing data will be determined by conducting a sensitivity analysis using a complete-case analysis. Lastly, two sensitivity analyses will be conducted to determine the robustness of covariance matrix structure when using LMEM, ie compound symmetry (CS) and unstructured (UNS) covariance structures.

3.1 Methods:

This is a statistical re-analysis of the data from the MOBILE trial to determine whether the statistical method had a major impact on the results.

3.1.1 Description of MOBILE trial: Total Knee Arthroplasty

In a single-centre, blinded, randomized, placebo-controlled study, 102 patients were randomized to receive either gabapentin or placebo, in addition to standard of care, 2 hours before undergoing total knee arthroplasty [2]. Morphine consumption, a continuous outcome, was recorded at four specific time points. The four time periods were: at post-anethesia care-unit (time 0), 24 hours after surgery (time 1), 48 hours after surgery (time 2), and 72 hours after surgery (time 3). In the trial, the statistical analysis was done by combining the morphine consumption at these four time points and using a t-test to compare the treatment group difference on the cumulative 72 hour morphine consumption.

3.1.2 *Description of MOBILE trial: Total Hip Arthroplasty*

In a single-centre, blinded, randomized, placebo-controlled study, 101 patients were randomized to receive either gabapentin or placebo, in addition to standard of care, 2 hours before undergoing total hip arthroplasty [3]. Morphine consumption, a continuous outcome, was recorded at four specific time points. The four time periods were: at post-anesthesia care-unit (time 0), 24 hours after surgery (time 1), 48 hours after surgery (time 2), and 72 hours after surgery (time 3). In the trial, the statistical analysis was done by combining the morphine consumption at these four time points and using a t-test to compare the treatment group difference on the cumulative 72 hour morphine consumption.

3.1.3 Data Analysis

Data from the MOBILE trial [2, 3] were obtained to conduct the analysis of the primary outcome of morphine consumption in its longitudinal form. The data of the two trials were kept separate and the morphine consumption of patients who underwent total knee or total hip arthroplasty were analyzed separately. A total of 203 (101 and 102 patients for knee and hip replacement, respectively) patients were included in this re-analysis and missing data were imputed using multiple imputation.

The multiple imputation (MI) process included a number of baseline variables from the trial and all these variables, including the morphine consumption outcome, was imputed in a similar process, using Markov chain Monte Carlo (MCMC) method. Five iterations of the MI process were included and a combined treatment estimate was generated at the end. The baseline variables included for the MI process were: gender, weight, height, ASA, blood pressure systolic, blood pressure diastolic, pain at rest at the four difference time points, pain with passive movement at the time points, pain with weight at different time points. These variables were included in the MI strategy to have the most complete imputation method without any bias.

The treatment estimate between the control and the experimental group was determined using LMEM, with a covariance structure of first-order auto-regressive. All statistical tests were performed using SAS (SAS 9.2 (32) English).

3.1.4 Sensitivity Analyses:

Three robustness tests were conducted for the secondary objective of the study, one for the two other longitudinal methods, one for handling of missing data, and the last one for the use of different covariance matrix structures when analyzing using LMEM.

i) Robustness of RM-ANOVA and GEE

The additional two longitudinal methods were used to test the robustness of the results obtained from LMEM, namely repeated measures analysis of variance (RM ANOVA) [45], GEE [46]. The sensitivity analysis was done by first analyzing the results with these two additional methods and qualitatively comparing the results in terms of direction of effect (sign), magnitude of effect (number), and precision of effect (p-value). The robustness of the method is then determined based on the above three criteria.

ii) Robustness of Complete-Case Analysis

To compare whether the results were affected qualitatively by the implementation of MI, the results obtained by using complete-case analysis for each of the longitudinal methods is used as the comparator. The robustness test is done by comparing the two types of data handling method based on a number of factors, including direction and magnitude of effect, precision of estimate. To further illustrate the difference between the two missing data handling methods, a forest plot of the 95% CI of each of the 6 analysis was used to provide a visual comparison.

iii) Robustness of Covariance Matrices, Compound Symmetry and Unstructured, in LMEM

To compare whether the results of the LMEM were affected qualitatively when using AR(1) compared to CS and UNS. For the purposes of this robustness test, both patient groups were used but only the MI results were used. The robustness test is done by comparing the different covariance matrix structures based on a number of factors, including direction and magnitude of effect, precision of estimate. To further illustrate the difference between the LMEM and the two analysis methods, a forest plot of the 95% CI of each of the 6 analysis was used to provide a visual comparison.

3.2 Results:

3.2.1 Analysis of Morphine Consumption with Four Repeated Time Points (Table 3.1):

Patients undergoing TKA (n=101): The mean effect size estimate obtained was 1.0 (95% CI -4.7, 6.7; p=0.73) between the groups, when analysis was performed with LMEM with MI. There was no statistical difference in morphine consumption between the gabapentin and control groups.

Patients undergoing THA (n=102): The mean effect size estimate obtained was -1.0 (95% CI -5.4, 3.3; p=0.63) between the groups, when analysis was performed with LMEM and MI. There was not a statistically significant difference in morphine consumption between the gabapentin and control groups.

3.2.2 Sensitivity Analyses (Figure 3.1 and 3.2):

i) Sensitivity to Method of Analysis

The primary method of analysis for the longitudinal data of morphine consumption was conducted using LMEM with MI as the method for handling missing data. Compared to the results of the TKA and THA patients generated from LMEM, the results from RM-ANOVA and GEE remained robust and the overall findings were consistent across methods (Figure 3.1 and 3.2). More specifically, the direction, magnitude, and precision were similar across all three methods. The two comparator methods had slightly tighter 95% CI and p-value.

ii) Sensitivity to Missing Data

There was a slight discrepancy in results between the TKA and THA patients and their results will be reported separately.

Patients undergoing TKA (n=101): There were 38 data points (9%) missing across the four time periods for the TKA patients, with only 2 patients missing all data points. The missing data did not exhibit a monotone pattern, where subsequent morphine consumption values were not necessary missing at subsequent time points. For the TKA patients, with LMEM, the complete-case analysis had a mean effect estimate of -0.85 (95% CI -7.1, 5.4; p=0.84). Comparing to the LMEM analysis with MI, the direction of the effect estimate from complete-case analysis was opposite, however, with such wide 95% CIs and the near zero effect estimates, it should not be said that the direction of effect was changed. The magnitude was similar, both being very close to 0. The complete-case analysis had a wider 95% CI and a larger p-value than the MI analysis.

Patients undergoing THA (n=102): There were 30 data points (7%) missing across the four time periods for the THA patients. The missing data did not exhibit a monotone pattern, where subsequent morphine consumption values were not necessary missing at subsequent time points. For the THA patients, complete-case analysis with LMEM generated similar results compared to MI. The direction, magnitude, and precision of the estimates were all similar to the MI analyses, and none of the comparisons attained statistical significance.

iii) Sensitivity to Covariance Matrix

Changing the covariance matrices in the LMEM analysis had no qualitative effect on the results generated. The results of the LMEM analysis remained robust and findings were consistent regardless of the covariance matrix used. Qualitatively, using the UNS covariance matrix provided with the highest precision via a tighter 95% CI and lower p-value, which was the case for both patient groups.

3.3 Discussion:

3.3.1 Key Findings

Clinical finding: In the present study, the comparison between morphine consumption in the gabapentin and control groups was made and the results suggested that there was no statistically significant difference in morphine consumptions following TKA or THA. Unlike the original analysis from the MOBILE trial [2, 3], this study analyzed the primary outcome of morphine consumption as a longitudinal outcome, where an additional time factor was incorporated in the statistical model. Nonetheless, the results were consistent with results of the MOBILE trial, where no statistically significant difference in morphine consumption was observed between the control and intervention groups. However, it should be noted that there have been a number of other studies suggesting the efficacy of gabapentin as an adjuvant for a multi-modal analgesia to provide post-operative pain management [24]. More trials should be conducted in order to evaluate gabapentin and create a coherent conclusion on the use, dosage, and timing of gabapentin for the management of post-operative pain.

Methodological findings: The current study addresses a number of different methodological issues related to longitudinal studies. The analysis of longitudinal data, which are repeated measures of the same subject over a period of time, has always been an important part of clinical research. Subject-specific and population-average methods exists for the analysis of longitudinal data and in the present study, three methods, 2 subject-specific methods and one population-average method, were used to analyze morphine consumption. Across the three methods, morphine consumption was not statistically different between the two treatment arms, across the two patient populations. Moreover, the results generated from RM-ANOVA and GEE had the same direction, and magnitude of effect compared to LMEM. Also, the precision of the results remained similar across methods.

The robustness of any statistical test can be used as a measure of how the results differ under changes in statistical assumptions, parameters, and other study factors. For instance, the robustness to missing data can be measured by changing the method of handling missing data (such as complete case analysis, multiple imputation, single imputation, etc.). If results remain the same, it can be said that the results are robust and consequently the conclusion drawn from the result is strengthened.

The proper handling of missing data is important as missing data can potentially affect the conclusion drawn from the analysis. The analysis of longitudinal data with a classical linear model restricts the analysis to only participants with complete data of all time points. When the missing data are not MCAR, the results from complete-case analysis may be biased because the complete case can be unrepresentative of the full population. An effective method of dealing with missing data is by conducting MI. All of these methods require the data to be missing at random (MAR). In our study, results from both method of handling missing data, by either MI or complete-case analysis, yielded similar conclusions that there was no difference between the two treatment groups. Moreover, this conclusion about the robustness of different methods of handling missing data was consistent between the two patient populations who underwent TKA and THA. Numerous studies have been conducted to compare parametric models using likelihood functions and semiparametric models using GEE, both with and without MI, in the context of incomplete longitudinal data [47, 48]. The results of this study were consistent with previous research, although not perfectly comparable since a qualitative approach was employed in the present study.

Our primary analysis used LMEM with AR(1), due to the hypothesis that morphine consumption will likely decrease on a daily basis and readings separated by a longer temporal period are less correlated with each other. Although AR(1) implies that observations on the same patient far apart in time would be essentially independent and this may not be truly realistic, with only four repeated measures in this study, we still thought AR(1) represented the most appropriate covariance structure for the model. The change in the covariance structure in the LMEM did not have much effect as the results remained robust across the other two covariance structures. Although the literature suggests the use of AR(1) since this covariance model provides a good fit compared to UN [23], the present study did not show any quantifiable differences. Nevertheless, the present results agree with previous studies, where the estimate of fixed effect, in this case difference between the two treatment groups, remains the same for different covariance structures.

3.3.2 Key Limitations

One of the key limitations of this study is the assumption of MAR. The mechanism of missingness plays an important role in determining the most appropriate statistical method and imputation method. GEE treats covariance structure as a nuisance and GEE is not concerned about variance of each data. However, GEE often performs poorly unless the mechanism of missing data was MCAR. Similarly, MI assumes that the mechanism of missingness is MAR.

However, no test was conducted in this study to determine the mechanism of missing data and these assumptions may not hold true. Testing for the type of missing data mechanism is often difficult, especially with a lack of auxiliary information, such as demographic, social characteristics of the participants. Missingness was assumed to be MAR in this study because of the low percentage of missing data (7% and 9% in THA and TKA patients, respectively) and the lack of monotone pattern in the missing data. Moreover, the robustness of the complete-case analysis further suggests that there was not a substantial amount of missingn of data.

3.4 Conclusions

The study compares three statistical methods of analyzing longitudinal data by applying the methods to an empirical dataset. Using morphine consumption taken at 4 different time points, we were able to strengthen the conclusion from the MOBILE trial that there was not a statistically significant difference between post-operative morphine consumption between the two treatment groups. The study did not suggest that on a qualitative level, using LMEM was superior to GEE or RM-ANOVA in terms of statistical power. Moreover, the results remained robust even when complete-case analysis was done and the misspecification of the covariance structure did not affect the results.

References

1. Concato J, Shah N, Horwitz: **Randomized, Controlled Trials, Observational Studies, and the Hierarchy of Research Designs**. N Engl J Med 2000, 342: 1887-1892.

2. Paul JE, Nantha-Aree M, Buckley N, Cheng J, Thabane L, Tidy A, DeBeer J, Winemaker M, Wismer D, Punthakee D, Avram V: **Gabapentin does not improve multimodal analgesia outcomes for total knee arthroplasty: a randomized controlled trial.** Can J Anaesth 2013, **60**(5):423-431.

3. James Paul - MOBILE: **MOBILE Trial Total Hip Replacement Surgery;** 2011 data on file. Manuscript is currently under preparation.

4. Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, Pitkin R, Rennie D, Schulz KF, Simel D: **Improving the quality of reporting of randomized controlled trials.** JOURNAL-AMERICAN MEDICAL ASSOCIATION SOUTH EAST ASIA 1996, **12**:33-35.

5. Moher D, Schulz K, Altman D: **The CONSORT statement: revised recommendations for improving the quality of reports of parallel group randomized trials.** BMC Medical Research Methodology 2001, **1**(1):2.

6. Faulkner C, Fidler F, Cumming G: **The value of RCT evidence depends on the quality of statistical analysis.** Behav Res Ther 2008, **46**(2):270-281.

7. Whiting-O'Keefe Q, Henke C, Simborg DW: **Choosing the correct unit of analysis in medical care experiments.** Med Care 1984, :1101-1114.

8. Tu YK, Blance A, Clerehugh V, Gilthorpe M: **Statistical power for analyses of changes in randomized controlled trials.** J Dent Res 2005, **84**(3):283-287.

9. Assmann SF, Pocock SJ, Enos LE, Kasten LE: **Subgroup analysis and other (mis) uses of baseline data in clinical trials.** The Lancet 2000, **355**(9209):1064-1069.

10. Greenland S: **Modeling and variable selection in epidemiologic analysis.** Am J Public Health 1989, **79**(3):340-349.

11. Etches RC, Warriner CB, Badner N, Buckley DN, Beattie WS, Chan V, Parsons D, Girard M: **Continuous intravenous administration of ketorolac reduces pain and morphine consumption after total hip or knee arthroplasty.** Anesthesia & Analgesia 1995, **81**(6):1175.

12. Canadian Joint Replacement Registry: Jun 16, 2009, **:**.

13. Kehlet H, Jensen TS, Woolf CJ: **Persistent postsurgical pain: risk factors and prevention.** The Lancet 2006, **367**(9522):1618-1625.

14. Buvanendran A, Kroin JS: **Multimodal analgesia for controlling acute postoperative pain.** Current opinion in Anesthesiology 2009, **22**(5):588.

15. Dahl JB, Kehlet H: **Preventive analgesia.** Current Opinion in Anesthesiology 2011, **24**(3):331.

16. White PF: **Multimodal analgesia: its role in preventing postoperative pain.** Curr Opin Investig Drugs 2008, **9**(1):76-82.

17. Buvanendran A, Kroin JS: **Useful adjuvants for postoperative pain management.** Best Practice & Research Clinical Anaesthesiology 2007, **21**(1):31-49.

18. Palmer PP, Miller RD: **Current and Developing Methods of Patient-Controlled Analgesia.** Anesthesiology Clinics 2010, **28**(4):587-599.

19. Walder B, Schafer M, Henzi I, Tramer M: **Efficacy and safety of patient‐controlled opioid analgesia for acute postoperative pain.** Acta Anaesthesiol Scand 2001, **45**(7):795-804.

20. HORLOCKER TT: **Pain management in total joint arthroplasty: a historical review.** Orthopedics 2010, **33**(9):14-19.

21. Woolf CJ: **Central sensitization: implications for the diagnosis and treatment of pain.** Pain 2011, **152**(3):S2-S15.

22. Gilron I, Bailey JM, Tu D, Holden RR, Weaver DF, Houlden RL: **Morphine, gabapentin, or their combination for neuropathic pain.** N Engl J Med 2005, **352**(13):1324-1334.

23. Littell RC, Pendergast J, Natarajan R: **Modelling covariance structure in the analysis of repeated measures data.** Stat Med 2000, **19**(13):1793-1819.

24. Seib RK, Paul JE: **Preoperative gabapentin for postoperative analgesia: a meta-analysis.** Can J Anaesth 2006, **53**(5):461-469.

25. Beach ML, Meier P: **Choosing covariates in the analysis of clinical trials.** Control Clin Trials 1989, **10**(4):161-175.

26. Bonate PL: *Analysis of pretest-posttest designs:* Chapman & Hall/CRC; 2000.

27. Dimitrov DM, Rumrill PD: **Pretest-posttest designs and measurement of change.** WORK-ANDOVER MEDICAL PUBLISHERS INCORPORATED THEN IOS PRESS- 2003, **20**(2):159-165.

28. Schafer WD: **Analysis of Pretest-Posttest Designs.** Measurement and Evaluation in Counseling and Development 1992, **25**(1):2-4.

29. Rossi PH, Freeman HE, Lipsey MW: *Evaluation: A systematic approach:* Sage Publications, Incorporated; 2003.

30. Campbell DT, Stanley JC, Gage NL: *Experimental and quasi-experimental designs for research:* Houghton Mifflin Boston; 1963.

31. Pocock SJ, Assmann SE, Enos LE, Kasten LE: **Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practiceand problems.** Stat Med 2002, **21**(19):2917-2930.

32. Raab GM, Day S, Sales J: **How to select covariates to include in the analysis of a clinical trial.** Control Clin Trials 2000, **21**(4):330-342.

33. Senn S: **Covariate imbalance and random allocation in clinical trials.** Stat Med 2006, **8**(4):467-475.

34. Wei L, Zhang J: **Analysis of Data with Imbalance in the Baseline Outcome Variable for Randomized Clinical Trials.** Drug Information Journal 2001, **35**(4):1201-1214.

35. Overall JE, Magee KN: **Directional baseline differences and type I error probabilities in randomized clinical trials.** J Biopharm Stat 1992, **2**(2):189-203.

36. Oakes JM, Feldman HA: **Statistical Power for Nonequivalent Pretest-Posttest Designs The Impact of Change-Score versus ANCOVA Models.** Eval Rev 2001, **25**(1):3-28.

37. Tu YK, Baelum V, Gilthorpe MS: **A structural equation modelling approach to the analysis of change.** Eur J Oral Sci 2008, **116**(4):291-296.

38. Winkens B, van Breukelen GJP, Schouten HJA, Berger MPF: **Randomized clinical trials with a pre-and a post-treatment measurement: Repeated measures versus ANCOVA models.** Contemporary clinical trials 2007, **28**(6):713-719.

39. Vickers AJ, Altman DG: **Statistics notes: Analysing controlled trials with baseline and follow up measurements.** BMJ 2001, **323**(7321):1123-1124.

40. Van Breukelen GJ: **ANCOVA versus change from baseline: more power in randomized studies, more bias in nonrandomized studies.** J Clin Epidemiol 2006, **59**(9):920-925.

41. Chan SF, Macaskill P, Irwig L, Walter SD: **Adjustment for baseline measurement error in randomized controlled trials induces bias.** Control Clin Trials 2004, **25**(4):408-416.

42. Schafer JL, Graham JW: **Missing data: our view of the state of the art.** Psychol Methods 2002, **7**(2):147-177.

43. Robinson LD, Jewell NP: **Some surprising results about covariate adjustment in logistic regression models.** International Statistical Review/Revue Internationale de Statistique 1991, :227-240.

44. Chu R, Thabane L, Ma J, Holbrook A, Pullenayegum E, Devereaux PJ: **Comparing methods to estimate treatment effects on a continuous outcome in multicentre randomized controlled trials: a simulation study.** BMC Med Res Methodol 2011, **11**:21-2288-11-21.

45. Cole JWL, Grizzle JE: **Applications of Multivariate Analysis of Variance to Repeated Measures Experiments.** Biometrics 1966, **22**(4):810-828.

46. Liang KE, Zeger SL: **Longitudinal data analysis using generalized linear models.** Biometrika 1986, **73**(1):13-22.

47. Beunckens C, Sotto C, Molenberghs G: **A simulation study comparing weighted estimating equations with multiple imputation based estimating equations for longitudinal binary data.** Computational Statistics & Data Analysis 2008, **52**(3):1533-1548.

48. DeSouza CM, Legedza ATR, Sankoh AJ: **An Overview of Practical Approaches for Handling Missing Data in Clinical Trials.** Journal of Biopharmaceutical Statistics 2009, **19**(6):1055-1073.

49. Tariot PN, Solomon PR, Morris JC, Kershaw P, Lilienfeld S, Ding C, the Galantamine USA-Study Group: **A 5-month, randomized, placebo-controlled trial of galantamine in AD.** Neurology 2000, **54**(12):2269-2276.

50. Vickers AJ: **Statistical reanalysis of four recent randomized trials of acupuncture for pain using analysis of covariance.** Clin J Pain 2004, **20**(5):319-323.

51. Cribbie R, Jamieson J: **Decreases in Posttest Variance and The Measurement of Change.** Methods of Psychological Research Online 2004, **9**(1):37-55.

52. Liu GF, Lu K, Mogg R, Mallick M, Mehrotra DV: **Should baseline be a covariate or dependent variable in analyses of change from baseline in clinical trials?** Stat Med 2009, **28**(20):2509-2530.

53. Wright DB: **Comparing groups in a before-after design: when t test and ANCOVA produce different results.** Br J Educ Psychol 2006, **76**(Pt 3):663-675.

Appendix:

Table 2.1 summarizes a highlight of published studies of descriptive, empirical and theoretical studies that looks at various baseline adjustment methods in studies with a baseline/post-treatment design.

Design	Studies	Methods Compared	Results/Findings
Descriptive	Assmann et al. 2000 [9]	Current baseline covariate adjustment methods in clinical-trial reports	In general, unadjusted method of analysis is used. However, for trials with baseline factors that are known to have strong relation to the outcome, ANCOVA is the recommended primary analysis since strong correlation between the baseline variable and the outcomes variable is expected.
Descriptive	Pocock et al. 2002 [31]	Covariate-adjusted analysis from the survey of 50 trial reports in four major journals	In the survey of trials in this study, only a few used the covariate adjusted analysis as the primary analysis. Moreover, substantial variation exists with regards to the number of covariates used in the analysis, ranging from zero to ten or more. In trials with strong correlation between the baseline and outcome variables, ANCOVA is the most appropriate choice analysis.
Empirical	Tariot et al. 2000 [49]	ANCOVA ANOVA for change score	The ANCOVA and ANOVA for changes from baseline measures analyses produced similar conclusions, and therefore the results based on the ANOVA model are reported here
Empirical	Tu et al. 2005 [8]	Post-treatment score Change score Percentage score ANCOVA	Due to the variability of the correlation between pre- and post-treatment, ANCOVA should be used in preference to change score or percentage change score, as it was the method that reduces Type II error rates
Empirical	Vickers et al. 2004 - [50]	Unadjusted (Post-treatment score) ANCOVA	For analysis of trials in the pain literature, typically there are no interaction between baseline score and treatment. Therefore, ANCOVA was concluded as the more appropriate method of analysis with higher statistical power compared to the unadjusted analyses.
Simulation	Breukelen et al. 2006	ANCOVA	Randomized trials and studies where treatment assignment is based on

	[40]	ANOVA for change score	a baseline variable, ANCOVA is the more appropriate method. On the other hand, for nonrandomized studies where there are more than one control groups and multiple baseline measurements, ANOVA of change scores seems less biased than ANCOVA.
Simulation	Cribbie et al. 2004 [51]	ANCOVA Change score with ANOVA	For studies conducted to detect predictors of change in two-wave design, the post-test variability has a major effect on the choice of the appropriate statistical method. ANCOVA is superior to change score with ANOVA when the variability decreases.
Simulation	Liu et al. 2009 [52]	Constrained Longitudinal Data Analysis ANCOVA	The study looks at two methods to determine the treatment difference with respect to mean change from baseline In this paper, we have considered the parameter of interest to be the mean change from baseline effect at a given time point such as the last visit time point *T* In general, under similar modeling conditions, the cLDA model is more efficient than the longitudinal ANCOVA model. The efficiency loss of the ANCOVA model is partially from treating the baseline values as fixed
Simulation	Oakes et al. 2001 [36]	ANCOVA Change score	In randomized studies, the ANCOVA method gives unbiased treatment estimates and typically has superior power to analysis with change score. On the other hand, I nonrandomized studies, where baseline differences between treatment groups exists, the change score model yield less biased estimates.
Simulation	Wright et al. 2006 [53]	ANCOVA T-test	Results from ANCOVA and t-test will not differ when appropriate measures have been taken to ensure random allocation. In situations where allocation is based on a baseline score, ANCOVA would yield an unbiased result and should be the method of choice.

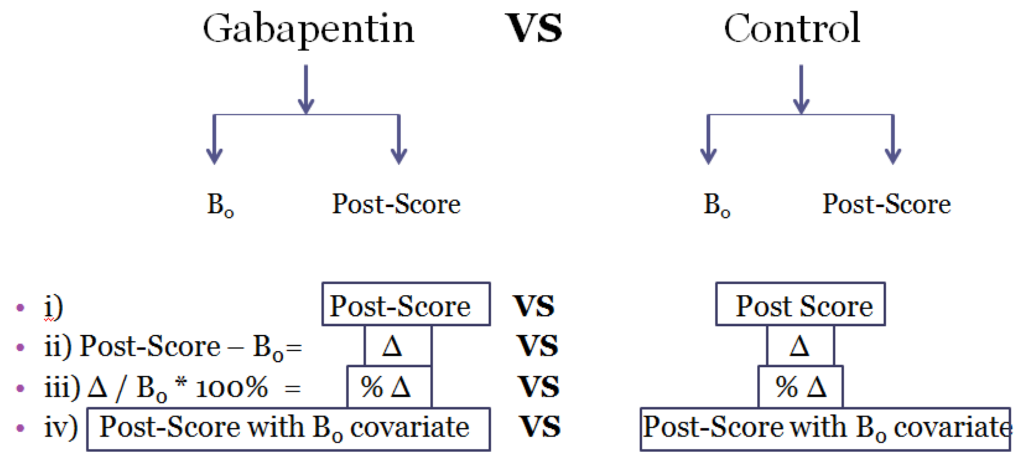

Figure 2.1 shows the schematic depiction of the four baseline adjustment methods. Post-score refers to post-treatment score, the outcome of the study. B_o refers to the baseline covariate used to adjust the score. Δ is the change score, calculated by subtracting the post-treatment score by the baseline score. The four methods depicted in Figure 2.1 are referred to, in this paper, as post, change, percent change, and analysis of covariate (ANCOVA).

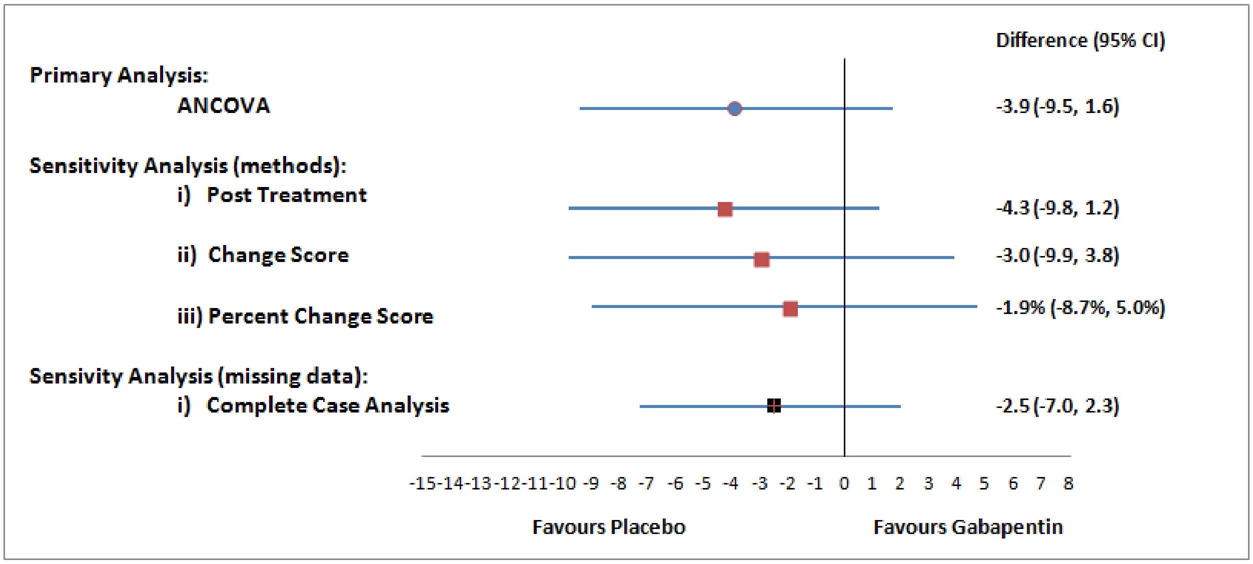

Figure 2.2 summarizes the results from the first part of the study. The difference between the treatment groups was not statistically significant for the knee flexion score. Furthermore, the results were robust across statistical methods and across methods of handling missing data. More specifically, the magnitude, direction, and precision of effect were qualitatively similar; although two of these methods (ANCOVA and post treatment, p=0.15 and 0.12, respectively) demonstrated trend towards lower scores in the treatment group (i.e. control group had better outcomes).

Table 2.2 summarises the results of the sensitivity analyses for the different baseline adjustment methods.

Multiple Imputation (m=5)

Statistical Method	**Mean Group Difference**	**95% CI**	**P-value**
ANCOVA with baseline as covariate	**-3.9**	**-9.5, 1.6**	**0.15**
Post treatment	**-4.3**	**-9.8, 1.2**	**0.12**
Change score	**-3.0**	**-9.9, 3.8**	**0.38**
Percent change score	**-0.019**	**-0.087, 0.050**	**0.58**

Table 2.3 summarises the sensitivity analysis of the method for handling missing data. Using complete case analysis, each of the four baseline adjustment methods were employed to provide treatment effect estimates.

Statistical Method	**Mean Group Difference**	**95% CI**	**P-value**
ANCOVA with baseline as covariate	**-2.5**	**-7.0, 2.3**	**0.27**
Post treatment	**-2.0**	**-6.6, 2.4**	**0.39**
Change score	**-1.7**	**-8.2, 3.3**	**0.60**
Percent change score	**-0.0052**	**-0.071, 0.034**	**0.88**

Table 3.1 summarizes the results of the longitudinal analysis of morphine consumption in the two studies [2, 3] using LMEM, with AR (1) and MI. The differences in treatment effect between the placebo group and the intervention group was not statistically significant, for both TKA patients and THA patients.

Trial/Patients	**Effect Estimate**	**95% CI**	**P-value**
Total Knee Arthroplasty (n=101)	**1.0**	**-4.7, 6.7**	**0.73**
Total Hip Arthroplasty (n=102)	**-1.0**	**-5.4, 3.3**	**0.63**

LMEM - linear mixed effects model; MI - multiple imputation, CI - Confidence interval

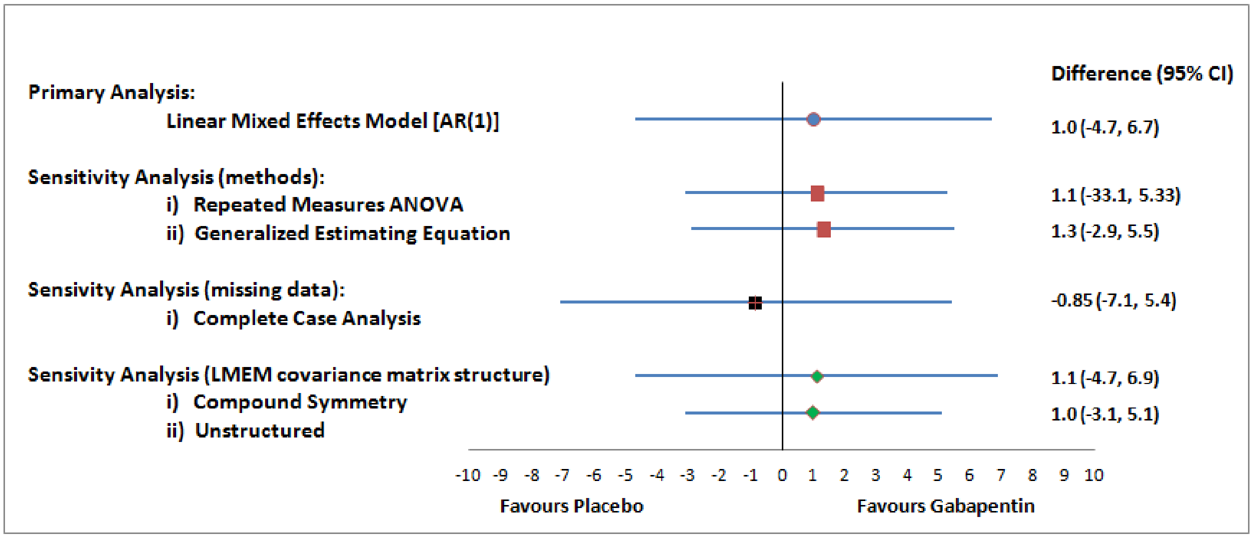

Figure 3.1 summarizes the results from the three sensitivity analyses conducted for the secondary objective of the study for patients undergoing total knee arthroplasty [2]. Specifically, three analyses were the robustness of longitudinal method (RM-ANOVA and GEE) with multiple imputation, the robustness to method of handling missing data (complete-case analysis), and the robustness to covariance matrix structure in LMEM with multiple imputation. The difference between post-operative morphine consumption was not statistically significant in patients undergoing total knee arthroplasty. Furthermore, the results were robust across statistical methods, methods of handling missing data, and LMEM covariance matrix structures. More specifically, the magnitude, direction, and precision of effect were qualitatively similar.

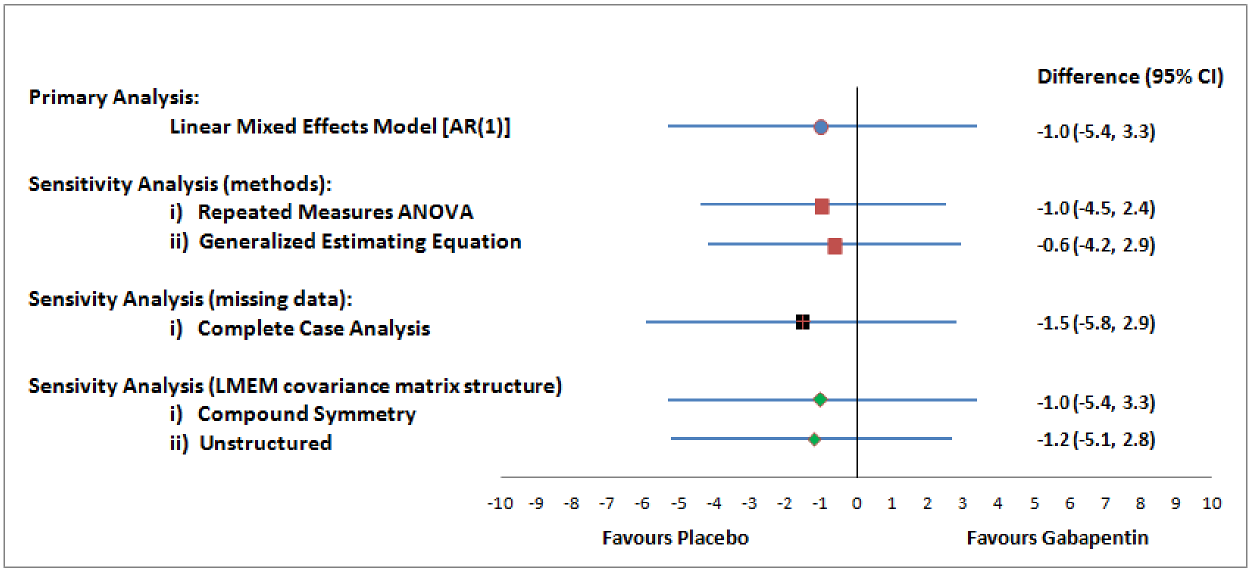

Figure 3.2 summarizes the results from the three sensitivity analyses conducted for the secondary objective of the study for patients undergoing total hip arthroplasty [3]. Specifically, three analyses were the robustness of longitudinal method (RM-ANOVA and GEE) with multiple imputation, the robustness to method of handling missing data (complete-case analysis), and the robustness to covariance matrix structure in LMEM with multiple imputation. The difference between post-operative morphine consumption was not statistically significant in patients undergoing total knee arthroplasty. Furthermore, the results were robust across statistical methods, methods of handling missing data, and LMEM covariance matrix structures. More specifically, the magnitude, direction, and precision of effect were qualitatively similar.