Statistical analysis of randomised controlled trials: a simulation and empirical study of methods of covariate adjustment

Bolaji Emmanuel Egbewale

A thesis submitted in fulfilment of the requirements

of the degree of Doctor of Philosophy

June 2012

Arthritis Research UK Primary Care Centre

Keele University

Table of Contents

LIST OF TABLES

LIST OF FIGURES

Abstract

Randomised controlled trials (RCTs) are widely accepted as the optimum design for comparing two or more medical therapies. Chance baseline imbalance (BI) through randomization opens the estimate of effect to bias. Statistical methods such as change score analysis (CSA) and analysis of covariance (ANCOVA) - are commonly used to deal with BI. However, unadjusted analysis by analysis of variance (ANOVA) is still common.

This study examined precision, power, efficiency and bias of estimates of effect associated with ANOVA, CSA and ANCOVA in RCTs with a single post-treatment assessment of a continuous outcome variable. A total of 210 hypothetical trial scenarios were evaluated; each was simulated (1000 iterations per scenario) using combinations of specific levels of treatment effect, covariate-outcome correlation, direction and level of BI. Evaluation was also performed on three empirical trial datasets, which showed different baseline-outcome correlations.

Precision and efficiency of CSA were not better than those of ANOVA unless baseline-outcome correlation exceeded 0.5. Depending on the level of baseline-outcome correlation, the sample size required at a given level of nominal power can be reduced by up to 80% using ANCOVA. Conditionally, both ANOVA and CSA are prone to false-negative or false-positive error. When BI exists in the same direction as treatment effect, the conditional power to detect the unbiased effect by ANCOVA falls below the nominal power in most trial scenarios. Also the review of current practices regarding covariate adjustment shows that whereas

over 60% adopt appropriate (modelling and stratified analysis) statistical adjustment, the most widely used single analytical approach is change. Overall, minimum covariate-outcome correlation of 0.3 is necessary but not sufficient to consider a covariate for inclusion in the model for adjustment. Appropriateness of CSA depends largely on baseline-outcome correlation and direction of imbalance. ANOVA is reasonable if the prognostic strength of the covariate is low. ANCOVA is the optimum statistical strategy regardless of BI.

Chapter 1: Introduction

1.1 Background information

Randomised clinical trials (RCTs), also referred to as randomised controlled trials, have been described as the gold standard and widely accepted as the best trial design for comparing two or more medical therapies or health care interventions (McLeod et al, 1996; Grimes & Schulz, 2002a). This claim, though, is correct only if the trial is appropriately designed, conducted and reported (Schulz et al 2010 – CONSORT statement). RCTs offer solutions to some of the issues that have been raised against observational studies. Kang et al (2008) argue that treatment differences identified from observational designs, rather than from experimental clinical trials are subject to methodological weaknesses, including confounding, cohort effects[1] and selection bias. The simplest and perhaps the most popular type of clinical trial is the two-group parallel design, in which the study participants on recruitment to the trial are randomised to either one of the two treatment groups (Altman, 1991; Overall & Doyle,1994; Tu et al, 2000; Cook & DeMets, 2008; Schulz et al 2010 – CONSORT statement).

A major principle of RCTs is the random allocation of participants to treatment groups. The essence of a randomization exercise is to bring about comparable treatment groups. The groups are expected to be similar in factors – known or unknown – that are related to the prognosis of the outcome or condition of

[1] Variations in characteristics of an illness over time among individuals who are defined by some shared temporal experience, such as year of birth (Maggio et al, 2001).

interest. These factors are referred to as prognostic factors. For the investigator to draw a valid inference on a treatment effect it is assumed that potential covariates are evenly distributed between the treatment groups. In other words, it is necessary that the treatment groups be balanced at baseline in terms of prognostic factors for the investigators to be able to correctly attribute change in outcome to the treatment intervention (Altman & Dore, 1990).

However, practical experience has suggested that such balance in covariates between groups is often not attained with randomisation (Tu et al, 2000; Altman & Dore, 1990). The resultant imbalance subtly opens the trial intervention to a degree of misrepresentation of estimates of its effect. Thus, the need for a correct and more reliable inference on the effect of interventions under trial has led to efforts to try to ensure that balance is achieved in the distribution of covariates across the treatment groups. It was observed by Kent et al (2009) that even when the treatment groups received the same treatment and in the absence of allocation bias, treatment outcome is not likely to be same for the groups when there is covariate imbalance. They argue that for this reason, unadjusted analysis that does not take this imbalance in groups' covariate distribution into account, may be both inefficient and yield an inaccurate estimate of the treatment effect.

The issue with covariate imbalance between treatment groups is partly design and partly statistical in nature. It is regrettable to note that several studies with great prospect at onset have failed to maximise their potential owing to issues related to improper design; especially in relation to imbalance in risk factors between treatment groups. In their article, Rosenberger and Sverdlov (2008) recall the termination at an early stage of a trial of the role of erythropoietin in maintaining normal haemoglobin concentrations in patients with metastatic cancer. This was intended to be a major study involving 139 clinical sites and 939 patients. In the words of the investigators, (Rosenberger & Sverdlov 2008):

> ...drawing definitive conclusions has been difficult because the study was not designed to prospectively collect data on many potential prognostic survival factors between treatment groups…. *The randomization design of the study may not have fully protected against imbalances* because the stratification was only done for one parameter, and was not at each participating centre… It is extremely unfortunate that problems in design… have complicated the interpretation of this study. Given the number of design issues uncovered in the post hoc analysis, the results cannot be considered conclusive.

It is clear that covariate distribution and balance between treatment groups warrants careful consideration by researchers prior to starting the trial.

However, it still appears that the statistical community is unclear on how to deal with covariates at the design stage, especially on the first line strategy for balancing important prognostic factors in the design of RCTs (Kernan & Makuch 2001; Scott et al, 2002; Hagino et al, 2004; Taves, 2004; Rosenberger & Sverdlov, 2008). The general consensus seems to be that, whichever method is employed at the design stage to attempt balance in covariate distribution, an adjusted statistical analysis that takes into account important covariate imbalance should take precedence over the unadjusted analysis (Altman & Doré 1990; Scott et al 2002; Hagino et al, 2004; Hernandez et al, 2004; Moore & Vanderlaan, 2007; Kent et al, 2009).

Various methods used at the design stage to attempt balance in prognostic factors between treatment groups include: blocking, stratification and minimisation. Also commonly used is the basic simple randomisation; the principle being that between-group inequalities are reduced through chance correction with increased sample numbers. However, each of these design methods has certain drawbacks. For example, with simple randomisation one may end up having unbalanced treatment groups, especially when the design is implemented on a study with a small sample size.[2] Similarly, stratification breaks down with small sample sizes per unit of stratification. Also, it is limited in the number of stratification factors that can be included. When there are large numbers of covariates each presenting with multiple levels the stratification procedure requires that separate allocation lists be prepared at each level of an identified covariate (Matthews, 2000). Inevitably, such multi-stratification may pose logistical problems, making the whole exercise almost impracticable. For example, in a trial with four prognostic factors at levels 2, 2, 3 and 4 respectively, 48 separate allocating lists have to be prepared and maintained for as long as the study lasts. Minimisation is not a pure randomisation procedure, as the next patient is allocated to a group based on an already determined attribute. Thus, the risk of selection bias, and subsequent applicability of crude statistical methods is in question. This explains the reason for reporting together both the

[2] For a two arm trial, the chance of pronounced imbalance becomes negligible with n>200 (Lachin, 1988)

unadjusted analysis and the adjusted analysis that takes into account the minimisation factors.

There is evidence indicating that adjusted analysis is more efficient than unadjusted analysis i.e. adjusted analysis gives an increase in statistical power (Hernandez et al, 2004; Wang & Hung, 2005; Moore & Vanderlan, 2007; Kent et al, 2009), improved type I errors (Hagino et al, 2004), provides increased precision of estimates of treatment effect (Tsiatis et al, 2007; Wang & Hung, 2005), and reduced bias, giving more accurate estimates of the true value (Altman & Doré; 1990). It is important to note that covariate imbalance is not important for consideration unless it is in some way related to the outcome variable. Unless covariates have an established relationship clinically or physiologically with the outcome variable, it is unnecessary to attempt to balance them across the treatment groups (Assmann et al, 2000). An investigator therefore need not be concerned when treatment groups are not balanced in factors or attributes that are not related to the outcome measure of interest.

1.2 Baseline imbalance

A particular class of covariate imbalance between treatment groups is that which involves baseline difference in an outcome variable of interest. In the primary care setting, randomised controlled trials often involve quantifying a numerical outcome variable at baseline and repeating the same after treatment. Measurement of treatment effect often depends on the observed changes from the baseline value within a designated period of time after treatments have been administered. The period of time allowed to monitor treatment effect of such therapies under investigation is referred to as the follow-up period. At such time, it is the change in the mean value of an outcome variable that is under investigation.

For example, two or more diets may be compared for the mean change in body weight they produce Sacks et al (2009); two or more treatments for hypertension may be compared for the mean changes in diastolic or systolic blood pressure which they produce; two or more cancer therapies may be with respect to the mean changes in tumour size they produce Wieder et al (2005) and finally, the effect of exercise and diet on obesity may also be compared in osteoarthritis patients in terms of mean change in body mass index. In all of these empirical examples, the difference in baseline score of the outcome variable between treatment groups has a direct influence on the treatment effect. The fact that one group has a higher mean score at baseline reflects an unfair advantage/disadvantage for that group in relation to the other. Accordingly, a primary concern of this thesis is with the implications of the directions and magnitudes of chance baseline differences for the analysis of randomised controlled trials.

1.3 Methods for statistical analysis of RCTs with baseline imbalance

The choice of statistical methods for handling baseline imbalance is dependent on a number of factors. In the first instance, there are four such methods that are commonly used in the case of a single post-treatment outcome assessment. They are: analysis of variance (ANOVA) for direct comparison of average post-treatment scores between groups; analysis of covariance (ANCOVA); change score analysis (CSA) – essentially an analysis of variance based on comparison of change values, and percentage change. The last of these, however, is believed to be inefficient and inappropriate for the purpose of baseline adjustment, as it presents with poor power and yields estimates of effect with high variance (Vickers, 2001; Overall and Magee, 1992). For this reason an analysis of percentage change will not be presented in this study.

Generally, the nature of the outcome variable plays a major role in determining the statistical method for covariate adjustment; for example, multiple linear regression for a quantitative outcome variable, logistic regression for a binary response, or Cox's proportional hazard models for time-to-event (e.g. survival) data (Assmann et al, 2000). Another method that has been used for covariate adjustment for a quantitative outcome variable is: hierarchical linear modelling (multi-level modelling), particularly when the post-treatment score is assessed at more than one follow up time point (Overall & Doyle, 1994). In view of the foregoing, it is important to restate that the evaluations in this thesis focus on methods that are used for statistical adjustment of baseline imbalance of a quantitative-numerical outcome variable for a single post treatment score (though similar implications may be drawn for non-numerical outcomes). It shall critically examine, under certain experimental conditions, the strengths and weaknesses of known statistical methods for adjustment for baseline imbalance and compare these with the unadjusted reference analysis.

1.4 Statistical methods of interest to the study

In this study, specifically, analysis of variance ANOVA will be the reference unadjusted analysis for the study while ANCOVA is the model-based adjustment and CSA is a basic baseline statistical adjustment. The results from these three statistical methods shall be compared at various pre-determined experimental conditions mimicking a randomised controlled trial using simulated data. In addition, a further aim of this research thesis is to evaluate these methods of statistical analysis using empirical data from a number of primary care trials of musculoskeletal disorders within the Arthritis Research UK Primary Care Centre. Information shall be drawn on such trial attributes as: statistical power, efficiency, precision and bias of estimate of effect. ANCOVA has been used widely for covariate adjustment, especially when its underlying assumptions are met. Basic assumptions that underlie the use of ANOVA and ANCOVA models shall be established in the datasets before further statistical analyses and comparisons are performed.

These three statistical methods have been variously used for a single post-treatment assessment of a continuous outcome variable in randomised controlled trial settings. Although differences in effect estimates are evident among the three methods (Christensen, 1985), their comparative effectiveness has not been completely explored under several combinations of levels of various experimental conditions. In this thesis the appropriateness of each method is explored across a range of different experimental conditions typical of clinical trial settings. This information is crucial to informing analysts on which statistical approach they should adopt for their statistical evaluation of the treatment effect in their clinical trial.

1.5 Rationale

Previous authors, though few in numbers (Wang & Hung, 2005; Tsiatis et al, 2008) have reported the precision benefits of adjusted analysis over the unadjusted. They also reported that the benefits of adjusted analysis in randomised controlled trials extend to reduction of bias, thus guaranteeing an estimate of treatment effect that is close to the true value. However, it remains unclear the extent of the benefit and loss of precision and bias across different potential experimental conditions. Since both adjusted analyses in the context of this study, CSA and ANCOVA, operate on a different underlining principle, it is necessary to assess the impact of such difference on the various attributes of treatment effect estimate under the same trial conditions (for varying levels and directions of baseline imbalance, prognostic indication and various levels of treatment effect. By this means, the comparative strengths and weaknesses of these statistical methods under the same trial scenario will be made clear.

There is no known study that has used the same datasets to assess the performance of these statistical methods (ANOVA, CSA and ANCOVA) in respect of all of the following attributes: precision of estimate, bias, statistical power and trial efficiency. Studies that have been carried out are characterized by having considerable mathematical notation and expressions that render the findings inaccessible to non-mathematicians (Twisk & Proper, 2005). This study attempts to present findings in a simple and clear way within a clinically meaningful context that makes it accessible to non-specialists without compromising the theoretical framework on which it is based.

It is not only the method of analysis used – unadjusted or adjusted – that can have a profound effect on the conclusion of a trial; the variable chosen for adjustment can also have marked influence (Altman, 1985; Beach and Meier, 1989). Thus, there is a need to determine under what circumstance or at what time an adjusted analysis should be preferred to an unadjusted or vice versa, which covariates are to be selected for adjustment, and what method of adjustment should be used, given certain experimental conditions. Contrary views have been reported on the implication of using CSA as a method for statistical adjustment in an RCT (Senn, 1989b; Altman & Dore, 1990; Senn, 1990) and this needs further investigation, for purposes of clarity and to facilitate informed decisions on when to use and when not to use particular methods. Trial situations in which model-based adjustment (ANCOVA) and basic adjustment by CSA will possibly yield different estimates need be well understood to provide guidance on future analysis of a randomised controlled trial with a continuous outcome variable.

Previous authors on this subject have not yet fully explored various levels of experimental conditions in clinical trial settings and the impact they have on precision, efficiency, statistical power and associated bias of estimates of effect of each of the statistical methods being studied. For example, the effect of direction and size of covariate imbalance, various levels of covariate-outcome correlation at different levels of anticipated treatment effect – small, medium and high (Cohen 1982) – need to be investigated. When imbalance at baseline is evident, the credibility of the crude estimate of effect becomes a matter of concern. At such time, researchers sometimes select those covariates with large imbalance for statistical adjustment. The practice whereby baseline scores between treatment groups are assessed using tests of significance has been variously criticised and condemned by different authors (Altman 1985; Schulz et al, 1994; Schulz 1995; Senn 1997). An attempt shall be made to evaluate the correctness of this practice in the context of this study and to provide information that may guide what, when and how to adjust in future clinical trials.

Previous studies have performed comparative evaluations of different statistical approaches (ANOVA, CSA and ANCOVA) under a range of experimental conditions. Vickers (2001) in studying statistical power of these methods, with a similar design, considers a single level of treatment effect, equal allocation, different levels of correlation, and balanced treatment groups; no mention was made of direction and size of imbalance. In his master's thesis, Sim (2003) considers only two levels of covariate-outcome correlation (0.1 and 0.5) and a single effect size, and did not consider statistical power and the relative efficiency of these methods. This PhD study proposes a more elaborate comparative study of the statistical methods involved and hope to use a different approach that ensures efficient data simulation procedure in several combinations of factors: levels of baseline-outcome correlation, levels and direction of baseline imbalance, levels of treatment effect and statistical power. This PhD study is based on different simulation syntax from that which Sim used in his master's thesis. The simulation program is similar however to Vickers' (2001), in that it creates a shift on an existing baseline variable to represent treatment effect. Beach and Meier (1989) had previously used an algebraic procedure to compare certain attributes of the unadjusted and adjusted analysis; they also recommended a computer simulation approach to investigate the issue.

1.6 Design methods for baseline imbalance

Measures to create balance at the design stage include: stratified-blocking, minimization or dynamic allocation (covariate-adaptive randomization). There has been controversy among statisticians on whether or not a covariate-adaptive procedure should be used and if so, how. Against this background, a relatively new procedure known as covariate-adjusted response-adaptive (CARA) randomisation for ensuring balance in the distribution of covariates between treatment groups at the design stage was recently introduced (Rosenberger & Sverdlov, 2008). It is worth noting, however, that another popular measure to deal with covariate imbalance in the interest of gaining a more reliable estimate of treatment effect across groups is subgroup analysis (Assmann et al, 2000; Hernandez et al, 2003); though some authors have also warned against mishandling and erroneous reporting of subgroup analyses (Altman, 1991;

Matthews, 2000). Subgroup analyses suffer from inevitable loss in statistical power (Pocock, 2002; Cook & DeMets, 2006) and are therefore practicably only useful in terms of gaining better insight into which group may benefit most from the treatment. This concept is handled in chapter 2 section 2.5.5.

1.7 Aims and objectives

The primary aim of the study is therefore to examine, through simulated and real-life data, the effect of combinations of different degrees of covariate-outcome correlations, different degrees of covariates imbalance, and different degrees of treatment effect, on:

a) Bias in estimates of treatment effect

b) Precision

c) Statistical power of the methods;

   For this simulation study, the actual (conditional) power of the statistical methods being studied is computed and defined for a particular scenario as the number of times the null hypothesis of no significant treatment effect was rejected in the simulations by each of the statistical methods multiplied by 100% (Vickers; 2001; Tu et al, 2005).

d) Relative efficiency - for any given trial scenario, a more efficient trial will require fewer patients to have a stated level of power (usually 80 or 90%) to detect an important difference between two treatments (Kernan et al 1999; Vickers, 2001; Kent et al, 2009).

Conclusions will also be drawn regarding the trade-offs between these properties (e.g. between the power of an analysis to detect an effect as statistically significant and its ability to provide an unbiased estimate of this effect). Results from ANCOVA in these scenarios will be compared with those from:

1) An unadjusted one-way ANOVA (as the reference unadjusted analysis)

2) CSA (as an alternative method of adjustment).

1.8 Research methods to be used

This study requires that several levels of experimental factors – such as the degree of baseline-outcome correlation, the levels of treatment effect, levels and direction of baseline imbalance – be variously combined. Each combination of levels of experimental conditions represents a hypothetical trial scenario. The only practical way these statistical methods (ANOVA, CSA and ANCOVA) can be studied at the same time in several trial scenarios in respect of the trial attributes mentioned earlier is through a program-driven computer simulation. Also the requirement to study several hypothetical trial scenarios and the need to study pre-specified levels of experimental conditions can only be guaranteed by a simulation procedure. For the purposes of this PhD, computer simulation of data will be carried out using STATA statistical software (version 10). The statistical program that shall be developed during the course of the study will be executable in the 'do file' facility of the STATA package.

Data will be simulated for a two-group RCT to represent a specific treatment effect in the outcome variable, and for small, medium and large correlations between the outcome variable and the covariate. Numerous scenarios representing different combinations of levels of parameters or experimental conditions (covariate difference, correlation with outcome, effect differences, sample size, nominal power, direction and size of baseline imbalance) will be investigated. These will be considered alongside parameter values that are observed in the empirical datasets (of the Centre's musculoskeletal trials) to assess the impact of the imbalances and correlations on the work in our health field. Simulations shall be repeated 1000 times for each scenario and the evaluation estimates (of the ANOVA, CSA and ANCOVA models) averaged across the multiple outputs generated. In terms of scope and extent of experimental conditions involved, previous studies have not been as extensive as this study, which will consider 210 hypothetical trial scenarios – involving seven levels of covariate-outcome imbalance in both directions of treatment effect, five levels of baseline-outcome correlation, three levels of effect, and two levels of power (80% & 90%). Thus a total of 210,000 simulated datasets and 3 empirical trial datasets will be studied, comparing the three statistical methods each time.

Through a series of comparative analyses, this project will examine different situations involving adjustment for a continuous covariate in the analysis of a two-group parallel RCT. The focus of the study will be on the effects of a covariate-adjusted analysis on important aspects of a statistical hypothesis test: precision, bias, power and statistical efficiency (sample size reduction).

1.9 Data simulation

Scientific data simulations are increasingly being employed to solve a wide range of problems. They have been widely adopted for studying concepts and issues that would have remained ordinarily difficult and almost impossible in a wide range of disciplines, for example; astrophysics, engineering, chemistry, environmental study (Abdulla et al, 2004). Data simulations have also been widely employed in medical sciences to facilitate informed decision. For example, researchers at Massachusetts Institute of Technology-Harvard MIT Division of Health Sciences and Technology were able, using mathematical models from simulated data, to study cardiovascular function, in particular orthostatic intolerance, among astronauts, and thus overcome the difficulty associated with interpretation of limited experimental data (Mark, 2007). Various medical researchers have also availed themselves of the opportunities data simulation offers to investigate patterns and distribution of disease for the purpose of informed preventive and curative recommendations. Mellor et al (2007) study of targeted strategies for tuberculosis in areas of high HIV prevalence was based on simulation. In another study, Hughes et al (2006) modelled tuberculosis in areas of high HIV prevalence based on simulated data.

Various authors have also used simulated datasets when interest was in addressing a statistical problem such as varying parameters to illustrate a statistical concept or deepen understanding of a statistical process; for example, Vickers (2001) used a simulation study to show how statistically inefficient the use of percentage change from baseline could be in an attempt to adjust for baseline. Rosenberger and Sverdlov (2008) used simulation to examine various ways by which covariates are handled in the design of clinical trials and came up with a new proposal called covariate-adjusted response adaptive (CARA) randomisation procedures. Hagino et al, (2004) used a simulation-based study to justify the use of minimisation over simple randomisation and stratified randomisation. Overall and Magee (1993) used a simulation procedure to study directional baseline differences and type I error probabilities in randomised controlled clinical trials.

Data simulation procedures usually make use of specialized commands or syntaxes. The simulation work in this study was carried out in STATA version 10 (see Appendix 1 for the linked STATA 'do file').

1.10 Empirical trial data

The study shall also investigate and compare the statistical approaches under real-life pragmatic conditions using a number of Centre datasets of randomised trials in primary care musculoskeletal disorders. The correctness of the usual practice of including a priori selected covariates (e.g. age and sex) in the model – without observable information on the levels of their prognostic importance – will be scrutinised. Information on what level of covariate-outcome correlations exist in musculoskeletal trial settings will be sought, along with typical treatment effect sizes based on nominated outcomes of interest in this setting. The effect on bias and precision of including more covariates other than the baseline of the outcome variable in the model-based adjustment will be explored with respect to these empirical clinical trial datasets.

Chapter 2: Literature review

2.1 Introduction

This chapter is devoted to review of previous works on various issues and concepts related to design and statistical methods in clinical investigations especially randomised controlled trials. Although evidence abounds that randomised controlled trials present the desirable qualities expected of a design for medical investigations, it is sometimes not possible to carry out an RCT. Earlier trials focus on in vivo and animal-based research before rolling out into tests of human efficacy on a small number of subjects (e.g. phase 1 and 2 trials). These are largely concerned with determining whether a new intervention is safe (as well as potentially producing beneficial effects) and are carried out using adaptive sequential allocation approaches. If an experiment is not possible then medical research studies are concerned with observational evaluations (which are briefly reviewed, below). However, since this study is primarily concerned with randomised controlled trials, focus shall mostly be on concepts and issues related to this design method.

2.2 Basic design in clinical research

Medical research is carried out with set of objective(s); this to a very large extent determines how such studies should be designed. An investigator could have an idea of what to do in some cases but more often there is a choice of reasonable ways of designing a study. Clinical research occurs in two main arenas: experimental research and observational research. This classification is based on whether the investigator assigns the exposures (treatments) or whether a clinical practice or a population characteristic was just observed (Grimes & Schulz; 2002b). Funai et al (2001) observed that observational studies are the more popular studies and dominate medical literature; perhaps because observational studies often guarantee a quicker result, are methodologically less demanding, and may also be relatively cheaper (most notably in the case of surveys and case-control designs). They are used to investigate factors or exposures that cannot be controlled. For example, in investigating possible association between passive smoking and lung cancer, individuals cannot be randomised into smoking or not smoking (Rothwell & Bhatia, 2007).

Von et al (2007) argue that observational studies are also essential for effective clinical practice; although, to make the most of the enormous potential of observational epidemiology to transform clinical practice and improve public health, studies must be designed and reported as rigorously as possible. This explains the reason for the development of recommendations by the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) project on what should be included in an accurate and complete report of an observational study. Observational study design is basically divided into two groups based on whether there is a comparison or control group. An observational study design in which there is no control or comparison group is called a descriptive study, and when control group is present it is called an analytical study (Grimes & Schulz; 2002b). While analytical studies include case control and cohort studies, descriptive observational study designs include case reports, case series and cross-sectional studies.

The other arm of clinical research is the experimental design. Here, the investigators deliberately change or manipulate one or more variables (intervention variables) in an organised manner in order to examine the effects of so doing on one or more other variables (outcome variables). Experimental studies can either be a randomised controlled trial or a non-randomised controlled trial, depending on whether or not there is a random element to the allocation of participants to treatment groups (Grimes & Schulz; 2002b). A non-randomised experimental design is also called a quasi-experimental design (Sim and Wright, 2000). Whereas quasi-experimental designs are discussed briefly in section 2.2.2, issues and concepts related to randomized controlled trials are treated afterwards following this section and in the remainder of this thesis.

2.2.1 Descriptive and analytic observational studies

Descriptive studies are the set of studies designed to elicit information on individuals' health profiles with the aim of formulating hypotheses. They are sometimes a report of an unusual occurrence of clinical importance, which may be a disease condition or side effect of a treatment; for this reason, they are popularly regarded as opportunistic and unplanned. Descriptive studies also involve the assessment of a sample at one time with the aim of making statements on the distributional pattern of the event of interest. This does not however include inferring a causal relationship between exposure and the disease condition or the health related event (Grimes & Schulz, 2002b); a control group is usually needed for such a purpose. This thus implies that when a study design does not incorporate a control group then it might be difficult to draw conclusion on the causal relationship between the exposure and the disease condition - no control no conclusion. Descriptive studies can be categorised into two main classes: those that deal with individuals and those that are concerned with populations. While case reports, case series report, cross- sectional studies, and surveillance relate with individuals, ecological deals with populations (Hennekens & Buring, 1987). Generally descriptive studies are used for planning; to generate information on trend analysis as obtained in surveillance, and to develop hypotheses about cause (Grimes & Schulz, 2002b).

2.2.1.1 Case report

A case report is a detailed report of an individual diagnosed with a specific condition**.** It is about the most elementary way of seeking and preparing information on the distribution of a health related event of interest. The simplest studies are descriptions of a single case (case report) or a number of cases (case series) that were encountered in clinical practice or routine disease surveillance (Callas, 2008). A classic example is the case series study of 5 homosexual males who developed a rare pneumonia. This case series study led to the eventual discovery of HIV (The National Emergency Medical Services to Children Data Analysis Resource Centre – 04/04/09). The major weakness of case report and case series as observed is that without any frame of comparison for the cases, the meaning of any observed association is unclear (Callas, 2008).

2.2.1.2 Cross-sectional study

This is also called a frequency survey or a prevalence study. Here researchers attempt to investigate both exposure and outcome at the same time in a predefined population. Usually a sample of the population is taken and studied for the population characteristics of interest. The researcher provides information simultaneously on the frequency both of exposure to factor(s) of interest and of the disease condition. However, such information is not essentially meant to indicate causality, as the temporal sequence is often impossible to work out, given that exposure and outcome are measured at the same time (Bamgboye, 2006).

2.2.1.3 Case-Control study

This study design starts with a group of people known to have the outcome of interest – usually disease condition – who are referred to as "cases". A comparison group of people without the outcome is then assembled – these are referred to as "controls". The two groups are usually matched on certain characteristics e.g. age, sex but not in terms of disease status (Sim & Wright, 2000). Finally, the exposure history of both groups is compared and the results analyzed for any association between the exposure and the outcome. Past exposure is determined through interviews, questionnaires, medical record reviews, laboratory tests for biomarkers, or similar methods (Callas, 2008).

Case-control studies are especially useful for outcomes that are rare or that take a long time to develop, such as cardiovascular disease and cancer. These studies often require less time, effort, and money than cohort studies (Grimes & Schulz, 2002b). A drawback with this study design is that it can be very challenging to create a valid control group. The controls should represent the population from which the cases arose with regard to past exposure history, but in practice they are often not entirely representative. Another disadvantage of the case-control study design is the possibility of information bias through incorrect recall or memory of exposure data which is imbalanced between case and control groups, leading to bias in the estimation of the exposure-outcome association, usually weighted towards a more positive association or an overestimated association (Grimes & Schulz, 2002b; Callas, 2008; Bamgboye,2006).

2.2.1.4 Cohort study

These are studies in which a group of individuals with an exposure of interest (exposed group) is identified and studied simultaneously with another group of individuals who do not have the exposure (unexposed group) over a period of time, so as to study the development of the outcome of interest (e.g. disease incidence). Since the development of the outcome variable usually entails looking forward; this design is also called a prospective cohort design. A retrospective cohort involves identifying a historical cohort (or past sample) for which, in its simplest form, two groups are defined – one has past exposure of interest, the other does not. The investigator looks considers the comparative development/incidence of disease or 'outcome' of interest between the two groups. It is important that the two groups are similar - differing only in the factor or exposure of interest (Bamgboye, 2006). If the exposed group shows a higher incidence of outcome than the unexposed, this is evidence of an association between exposure and disease. Here, there is a logical sequence from exposure to outcome. A cohort study enables calculation of true incidence rates, relative risks and attributable risks (Grimes & Schulz, 2002b). However, it can be very expensive and time-consuming to conduct, particularly for chronic diseases that may require many years to develop or rare conditions that require a large number of subjects to obtain enough cases of the outcome to be able to compare the exposed and unexposed groups (Callas, 2008). A cohort study design is appropriate when randomisation to exposure is not possible (Chan et al, 2007).

2.2.2 Quasi-experimental design

These are experimental designs in which there is no true random assignment to treatment groups (Sim & Wright, 2000). Here, investigators study treatment and control groups that are as similar as possible. However, the similarity cannot compare with that obtained between randomly selected group in which randomisation provides balance on both known and unknown factors. The analysis follows that of a cohort study after a pre-specified follow-up period has been completed. This method can also be used to investigate the effect of an intervention in community-based studies whereby two similar communities are identified; one is exposed to the intervention and the other not exposed. The treated and untreated groups are followed up for a period of time for the development of outcome of interest. An advantage of this design method is the ease with which the study is conducted and results obtained. Another advantage is the use of a concurrent control group and uniform ascertainment of outcomes for both groups; however, with no proper random allocation procedure there is usually difficulty in applying statistical methods correctly and in interpretation of the findings (Grimes & Schulz, 2002a)

2.3 Random allocation in controlled clinical trials

Randomisation is a major principle and the hallmark of randomised controlled trials. It is a procedure that allows for chance placement of trial participants in the treatment groups. The procedure in principle ensures that treatment groups are balanced with respect to baseline characteristics and thus provides the basis for post-treatment crude comparison of treatment effect in the groups. Ever since its first usage in agricultural experiments by Fisher in 1920s and its introduction to medical research by Hill in the 1940s, it has generated a lot of methodological interest and controversies among medical researchers – particularly statisticians (Grimes & Schulz, 2002). This perhaps has led to the methodological advances and developments of strategies by which clinical controlled trials are not only analysed but designed.

2.3.1 Random allocation and how not to do it

The randomised controlled trial has been widely described and accepted as the best trial design in the investigation of medical therapies (McLeod et al, 1996). The reason for this is that conscious efforts are made to reduce bias in estimates of effect, by nullifying potential allocation bias. Randomisation is about chance allocation of participants to treatment groups; thus, in addition to reducing allocation bias, it also permits a valid test of significance Hall (2007) since such tests are based on the assumption of random assignment to both treated and control groups. The successful implementation of this all-important procedure depends on the generation of an unpredictable random allocation sequence and concealment of that sequence until assignment occurs (Grimes and Schulz, 2002).

The benefits of randomisation can be greatly undermined if allocation sequence is not properly concealed and implemented. For this reason, researchers have suggested that the person who generates the sequence should not be the person who determines eligibility and enters patients into the trial; they also advocate the use of people not involved in the trial for treatment allocation – i.e. 'third party randomisation' (Altman and Schulz, 2001). Allocation concealment is thus a procedure by which measures are taken to conceal allocation sequence or study group assignment from those responsible for assessing patients for entry in the trial. Such measures include; central randomisation, sequentially numbered opaque sealed envelopes, numbered or coded bottles or containers, drugs prepared by the pharmacy, or other descriptions that contain elements convincing of concealment. Allocation concealment is not the same as blinding; as while it is possible to conceal the randomised sequence in all randomised controlled trials, blinding is only possible in some (e.g. in many trials of surgical procedures). Whereas concealment guards against confounding through inappropriate selection, related to the way patients are selected and allocated to treatment groups, blinding guards against confounding through inappropriate recording of information – assessment bias.

There have been some erroneous opinions and misrepresentations of the concept of randomisation, both in principle and practice (Liu et al, 2002), and although randomisation is supposed to be fundamental to the success of any controlled clinical trial, it remains perhaps the least understood element thereof. Many researchers often confuse non-random techniques such as haphazard and alternating assignment with random (Grimes and Schulz, 2002b). Chalmer (1999), while commenting on the allocation technique of the trial involving streptomycin in treating tuberculosis, which is generally accepted as the first randomised clinical trial believed that randomisation could not have been a better procedure in controlling selection bias than assignment based on strict alternation. He further reckons that the sole reason for the preference given to the former procedure was its tendency to conceal allocation sequence. It has been suspected that the reason for not reporting allocation techniques in some controlled clinical studies by researchers was because the methods used by such authors for such purpose could be short of a true random process.

In separate reviews of medical journals for the adequacy of randomisation procedures in trials, Hewitt and Torgerson (2006) and Grimes and Schulz (2002) respectively found that 79/232 (34%) and 129/206 (63%) authors did not specify the method used to generate an allocation sequence, despite the CONSORT statement Moher et al (2001) which stipulates that authors should make clear how randomisation was conducted. The result of their reviews also shows that non-random methods, such as using case record number, date of birth and date of presentation, are still being confused and presented as random methods by some researchers. However, it is worth noting that a considerable reduction in the proportion of medical researchers that did not comply with the CONSORT statement was observed between the two reviews – four years apart. This implies that awareness and compliance to the CONSORT regulatory statement grew between the times of the reviews.

Despite all the benefits randomisation confers on controlled clinical trials, its applicability in some studies has been challenged; Cotton (2000) argues that a well-conducted observational study could be better than a randomised controlled trial in the investigation of endoscopic therapy. The reason for his argument was that the randomised controlled trial is accepted as the gold standard in the design of medical studies with randomisation a fundamental activity; it becomes dangerous therefore to put together a randomised controlled trial with distorted randomization. A distorted randomisation in a trial can actually show a treatment effect that is biased through bias in allocation, such trials are more dangerous than observational studies as statistical considerations and overall interpretations usually take into account bias in non-experimental studies (Torgerson and Roberts, 1999).

2.3.2 Simple randomisation

This is a procedure that ensures that trial participants have an equal chance of being allocated to any treatment group. Different authors have observed that this elementary and basic allocation procedure surpasses all other sophisticated and complex allocation techniques in its unpredictability of sequence and control of bias (Lachin, 1988; Grimes and Schulz, 2002a). The techniques of sequence generation by simple randomisation in a controlled clinical trial setting can easily be facilitated, especially with a small trial. Such techniques include tossing a coin, throwing a die, card shuffling, and using a table of random numbers.

However, a major drawback of simple randomisation is that, treatment groups can by chance end up being dissimilar both in size and in composition as regards prognostic factors. Such dissimilarity may be very pronounced in small sample trials. The implication of this is that post-treatment crude comparison of effect between the groups may produce a biased estimate and hence a misleading trial result (Hewitt and Torgerson, 2006) – notably in small trials as randomisation ensures less balance in small trial than in large trial. This explains the reasons for using some sort of mechanisms both at the design and statistical analysis stages to attain balance in treatment groups prior to a final comparison of effect. Researchers have different views on the subject of handling chance imbalance by simple randomisation, especially at the design stage (Rosenberger & Sverdlov, 2008; Scott et al, 2002; Kernan & Makuch, 2001; Hagino et al, 2004).

In their article: “Is restricted randomisation necessary”, Hewitt and Torgerson, (2006) while commenting on the reason for stratification on important covariates, argue that simple randomisation is safe and there is no need for stratified randomisation since covariate imbalances can be adjusted statistically. In support of their claim, they refer to a previous study by Grizzle (1982) where the author noted that stratification followed by an adjusted analysis does not add much to the statistical power to detect treatment effect. The effect of the level of baseline-outcome correlation and other experimental factors on certain attributes of treatment effect following adjusted analysis is handled in chapters 4 and 5. Grimes and Schulz (2002a) observe that the inherent baseline imbalance with simple randomisation becomes negligible in large sample trials; here, n ≥ 200. These authors appear not to see any need for design effort at attaining balanced treatment groups especially if there is a plan to account for important covariates during statistical analysis or when the trial is large. Given this submission, one wonders what ‘important covariates’ are and in what context is the claim that the design effort such as: stratification or minimisation is not necessarily needful. Also another issue to consider from the above arguments is whether or not large sample trials should preclude the use of statistical methods that appropriately account of covariates. These issues are handled in the results chapters 4, 5 and 7, here, the pros and cons of statistical adjustment at different levels of imbalance and the influence of levels of prognostic variables on the attributes of treatment effect are treated using both real-life and hypothetical trial datasets.

2.3.3 Blocking

In view of the potential failure of a simple randomisation procedure to ensure balance in treatment groups, especially regarding the number of participants, the first design method that seeks to equalise group sizes is blocking. Blocked randomisation, also called random permuted blocks, belongs to the family of restricted random allocation procedures. It is the most frequently used method for achieving balanced randomisation at the design stage (Grimes and Schulz, 2000a). Stratified blocking not only strives for a balanced randomisation and thus, for an unbiased groups' comparison it also strives for comparison groups of about the same size throughout the trial. This attribute becomes helpful when investigators plan interim analyses; it makes meaningful treatment effect comparison possible between groups at such time when there are indications that the trial might be terminated before the final recruitment target. Blocking makes use of short sequences of assignment (blocks) randomly generated to ensure balance in the treatment groups. For example, with a block size of 4, there is an assurance that the group is balanced each time the 4th patient is enrolled. With a block of size 4, there are six ways in which treatments are allocated such that two subjects get A and two get B (Altman and Bland 1999):

1. AABB 2.ABAB 3.ABBA 4.BBAA 5.BABA 6.BAAB

A note of caution while using this method, however, is that the blocking arrangement used should not be revealed to the clinic personnel until it is appropriate to do so. However, because of the possibility of sequence prediction, especially in an unblinded trial and when the block size is small, the use of two or more block sizes (or random permuted blocks) has been advocated (Doig & Simpson, 2005).

2.3.4 Stratified random allocation technique

This is a procedure that divides the trial participants into strata according to important outcome-related prognostic factors. A separate randomization allocation schedule is used within each stratum by which participants are assigned to treatments groups. During this process it has been advised that the investigators use some form of restricted randomisation – such as blocking – to generate the allocation sequence in order to ensure balanced numbers per treatment group per stratum (Hewitt and Torgerson, 2006). Often stratification is done in combination with blocking (Altman and Bland, 1999; Grimes and Schulz, 2002a). This is to ensure equal distribution of covariates as well as equal allocation sizes between study groups.

As a note of caution, however, it has been argued that the usefulness of stratified blocking can be reduced by the use of too many allocation strata. Given that, as previously noted, simple randomization potentially fails to produce balance in baseline prognostic factors between treatment groups, especially when the sample size is small, stratified block random allocation is a design that attempts to overcome this limitation. Kernan et al (1999) argue that the estimate of treatment effect can be more precise and has more power following a stratified block random allocation procedure in a trial with small sample size than simple randomisation since stratification makes groups similar on the outcome variable at baseline. Both power and precision are inversely related to the variance of the estimate. These authors further identified benefits of stratified randomisation to include; increased efficiency, facilitation of subgroup analysis and protection against type I error. Kernan & Makuch (2001) explains that the greatest benefit of stratified block randomisation was to be observed for therapies with large treatment effect. This supposes that the trial sample will be relatively smaller compare with when the anticipated effect is small and thus, then, simple randomisation has a higher tendency to have treatment groups not balanced.

However, the potential complexity of stratified randomisation has limited its use and wide application in the design of randomised controlled experiments. It has been observed that stratification breaks down – too complex and almost impossible to manage when there are several important prognostic factors to account for at the design stage (Rosenberger & Sverdlov, 2008). In addition, it has been argued that, in large trials, stratified randomisation confers little or no benefit; given its potential complexity, some researchers have argued for the use of simple randomization for allocation of treatment (Lachin, 1988; Grimes and Schulz, 2002a). These authors claim that the gain from stratification becomes minimal when participants in each of the treatment groups are more than 50 in number, given a minimum sample size of 100. It should also be noted – and this can perhaps be regarded as a rule of thumb – that when stratified randomisation is used the stratification factor should be in the adjusted statistical analysis (Hagino et al, 2004). This is important as stratification or minimisation as the case may be somehow imposes a constraint on the simple randomisation procedure.

Another issue of much importance in stratified randomisation is the choice of an appropriate number of strata; the popular view on this is to make the number of strata few, so as to keep the trial manageable (Grimes and Schulz, 2002a; Altman and Bland, 1999). However, it has been observed by various authors that, in practice it is rarely possible to stratify for more than two factors, especially in small trials (Roberts and Torgerson, 1998; Altman and Bland, 1999). Hallstrom and Davis, (1988) suggest that the number of strata should be appreciably less than n/B, where n is the total sample size and B is the block size for trials that make use of blocking. Regarding appropriate number of strata in stratified randomization Kernan et al, (1999) suggested that number of strata should not be more than n/(Bx4), where n is the sample size at the first planned interim analysis and 4 is a safety factor that accounts for unequal distribution of patients among strata. They illustrate this by example of a trial of 1000 patients with random allocation of block size of 4 and with the first interim analysis planned after 500 are enrolled. Following their formula on calculating the optimum number of strata there would be a maximum of 31 strata (Kernan et al, 1999).

2.3.5 Minimisation

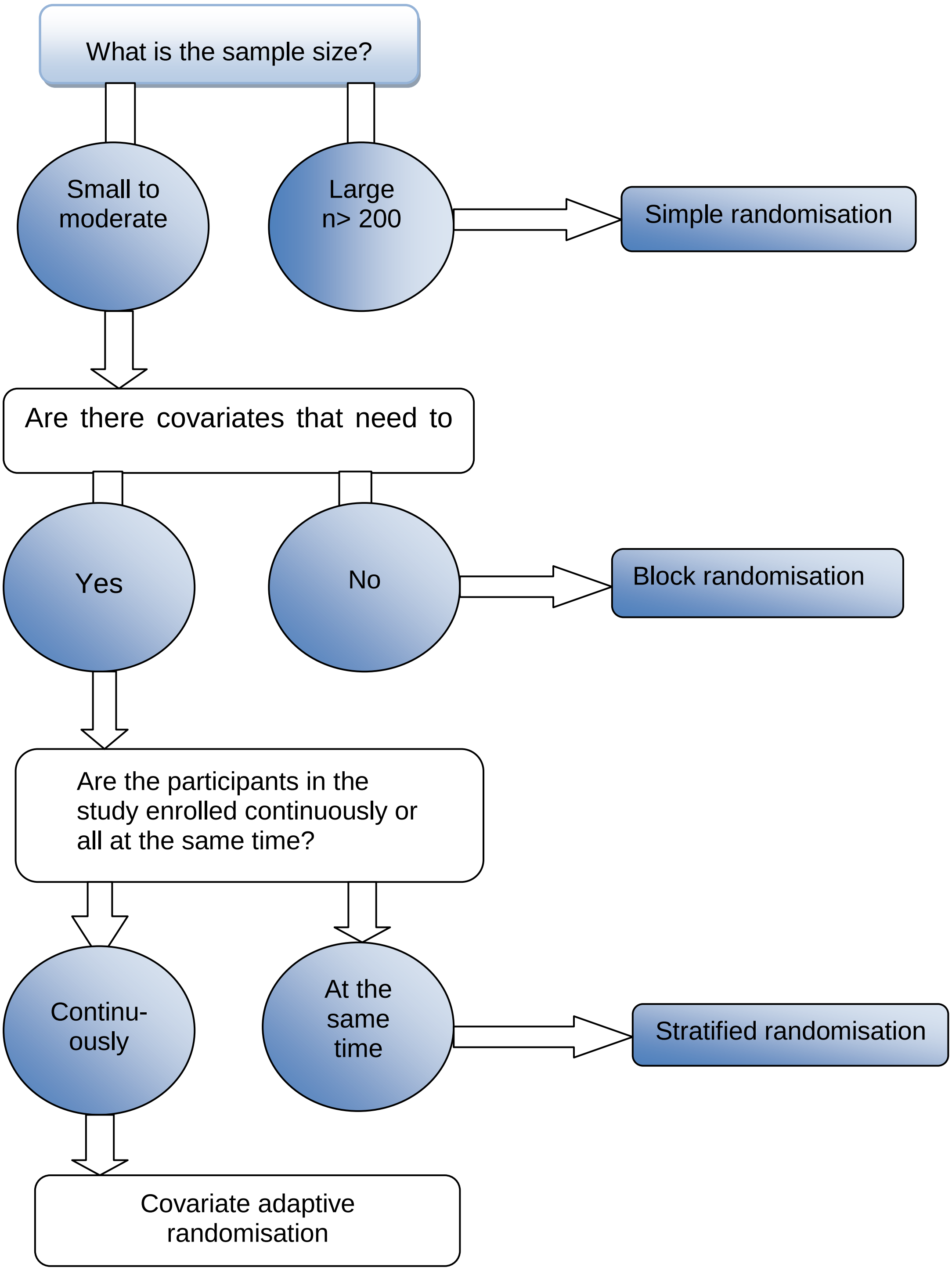
This is the first covariate-adaptive randomisation procedure to be developed and was first proposed in the 1970s by Taves (Cai et al, 2006). It is a covariate adaptive allocation procedure because a conscious effort is made to create balance in covariate distributions between groups. (Rosenberger & Sverdlov, 2008). Here, as was observed by Everitt & Pickles (1999), imbalances in the distribution of prognostic factors are minimised according to a certain criterion – thus the term minimisation. Minimisation has been described as an alternative treatment allocation procedure to stratified block randomisation (Roberts and Torgerson, 1998; Kernan et al, 1999).

A random allocation technique is used to assign treatment to the first patient and subsequently the decision is made for each patient on which allocation would result in a better balance in the groups, with respect to a particular prognostic factor (Altman & Bland, 2005). Such important prognostic factors are identified before the trial starts. Minimisation thus uses information about patients who are already in the trial to determine treatment assignment for the incoming participant, such that differences between groups are minimised (Kernan et al, 1999). Taves's minimisation method makes use of the information on the previous participant's group assignment to decide on the next assignment until all the enrolled participants have been completely assigned (Minsoo et al, 2008). Thus, the next patient is usually assigned to the treatment group with the lower covariate marginal total. This feature makes the allocation of the next subject to treatment group dependent on the balance of the groups is sometimes referred to as dynamic allocation (Cai et al, 2006). As was observed, minimisation though applies to any of Pocock and Simon methods, it is most commonly used to refer to the special case described by Taves, which is less complicated to employ in practice (Scott et al 2002).

Medical researchers have variously submitted that minimisation proffers a solution to the limitations of stratification in balancing for multiple prognostic factors in small trials, as the procedure makes treatment groups similar in several important features even with small samples (Roberts and Torgerson, 1998; Grimes and Schulz, 2002a; Altman and Bland, 1999; Minsoo et al, 2008). Scott et al, (2002) observed from the results of a simulation study that minimisation provides better balanced treatment groups when compared with unrestricted randomisation and that it can incorporate more prognostic factors than stratified randomisation methods such as permuted blocks within strata. They added, however, that adjustment should always be made for minimisation factors when analysing trials that used minimisation as an allocation method. This recommendation is also in order since minimisation is not entirely a random method and statistical tests or statistical inference are based on the assumption of random assignment to treatment and control groups, a property that is satisfied by only simple randomisation . The view that minimisation is not entirely a probabilistic procedure was also shared by (Grimes and Schulz, 2002; Rosenberger & Sverdlov, 2008).

However, it has been argued that minimisation is open to predictability of assignment (Hewitt and Torgerson, 2006; Minsoo et al, 2008) and researchers can therefore add a random element to the procedure at least to reduce prediction of assignment (Hewitt and Torgerson, 2006), perhaps by randomly allocating the first few participants. Another drawback with minimisation is the complex computation process involved; however, a user-friendly program that manages this has been developed (Minsoo et al, 2008).The flow chart below (Figure 2.1), adapted from Minsoo et al (2008), provides some guidance on selecting an appropriate randomisation technique.

Figure 2.1: A flow chart for selecting appropriate randomization techniques



2.4 Selected design issues in randomised controlled trials

2.4.1 Pragmatic and explanatory trials

According to Alford (2006), Schwartz and Lellouch (1967) were the earliest authors to publish on the differences between explanatory and pragmatic RCTs. Explanatory RCTs are intended to assess the underlying effect of a therapy carried out under optimal conditions, whereas pragmatic trials are intended to assess the effectiveness to be expected in normal medical practice (Cook & DeMets, 2008). In normal medical practice, patients are sometimes seen not to comply with treatment prescriptions one way or the other; they default, take some other treatments not prescribed, or do not take the treatment when it is due. Explanatory RCTs are usually associated with treatment efficacy in drug trials, and are limited in relation to generalisability of results. This is because of the tight inclusion criteria that are inherent with this randomised controlled trial method, which places artificial constraints upon participation that limit the applicability of the findings.

It was however noticed that while this is a particular concern for efficacy (explanatory) studies of drugs, it is likely to be less of a problem in quality improvement evaluations that are likely to be inherent with pragmatic trials that allows for what obtains in normal medical practice, for example switching of treatment by the patients. Also, efficacy studies assess differences in effect between two or more conditions under ideal, highly controlled conditions, while effectiveness studies assess differences in effect between two or more conditions when used in normal real-world clinical circumstances (Alford, 2006).

As was observed, a particular treatment approach might be shown to be efficacious, but may prove not to be clinically effective (Helms, 2002). It has been argued that since pragmatic studies aim to test whether an intervention is likely to be effective in routine practice by comparing the new procedure against the current regimen, they are as such the most useful trial design for developing policy (Eccles et al, 2003). While the explanatory approach recruits homogeneous populations and aims primarily to further scientific knowledge on, for example, underlying pharmacological effects, a pragmatic trial reflects variations between patients that occur in real clinical practice and aims to inform choices between treatments. Pragmatic trials are normally conducted on patients who represent the full spectrum of the population to which the treatment might apply. These patients may demonstrate variation in compliance, have a number of co-morbid conditions, and use other medication (Roland & Torgerson, 1998; Godwin et al, 2003). Another important point of difference between these two trials, as observed by Macpherson (2004), is the use of placebo in explanatory trials. Pragmatic trials would not compare placebo with an active treatment since a placebo is never administered in real life clinical practice; instead an existing treatment is compared with a new intervention.

Various authors have argued that in a pragmatic trial, it is neither necessary nor always desirable for all subjects to complete the trial in the group to which they were allocated; so as to have a good representation of the population to which treatment may apply. However, patients are always analysed in the group to which they were initially randomised even if they drop out of study. Application of intention-to-treat analysis (ITT) is considered to be synonymous to the pragmatic approach (Cook & DeMets, 2008; Roland & Torgerson, 1998). The concept of intention to treat analysis (ITT) is further treated in section 2.5.3. Pragmatic trials are well suited to situations where blinding is difficult or impossible (Helms, 2002). Roland & Torgerson, (1998) claim, somewhat controversially, that in pragmatic trials the biases of both clinicians and therapists can be accepted. This, they argue, reflects a normal clinical environment where the expectations of the patient and the therapist may influence the size of treatment effect. Even in this circumstance, Hotopf (2002) and Herbert et al (2005) have warned that concealment of randomisation is still important, as is blinding the assessor of outcomes so as to minimise the risk of selection, information or measurement bias by the researchers. In conclusion, with a pragmatic study, it has been observed that if an intervention is shown to have a beneficial effect, then it has been shown not only that it *can* work, but also that it *does* work in real life. (Godwin et al, 2003).

2.4.2 Blinding

Blinding is a procedure by which groups of individuals involved in a trial are made unaware of which treatment the participants are assigned. These groups of individuals may include some or all of: the participants, trial investigators or assessors, and data analysts. Some individuals and research organisations prefer the term masking to blinding to describe the same procedure. It has been argued that masking might be more appropriate in trials that involve participants who have impaired vision, and could be less confusing in trials in which blindness is an outcome (Schulz et al, 2002). In a trial, knowledge of treatment allocation can bring about subjective bias by both patient and the investigator. This can influence the reporting, evaluation and management of data and the statistical analysis of treatment effect can also be influenced (Pocock, 1983; Chow & Liu, 1998). Knowledge of treatment allocation can also affect compliance and retention of trial participants.

There are some instances in which it may be relatively difficult to achieve blinding. For example, if the new intervention under consideration is a surgical procedure and this is being compared with a chemotherapy delivered by tablet, here the difference between the two is clear and the trial needs be carried out unblinded as far as patients and caregivers are concerned. Such studies are known as open or unblinded. Open or unblinded studies have the advantages of being simple, relatively inexpensive and a true reflection of clinical practice. Single blind usually means that one of the three groups of individuals aforementioned remains unaware. In a double-blind trial, participants, investigators and assessors usually all remain unaware of the intervention assignments throughout the trial (Schulz et al, 2002). Here, three groups are kept ignorant thus, double blind is sometimes misrepresented. It should be noted that in medical research, the investigator frequently assesses, so in this instance there are actually two groups. Triple blind usually refers to double blind trials that also maintain a blind data analyst (Pocock, 1983).

As suggested in the CONSORT guidelines, it is no longer sufficient for investigator to use the terms single blind, double blind or triple blind; authors must show who was blinded and how and also provide information about the procedure on how it was carried out (Schulz et al, 1996).

2.4.3 Outcome measure/end-point

The choice of appropriate outcome measures, which can be primary or secondary, has been identified as one of the most challenging activities in the design of a randomised controlled trial (Wang & Bakhai, 2006). A primary endpoint is used to address the primary objective of the clinical trial, whereas a secondary endpoint addresses a secondary objective of a study. It is necessary that a clear definition of the two be stated. The choice of the most appropriate outcome measures has implications for the cost of the trial, the sample size, the burden that the trial will place on patients and clinicians taking part, and the likelihood that the result of the trial will influence clinical practice; therefore, whichever outcome is chosen it is important that it has been properly validated in a representative sample of patients for the disease under study (Rothwell, 2000).

Information about the effect of a treatment is often gathered in relation to many variables, thus, there is a temptation to analyse each of the variables and look to see which difference is significant between groups. It should be noted that such an approach leads to misleading results. Presenting only the most significant results as if they were the only analyses performed has been described as fraudulent (Altman, 1991). Thus it was suggested that the best practice would be to decide in advance of analysis on the main outcome variable of interest for particular trial; data could be analysed for other emerging variables but this should be considered as of secondary importance. Results of such secondary outcome variables should be interpreted cautiously and should be seen as ideas for further rather research than as definitive result. The major reason for this is that the study might not have been powered to detect difference in respect of such secondary variables. It is also important to note that even when the major or the primary outcome variables number more than one, the sample size calculation is usually based only on one variable (Altman, 1991).

2.4.4 Study population in RCT setting

Patient selection basically hinges on two opposing principles: homogeneity and heterogeneity of the study population. Both have pros and cons. The more homogeneous the population is the narrower the population on which the results apply (internal validity) and hence the smaller the number of patients needed to detect a given difference. On the other hand the greater the heterogeneity, the broader the basis for generalising findings at the end of the study – external validity (Curtis, 1986). In the spirit of a large and simple trial some authors recommend that eligibility criteria be kept to a minimum (Peto, et al, 1976; Peto et al, 1977). They are not to be too restrictive; otherwise, they undermine the external validity of the trial. However, some valid reasons exist for exclusion of certain participant, for example contraindication to intervention.

Eligibility criteria for a trial should be clear, specific, and applied before randomization. Trialists should endeavour to minimise exclusion after randomisation. For the primary analysis, all participants enrolled should be included and analysed as part of the original group to which they were assigned – intention-to-treat analysis. Mishandling of exclusions causes serious methodological difficulties and undermines trial validity (Schulz & Grimes, 2002a).The aim of the trial is to generalise its results to all patients who are like those randomised and treated in the trial. Without a strict set of eligibility criteria it is more or less impossible to describe which types of patients the results of the study can be applied.

2.4.5 Single and multicentre trials

'Centre' in a clinical trial sense refers to an autonomous unit that is involved in the collection, determination, classification, assessment or analysis of data or that provides logistical support for the trial (Meinert, 1986). For a trial to be multi-centre, it must consists of two or more centres and must involve a common treatment and data collection protocol, with each centre to receiving and processing study data (Meinert, 1986). Centres are treated as a stratifying variable in a multicentre trial and as such patients need to be randomised independently unless there is a central coordinated randomizing service (Altman, 1991).

A multi-centre study, unlike a single-centre study, allows a large number of patients to be recruited in a shorter time as recruitment can take place in each of the centres at the same time. The results are more generalisable since the scope of recruitment is generally wider than that obtained in a single centre trial and the participants are like going to be more diverse in their attributes. Multi-centre trial studies are critical in trials involving patients with rare presentations or diseases

(Wang & Bakhai, 2006; Hedman et al, 1987). Large trials are usually studied when investigating rare conditions. However, when the number of centres is too large, multicentre trials can pose administrative and logistic challenges. In contrast, a single centre trial demonstrates homogeneity of the study population since patients enrolled for the trial usually come from the same area (Meinert, 1986). It has been noted that the analysis of data collected in multicentre trials offers challenges because the data from the individual centres must be combined in some way to give an overall evaluation of the differences between the treatments in the trial (Fedorov & Jones, 2005). However, the practice of combining together all the data and ignoring the centres is not theoretically sound and should therefore be avoided. Ideally, the centre variable ought to be accounted for in the analysis as it is often treated as a stratifying variable and thus places some constraints on randomisation.

2.5 Selected statistical analysis issues in randomised controlled trials

2.5.1 Baseline comparability

This is a single concept that has generated much controversy among trialists, statisticians, and clinical investigators who have the responsibility of interpreting treatment effects (Altman, 1991).It has been observed by various authors that randomisation guarantees unbiased allocation of treatments to patients, but does not ensure for a particular trial that the patients in each treatment group have similar characteristics (Altman, 1991; Senn,1989). This then suggests that randomisation at best secures unbiased treatment allocation and not necessarily balance. This view was shared by other researchers; for example, Tu et al

(2000) noted that randomisation prevents biased allocation of subjects to treatment groups and provides foundation for statistical tests in practice, but that some important covariates may not be balanced at the end of the study, especially when the sample size is small. Such imbalance in baseline characteristics in treatment groups will either have a masking or exaggerated influence on treatment effect depending on the direction of the proposed intervention. Thus, it is usual practice in trials for researchers to compare treatment groups for their similarities in prognostic factors at baseline (Begg, 1990; Raab, et al; Altman, 1985).

However, the way in which such comparison is done differs between clinical researchers; while some adopt a practice of using tests of significance by using p-values to justify their choice of covariates to control or adjust in the mainstream analysis, others renounce such practice and regard it as unnecessary. It is advocated that researchers should present distributions of baseline characteristics by treatment group in a table as such information describes the hypothetical population from which their trial arose and allows readers to see the extent of similarities of the groups (Lachin, 1998; Burgess et al, 2003). Furthermore, this practice allows physicians to infer the results to particular patients (Pocock, 1982). Many authors disapprove of the use of hypothesis tests,e.g. p-values – in such tables as a means of comparing baseline characteristics across groups (Schulz et al, 1994; Schulz,1995; Senn, 1997; Matthews, 2001). They contest the practice whereby tests of significance are used to assess the magnitude of baseline imbalance as to whether or not to include the covariate concerned in the model for statistical adjustment. It has also been argued that there is no need for such tests as a proper randomisation procedure ensures that groups' difference is entirely due to chance and all of such tests seek to establish that the observed difference could or could not have been due to chance. They also argue that researchers who use hypothesis tests to compare baseline characteristics report fewer significant results than expected by chance – thus suspecting a foul play in reporting. The procedure of hypothesis tests on baseline characteristics has been described as not only clearly absurd but also as unnecessary and might also be harmful (Altman, 1985; Schulz et al, 1994; Senn, 1997). It has been stated that a significant imbalance will not matter if a factor does not predict outcome, whereas, a non-significant imbalance can benefit from covariate adjustment (Assmann et al, 2000).

2.5.2 Selection of covariates for adjustment

Irrespective of the method adopted at the design stage to bring about balance at baseline, the view shared by most clinical investigators (especially statisticians) is to further account for baseline imbalance by applying relevant statistical analysis methodologies. However, a major bone of contention is which covariates to include in the statistical adjustment model. A review of previous literature shows that there are basically three different views on this issue of covariate selection for statistical adjustment. Perhaps the most popular of these is the use of baseline tests of significance to determine which covariate to include in the model for adjustment. In this case, study groups are compared on a wide range of baseline variables and any that are significantly different between groups are automatically accounted for in the analysis (Meinert, 1986); those that are non-significant are ignored. Assmann et al (2000) observed that about 50% of clinical trial experiments published in four leading medical journals adopted this method. However, this idea has suffered major criticism over the years, and its use has been greatly discouraged among methodologists (Fayers & King, 2009).

The basic argument of those that disagreed with this method is that, since the patients were randomly allocated to treatment groups in the first instance, then it must be that any observed difference must have been due to chance. It then appears absurd testing whether the observed difference is purely by chance or not which is what the test of significance does. In addition, the case of baseline covariates that have prognostic influence but are not significantly different between groups is also an argument against the correctness of the use of hypothesis testing for covariate selection for adjustment (Schulz et al, 1994; Schulz, 1995; Senn, 1997; Matthews, 2000; Assmann et al, 2000; Senn, 1993).

The second view reflects the importance attached to the prognostic strength of covariates. However, there are two variants of this idea. The first bases the covariate prognostic importance on the level of correlation between the particular covariate and the outcome variable. The usual practice here is that, if there is a weak correlation, say $r \leq 0.1$, adjusting for the imbalance in such a covariate is not necessary even with a significant baseline difference in the covariate between the treatment groups. This seems also to support the idea that non-significance does not matter if the covariate-outcome correlation is strong. Just like significance testing, examining strength of correlation between the baseline and the outcome variables is a data-driven procedure, indicating that analysts should examine the correlation between the covariates and the outcome of interest before deciding on selection of such covariates for adjustment. A classic example by Christensen et al (1985) is a trial of primary biliary cirrhosis that had a non-significant imbalance in a strong prognostic variable, serum bilirubin: unadjusted and adjusted analyses yielded p=0.2 and p=0.02, respectively, for the treatment differences in survival. This example touches on the importance of recognising the prognostic strength of baseline variables rather than the statistical significance of imbalances.

The third known principle that guide covariate selection for the purpose of adjustment of baseline imbalance appears to be a variant of the second with selection being on the basis of covariates that have been found a priori to be prognostic in relation to the outcome variable. This includes evidence of suitable covariate-outcome correlation ($r \geq 0.3$) from previous research or pilot studies (Altman, 1985; Senn, 1994). The decision on which covariate is selected for adjustment is taken before the trial starts and usually specified in the protocol. This agrees with the recommendation of the International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for human use (ICH) guideline (Tu et al, 2000). The idea of using covariates identified a priori would also cover statistically adjusting for stratification or minimisation factors.

2.5.3 Covariate adjustment

In practice, simple randomisation may not ensure balance in some important covariate. If any unbalanced covariates are strongly correlated with the study outcomes, their presence may make it difficult to interpret the results of statistical tests for the treatment effect (Tu et al, 2000). Thus it is important that such imbalances are corrected or adjusted. Other studies have recorded a beneficial effect of a covariate adjustment over the unadjusted even for moderate correlation of covariates with outcome (Canner, 1991; Senn, 1989). The procedures for controlling the covariate imbalance can either be at the design stage or during statistical analysis; adjustment at the design stage includes the use of such techniques as minimisation and stratification. They are all employed at the design stage to prevent imbalance in prognostic factors in the first instance.

Stratified randomisation reduces groups' variability and thus enhances precision of the estimate; however, it has been argued that no amount of stratification can achieve balance for all covariates (Cook & DeMets, 2008). The procedure for adjustment during statistical analysis accounts for covariate imbalance at the analysis stage by using a statistical basis for the purpose. In the context of this study, methods for adjustment at the statistical analysis stage are: CSA that determines group effect base on the difference between the baseline and the post treatment score (basic adjustment) and ANCOVA, which is a model-based adjustment that includes the covariate in the model. Statistical adjustment can also be by pooling the stratified analyses, by using for example a Mantel

Haenszel test. In many clinical trials, both design methods that reduces covariates imbalance and statistical adjustment during analysis are used simultaneously. Raab et al, (2000) observed that for a given set of covariates, even though the stratification or minimisation methods will make the treatment groups comparable in these variables, they do not completely remove the effect of imbalance unless the stratification or minimisation factors are incorporated in the model for adjustment. Statistical adjustment can have a profound effect on effect estimates and tests of significance. For example, it has been observed that covariate-adjusted estimates are not only more precise, but the odds ratio or hazard ratio for logistic regression analyses and hazards models becomes further away from the null and that adjustment for strong predictors of outcome achieves more valid treatment effect estimates and significance tests (Assmann et al, 2000).

In addition, with respect to chance imbalance between treatment groups in a baseline covariate (especially when the baseline covariates is strongly correlated with the outcome) an adjusted estimate of the treatment effect accounts for this observed imbalance while an unadjusted analyses does not (Assmann et al , 2000; Pocock et al, 2002). A further benefit of the covariate adjusted analysis can be the creation of a predictive model which combines the influences of treatment and prognostic covariates in estimating the expected outcome for individual patients (Pocock et al, 2002). This allows for projection and informed decisions about the expected treatment outcome in relation to certain prognostic variables. It has been observed that the direction of imbalance is a factor that affects treatment outcome, for example, if the imbalance is such that the experimental group has a better prognosis than the control group, then adjusting for the imbalance is particularly important (The European Agency for the evaluation of medicinal product, 2003).

The validity of an unadjusted analysis relies on the assumption that there are no important imbalances involving measured and unmeasured baseline covariates across treatment groups. When imbalances occur on measured predictors of outcome variables, adjusted analyses should be performed (CPMP, 2003; Wang & Bakhai, 2006). It should be added that even if the groups have similar characteristics, it might still be desirable to adjust for another variable if we know in advance that the variable is strongly related to prognosis. Age is often such a variable (Altman, 1991)· The relevance of adjusting for age is further investigated in chapter 7. Although, a primary reason for adjustment for imbalance in one or more covariates is the removal of chance bias, adjusting for a prognostic variable may also lead to greater power for the trial (Altman, 1991).

2.5.4 Intention-to-treat (ITT) analysis

Intent-to-treat analysis is the strategy for the analysis of randomised controlled trials that compares patients in terms of the groups to which they were originally randomly assigned (Hollis & Campbell, 1999). This implies that patients are always analysed in the group to which they were initially randomised even if they drop out of the study (Hollis & Campbell, 1999; Roland & Torgerson, 1998; Wright and Sim, 2003). This principle is foundational to the experimental nature of randomised controlled trials as it ensures that the ideal structure for comparison created by random assignment of participant into treatment group is not distorted. There is wide agreement that the most appropriate analysis set for the primary effectiveness analyses of any confirmatory (phase III) clinical trial is the intent-to-treat analysis; it could be argued that an ITT analysis assesses the overall clinical effectiveness most relevant to the real-life use of the therapy (Cook & DeMets, 2006).

It is a recommendation of the CONSORT statement that authors should indicate whether analyses were performed on an intention to treat basis (Begg et al, 1996). The only safe way to deal with all forms of protocol violation is to apply intention-to-treat analysis; included here are patients who actually receive a treatment other than the one allocated, and patients who do not take their treatment (known as non-compliers). However, whether ITT principle also applies if it is discovered after the trial has begun that a patient was not after all eligible for the trial is opened to debate.

A different analysis strategy commonly used (as a secondary evaluation) is to exclude patients who have not adhered to the allocated management strategy for whatever reason. This form of analysis is called per protocol analysis, efficacy analysis, explanatory analysis, or analysis by treatment administered; this form of analysis only describes the outcomes of the participants who adhered to the research protocol. Montori & Guyatt, (2001) observed that per protocol analysis becomes a problem especially when the reasons for non-adherence to the protocol are related to prognosis. Empirical evidence suggests that participants who adhere tend to do better than those who do not adhere, even after adjustment for all known prognostic factors and irrespective of assignment to active treatment or placebo (Howitz et al, 1990). Thus, excluding non-compliers participants from the analysis leaves those who are destined to have a better outcome and destroys the unbiased comparison afforded by randomization. However, a relationship between a higher methodological quality of the trials and the reporting of the intention to treat has also been established (Miguel & Miguel, 2000).

2.5.5 Subgroup analysis

One of the reasons for collecting substantial baseline data from patients in a randomised controlled trial is that subgroup analyses (treatment outcome comparisons for patients subdivided by baseline strata) may be carried out (Assmann etal, 2000). This is to assess whether treatment differences in outcome or lack of it depends on certain characteristics of patients. The results from such group-specific assessment can be used to generate hypothesis for future study (Assmann et al, 2000; Wang & Bakhai, 2006). Subgroup analyses are important if there are potentially large differences between stratified groups in the risk of a poor outcome with or without treatment; if there is potential heterogeneity of treatment effect in relation to pathophysiology, if there are practical questions about when to treat, or if there are doubts about benefit in specific groups such as elderly people which are leading to potentially inappropriate over- or under-treatment (Rothwell, 2005).

Since patients recruited into a clinical trial are not a homogeneous sample, their response to treatment and the differing impact on them of different treatments may well vary in ways that affect the choice of which treatment is best for which patients. Pocock, (2002) argues that if in truth, there are specific subgroups of patients for which a new treatment is more or less effective or harmful than is indicated by the overall comparison with standard treatment in the trial as a whole, there is a scientific and ethical obligation to try and identify such subgroups.

However, most trials only have sufficient statistical power to detect the overall main effect difference in response between treatment groups, so that if subgroup effects do exist, they may well go undetected because the trial was not large enough (Pocock, 2002; Cook & DeMets, 2006).Smaller sample sizes within subgroups lead to greater standard errors and reduced power relative to the overall clinical trial resulting in an increased risk of a false-negative result, whereas, the multiplicity of hypotheses tests that results from examining multiple subgroups will lead to an increased risk of a false positive result-inflation of type I error (Altman, 1991). Altman (1991) further reckoned that to look for effects in subgroups is never a good way to rescue a study in which the primary ITT analysis fails to show an overall effect. The suggested approach to a sub-group analysis is to compare the difference between the treatments for the sub-groups of interest. The interaction can be examined within an appropriate multiple regression model, whether the outcome is continuous, binary or survival time (Altman, 1991).

The unadjusted strategy yields an average treatment effect without any consideration of heterogeneity in prognosis among patients. Although covariate adjustment and subgroup analyses both consider heterogeneity and attempt to provide more individualized estimates of treatment effect they are, however, substantially different (Hernandez et al, 2004). The difference is that, while covariate adjustment obtains a single more individualised treatment effect estimate, which is assumed to be applicable to all patients (Pocock et al, 2002; Robinson & Jewell, 1991), subgroup analyses provide multiple treatment effect estimates, assuming that treatment effects differ between particular groups of patients (Parker & Naylor, 2000). For the reason aforementioned, though subgroup analyses are sometimes performed, they rarely have enough power to detect differential treatment effect. It has however been variously observed that tests of interaction are underused and subgroup analyses are commonly over-interpreted (Altman 1991; Robinson & Jewell, 1991). Researchers should therefore be wary of this.

Chapter 3: Measuring bias, precision of estimate, power and efficiency in RCTs – the methods

3.1 Introduction

Even though there have been concerted efforts to improve the quality of reporting of RCTs (Hernandez et al 2004), dealing with covariates has remained an issue not only at the design stage of a controlled clinical trial but also during statistical evaluation of the treatment effect. The usual statistical approaches to the analysis of RCTs are either a crude comparison of post-treatment scores or a statistical evaluation of treatment effects in a way that chance imbalance in the baseline characteristics of the treatment groups is taken into consideration. This alternative approach is widely referred to as the adjusted analysis. If treatment comparison is to take into account the distribution of covariates between the treatment groups, or when heterogeneity in treatment groups as a result of chance imbalance of patients is of any importance, then, adjusted analysis is done. Thus, for a given RCT scenario, researchers are presented with various approaches to statistical analysis. While the practice of crude comparison of post treatment scores is still popular in some circles, some researchers will prefer statistical adjustment. Those that adjust traditionally fall into two groups: those that do so on the basis of covariates already pre-specified in the protocol, and those who opt for covariates that show large disparity between groups (Beach & Meier, 1989).

As mentioned in section 1.2, methods for statistical adjustment of baseline imbalance for a single post-treatment assessment of a continuous outcome variable are change score analysis (CSA), percentage change score and analysis of covariance (ANCOVA). However, the use of percentage change score for the evaluation of treatment effect in a clinical trial setting has been shown not to be statistically efficient (Vickers, 2001). Percentage change score analysis presents large error variance of the estimator and as a result has poor power to detect a difference in treatment effect when one exists. Van Breukelen (2006), when comparing analysis of change from baseline with ANCOVA, argued that only ANCOVA should be used if chance imbalance in treatment groups is to be taken into consideration since ANCOVA takes account of regression to mean whereas CSA does not. It has been argued that crude comparison of post treatment score by unadjusted analysis – ANOVA – will usually fail to detect a bias in an effect estimate since there is no term in the ANOVA model that takes account of the baseline difference in the treatment groups (Camilli & Shepard 1987).These considerations suggest that various methods used for the analysis of clinical trials can have a very profound effect on the estimate of treatment effect. In fact, under the same experimental conditions, ANOVA, CSA and ANCOVA have been observed to yield estimates of effect that are conspicuously different in size and precision (Van Breukelen 2006; Christensen et al, 1985, Piantodosi, 1997).

Furthermore, it is not only the method of analysis used – unadjusted or adjusted – that can have a profound effect on the conclusion of a trial; the variable chosen for adjustment can also have marked influence (Beach and Meier, 1989; Altman, 1985). Thus, there is a need to determine under what circumstance or at what time an adjusted analysis is preferred to an unadjusted, which covariates are to be selected for adjustment, and what method of adjustment should be used, given certain experimental conditions. However, the pros and cons of the adjusted analyses have not been fully understood as there remains a dearth of information in certain areas. Pocock et al (2002) asserted that the statistical properties of covariate adjustment are quite complex and often poorly understood. There is still need for systematic comparison of the precision, bias-reduction, and statistical power of these methods of analysis of RCT across various experimental conditions. Indeed, previous authors on this subject have not explored the full set of parameter changes (notably the simultaneous influences on precision, efficiency, statistical power and associated bias of estimates) across a diversified set of experimental conditions: notably, the effect of direction and size of covariate imbalance across various levels of covariate-outcome correlation at different levels of anticipated treatment effect – small, medium and high (Cohen, 1988).

When imbalance at baseline is evidenced, the credibility of the crude estimate of effect becomes a matter of concern. At such time, researchers usually select those covariates with large imbalance for statistical adjustment. The practice whereby baseline scores between treatment groups are assessed using tests of significance has been variously criticised by different authors (Schulz et al, 1994; Schulz 1995; Senn 1997; Altman 1985). An attempt shall be made to evaluate the correctness of this practice in the context of this study and to provide information that may guide what, when and how to adjust in future clinical trials.

Despite the dearth of information that exist in this area, in the few existing studies results and findings are often presented in ways that make them inaccessible to non-mathematicians owing to the level and number of mathematical notations, expressions and formulae used. This study, despite the fact that it shall attempt to simultaneously compare the statistical methods aforementioned under a wide range of levels of experimental conditions for pattern of bias, precision, power and efficiency, will also attempt to present the results in a way that will make them accessible to a broader readership without compromising the theoretical basis on which they rest. Subsequent sections in this chapter are devoted to the fundamental principles of the statistical methods of randomised controlled trials that are of concern to this study. An effort is also made to describe the methodology adopted for gathering the study datasets and analysis.

3.2 ANOVA – the reference unadjusted analysis

It might be asked why analysis of variance of post-treatment scores has been chosen in preference to the t test in this study as the reference unadjusted analysis, especially when the design focus is on two treatment arms. In this section, an attempt shall be made to justify this choice and also to outline the basic principles of this popular statistical test.

ANOVA and the t test are well known statistical methods in the assessment of groups average scores on a continuous outcome variable. Both are equivalent analyses (Porter and Raudenbush, 1987), except that ANOVA has more generality, as it can be used for two or more groups. Apart from this, it produces more useful output, such as mean square error (MSE). It is also closer to the generalised linear modelling GLM that includes ANCOVA, and which generates the regression coefficients that are to be used in this study. Since the idea of covariate or baseline imbalance in relation to appropriate statistical strategies is central to this study, conceptually it is more appropriate to think in terms of ANOVA than the t test, as the principle is based on explaining variability in scores between treatment groups. One of those factors in ANOVA that explains variability between groups is the 'individual differences' within each of the groups (Roberts and Russo, 1999).

In the analysis of a clinical trial, information on other sources of variation other than treatment effect may also be of particular interest to researchers, and so it is in this study. For example, in addition to providing information on estimates of different treatment effects by various methods of statistical analysis of RCTs, this study is also interested in capturing the associated 'noises' (variances). The study is interested in showing how such noise changes with levels of experimental conditions in each of the statistical methods that are of interest to the study. With ANOVA as the reference unadjusted analysis, the effect of statistical adjustment on the experimental error will be apparent for comparison purposes, especially in the context of precision and bias of these various statistical methods.

3.2.1 ANOVA – the principle

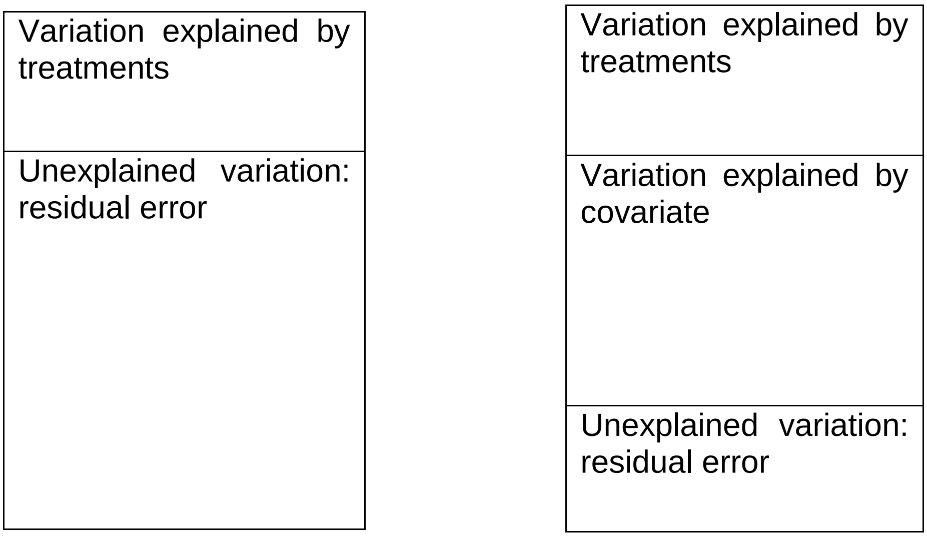
Since analysis of variance is the reference unadjusted analysis in this study, it is important to give a clear description of how it works – its principles. This will provide a basis for some of the arguments that follow in the section that compares the selected attributes of these various statistical methods of ANOVA, CSA and analysis of covariance (ANCOVA). The description that is given in this section follows Roberts & Russo (1999). As noted earlier, the basic task that ANOVA seeks to perform is to determine the sources of variation in an experiment; it explains why scores of any pair of individuals is likely to differ within a group of people that have similar exposure or treatment – what are the causes of within-group variability? A similar question, which is more fundamental to a clinical trial, is why is it that between two groups of people, the scores of any pair of individuals are also likely to differ? In other words, what are the causes of between-group variability? The answers to these questions will of course present in clear terms inherent limitations of ANOVA as the statistical method of analysis of RCT. All three methods to be examined operate on the principle of ANOVA, with only slight modifications in ANCOVA modelling.

Whereas, Porter & Reudenbush (1987) identified two potential sources of variability between treatment groups in ANOVA; variation due to treatment and that due to residual error, more recent authors Roberts & Russo (1999), have partitioned the overall variability in posttreatment outcome into three; variation due to treatment effect, individual characteristics and the residual or unexplained variation. Treatment effects – the effect the researcher is looking for and which is a result of the treatment or intervention of interest; under the alternative hypothesis (of a superiority trial) treatment groups are expected to be different since they are treated differently. For example, the outcome in a treated group is expected to be different from that of a control group, or people who receive treatment A are expected to have a different outcome score compared to people who receive treatment B.

Individual differences – despite the fact that subjects in the same group are treated in exactly the same way, their level of response to treatment can be and usually will be different. This might be due to difference in individual tendency to react to a stimulus, level of an underlying factor in participants that is related to the treatment in some way and also because they are simply different individuals. The basis for comparison of unadjusted post treatment scores is the assumption that groups are similar in the attributes of individuals that are related to outcome Matthews (2000) otherwise, resultant estimates of effect would be biased. Practically, randomization does not guarantee that treatment groups will be balanced at baseline. Authors have always noticed a chance imbalance in individual attributes despite randomization. This becomes important when such imbalance occur in individual characteristics that have prognostic relationship with the outcome variable. The effect of this chance imbalance on certain attributes of the estimate of effect is examined within the thesis.

Porter and Reudenbush (1987) and Roberts and Russo (1999) shared similar views, except that the former interpreted that the variation resulting from the individual differences within the groups and that of residual variation are inseparable in the context of ANOVA modelling. They argue that only ANCOVA provides further information on the between groups variability by identifying and using the variation due to the covariate or individual characteristics.

In the view of Porter and Raudenbush, the breaking down of variability is represented as follows;


Variation explained by treatments
Unexplained variation: residual error
Variation explained by treatments
Variation explained by covariate
Unexplained variation: residual error

Without ANCOVA **With ANCOVA**

The above model provides quick information on how the adjusted analysis by ANCOVA yields a more precise treatment effect estimate than that of ANOVA as a result of the reduction in the unexplained variance or the residual error. The effect of baseline-outcome correlation on the precision of estimates by these statistical methods is considered in chapter 4.

3.2.2 Within-group and between-group variability

It is a common occurrence to observe differences in individuals within the same group despite all the precautionary efforts made to ensure that members are treated the same way. Even in well-designed RCT settings, individual differences both within and between groups will occur. A well designed RCT will particularly ensure that no two members of the same group are treated differently; all members receive exactly the same treatment (within the latitude permitted by the trial protocol). The extent of the observed difference within the same group or sample is a measure of within–group variability or the residual error, Doncaster & Davey (2007). This phenomenon, essentially caused by differences in individual level characteristics and by random variation or chance explains why members of the same treatment group end up having different treatment scores.

On the other hand, treatment groups are also expected to differ as a result of difference in treatment they received; the extent of overall difference between treatment groups is also a measure of between-group variability. Obviously, this will comprise the difference due to treatment effect and that due to the chance or random variability in the groups – residual error. Expectedly, if the treatment under trial is effective, the within-group variability should be small compared to the between-group variability. Since difference by treatment is not expected among members in the same group, the only source of variability here, therefore, is the random error. Whereas within-group variability depends on the distribution of data in the particular group, the between-group variability depends on the group means, the higher the difference the higher the variability. The difference in overall means of the treatment groups is an indication of the between-group variability; if this difference is zero – that is, where there are two identical group means – the between group variability is zero. If the difference between means were to be increased, the between-group variability would also increase.

The adjusted statistical techniques attempt to account for the chance imbalance in individual characteristics and the random error in the measurement between the groups by some mechanisms, so that similar treatment groups are compared, allowing for a possibility of a more accurate estimate of treatment effect. So therefore, there is an expectation of a reduction in the residual error when adjusted analysis is done. The extent of this reduction and the accuracy of effect estimate under various levels of experimental conditions shall be investigated in chapter 4. Unless the baseline score or the covariate in question can together with the intervention completely explain the between group variability for a particular method of statistical adjustment, there will still be a minimised error term (residual error) after adjustment. Porter & Raudenbush (2007) observed that if the covariate has a strong correlation with the outcome, the residual variation will be small, and the statistical power will be substantially improved. Similarly, Rutherford (2000) observed that the unexplained error term, though minimal, will still persist despite adjustment.

3.2.3 Analysis of variance – the F test

When an analysis of variance test is conducted, the common statistic which gives a description of the groups' variability is the *F* ratio. The size of the observed F statistic is a reflection of the extent to which differences in the treatment groups diverge from that expected on the basis of chance (Rutherford, 2000). The practice is to compare the obtained F ratio values against a critical value from the F table at appropriate degree of freedoms; a significant F value means that random error can be ruled out (at a stated level of confidence) as the reason for the observed difference in a set of means (Roberts & Russo, 1999). Then, the observed difference in means is attributable to a treatment effect.

$$F = \frac{\text{between-group variance}}{\text{within-group variance}}$$

The F value can be influenced by individual differences and measurement error both brought about by chance as represented in the equation that follows. The value of F provides very useful information on the size of the treatment effect, the further away from 1 the F value, the larger the effect size.

The denominator of the equation is known as the error term. It is a measure of the extent to which residual error and not the treatment effect causes the scores to differ. It comprises both the individual differences within the groups and the measurement error.

Thus,

$$F = \frac{\text{treatment effects} + \text{residual error}}{\text{residual error}}$$

showing all sources of variability.

As mentioned earlier, since residual error has two sources of variability for a between-subjects design, the equation can also be expressed thus:

$$F = \frac{\text{treatments effects} + \text{individual differences} + \text{measurement error}}{\text{individual differences} + \text{measurement error}}$$

Thus, in order to conclude that a treatment has worked, the variance due to the treatment effects, i.e. the effect due to the intervention variable, must be sufficiently large to stand out from the ‘noise’ that constitutes the residual error.

Given the above equation, if it is imagined that a treatment has no effect on the outcome variable, then the amount of variability between groups caused by the treatment effect is zero. Hence, any variability in subject scores and overall means is due only to random error; and the equation will look like this:

$$F = \frac{0 + \text{residual error}}{\text{residual error}} = 1$$

Hence, if F is 1 or approaching 1, it is an indication that the treatment has little or no effect.

This does not however mean that each time F is large there is a significant treatment effect as with small numbers of subjects, one will obtain F values larger than 1 by chance more often than 1 time in 20. The usual practice is to calculate the probability of finding a particular value of F for a given number of degrees of freedom. The degrees of freedom take account of both the number of subjects and the levels of the intervention variable. It is also important to note that the F ratio is similar to a t test in that it is more likely to detect a given effect more statistically significant with a larger sample than a smaller sample; F tables must therefore be used in order to determine whether or not an effect is significant. The table gives critical values –- hurdles – for given degrees of freedom that must be equalled or exceeded in order to conclude that there is a significant effect. If however, the within-group variance is low – members in the same group have similar scores and there is variability in the overall level means – then the value of F will be large. The larger the value of F, the more the variability explained by the treatment effect, and thus the less likely that the observed difference in the overall means is caused by residual error, the more the treatment effects are standing out from the random error, and the less likely that differences in treatment effect were caused by chance.

3.3 Residual error, standard deviations and standard error

In an experimental design setting involving two or more groups, residual variation defined by the variation amongst sampling units within each sample or group is caused by difference from the individual characteristics and measurement errors. These errors will usually either mask or exaggerate treatment effects and hence, the aim of any experiment is to minimize them. Residual errors are known normally to disperse data away from the group means, thus increasing the variability from the mean score – i.e. increase the standard deviations. Similarly, the values of the group means may be pushed away from the mean of the population from which they are drawn – standard error is a measure of spread of sample means in relation to the population mean. Either of the standard deviation or standard error increases with residual error; the greater the standard deviation or standard error the less confident we can be that a difference between a pair of means has been accurately measured. Standard error is a measure of how precisely the population mean can be estimated using sample data, the smaller the value the greater the precision. Wang & Hung (2005) observe that a major advantage of the covariate adjustment analysis is the reduction of (residual error) variance in the estimation of treatment effects and the production of a more powerful statistical test for detecting the treatment effect.

Thus, in relation to the focus of this study, associated standard error is taken to be the measure of precision of the estimate of treatment effect. This statistic is a direct reflection of the mean square error; for example, when the residual post-test variance is minimised, standard error of the estimate of treatment effect is also minimised (Van Breukelen 2006). Standard error also determines the width of the confidence interval around the estimate as a correct representation of the population parameter. For example, the 95% CI of a sample mean ($\bar{x}$) is given as:

$$\bar{x} \pm (1.96 \times s / \sqrt{n})$$

Any statistical method that minimises the mean square error, and consequently standard error, will be adjudged as most precise. The width of the associated confidence interval measures the precision of an estimate; this ultimately depends on standard error for its computation. At a given confidence level, the smaller the standard error the narrower the width of the interval, the higher the precision of the estimate and the more confidence we have that the estimate is close to the true value.

3.3.1 Minimum variance unbiased estimator and bias

The deviation between the average value of an estimator and the population parameter to be estimated is termed bias or systematic error. It is a measure of how close, on average, that estimator comes to the true value. It is the difference between the true value and actual measurement. Millard & Neerchal (2001) express this as:

$$\text{Bias}(\hat{\theta}) = E(\hat{\theta}) - \theta \text{ ....(3.1)}$$

(where θ is the population parameter to be estimated and $\widehat{\theta}$ is the estimator)

From the above equation, if the average value of an estimator equals the population parameter, then the bias is equal to 0 and the estimator is said to be unbiased. An estimator can be negatively or positively biased. From the above equation, when the average value of an estimator is less than the value of the population parameter to be estimated a negatively biased estimator results; otherwise, the bias is positive. Thus, in an RCT setting, where the interest is to estimate treatment effect, it becomes imperative to seek information on how well the resultant estimate represents the true treatment effect by quantifying the associated systematic error or bias, which also contributes to the overall variability.

At certain experimental conditions – when there is baseline imbalance – the estimate of effect by both ANOVA and CSA cannot be unbiased. The percentage of the associated bias using either of these methods is quantified by using the formulae adapted from Austin, (2008);

$$\%\,\text{bias for ANOVA} =$$

$$100\,\text{X}\,\frac{\text{ABS(regres sion coef. ANCOVA) - ABS(regres sion coef. ANOVA)}}{\text{ABS(regres sion coef.ANOVA)}}\text{.....3.2)}$$

% bias for CSA=

$$100\,\text{X}\,\frac{\text{ABS(regres sion coef. ANCOVA) - ABS(regres sion coef. change)}}{\text{ABS(regres sion coef. change)}}\text{.... (3.3)}$$

(Where ABS represents the absolute value of the estimates)

3.3.2 Precision, mean square error (MSE) and standard error of estimate

An estimator is precise if it has a small variability and imprecise if the variability is large. Given that $\hat{\theta}$ is an estimator of some parameter θ: since there are various ways by which the parameter of a probability distribution can be estimated, it is important to know which method will give the best estimate. A commonly used approach is to quantify the associated mean square error (MSE) (Matthew, 2000; Millard & Neerchal, 2001; Porter & Reudenbush, 1987). Mean square error is a measure of variability that comprises the variance of the estimator (random error) and bias (systematic error):-

$$\text{MSE}(\theta) = \text{E}[(\hat{\theta}-\theta)^2] = \text{var}(\hat{\theta}) + [\text{Bias}(\hat{\theta})]^2 \text{………(3.4)}$$

= random error + the square of the systematic error

The MSE of an estimator is the average or mean of the squared distance between the estimator and parameter it is trying to estimate. It measures how much the estimator represents the parameter. As noted by Millard & Neerchal (2001), the estimator that has the smallest associated variance is the most efficient estimator. Thus, a statistical methodology that gives an estimate which minimises MSE as a result of reducing either the associated variance of the estimator (random error) or bias (systematic error) or both will have a higher precision of the estimate. As mentioned earlier, associated standard error of an estimate which is more relevant given the context of this study is a direct reflection of the mean square error.

The ratio of the standard error of each of the methods of adjusted analysis (ANCOVA and CSA) to that of the unadjusted provides an indication of the relative magnitude of the MSE for the adjusted analysis compared to the unadjusted analysis: this will be inversely related to the measure of precision.

3.4 Adjusting for baseline imbalance – the analogy

Not adjusting for baseline covariate imbalance in an RCT setting could be likened to two athletes who prepare to run a 100m race but start at different points on the track – this gives one an unfair advantage over the other, and the winner of the race may not be the truly faster runner over 100m; the level of unfairness (and implication on the result) ties in with the size of the difference in starting position. Clearly, the result or outcome would not be a precise or correct reflection of their true performance because the baseline difference is a factor that has a direct relationship with the outcome, – in this case time to finish. So therefore, it is only expected in the interests of a fair result, and correct measurement of performance, that the baseline difference at the starting point should be ‘accounted for’ in the system so that the measure of true performance (time to finish) of the two athletes would be a valid measure. Nobody bothers of course if the two athletes differ at baseline in some respects that do not affect the outcome; for example, colour of their outfit.

In a two parallel-arm RCT setting, the two athletes in the above scenario represent the two treatment groups (treated and control), while difference in starting point is analogous to baseline imbalance as it has an established relationship with the outcome variable, and difference in time-to-finish the race represents treatment effect. This analogy of course, may not completely represent what transpires in a trial setting, where the average scores of various responses (rather than individual responses) within each treatment group is compared. The mechanisms of adjustment by CSA and ANCOVA differ; the effect of this difference on the estimate of treatment effect shall be considered in chapter 4.

3.4.1 Issues with covariate-outcome joint distribution

If there are two random variables that are jointly normally distributed, for example; the baseline and the outcome scores are denoted here as Z and Y, respectively. As Z and Y are normally distributed, they have means and variances of;

$$\mu_Z, \sigma_Z^2 \text{ and } \mu_Y, \sigma_Y^2 \text{ respectively.}$$

Then,

$E(Y) = \mu_Y$ ....(3.5)

If Z and Y have a correlation ρ. then for every Z=z, there is a corresponding expected value of Y=y.

The distribution of Y conditional on this value of Z is not the same as the unconditional distribution of Y. This implies that since both Y and Z are related, the value of Y adjusting for Z is not the same as the value of Y unadjusted for Z.

Hence, the expectation of Y conditional on Z=z, written $\mathrm{E}(\mathrm{Y} \mid \mathrm{Z} = \mathrm{z})$, is not the same as E (Y), this is because Z and Y are not independent.

This fact explains why there is an expectation of difference between the estimated effect given Z, that is, Y conditional on Z, and the crude estimate of effect, that is Y unconditional on Z. The results of a series of simulations in this study explain this further in subsequent chapters. So, a statistical method that does not reflect a corresponding difference in Y as a result of the difference in Z is theoretically biased. If the covariate and outcome are prognostically related, the estimated value of outcome at a level of imbalance that was not taken into consideration is not expected to equal the observed simulated effect if the imbalance was taken into consideration.

However, consider another random variable M that is independent of Y, then, specifying a value for M does not affect the distribution of Y.

Thus,

$E(Y \mid M=m)=E(Y)$ since Y and M are not related……………………(3.6)

3.4.2 ANOVA of post treatment scores theoretically yields unbiased estimate of treatment effect

The simplest and perhaps the most common approach to estimating treatment effect between treatment groups is the crude comparison of post-test scores using statistical tests, such as t test or ANOVA for quantitative outcome variables.

The underlying model representation for a two-group trial is given as:

$Y_{ij}=\beta_o+\tau_i+\varepsilon_{ij}$...(3.7)

i =1,2; j = 1…n, or

$Y_{ij}=\mu+\tau_i+\varepsilon_{ij}$...(3.8)

where Y_{ij} is the posttreatment score for the $j\mathrm{th}$ patient in the $i\mathrm{th}$ group, β_O or μ is the common mean value of the outcome variable, τ_i is the treatment effect in the $i\mathrm{th}$ group and ε_{ij} is the error term. There is clearly no term in the model for ANOVA of posttest to accommodate any systematic variation in the groups that is related with the outcome as ANCOVA does and this explains the larger error term associated with the estimate from ANOVA.

Essentially, with respect to ANCOVA the model extends to:
$Y_{ij}=\beta_o+\beta_1 G_{ij}+\beta_2 Z_{ij}+\varepsilon_{ij}$..(9)

G_{ij} is a treatment indicator, β_1 is the group difference in Y adjusted for differences on Z.

When β_2 is close to 0, then it approximates the ANOVA model. It becomes obvious, therefore, that the difference between the statistical methods under investigation in this study actually lies in the different ways in which each of them responds to the presence of baseline imbalance. For example, as mentioned earlier, with ANOVA of post-test, $\beta_2 = 0$, for ANOVA of change $\beta_{2\,=}\,1$, and with ANCOVA β_2 is computed such that the residual post-test variance is minimized, thereby minimizing the standard error of the treatment effect estimate (Van Breukelen 2006).

The basis for the statistical procedure of ANOVA on posttest is that baseline scores of the outcome are comparable between the treatment arms by randomization. In other words, the statistical procedure assumes that baseline data for the groups to be compared are sufficiently similar and thus only the post treatment score is entered into the analysis.

In an RCT, let the baseline measurement from the control group be represented by random variable Z_C and the outcome variable by Y_C; the corresponding measurements for the intervention group are Z_T and Y_T for baseline and outcome respectively, following Mathews (2000).

Thus,

$E(Y_C) = \mu$ and $E(Y_T) = \mu + \tau$ ....(3.10)

and since by randomization the baselines have a common mean value μ_Z:

$E(Z_C) = E(Z_T) = \mu_z$ ....(3.11)

$E(\overline{Y}_T) - E(\overline{Y}_C) = \tau$ ....(3.12)

From the above, the sample mean of the outcome in the control group $(\bar{y}_C)$ will have expectation of μ and in the intervention group it will be $\mu+\tau$; hence, the difference in means will have expectation τ as required. This shows that the analysis based on post score is unbiased yielding an unbiased estimate of treatment effect.

This can be a simple way of demonstrating the correctness of the simulation procedure and the bias or otherwise of other methods of statistical analysis change score and ANCOVA, at least when the treatment groups are balanced at baseline. Both other methods are expected to yield the same unbiased estimate as ANOVA when the treatment groups are comparable in baseline scores. This shall be examined in the simulations.

However, when treatment groups are not comparable at baseline and such that there exist a correlation between baseline and outcome scores, then,

$E(Z_T) \neq E(Z_C)$ ....(3.13)

Direct comparison of outcomes from the groups becomes invalid and the resultant estimate is not unbiased.

Thus the true effect is modelled as,

$E(\bar{Y}_T - \bar{Y}_C \mid \bar{Z}_T, \bar{Z}_T) = \tau + \rho(\bar{Z}_T - \bar{Z}_C)$....................................(3.14)

However, since the ANOVA model does not have such a term that accounts for the baseline imbalance, its estimate of treatment effect will not respond to baseline-outcome correlation, direction and magnitude of baseline imbalance as in the last equation. The results of the simulation exercise highlight the non-responsiveness of ANOVA to various degrees of baseline imbalance and prognostic strength.

3.4.3 Change score analysis - CSA

If the analysis is based on ANOVA of change from baseline, there is a conscious effort to bring about balance in baseline data in the treatment groups by analysing the absolute difference between the baseline and the posttest score in the groups; however baseline scores are not included in the analysis as independent variables.

Here, analysis concerns (Y-Z)s in the two groups.

The underlying model is given as:

$$Y_{ij} = \beta_o + \beta_1 G_{ij} + Z_{ij} + \varepsilon_{ij} \text{..(3.15)}$$

Where Z_{ij} is the baseline value for the jth patient in the ith group. For change score analysis, the regression coefficient for the covariate is equal to 1.

Again, supposing treatment groups are comparable by randomization, the expectation will be,

$$E(\bar{Y}_T - \bar{Z}_T) - E(\bar{Y}_C - \bar{Z}_C) = (\mu + \tau - \mu_z) - (\mu - \mu_Z) = \tau \text{.....................(3.16)}$$

and this demonstrates that CSA yields unbiased estimate when treatment groups are comparable.

However, the associated variance differs from the variance of the unadjusted analysis. The variance of the unadjusted analysis is completely independent of the baseline outcome correlation, furthermore, if we assume randomisation makes groups similar, according to Matthews, (2000) it is convenient to assume in practice that:

$$\mathrm{var}(Y_T) = \mathrm{var}(Y_C) = \mathrm{var}(Z_T) = (Z_C) = \sigma^2 \text{.......... ....(3.16b)}$$

(where Y_T, Y_C are the outcome variables for both treated and control groups and Z_T and Z_C are the baseline variables for the both treated and control groups)

but the variance of the CSA is given as;

$$\mathrm{var}(Y - Z) = \mathrm{var}(Y) + \mathrm{var}(Z) - 2\mathrm{cov}(Y, Z)$$

$$= \sigma^2 + \sigma^2 - 2\rho\sigma^2 = 2\sigma^2(1-\rho)\text{.................................(3.17)}$$

where ρ is the correlation between Y and Z, assumed to be same for both groups.

The above presentation shows that the analysis of change scores from baseline has an entirely different variance structure compared with analysis from post-score comparison, and this has implications for the precision of the effect estimate. For example, if the correlation ρ exceeds 0.5 then a small variance (standard error) results and the analysis becomes more powerful than the comparison of post-test outcomes. However, if the correlation is below 0.5, using analysis of change from baseline (CSA) will bring about increased variance – a large standard error and less power to detect a real difference between groups. This fact was observed by Fleiss (1986), who argued that the estimate by analysis of change would not always have a lesser magnitude of associated variability compared with that from an unadjusted analysis – crude comparison of post-treatment scores. He states that precision will be lost by change score analysis if the baseline-outcome correlation is less than θ/2,

where,

$$\hat{\theta} = \frac{s_Z}{s_Y} \text{...(3.18)}$$

s_z, s_y being the standard deviation for pre and post treatment scores, respectively.

If, however, there is a correlation (ρ) between the baseline and outcome variables with random imbalance at baseline between the treatment groups despite randomization, then the expectation of the treatment effect by change score analysis cannot be unbiased.

Thus,

$$\mathrm{E}[(\overline{\mathrm{Y}}_{\mathrm{T}} - \overline{\mathrm{Z}}_{\mathrm{T}}) - (\overline{\mathrm{Y}}_{\mathrm{C}} - \overline{\mathrm{Z}}_{\mathrm{C}}) \mid \overline{\mathrm{Z}}_{\mathrm{T}}, \overline{\mathrm{Z}}_{\mathrm{C}}] = \tau + (\rho - 1)\overline{\mathrm{Z}}_{\mathrm{T}} - \overline{\mathrm{Z}}_{\mathrm{C}}) \text{..........(3.18b)}$$

This implies that the estímate of effect will be biased for any value of (ρ) that is not 1 and will also be related to both direction and magnitude of imbalance. The combined effect of baseline-outcome correlation, magnitude and size of imbalance on bias in treatment effect is shown in chapter 4.

Matthews (2000) observed that a common fallacy in practice involving baseline analysis is to conduct a separate comparison of the change in baseline score for the two groups and drawing conclusion on the treatment effect based on the two separate p values. For example, if the p values arising from comparing the pre-

and post-intervention scores for the treatment and the control groups respectively are 0.002 and 0.18, it would be erroneous to conclude that the treatment was effective. This is because randomised controlled trials require that comparisons be carried out between groups (usually including a control group and not within groups) for a valid inference on treatment effect.

3.4.4 Analysis of covariance – ANCOVA

Analysis of covariance (ANCOVA) is a statistical technique that makes use of the distribution of baseline scores and disparity in this between treatment groups to explain the overall treatment effect. ANCOVA conspicuously features baseline score as a covariate in its model equation and thus accounts for the imbalance during the analysis. Thus, because the model incorporates additional information, there is already an expectation of efficiency in the estimation of the effect. This extra or ancillary information accounts for the reduction in residual variance by ANCOVA (Mathews, 2000).

Similar to other authors on this subject, Van Breukelen (2006) presents ANCOVA models as;

$$Y_{ij} = \beta_o + \beta_1 G_{ij} + \beta_2 Z_{ij} + \varepsilon_{ij}$$

equivalently as,

$$Y_{ij} - \beta_2 Z_{ij}) = \beta_o + \beta_1 G_{ij} + \varepsilon_{ij} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.19)$$

This, though, presents the method as removing all the effect of the covariate from the outcome. However, Rutherford (2001) argues that outcome variables are not adjusted to completely remove the effect of the covariate but rather, adjustment is done such that all patients obtain a covariate score equal to the general covariate mean. In other words, ANCOVA uses the general covariate score to equalize the covariate distribution in the treatment groups. Thus, if a treatment group has a group mean at baseline that is greater than the grand or general covariate mean, the average treatment outcome for that group is adjusted downward. On the other hand, if a group has a mean score at baseline that is lower than the grand mean, then, the group average treatment outcome will be adjusted upward. The issue here is more of semantic (language) than concept. When ANCOVA equalizes the covariate distribution in the treatment groups by using the grand covariate mean, baseline imbalance is inevitably removed and thus offers a platform for a justifiable comparison of groups' treatment effect.

Thus, Rutherford (2001) expresses the ANCOVA model following adjustment as:

$$Y_{ij} = \beta_o + \beta_1 G_{ij} + \beta_2(Z_{ij} - \overline{Z}) + \varepsilon_{ij} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(3.20)$$

Equivalently as;

$$Y_{ij} - \beta_2(Z_{ij} - \overline{Z}) = \beta_o + \beta_1 G_{ij} + \varepsilon_{ij} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(3.21)$$

$\beta_{2,}$ represents the degree of linear relationship between the covariate and the outcome and is empirically determined from the data – again, in ordinary language, this represents the portion of the post treatment outcome that is explained by the baseline difference. This must be separated from the main effect otherwise it biases the estimate of effect. $\overline{Z}$ represents the grand covariate mean (average of all the baseline score).

Thus, the adjusted dependent variable score based on the difference between the recorded baseline score and the grand or general covariate mean, that is, the adjusted score for person j in treatment group i is given in algebraic notation as:

$${}_a y_{ij} = y_{ij} - \hat{\beta}(z_{ij} - \bar{z}) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.22)$$

In terms of the groups' adjusted effect,

$${}_a \bar{y}_i = \bar{y}_i - \hat{\beta}(\bar{z}_i - \bar{z}) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.23)$$

for ith group and

$${}_a \bar{y}_T = \bar{y}_T - \hat{\beta}(\bar{z}_T - \bar{z}) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.24)$$

for the treatment group

$${}_a \bar{y}_C = \bar{y}_C - \hat{\beta}(\bar{z}_C - \bar{z}) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.25)$$

for the control group

Thus for ANCOVA by taking the difference, this translates to having the adjusted estimate of effect τ as:

$$\tau = (\bar{y}_T - \bar{y}_C) - \hat{\beta}(\bar{z}_T - \bar{z}_C) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.26)$$

Only ANCOVA yields an unbiased estimate of effect (with respect to a covariate) when baseline imbalance in the prognostic baseline variable is accounted for.

This then suggests that the estimate of treatment effect by ANCOVA approximates that of ANOVA if the mean baseline score for the two groups is similar. Alternatively, both analyses are equal if $\rho = 0)$irrespective of the size and direction of imbalance. If, however, the baseline score for the control group is greater than the baseline score for the treated group in absolute value, then the overall treatment effect by ANCOVA is expected to be greater than that of the ANOVA (also in absolute value). Similarly, if the treatment group is higher in baseline absolute score than the control group, then the overall treatment effect by ANCOVA will be smaller in absolute score compared to that from ANOVA of post treatment scores. The results from the simulation exercise (in chapter 4) will demonstrate this.

As earlier mentioned from a review of previous studies, randomisation in a practical sense does not ensure that treatment groups are always balanced in prognostic factors. This creates a random variation that needs be accounted for in the interests of a valid estimate of treatment effect. ANOVA does not have a term that takes account of random variation in prognostic factors between the treatment groups. This explains why direct comparison of treatment outcomes between groups with ANOVA presents with a larger error term and, consequently, a less powerful analysis. The lack of such an extra term in ANOVA has also been observed to be responsible for the exaggeration or masking of treatment effect by this method, depending on the direction of imbalance. For example when the imbalance in prognostic factor is in the same direction as the treatment, ANOVA fails to detect such imbalance irrespective of its size and the prognostic importance of the factor. At such time, Camilli & Sheperd (1987) argue that the imbalance would contribute to the group treatment effect rather than being detected as bias by this method. This implies that some of the difference in means after treatment will be as a result of the random difference in means before treatment – the group difference in the characteristics of the patients that is responsible for extra gain or loss in treatment effect depending upon the direction of imbalance.

There is directionality in how treatment outcome is affected by covariate imbalance between the treatment groups. For example, if the treated group has a lower mean value at baseline and reduction in baseline score implies that treatment is effective, then the unadjusted treatment effect will fail to identify the possible exaggeration on the overall treatment effect. The overall treatment effect will not reflect the undue advantage of a better prognosis that the treated group had at baseline (That is, when the effect is in the same direction as the baseline imbalance). Conversely, if a lower mean value is recorded at baseline for the control group, then the masking effect on the overall treatment effect will still not be identified and the unadjusted analysis will yield an overall under-estimate of effect (analysis being carried out as if baseline prognosis of both treatment groups is the same). Thus, whether imbalance is in the same direction as treatment or opposite the crude unadjusted analysis will give the same (biased) estimate of effect. With respect to the direction of baseline imbalance, change score analysis will yield an exaggerated treatment effect when baseline imbalance is in the opposite direction of the treatment, that is, the control group has a better prognostic status (lower baseline score) than the treated group. The overall treatment effect however, will be masked by using change score analysis if the imbalance is in the same direction as treatment.

This situation may be overcome by ANCOVA accounting for the imbalance at baseline, thus reducing the systematic variation in the interests of a less biased and more precise estimate of treatment effect. ANCOVA does not crudely compare the treatment groups' outcomes, but first adjusts the outcomes in relation to the covariate level in the groups. Thus, the procedure of covariate adjustment by ANCOVA, as explained by Rutherford (2000), usually involves two stages: 1) ANCOVA determines the co-variation between the covariate(s) and the outcome variable, that is, the influence that the group imbalance has on the treatment outcome for that group, and 2) it removes that variance associated with the covariates from the outcome variable scores (adjusts in a way that the covariate mean value is made equal between the groups). These two stages occur prior to determining whether there is difference in outcome. So, essentially ANCOVA compares two adjusted outcome values. Wang and Hung (2005) observe that the precision of the adjusted estimate of treatment effect increases as a function of the correlation between the response variable and the covariate. This implies that as correlation between the covariate and outcome variable increases, the precision of the estimate by ANCOVA also increases; this proposition will be investigated further in the simulation results (see chapters 4 and 5).

3.5 Regression to mean

The mechanism of adjustment by ANCOVA is quite different from that of change score analysis; this explains the reason for the differences in their estimate of effect and in the standard errors that they yield.

For ANCOVA, the baseline score for an individual patient, irrespective of the treatment arm, is adjusted by the overall mean of the baseline scores. The implication of this is that the treatment arm that has a lower mean score at baseline by chance will by this mechanism of adjustment have its new adjusted mean jacked upward, and the other treatment arm with the higher mean score at baseline will have its mean score adjusted downward, compared to the original observed value. This mechanism of adjustment by ANCOVA is based on the phenomenon that has been described as regression to the mean (Fleiss 1986; Van Breukelen2006). Assuming treatment leads to reduction in baseline score as being the case in some empirical trial settings, those participants with low baseline score on the outcome variable tend to show less change than the average patient, similarly those with high baseline scores tend to show more changes than the average patients. This occurs when patients are measured as being severe or extreme on a variable that is subject to random fluctuation over time. The patients that are measured as being extreme, for example, at baseline on such variables are generally going to end up as having highest improvement over time with or without the treatment (Fleiss, 1986; Van Breukelen, 2006).Of the three methods under consideration in this study, only ANCOVA takes regression to mean into account while estimating the overall treatment effect.

3.6 Basic ANOVA model assumptions

Three basic assumptions underlying the validity of the estimate of treatment effect by ANOVA or a regression model were taken into consideration in this study. Executable commands which assess whether these assumptions are met or not have been included in the STATA statistical program developed. The assumptions are:

1. Normality of the residual
2. Linearity of the baseline (covariate) and outcome relationship
3. Parallel regression lines (homogeneity of regression slope). Interaction terms between the covariate and the outcome variable were tested to see if terms were sufficiently dispersed from zero.

   Other assumptions that were assumed met by virtue of the simulation exercise include:

4. Covariate is measured without error
5. Random sampling of participants
6. Outcome variable is at least interval
7. Residuals are identical and independently vary of each other
8. Variance of the outcome is equal in each group
9. Covariate is independent of the treatment variable, since baseline score (covariate) are generated before the treatment effects were added.

3.7 The study statistical program and simulation procedure

The nature of this study requires a statistical program that will generate hypothetical trials involving certain levels of experimental conditions, simultaneously run the regression models for the statistical methods being studied, and then post selected results into a file. Each hypothetical trial scenario will be repeated a thousand times, so as to generate a highly robust estimate of effect each time. Thus, the success or otherwise of the study depends to a large extent on the ability to assemble a series of executable and logical commands for this purpose. The program is such that the original simulated data is reproducible, this is made possible by setting the seed – the first line in the program was set to a constant value for the entire simulation. Details of the syntax used are available in the appendix 1. The body of the program contains several lines of executable command statements; each statement targets a specific task that is fundamental to the objective of the study. For example, the loop commands ensure that each simulation is repeated 1000 times for each hypothetical trial generated. The same number of iterations was used by (Vickers 2001). The essence of the repetition is to have a very robust estimate of the treatment effect by each statistical method, as the 1000 datasets obtained on a single trial scenario give a more accurate view of the population distribution in respect of the specific trial scenario. Any random fluctuation associated with the number of iterations used in this study is common to the estimates from all the three statistical methods and so does not affect the comparison.

A portion of the program randomly generates the scores for the treatment groups with specified levels of experimental conditions, such as: levels of treatment to be detected, levels of covariate-outcome correlation, sample size, and baseline imbalance.

An important aspect of the program is the group of commands that tests three major assumptions regarded as fundamental to the statistical analyses in this study. The assumptions are: the normality of the residuals, the linearity of baseline-outcome relationship, and the homogeneity of regression slopes assumption (which specifies that there is no interaction between the treatment group variable and the covariate). The tests on the assumptions have been mainly by inspecting the graphical presentations for normality and linearity assumptions and the inspection of the test of significance for the homogeneity assumption. If the p value for the interaction group variable in a separate ANCOVA model is greater than 0.05, for the first trial sample in each scenario, then there is no group-covariate interaction.

The other group of commands performs the statistical analysis by writing the three statistical methods in a regression model for information on respective regression coefficients and the associated standard error in each case. The last group of commands ensures that certain estimates, specifically the regression coefficients and the associated standard error, are collected for the 1000 datasets on each scenario and then saved in a postfile. The simulation was run at least two hundred and ten times to generate the data for the study. Each round of simulation, which represents a hypothetical trial, is unique in terms of the levels of experimental conditions it comprises. Hypothetical trials that have the same level of power for detecting a particular level of treatment effect (in standardized form) between groups have the same sample size. A population standard deviation of 1 (σ=1) was assumed in each trial and the same allocation ratio (i.e. equal sample size) was nominated within each treatment group. At each round of the simulation, selected output results were saved into a postfile and later collected into a SPSS data editor and re-analysed to meet the objectives of the study. The results of the re-analysis are presented in the results sections.

3.8 Levels of experimental conditions

It should be mentioned here that one of the things this study does differently from most other simulated studies that have similar objectives is not to choose levels of experimental conditions arbitrarily. Each of the levels that were chosen follows the principles that guide the design of a randomised controlled experiment. The reason for following this rather more thorough pathway is to be able to draw clear statements on the relative merits of each of the statistical methods under investigation to certain experimental conditions typically found in clinical trial settings. For example, at different levels of power of 80% and 90%, the required sample size was studied for each of the standardized effect sizes of 0.2, 0.5 and 0.8. These effect sizes have been previously specified and classified as low, medium, and large (Cohen 1988). For each sample size, the correlation between pre and post treatment scores varied from 0.1 to 0.9 at intervals of 0.2, as previous studies reveal that correlation between pre and post treatment scores in trial settings could range between 0.1 and 0.9 (Tu et al, 2005). Each level of correlation studied does not influence the number of patients studied in each group, but relates the baseline scores Z_1 and Z_2 to the treatment outcome scores Y_1 and Y_2, respectively.

As often done in the empirical trial settings, preliminary computations of certain parameters were made before embarking on the generation of the hypothetical trial datasets for this study. The reason for this is to ensure that this simulation study is conducted under such conditions that are typical of an empirical trial scenario. For example, the sample size was calculated in relation to 80% power of detecting each of the effect differences of: 0.2, 0.5 and 0.8 used in this study. This ensures that adequate sample size that correspond to the size of the effect is studied; small sample size for large effect size and large sample for small effect size.

For each hypothetical trial, the absolute imbalance was computed from a standardized score for the imbalance at baseline and the pooled standard error of the baseline imbalance (2/√n in this study, since the variance is 1 in each group and the two groups have equally balanced size), following the standard formula of:

$$Z = \frac{\text{absoluteimbalance(b)}}{\text{standarderror (simulatedeffect)}} \text{..(3.27)}$$

The absolute imbalance is therefore the product of the z-score and the standard error in each case.

With this arrangement, realistic values of imbalance were derived in relation to the sample size; thus avoiding large absolute imbalance for large sample sizes – which could have contradicted the principle of randomization. In the statistical program, the baseline scores were simply scaled by the absolute imbalance in each trial scenario to generate the trial dataset for that hypothetical trial. However, for trial scenarios that assume baseline comparability, the baseline imbalance was simply ignored.

The predetermined standardized normal deviates (standardized imbalance) for this study are; +/- 1.28, +/- 1.64 and +/- 1.96 (representing 20%, 10% and 5% two tailed probabilities of the standard normal distribution). This indicates that the various levels of imbalance in this study have a predetermined probability of occurring whatever the sample size and on whatever scale the covariate or outcome variable is scored. The imbalance was assumed on both direction of the treatment effect in the simulation exercise, it either occurs in the treated group (same direction) or in the control group (opposite direction).

This also represents another major difference between this study and most others, as pre-treatment scores are usually assumed identical for the treatment groups in a randomised control setting (Tu et al, 2005; Porter and Raudenbush, 1987). Tu et al, while investigating the statistical power for analysis of changes in RCT used arbitrary sample sizes of 10, 20 and 30 per treatment group in their simulation study. Even though their study produced certain power values to compare across a range of statistical methods, since the nominal power was not computed, it is rather difficult to associate important meaning to either low or high value of statistical power in their study. Markham and Rakes (1998) associate sample size and the robustness of regression analysis and artificial neural network.

Another advantage of this approach is that sample size, unit or scoring on which the covariate or outcome variables are measured does not really matter; the findings of the study will still be applicable. For example, in a study of comparison of osteopathic spinal manipulation with standard care for patients with low back pain (Anderson et al, 2010), a variety of outcome measures, all scored differently, were used; the visual analogue pain scale was scored from 0 to 10, the Roland Morris questionnaire was scored from 0 to 24 and the Oswestry questionnaire was scored from 0 to 50. The use of standardized scores has already solved the problem of inconsistency that could have arisen in the application of the findings of the study to different trial settings if absolute scores had been used for both imbalance and treatment effect. Standardized scores are particularly useful if several covariates are accounted for in the statistical model and the investigator is interested in ranking them according to their level of importance in the model. The differences in the original units of the covariates are already taken care of by the standardization of the coefficients.

For any given covariate, irrespective of the scores, the standardized normal deviate of 1.96 (p=0.05) represents the threshold at which some researchers would want to select the particular covariate for adjustment. Their claim is that at this point the imbalance between groups is large enough, despite randomization, to consider the covariate for statistical adjustment. Though the argument that this difference is unlikely to have occurred by chance would hold no logic if true randomisation is assumed to have been implemented.

So in the context of this study, the standardized imbalances used of +/-1.28, +/- 1.64 and +/- 1.96 represent low, medium and large imbalances (Cohen, 1988) for any covariates, irrespective of how they are scored. The table of sample sizes at various standardized effects, given nominal power levels of 80% and 90%, and the corresponding imbalances is included in the appendices 4.

Thus, specifically, the number of combinations of hypothetical trial scenarios simulated was 210 (evaluated per each of the three statistical methods) based around:-

7 standardized baseline imbalances: -1.96; -1.64; -1.28; 0; 1.28; 1.64; 1.96

5 covariate-outcome correlations: 0.1; 0.3; 0.5, 0.7 and 0.9

3 standardized treatment effect sizes: 0.2; 0.5 and 0.8

2 nominal power values: 80% and 90%

3.9 Conditional versus nominal power

Clinical trials are expected to be powered so as to be able to detect a particular treatment effect size considered to be of clinical importance by the researchers; most frequently, 80% and 90% power have been used by researchers. The power of a study has a profound effect on the importance which the wider community attaches to a study, a study that has a power of less than 80% is unlikely to have a wide acceptability. However, the power of a study also determines, to a large extent, the cost and the burden of a study on the investigators. Investigators often consider the efficiency of the methods; a statistical method that requires relatively fewer numbers of patients to provide a certain level of statistical power is said to be efficient (Vickers, 2001; Kernan 1999).

This study considers two-arm clinical trials at nominal power levels of both 80% and 90%. The level of a nominal power of a trial is one of the factors that determine how many participants or patients should be included in that trial. Higher powered studies require that more patients be studied than lower powered ones, at a given effect size. For example, a study designed to detect a standardised effect size of 0.5 at 80% power under a two tailed test will need to study 64 patients in each arm of the treatment groups, whereas the same study but with a 90% power to detect the difference will need to study 86 patients in each treatment arm. Power is inversely related to the variance of the difference between two means (Kernan et al, 1999).

For this simulation study, the actual (conditional) power of the statistical methods being studied is computed and defined for a particular scenario as the number of times the null hypothesis of no significant treatment effect was rejected in the simulations by each of the statistical methods multiplied by 100%. It is also interpreted as the proportion of true positives in the simulation per trial by the methods or the percentage of the simulated study that show a statistical difference between the two groups (Tu et al, 2005; Vickers, 2001).

Thus,

$$\text{Power} = \frac{\text{numberof simulations with statistically significant treatmenteffect}}{\text{Total numberof simulations for each hypothtical trial}} * 100 \ldots (3.2 \quad 8)$$

In this study, the computation is rather more laborious and time consuming, as this necessitates the recoding or computation of at least eighteen new variables from the regression coefficients and the standard errors for each trial scenario.

For example, in a particular trial scenario, the new variables that were computed are: the critical values (say, Z1, Z2 and Z3) for each of the three methods, given as;

$$Z = \frac{\beta}{se(\beta)} \quad , \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots..(3.28b)$$

where β is the regression coefficient and the standard error of the regression coefficient is the denominator.

The power of the test (for each of the three statistical methods) across the 1000 simulations per trial scenario was generated by examining the percentage of occurrences of the absolute value of the z-score that exceeded the critical value of the absolute of 1.96 (denoting a 5% two-tailed significance level). This critical value of 1.96 is applicable for large samples e.g. obtained when the studied effect size was small to moderate (i.e. 0.2 or 0.5).

However, when the effect size to be detected was large (0.8) such that only 52 patients need be studied, the critical (absolute) value was set at 2.0009. This is to conservatively allow for the extra uncertainty in the sampling distributions of smaller samples.

Thus, a total of 1260 new variables were computed for the entire 210 trial scenarios.

The sample size calculation in this study was done using the sampsi command in the STATA statistical package.

The widely used sample size formula according to Matthews, (2000) is given as;

$$n \geq \frac{2\sigma^2 (z_{\alpha/2} + z_{\beta})^2}{\tau^2} \text{...} (3.29)$$

where, n is the sample size and τ is the standardized treatment effect size (difference in means divided by the pooled standard deviation), σ is the standard deviation of z in both groups. Clearly, the higher the variability in the measurement (σ^2), the larger the sample needed to detect a given effect. Also, when a treatment effect to be detected (τ) is very small a much larger trial is also needed for the purpose. However, the computation of sample size in this study using the sampsi procedure follows the formula proposed by Machin et al (1997), which imposes a correction factor (the correction factor only adds 1 to the size required from the above formula (for a nominated significance level of about 2)) and is given as;

$$n \geq \frac{2\sigma^2 (z_{1-\alpha/2} + z_{\beta})^2}{\tau^2} + \frac{z^2{}_{1-\alpha/2}}{4} \text{..} (3.30)$$

3.10 Efficiency (relative sample sizes) of the adjusted and unadjusted analyses

As earlier mentioned in this chapter, the associated variance of the estimate of treatment effect by the three statistical methods differs and is dependent on the level of relationship between the baseline and the outcome in the case of the adjusted analysis. For example, the adjusted analysis by ANCOVA leads to reduction in the standard error as a result of the extra information – the baseline variable (with or without imbalance) – with which it tends to explain the variability in treatment groups. CSA provides an alternative means of accounting for a baseline imbalance, though, differently. On the other hand, ANOVA does not make use of the baseline variable at all in its explanation of the variability between treatment groups. Thus, for a given trial scenario, there is a contrast in the precision of the estimate of effect (measured by the standard error) for the three methods.

The standard error depends on the standard deviation of the scores and the sample size. Thus, this simulation study shall investigate the ratios of the standard error of each of the two methods of statistical adjustment with reference to the unadjusted analysis at different trial scenarios. The resultant ratio will thus provide information on the relative sample size requirement of the statistical methods under investigation. In this study, efficiency is seen in the context of using a reduced number of patients to detect a level of treatment difference at a stipulated level of power.

Pocock et al (2002) gave the following mathematical expressions for the ratio of the standard error for the ANCOVA against ANOVA as:

$\sqrt{1-\rho 2}$ ....(3.31)

and for the reduction in the original sample size for ANCOVA against ANOVA as:

$1-\rho^2$ ....(3.32)

Also, the ratio of standard error of ANCOVA to that of CSA is given as:

$\sqrt{(1+\rho)/2}$ ....(3.33)

and the relative reduction in sample size for using ANCOVA instead of CSA is given as:

$\frac{1+\rho}{2}$ ....(3.34)

The results of the simulation study will help demonstrate the application of the above formulae for different trial scenarios, and highlight the increased conditional power and efficiency savings for ANCOVA over ANOVA (as well as addressing the comparative power and efficiency statistics for CSA). Recall that for any given trial scenario, a more efficient trial will require fewer patients to have a stated level of power (usually 80 or 90%) to detect an important difference between two treatments (Kernan et al 1999).

Chapter 4: Directional pattern of precision and bias of statistical methods for RCTs – findings of the simulation exercise.

4.1 Introduction

This chapter presents a comparison of the pattern of precision and bias of the ANCOVA, change score and ANOVA in an analysis of an RCT with a single post treatment assessment of a continuous outcome variable. The comparison is in respect of various levels of experimental conditions typical of a clinical trial: levels of baseline-outcome correlation, levels of baseline imbalance in either direction relative to the treatment effect, levels of treatment effect (to be detected at 80% nominal power). The reference unadjusted analysis as mentioned previously is ANOVA and the adjusted analyses in this case are change score analysis (CSA) and ANCOVA. It should be noted that the same hypothetical trial data were used in both instances (with or without imbalance). The only difference is that, while the first part focuses on absolute balance of treatment groups as a result of randomisation – which is an unlikely scenario -, the second part does not and takes account of random differences. Throughout, results are presented on all three levels of standardized treatment effect (small – 0.2, medium – 0.5 and large – 0.8) to be detected. Results on trial scenarios in which there is baseline imbalance and treatment effect implies decrease in baseline score are presented in subsequent parts.

The two possible cases of defining an improvement in relation to baseline scores are considered in this study, when treatment effect or improvement implies increase or decrease in baseline scores. However, in this chapter results are presented on the latter possibility – treatment effect or improvement implies decrease in baseline score – and results relating to the other possibility are included as appendices. The reason for this choice is because most of the outcomes in musculoskeletal trials (and therefore of interest within the Arthritis Research UK centre) are such that treatment effect implies decrease in baseline score. For example, pain scales, depression/anxiety scales, disability scales are usually directional as low scores (often 0) denote no pain or no anxiety/depression or no disability whereas highest scale scores denote maximum pain or highest anxiety/depression or most severe disability. Hence, patients usually present with high scores (denoting increased severity of the condition) and the goal of treatment is to lower the scores. Results for situations in which treatment effect generally implies increase in baseline score (e.g. improved quality of life as shown by increased scores on the Short Form (SF-12/36) and EuroQoL questionnaire scales) and at 90% nominal power are attached as appendix 2. In all, there are six different graphs (Figures 4.1- 4.6) with each of them representing different aspect of the result in this chapter.

4.2 Bias and precision of estimate of effect associated with the statistical methods for groups with or without baseline imbalance

The first part of this section presents results on the precision and bias of estimates of treatment effect of RCTs in which treatment groups are homogeneous at baseline – in line with absolute balance due to randomisation. However, several authors have argued (Dougsheng et al, 2000; Altman et al,

1990) that the occurrence of comparable treatment groups at baseline is only possible in principle and not in practice. Hence, the second part of this section focuses on results for trial scenarios whereby treatment groups are heterogeneous to differing degrees at baseline.

4.2.1 Bias and precision of statistical methods of analysis of RCTs when groups are homogeneous

Table 4.1 presents the pattern of bias and precision of the three statistical methods across different levels of baseline-outcome correlations and different levels of treatment effect.

Table 4.2: Pattern of bias and precision of methods of analysis of RCTs with homogeneous treatment groups across differing levels of baseline-outcome correlations and treatment effect size.

Statistical Methods	**Levels of baseline-outcome correlations and estimates**				
	0.1 **{β, se(β)}**	**0.3** **{β, se(β)}**	**0.5** **{β, se(β)}**	**0.7** **{β, se(β)}**	**0.9** **{β, se(β)}**
ANOVA					
Small (0.2)	–0.20,0.07	–0.20,0.07	–0.20,0.07	–0.20,0.07	–0.20,0.07
Medium (0.5)	–0.50,0.18	–0.50,0.18	–0.50,0.18	–0.50,0.18	–0.50,0.18
Large (0.8)	–0.80,0.28	–0.80,0.28	–0.80,0.28	–0.80,0.28	–0.80,0.28
CSA					
Small (0.2)	0.20,0.10	0.20,0.08	0.20,0.07	0.20,0.06	0.20,0.03
Medium (0.5)	0.50,0.24	0.50,0.21	0.50,0.18	0.50,0.14	0.50,0.08
Large (0.8)	0.80,0.37	0.80,0.33	0.80,0.28	0.80,0.21	0.80,0.12
ANCOVA					
Small (0.2)	–0.20,0.07	–0.20,0.07	–0.20,0.06	–0.20,0.05	–0.20,0.03
Medium (0.5)	–0.50,0.18	–0.50,0.17	–0.50,0.15	–0.50,0.13	–0.50,0.08
Large (0.8)	–0.80,0.28	–0.80,0.27	–0.80,0.24	–0.80,0.20	–0.80,0.12

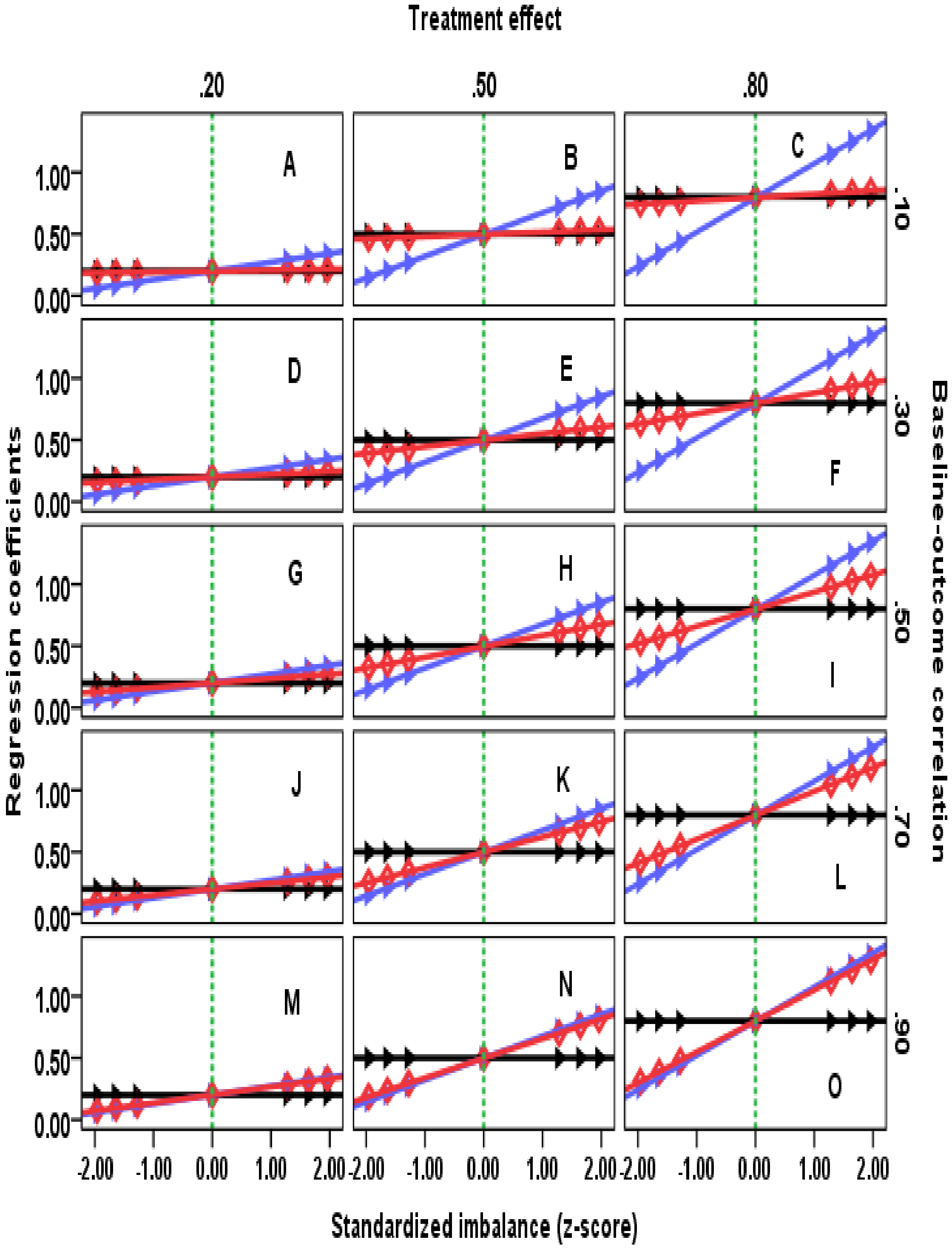
When Y = 0.2, n=788; Y=0.5, n=128, Y=0.8, n=52 (1000 iterations in each scenario)

The simulation results confirm the fact that all the statistical methods of interest yield unbiased point estimates of the treatment effect when the groups are comparable at baseline by randomisation; this is independent of baseline-outcome correlation. Whilst the CSA method yields an estimate of effect that is equal in magnitude to that of ANOVA and ANCOVA, the direction of its estimate is always opposite; this is connected with the way it is computed. To clarify, the interpretation of the difference in effect for CSA is in the opposite direction to that for ANOVA and ANCOVA such that a positive mean difference for ANOVA and ANCOVA implies that the treatment group has higher (i.e. 'worse') outcome than the control group, whereas, a positive difference for the CSA implies that the treated group has greater change (i.e. 'better') outcome than the control group. Thus, as way of measuring bias, absolute values of the regression coefficients were therefore taken in Figure 4.1 to illustrate a direct comparison of treatment effects between the three statistical methods. A green dotted line is drawn through the point at which baseline imbalance is zero for each hypothetical trial scenario – the Figure shows that the three methods have the same absolute effect estimate when the treatment arms are exactly balanced.

By contrast, also in Table 4.1, the precision of the estimate of effect by the methods varies dramatically with levels of correlation. Whereas precision of the estimate for ANOVA at a particular level of treatment effect to be detected is the same across all levels of covariate-outcome correlation, it is not so for both CSA and ANCOVA. The reason for this is because the variances of the estimates produced by both CSA and ANCOVA respond to baseline-outcome correlation. As shown in figure 4.2, a green dotted line is drawn through the point at which baseline imbalance is zero for each trial scenario, to further illustrate differences in precision between the three methods of statistical analysis when there is balance in baseline scores. Since ANOVA is independent on baseline-outcome correlation the precision of its estimate is given as in (Chapter 3, equation 3.16b).

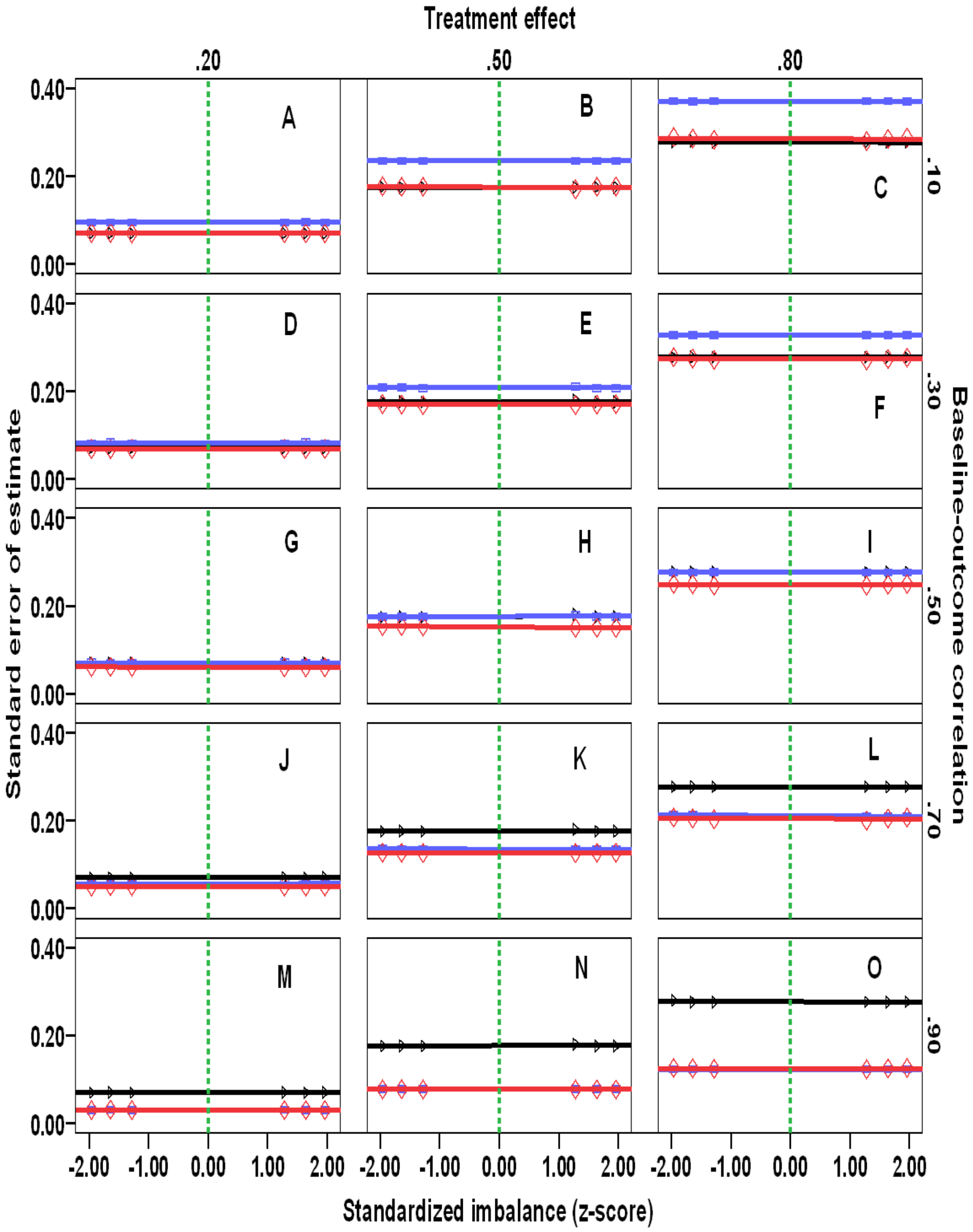
Mathematical relationships between the precision of effect estimate of each of the methods for statistical adjustment and baseline-outcome correlation was given in chapter three (equation 3.17). The results show that both ANOVA and ANCOVA are approximately equally precise in their estimate of treatment effect at a low correlation of 0.1, irrespective of the level of the simulated effect. However, an estimate of effect from ANCOVA becomes progressively more precise than that from ANOVA with higher levels of baseline-outcome correlation ($r \geq 0.3$). Also, at baseline-outcome correlation below 0.5, ANOVA presents with an estimate of effect that is more precise than that of CSA. Although estimates from both methods are equally precise at a correlation of 0.5, CSA has higher precision when baseline-outcome correlation is greater than 0.5. ANCOVA offers the benefit of a generally higher precision of estimate of effect than either ANOVA or CSA at most experimental conditions typical of a RCT. However, as shown in figure 2, the precision of estimate from CSA is somewhat comparable to that of ANCOVA at high baseline-outcome correlations (e.g. $r \geq 0.7$), and the precision of the ANOVA is somewhat comparable to the ANCOVA for low baseline-outcome correlations (e.g. $r \leq 0.3$). Note, for all three methods the

Figure 4.2: Directional pattern of bias of statistical methods for the analysis of RCTs at differing levels of baseline imbalance, baseline-outcome correlation and treatment effect sizes



KEY: __ ANOVA ___ CSA ___ANCOVA; When Y = 0.2, n=788; Y=0.5, n=128, Y=0.8, n=52 (1000 iterations in each case)

Figure 4.2: Directional pattern of precision of statistical methods for the analysis of RCTs at differing levels of: baseline imbalance, baseline-outcome correlation and treatment effect sizes



KEY ___ ANOVA ___ CSA ___ANCOVA; When Y = 0.2, n=788; Y=0.5, n=128, Y=0.8, n=52 (1000 iterations performed in each case)

standard errors are larger for greater treatment effect sizes since the required sample sizes are correspondingly greater for trials where the treatment effects are smaller (the simulations necessarily taking these differences in sample sizes into account). Hence, the standard errors are incremental across the increasing levels of treatment effect size shown in Figure 4.2.

4.2.2 Pattern of precision of statistical methods when groups are heterogeneous

The direction and size of imbalance does not affect the precision of estimate within a particular level of correlation. In fact, as illustrated in Figure 4.2, the precision of estimate of effect by these statistical methods is the same irrespective of baseline balance or heterogeneity at the same level of baseline-outcome correlation and same treatment effect size.

Table 4.2: Directional pattern of bias and precision of statistical methods at differing levels of: baseline-outcome correlation and baseline-imbalance in the same direction of effect [treatment effect size Y = 0.2; n=788]

	Levels of baseline-outcome correlations and estimates				
Methods (Z)	**0.1** **β, se(β)**	**0.3** **β, se(β)**	**0.5** **β, se(β)**	**0.7** **β, se(β)**	**0.9** **β, se(β)**
ANOVA					
1.28	–0.20, 0.07	–0.20, 0,07	–0.20, 0.07	–0.20,0.07	–0.20,0.07
1.64	–0.20,0.07	–0.20, 0,07	–0.20, 0.07	–0.20,0.07	–0.20,0.07
1.96	–0.20,0.07	–0.20,0.07	–0.20,0.07	–0.20,0.07	–0.20,0.07
Change score					
1.28	0.11, 0.10	0.11, 0.08	0.11,0.07	0.11,0.06	0.11,0.03
1.64	0.08,0.10	0.09, 0.08	0.08.0.07	0.08,0.06	0.08,0.03
1.96	0.06,0.10	0.06,0.08	0.06,0.07	0.06,0.06	0.06,0.03
ANCOVA					
1.28	–0.19,0.07	–0.17,0.07	–0.16,0.06	–0.14,0.05	–0.12,0.03
1.64	–0.19,0.07	–0.17,0.07	–0.14,0.06	–0.12,0.05	–0.10,0.03
1.96	–0.19,0.07	–0.16,0.07	–0.13,0.06	–0.10,0.05	–0.07,0.03

Y is the simulated effect size at 80% power and Z is the standardized imbalance
1000 iterations in each scenario

Table 4.3: Directional pattern of bias and precision of statistical methods at differing levels of: baseline-outcome correlation and baseline-imbalance in the same direction of effect [treatment effect size Y = 0.5; n=128]

	Levels of baseline-outcome correlations and estimates				
Methods (Z)	**0.1** **β, se(β)**	**0.3** **β, se(β)**	**0.5** **β, se(β)**	**0.7** **β, se(β)**	**0.9** **β, se(β)**
ANOVA					
1.28	–0.50,0.18	–0.50,0.18	–0.50,0.18	–0.50,0.18	–0.50,0.18
1.64	–0.50,0.18	–0.50,0.18	–0.50,0.18	–0.50,0.18	–0.50,0.18
1.96	–0.50,0.18	–0.50,0.18	–0.50,0.18	–0.50,0.18	–0.50,0.18
Change score					
1.28	0.27,0.24	0.27,0.21	0.27,0.18	0.27,0.14	0.27,0.08
1.64	0.21,0.24	0.21,0.21	0.21,0.18	0.21,0.14	0.21,0.08
1.96	0.16,0.24	0.15,0.21	0.15,0.18	0.15,0.14	0.15,0.08
ANCOVA					
1.28	–0.48,0.18	–0.43,0.17	–0.39,0.15	–0.34,0.13	–0.30,0.08
1.64	–0.47,0.18	–0.41,0.17	–0.35,0.15	–0.30,0.13	–0.24,0.08
1.96	–0.46,0.18	–0.40,0.17	–0.33,0.15	–0.26,0.13	–0.19,0.08

Y is the simulated effect size at 80% power and Z is the standardized imbalance
1000 iterations in each scenario

Table 4.4: Directional pattern of bias and precision of statistical methods at differing levels of: baseline-outcome correlation and baseline-imbalance in the same direction of effect [treatment effect size Y = 0.8; n=52]

	Levels of baseline-outcome correlations and estimates				
Methods (Z)	**0.1** **β, se(β)**	**0.3** **β, se(β)**	**0.5** **β, se(β)**	**0.7** **β, se(β)**	**0.9** **β, se(β)**
ANOVA					
1.28	–0.80,0.28	–0.80,0.28	–0.80,0.28	–0.80,0.28	–0.80,0.28
1.64	–0.80,0.28	–0.80,0.28	–0.80,0.28	–0.80,0.28	–0.80,0.28
1.96	–0.80,0.28	–0.80,0.28	–0.80,0.28	–0.80,0.28	–0.80,0.28
Change score					
1.28	0.44,0.37	0.44,0.33	0.44,0.28	0.44,0.21	0.44,0.12
1.64	0.34,0.37	0.34,0.33	0.34,0.28	0.34,0.21	0.34,0.12
1.96	0.25,0.37	0.25,0.33	0.25,0.28	0.25,0.21	0.26,0.12
ANCOVA					
1.28	–0.76,0.28	–0.69,0.27	–0.62,0.25	–0.55,0.20	–0.48,0.12
1.64	–0.75,0.29	–0.66,0.27	–0.57,0.25	–0.48,0.21	–0.39,0.13
1.96	–0.74,0.29	–0.63,0.28	–0.53,0.25	–0.42,0.21	–0.31,0.13

Y is the simulated effect size at 80% power and Z is the standardized imbalance
1000 iterations in each scenario

For example, under the same experimental conditions, a treatment effect estimate by ANCOVA when treatment groups are homogeneous at baseline will be as precise as an effect estimate for trial scenarios with heterogeneous groups.

Although level and direction of imbalance have little or no influence on the size of the standard error, the levels of covariate-outcome correlation are seen to have an important influence on the size of the standard error as it does if groups are balanced at baseline. With adjusted analysis, the precision increases with increase in levels of correlation; however, this increase in precision over the unadjusted analysis becomes noticeable only at a certain level of baseline-outcome correlation: $r \geq 0.3$ for ANCOVA and $r \geq 0.7$ for CSA. At a correlation of $r \leq 0.3$ the standard error estimate for ANCOVA is similar to that of ANOVA. Correspondingly, at low levels of correlation the effect estimate from the unadjusted analysis is more precise than that from CSA at all levels of baseline imbalance. Both ANOVA and CSA are equally precise (but less precise than ANCOVA) at trial scenarios where the baseline-correlation is 0.5.

Thus, when a suspected prognostic factor has little or no correlation with the outcome, statistical adjustment to account for any level of imbalance does not make the estimate of effect more precise than would have obtained if treatment comparison had been based on crude post-treatment comparison by ANOVA. The issue with precision of estimate of effect and the type of analysis to conduct hinges not on the levels of imbalance but rather on the degree of prognostic relationship – i.e. the correlation between covariate and outcome. Adjusting for a strong prognostic variable that appears balanced between treatment groups can lead to a more precise estimate of effect. Conversely, there is no gain in precision when adjusting for a covariate with large imbalance that is not prognostic ally related with the outcome. It is therefore clear from these results that the precision of estimate achieved through statistical adjustment depends on the level of prognostic relationship between the covariate and the outcome and not on the level of imbalance. Thus, researchers should endeavour to identify such baseline variables that are prognostic of outcome and account for them in the analysis.

4.2.3 Bias of statistical methods for RCT when groups are heterogeneous at baseline

In Tables 4.2- 4.4 it is observed that even at the lowest level of baseline-outcome correlation, when baseline imbalance exist, ANOVA and ANCOVA do not have the same estimates of effect, as denoted by the regression coefficients. The magnitude of this disparity in the regression coefficients of ANOVA and ANCOVA is dependent on the degree of baseline-outcome correlation and both level and direction of baseline imbalance. The observed difference in the regression coefficients by ANOVA and ANCOVA is as a result of the statistical adjustment of the baseline imbalance by ANCOVA. As earlier described in chapter 3, ANCOVA yields an estimate of effect that is conditional on the baseline imbalance for a prognostic variable, and in this case, baseline of the outcome variable. The size and direction of imbalance in a prognostic variable determine how much difference is observed between the estimate of effect by ANOVA that does not take such imbalance into account and the estimate of effect by ANCOVA that appropriately accounts for the imbalance. Indeed, the results of this simulated study agree with the algebraic equations and expressions on the subject. For example, as mentioned in (Chapter 3, Equation 3.26), the estimate of effect by ANCOVA, having taking into account the direction and size of a prognostic imbalance is given as;

$$\tau = (\overline{Y}_T - \overline{Y}_C) - \hat{\beta}(\overline{Z}_T - \overline{Z}_C)$$

In Table 4.4, this yields –0.74 and –0.31 respectively as effect estimates for a hypothetical trial scenario in which the treatment effect is ‘large’ (0.8), the baseline imbalance is in the same direction as the effect and is ‘large’ (1.96), and the baseline-outcome correlation is 0.1 and 0.9 respectively. Also the results of the simulation fit perfectly well with the above equation representing estimate of treatment for using ANCOVA. For example; at baseline-outcome correlation of 0.3 when absolute imbalance of (–0.09) (equivalence of 1.28 standardized imbalance) is in the same direction as treatment effect of (–0.2), the effect due to the covariate ($\hat{\beta}$) then is (0.322). By substituting these values in the above equation representing treatment effect by ANCOVA,

τ = (–0.2 – 0) – 0.322(–0.09 – 0)

= –0.2 + 0.0290

= – 0.171

this approximates the estimate of effect by ANCOVA in the simulation results in (Table 4.2)

Whereas, as expected theoretically, the regression coefficient is the same for ANOVA across various levels of baseline imbalance and baseline-outcome correlation, it varies at each level of these experimental conditions for ANCOVA. This variation is again closely related to the direction and size of imbalance. ANOVA naively yields the same simulated value as its estimates of treatment effect despite increasing levels of groups' heterogeneity at baseline; (Tables 4.2 – 4.4); these effect estimates are biased as they do not respond to or reflect the groups' baseline imbalance. The baseline scores are rather reckoned with by ANOVA as if they are not at all related with the outcome. The stronger the prognostic relationship between the baseline and the outcome the more bias the estimate given by ANOVA.

The estimate of effect by CSA does not change markedly with increase in baseline-outcome correlation as it yields approximately the same estimate across all observed levels of baseline-outcome correlation. It is important however, to point out that, with increasing magnitude in baseline-outcome correlation, and when all other factors remain the same, estimates of effect from CSA progressively approximate that of ANCOVA. In fact, for a baseline-outcome correlation of 0.9, and especially when the effect to be detected is low (0.2), irrespective of the size and direction of imbalance, the estimate from CSA and ANCOVA look so much alike. This result again is expected and supported theoretically. It was earlier mentioned in chapter 3 that, when there is imbalance in a prognostic factor at baseline, CSA cannot be unbiased unless the baseline – outcome correlation is 1. At that circumstance, the estimate of effect by CSA is the same as that of ANCOVA. It would be recall that the estimate of treatment effect by CSA is given as;

$$E[(\bar{Y}_T \quad \bar{Z}_T) \quad (\bar{Y}_C \quad \bar{Z}_C) \mid \bar{Z}_T, \bar{Z}_C] = \tau + (\rho - 1)(\bar{Z}_T - \bar{Z}_C)$$

This explains why estimates from both methods appear so close to each other at a correlation of 0.9. The difference in estimate however becomes more noticeable at lower levels of correlation. This then suggests that the lower the baseline- outcome correlation, the more biased are the estimates of effect by change score analysis. In this study, as evidenced from Tables 4.2–4.7, there is a substantial difference between the estimate of effect from ANOVA and that from change CSA at all levels of experimental factors when treatment groups are heterogeneous at baseline.

Even though the estimate of effect by CSA appears not to be markedly affected by the degree of baseline-outcome correlation, Tables 4.2 – 4.7 show that CSA regression coefficients are markedly influenced by both the magnitude and direction of imbalance. When imbalance is in the opposite direction to that of the treatment effect (signified by an increased positive value on the x-axis in Figure 4.1), that is, the control group have lower values (i.e. are better) at baseline, the absolute value of the effect estimate by CSA increases in relation to the underlying treatment effect. Here, the higher the level of imbalance the wider the distance between the estimated effect and zero, and the more likely it is to infer a significant result by change score. The reason for this seeming exaggeration is because the control group is treated by change score analysis as if it enjoys a level of treatment which was never assigned to it, giving rise to false positives. On the other hand, if the imbalance is in the same direction as the treatment effect, overall, there is a masking of the treatment effect by change score. This is a consequence of the way change is computed.

For example, assume that $\bar{Z}_T, \bar{Y}_T)$ represents the baseline and outcome score for the treatment group and $(\bar{Z}_C, \bar{Y}_C)$ represents the baseline and outcome for the control group.

Thus, with change (C) given as;

C = baseline – outcome

for an absolute imbalance of 0.09 in the same direction as an effect size of 0.2, the arrangement will be (note that reduction implies treatment effect and imbalance in same direction as treatment implies the treated group has a better prognosis at baseline):

$C= \bar{Z}_T - \bar{Y}_T) - (\bar{Z}_C - \bar{Y}_C)$

–0.09 – (–0.2) – (0 – 0) = – 0.09 + 0.2 = 0.11

Whereas, if the imbalance of 0.09 is in the opposite direction of effect size of (–0.2), then:

C= 0.09 – (– 0.2) – (0 – 0) = 0.09 + 0.2 = 0.29

Alternatively,

C = 0 – (-0.2) – (-0.09) – 0 = 0.29 + 0.09 = 0.29

These arrangements explain three points;

1) CSA yields estimates of effect in the opposite direction to the effect (improvement) to be determined. This is the reason for the positive sign on the estimate of effect that is expected to be negative.
2) Change score assumes the baseline-outcome correlation to be 1. Thus estimates of effect are the same across all levels of correlation.
3) Summarily, the computation of effect by CSA when imbalance is in the same direction as treatment effect is such that the magnitude of this imbalance is subtracted from the absolute value of the treatment group's effect. On the other hand, when imbalance is in the opposite direction of treatment, the computation of effect by CSA is such that the magnitude of imbalance is added to the absolute value of the treatment group's effect.

When imbalance is in the same direction as the treatment effect, the estimate from CSA is seen to converge to a zero value, indicating no effect. This phenomenon ultimately depends on the size of imbalance; the larger the imbalance the closer to zero is the estimate of effect by CSA. This tapering of effect size relative to size of imbalance is due to the deduction of the size of the imbalance from the treatment effect in the treatment group resulting in the loss of some effect. This then means that though some treatment effects exist, they will not be detected by CSA and thus, false negatives will result. Therefore, depending on direction, the larger the imbalance the larger the exaggerating or masking effect by CSA on its estimate of effect. It can also be observed from Tables 4.2 – 4.7 that the estimate of effect by ANCOVA for a corresponding level of imbalance is similar to that of change at correlation of 0.9.

Going by the absolute values, the imbalance in the opposite direction to the treatment effect in this study means that the baseline score of the treatment group is smaller and the group has a worse prognosis compared to the control group. Thus, the mean score for the treatment group at baseline is lower compared to the control group and by implication, given this method; the mean baseline score for the treatment is also lower than the grand mean on baseline covariate for the two groups. Under this condition, the mechanism of adjustment by ANCOVA which ensures the upward adjustment of the baseline score of the group with the lower mean score favours the treatment group. This explains the reason for the higher value of the regression coefficients that ANCOVA presents, in respect of the simulated score at this instance of baseline imbalance being in the opposite direction of the effect. It should be noted that, these regression coefficients which represent the estimate of true effect is the unbiased estimate, as it takes account of the baseline imbalance. While the results also show that these estimates of effect are highly dependent on levels of covariate-outcome correlation, they also change considerably with levels of imbalance.

The magnitude of the regression coefficient from ANCOVA is influenced by the levels of baseline imbalance and the degree of baseline-outcome correlation (as shown in Figure 1). For example, when imbalance is in the opposite direction to the treatment effect (positive direction in Figure 1), the size of the regression coefficient from ANCOVA increases as correlation increases; under the same conditions, the coefficient also increases as the baseline imbalance increases. On the other hand, when the imbalance is in the same direction with the treatment (reflected in negative values for the baseline imbalances in Figure 1), then the regression coefficient by ANCOVA decreases as the correlation increases and decreases as the imbalance also increases. The value of regression coefficient by ANCOVA would always be in excess of the estimate by ANOVA and by implication the simulated score, if imbalance is in the opposite direction of the effect. Thus, when heterogeneity in treatment groups is such that the imbalance is in the opposite direction of treatment group the likelihood of a significant result upon adjustment through ANCOVA (compared to ANOVA) is enhanced. This likelihood increases with an increase in the level of covariate-outcome correlation.

Conversely, when imbalance is in the same direction as treatment, the estimate of effect by ANCOVA will always be less than the unadjusted; this trend is also observed as correlation increases and thus a significant effect by ANOVA may not be significant by ANCOVA especially if the baseline and the outcome are very strongly correlated (Tables 4.2–4.4). For example, the estimate of effect by ANCOVA, represented by the regression coefficient;

$$\tau = \overline{Y}_T - \overline{Y}_C - \hat{\beta}_Z(\overline{Z}_T - \overline{Z}_C) = -0.17$$

If y = – 0.2, z= –0.09 at r=0.3; $\hat{\beta}_Z$ - regression coefficient for the baseline covariate is computed as 0.322

Conversely, when imbalance is in the opposite direction and with $\hat{\beta}_z$ computed as 0.229, the estimate of effect by ANCOVA is given as;

$$\tau = [(-0.2 - 0) - 0.229\ (0 - (-0.09))] = (-0.2) - 0.0206 = -0.221$$

This also equates the results of the simulation.

Table 4.5: Directional pattern of bias and precision of statistical methods at differing levels: of baseline-outcome correlation and baseline-imbalance in the opposite direction of effect [treatment effect size Y= 0.2; n=788]

	Levels of baseline-outcome correlations and estimates				
Methods (Z)	**0.1** **β, se(β)**	**0.3** **β, se(β)**	**0.5** **β, se(β)**	**0.7** **β, se(β)**	**0.9** **β, se(β)**
ANOVA					
1.28	–0.20, 0.07	–0.20, 0,07	–0.20, 0.07	–0.20,0.07	–0.20,0.07
1.64	–0.20,0.07	–0.20, 0,07	–0.20, 0.07	–0.20,0.07	–0.20,0.07
1.96	–0.20,0.07	–0.20,0.07	–0.20,0.07	–0.20,0.07	–0.20,0.07
Change score					
1.28	0.29,0.10	0.29,0.08	0.29,0.07	0.29,0.06	0.29,0.03
1.64	0.32,0.10	0.32,0.08	0.32,0.07	0.32,0.06	0.32,0.03
1.96	0.34,0.10	0.34,0.08	0.34,0.07	0.34,0.06	0.34,0.03
ANCOVA					
1.28	–0.21,0.07	–0.23,0.07	–0.25,0.06	–0.24,0.05	–0.28,0.03
1.64	–0.21,0.07	–0.24,0.07	–0.26,0.06	–0.28,0.05	–0.31,0.03
1.96	–0.21,0.07	–0.24,0.07	–0.27,0.06	–0.30,0.05	–0.33,0.03

Y is the simulated effect size at 80% power and Z is the standardized imbalance
1000 iterations in each scenario

Table 4.6: Directional pattern of bias and precision of statistical methods at differing levels of: baseline-outcome correlation and baseline-imbalance in the opposite direction of effect [treatment effect size Y = 0.5; n=128]

	Levels of baseline-outcome correlations and estimates				
Methods (Z)	**0.1** **β, se(β)**	**0.3** **β, se(β)**	**0.5** **β, se(β)**	**0.7** **β, se(β)**	**0.9** **β, se(β)**
ANOVA					
1.28	–0.50,0.18	–0.50,0.18	–0.50,0.18	–0.50,0.18	–0.50,0.18
1.64	–0.50,0.18	–0.50,0.18	–0.50,0.18	–0.50,0.18	–0.50,0.18
1.96	–0.50,0.18	–0.50,0.18	–0.50,0.18	–0.50,0.18	–0.50,0.18
Change score					
1.28	0.72,0.24	0.73,0.21	0.73,0.18	0.73,0.14	0.73,0.08
1.64	0.79,0.24	0.79,0.21	0.79,0.18	0.79,0.14	0.79,0.08
1.96	0.84,0.24	0.84,0.21	0.85,0.18	0.85,0.14	0.85,0.08
ANCOVA					
1.28	–0.52,0.18	–0.57,0.17	–0.61,0.15	–0.66,0.13	–0.70,0.08
1.64	–0.52,0.18	–0.59,0.17	–0.64,0.15	–0.70,0.13	–0.76,0.08
1.96	–0.53,0.18	–0.60,0.17	–0.67,0.15	–0.74,0.13	–0.81,0.08

Y is the simulated effect size at 80% power and Z is the standardized imbalance
1000 iterations in each scenario

Table 4.7: Directional pattern of bias and precision of statistical methods at differing levels of: baseline-outcome correlation and baseline-imbalance in the opposite direction of effect [treatment effect size Y = 0.8; n=52]

	Levels of baseline-outcome correlations and estimates				
Methods (Z)	**0.1** **β, se(β)**	**0.3** **β, se(β)**	**0.5** **β, se(β)**	**0.7** **β, se(β)**	**0.9** **β, se(β)**
ANOVA					
1.28	–0.80,0.28	–0.80,0.28	–0.80,0.28	–0.80,0.28	–0.80,0.28
1.64	–0.80,0.28	–0.80,0.28	–0.80,0.28	–0.80,0.28	–0.80,0.28
1.96	–0.80,0.28	–0.80,0.28	–0.80,0.28	–0.80,0.28	–0.80,0.28
Change score					
1.28	1.15,0.37	1.15,0.33	1.15,0.28	1.15,0.21	1.15,0.12
1.64	1.24,0.37	1.25,0.33	1.25,0.28	1.25,0.21	1.25,0.12
1.96	1.34,0.37	1.34,0.33	1.34,0.28	1.34,0.21	1.34,0.12
ANCOVA					
1.28	–0.83,0.28	–0.90,0.27	–0.98,0.25	–1.04,0.20	–1.12,0.12
1.64	–0.84,0.29	–0.94,0.27	–1.03,0.25	–1.12,0.21	–1.21,0.13
1.96	–0.85,0.29	–0.96,0.28	–1.07,0.25	–1.18,0.21	–1.29,0.13

Y is the simulated effect size at 80% power and Z is the standardized imbalance
1000 iterations in each scenario

4.3 Ratios of standard error of statistical methods for the analysis of randomised controlled trials at several hypothetical trial scenarios

Table 4.8 below presents the ratio of the associated variance of estimate of effect measured by its standard error for both ANCOVA and ANOVA over the range of study hypothetical clinical trial scenarios. This is analogous to a design effect, e.g. in the context of cluster randomised controlled trials and trials in which individuals are randomised. With the standard error of ANCOVA as the numerator, the results illustrate various levels of reduction in the standard error of estimate of effect by using ANCOVA instead of ANOVA for statistical analysis. A ratio of 1 indicates that both methods are equally precise at the given trial scenario, and this occurs when baseline-outcome correlation is low (0.1). From the table, it is clear that the precision benefit of ANCOVA over ANOVA is independent of the level and direction of imbalance and is solely driven by the level of baseline-outcome correlation. There is generally a reduction in the ratio of standard errors as the level of baseline-outcome correlation increases. The percentage reduction, therefore, in the standard error by using ANCOVA ranges from 0%, when correlation is 0.1, to 57%, when correlation is 0.9.

It can be seen from these results that, when treatment groups are balanced, there is only about a 4% reduction in the variance of estimates by using ANCOVA instead of ANOVA at baseline-outcome correlation of 0.3, but a 25-30% reduction is observed if the baseline-outcome correlation is 0.7. In fact, for a trial with balanced treatment groups, the percentage reduction in standard error

Table 4.8: Ratio of the standard error of ANCOVA and ANOVA

Effect Z	**Levels of baseline-outcome correlations**				
0.2	**0.1**	**0.3**	**0.5**	**0.7**	**0.9**
–1.96	1.01	0.97	0.86	0.71	0.43
–1.64	1.01	0.97	0.87	0.73	0.44
–1.28	1.00	0.97	0.87	0.73	0.43
0	1.00	0.96	0.86	0.71	0.43
1.28	1.00	0.97	0.86	0.71	0.43
1.64	1.00	0.97	0.86	0.71	0.43
1.96	1.00	0.97	0.86	0.71	0.43
0.5					
–1.96	1.01	0.97	0.88	0.72	0.44
–1.64	1.01	0.96	0.88	0.72	0.44
–1.28	1.01	0.96	0.87	0.73	0.44
0	1.00	0.96	0.86	0.71	0.43
1.28	1.03	0.93	0.83	0.70	0.43
1.64	1.01	0.97	0.86	0.72	0.44
1.96	1.00	0.97	0.86	0.72	0.44
0.8					
–1.96	1.04	1.00	0.90	0.75	0.45
–1.64	1.02	0.99	0.90	0.74	0.45
–1.28	1.02	0.98	0.89	0.73	0.45
0	1.00	0.96	0.86	0.71	0.43
1.28	1.01	0.97	0.89	0.72	0.45
1.64	1.03	0.99	0.90	0.74	0.45
1.96	1.04	1.00	0.90	0.75	0.46

When Y = 0.2, n=788; Y=0.5, n=128, Y=0.8, n=52, 1000 iterations in each scenario

using ANCOVA is the same for both large and small trials given a particular baseline-outcome correlation. Thus at correlation of 0.7, the respective ratios of the standard error of ANCOVA and ANOVA for large, medium and small trials is 0.71, implying a 29% reduction in the standard error of ANOVA at each instance. Thus, the stepwise precision pattern presented in Figure 2 above does not necessarily mean that there is a greater disparity in the standard error of ANOVA and ANCOVA across levels of effect or trial sizes, it is just a response to the size of effect being determined at each time.

Although size of imbalance may not matter when considering the precision of ANCOVA in a trial with a given sample size, the level of reduction in the standard error is lower at a small sample size, especially when the imbalance is large. For example, for a large trial (n=788), when the baseline-outcome correlation is 0.3 and 0.7 respectively, ANCOVA yields 3% and 29 % reduction in the standard error of ANOVA, whereas, for a small trial (n=52), given the same conditions, there is 0.0 and 25% reduction in the standard error of ANOVA respectively by using ANCOVA. This observation is not however consistent with change in the trial size and or size of imbalance. Here, a reason for the slight difference in the standard errors is the postulated random fluctuation inherent with simulated datasets and the differences in width of sampling distributions with different sizes of trials. Trials with small sample sizes are more likely to be prone to chance inconsistency. Generally, especially at low baseline-outcome correlation ($r \leq 0.3$), irrespective of the trial size and the size and direction of baseline imbalance, the precision of the unadjusted analysis approximates that of the adjusted analysis by ANCOVA. This is not so as correlation increases, the gain in precision for using ANCOVA instead of ANOVA is seen to be near 10% and 25% when baseline-outcome correlation is 0.5 and 0.7, respectively.

Furthermore, it is also interesting to note that the absolute value of the reduction in standard error is approximately equal on either side of zero (no imbalance). This appears symmetrical over the various levels of imbalance. This shows that direction of baseline imbalance does not matter in determining a comparative precision benefit of using ANCOVA against ANOVA. This fact has a very important implication at the design stage of RCTs, as researchers (statisticians) can confidently specify ANCOVA as the primary analysis even when they are completely blinded to random allocation procedure. They do not need to know which treatment group is favoured or disfavoured by randomization.

This further confirms the fact that, adjustment by ANCOVA for a baseline imbalance in the same direction of effect is as precise as for an imbalance in the opposite direction of the treatment effect. Essentially, these results provide illustration that, for randomised trials with balanced/unbalanced treatment groups, the ratio of the standard error of the adjusted analysis by ANCOVA against the unadjusted analysis by ANOVA is defined as equation (3.31) in chapter 3. The mathematical expression is

$$\sqrt{1-\rho^2}$$

When trials are large, the simulated ratios of the standard errors at any level of baseline imbalance approximate this algebraic expression; there is little or no deviation from the expression by the ratios. However, with small sample trials, in respect of the precision of estimate when imbalance is ignored, the observed slight deviation in the ratios at any level of imbalance does not exceed that which can be attributed to chance. These results thus confirm the fact that both the size and level of imbalance and the size of the trial do not matter when considering the precision of estimate if appropriate statistical adjustment is used. This result is further represented in Figure 4.3, though, some minor random fluctuations are observed in the data representing trial scenarios in graph labelled N..

Essentially, figure 4.3 provides a graphical illustration of table 4.8; both represent the ratio of the standard error (SE) of ANCOVA to that of ANOVA at different trial scenarios. The closer to 1 the ratio is the more the SE of the two methods resembles each other for that trial scenario. Figure 4.3 shows that the ratio of the SE of the two methods are very comparable at low baseline-outcome correlation (r<0.3) irrespective of the level of baseline imbalance. The implication of a low ratio for the SE of ANCOVA relative to ANOVA, as a result of a high baseline-outcome correlation, on relative sample size requirement is considered in chapter 5.

Table 4.9 gives the ratios of standard errors of both methods for statistical adjustment (CSA and ANCOVA) with ANCOVA as the numerator. The ratio is about 0.75 (signifying a 25% reduction in variance of estimates) when the baseline-outcome correlation is 0.1, and about 0.9 (10% reduction) when the correlation is 0.7. The estimate of effect by ANCOVA is more precise for most trial scenarios, and especially when correlation is less than or equal to 0.5. The precision benefit of ANCOVA over change score is minimal when the baseline–outcome correlation is greater than or equal to 0.7. Both direction and size of imbalance do not favour either of the two methods over the other.

Figure 4.3: Ratio of the standard error (SE) of the adjusted analysis (ANCOVA) to unadjusted analysis (ANOVA) at different hypothetical trial scenarios

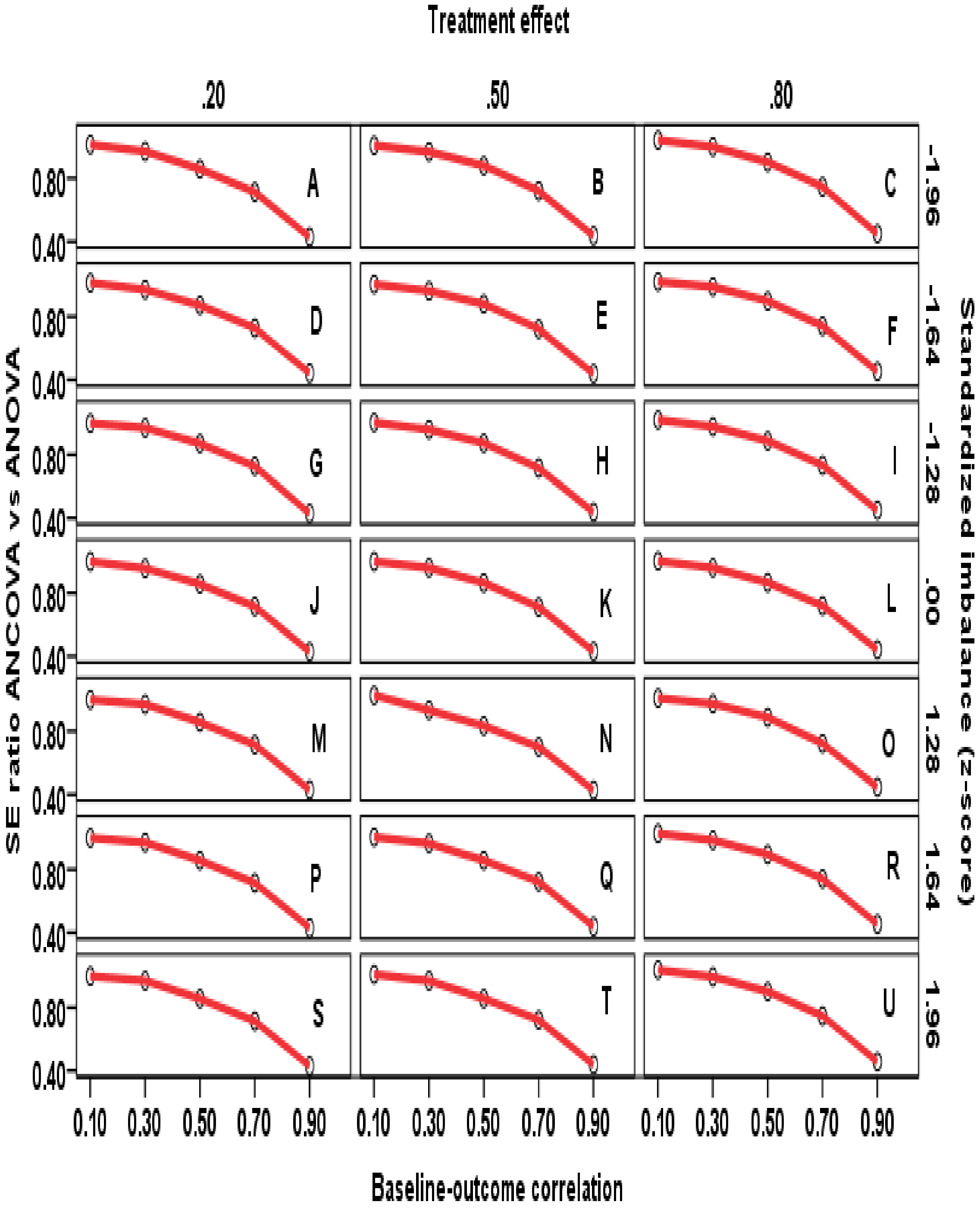


Table 4.9: Ratio of the standard error of baseline adjustment by ANCOVA and the change score analysis

Effect Z	**Levels of baseline-outcome correlations**				
0.2	**0.1**	**0.3**	**0.5**	**0.7**	**0.9**
–1.96	0.75	0.85	0.86	0.91	1.00
–1.64	0.75	0.81	0.87	0.93	1.00
–1.28	0.74	0.85	0.87	0.93	1.00
0	0.74	0.84	0.86	0.91	1.00
1.28	0.74	0.85	0.85	0.83	0.97
1.64	0.73	0.81	0.86	0.91	1.00
1.96	0.74	0.85	0.86	0.91	1.00
0.5					
–1.96	0.75	0.82	0.88	0.94	0.99
–1.64	0.75	0.81	0.88	0.94	0.99
–1.28	0.75	0.81	0.87	0.93	0.99
0	0.74	0.81	0.87	0.93	0.97
1.28	0.77	0.80	0.83	0.93	0.99
1.64	0.75	0.82	0.86	0.94	0.99
1.96	0.75	0.82	0.86	0.94	0.99
0.8					
–1.96	0.78	0.85	0.91	0.97	1.02
–1.64	0.77	0.84	0.90	0.96	1.02
–1.28	0.76	0.83	0.89	0.97	1.01
0	0.75	0.81	0.87	0.93	0.99
1.28	0.75	0.83	0.89	0.95	1.00
1.64	0.77	0.83	0.90	0.98	1.02
1.96	0.78	0.84	0.91	0.99	1.02

When Y = 0.2, n=788; Y=0.5, n=128, Y=0.8, n=52, 1000 iterations in each scenario

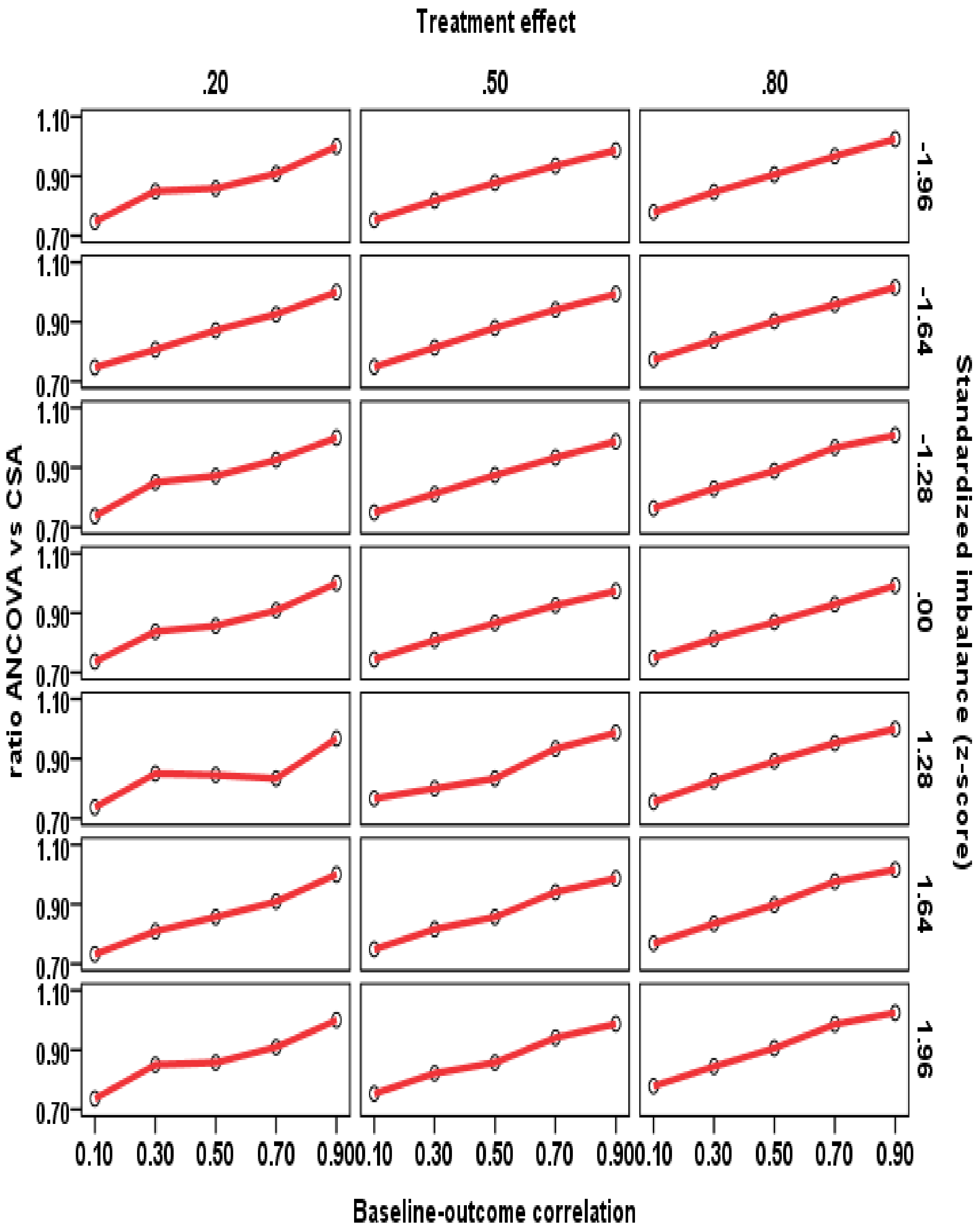
Also, reduction in standard error for using ANCOVA instead of CSA does not also show a distinct pattern across various trial sizes. For example, when there is large imbalance in a baseline covariate that has a correlation of 0.5 with the outcome; for large, medium and small trials, the percentage reduction in standard error for using ANCOVA instead of change score is 14%, 14% and 9 % respectively. However, if treatment groups are homogeneous at baseline, with the same baseline-outcome correlation of 0.5, the relative reduction in standard error is; 14%, 13% and 13% showing no definite pattern of deviation from that obtained when baseline imbalance is large.

Again, in all trial scenarios irrespective of the direction and size of imbalance and the level of prognostic relationship between the baseline and the outcome, the results of this simulation study provides illustration that the ratio of the standard error of ANCOVA to CSA is expressed as in (equation 3.33) in chapter 3 and the expression is reproduced below;

$$\sqrt{\frac{1+\rho}{2}}$$

The slight discrepancy that is however observed between the simulated result and the value of the algebraic expression can be attributed to random fluctuations in the simulated datasets and the degree of accuracy used in computing the ratios (2 decimal places instead of 8 decimal places of the simulated output). For example, at a correlation of 0.5, algebraically the ratio of the standard error is 0.87, whereas, for levels of imbalance, the range of the simulated results is between 0.83 and 0.91. This difference (+/– 0.04, 4%) could not have been due to any other source other than observed random error in the data and that which accrued from approximations. The results also show that for trial scenarios that have baseline-outcome correlation of greater than or equal to 0.7, the ratio of the standard error approaches 1 and thus, the tendency for the precision of the two methods to be equal. In fact, from Table 4.9 at a baseline-correlation of 0.9, the standard error of CSA is approximately equal to that of ANCOVA as the ratio of standard error for the two methods is approximately 1.

Figure 4.4: Ratio of the standard error (SE) of the two adjusted analysis (ANCOVA and CSA) at differing hypothetical trial scenarios



In table 4.10, with CSA as the numerator, the ratios of the standard errors for CSA relative to ANOVA are 1 at baseline-outcome correlation of 0.5 irrespective of the levels and direction of baseline imbalance and treatment effect or sample size, i.e this shows equality in precision of ANOVA and CSA at all trial scenarios for a correlation of 0.5. The precision benefit of ANOVA over CSA is about 35% ((1.35/1.00)*100%) at a small (≤0.1) correlation, whatever the level of imbalance. However, with baseline-outcome correlations exceeding 0.5, CSA shows a reduction in SE compared to ANOVA: about a 21% and 57% reduction in SE for a correlation of 0.7 and 0.9, respectively (and independent of other parameters).

Table 4.10: Ratio of the standard error of baseline adjustment by CSA and the unadjusted analysis by ANOVA

Effect Z 0.2	Levels of baseline-outcome correlations				
	0.1	**0.3**	**0.5**	**0.7**	**0.9**
–1.96	1.36	1.14	1.00	0.79	0.43
–1.64	1.36	1.20	1.00	0.79	0.44
–1.28	1.36	1.14	1.00	0.79	0.43
0	1.36	1.14	1.00	0.79	0.43
1.28	1.36	1.14	1.01	0.86	0.44
1.64	1.37	1.20	1.00	0.79	0.43
1.96	1.36	1.14	1.00	0.79	0.43
0.5					
–1.96	1.34	1.18	1.00	0.77	0.44
–1.64	1.34	1.18	1.00	0.77	0.44
–1.28	1.34	1.18	1.00	0.77	0.44
0	1.34	1.19	1.00	0.77	0.44
1.28	1.34	1.17	1.00	0.75	0.43
1.64	1.34	1.18	1.00	0.77	0.44
1.96	1.34	1.18	1.00	0.77	0.44
0.8					
–1.96	1.34	1.18	1.00	0.77	0.44
–1.64	1.32	1.18	1.00	0.77	0.45
–1.28	1.34	1.18	1.00	0.76	0.45
0	1.34	1.18	1.00	0.77	0.45
1.28	1.34	1.18	1.00	0.76	0.45
1.64	1.34	1.18	1.00	0.76	0.45
1.96	1.34	1.18	1.00	0.76	0.45

When Y = 0.2, n=788; Y=0.5, n=128, Y=0.8, n=52, 1000 iterations in each scenario

4.4 Percentage bias in the estimate by ANOVA and change score analysis with the unbiased estimate by ANCOVA as the reference

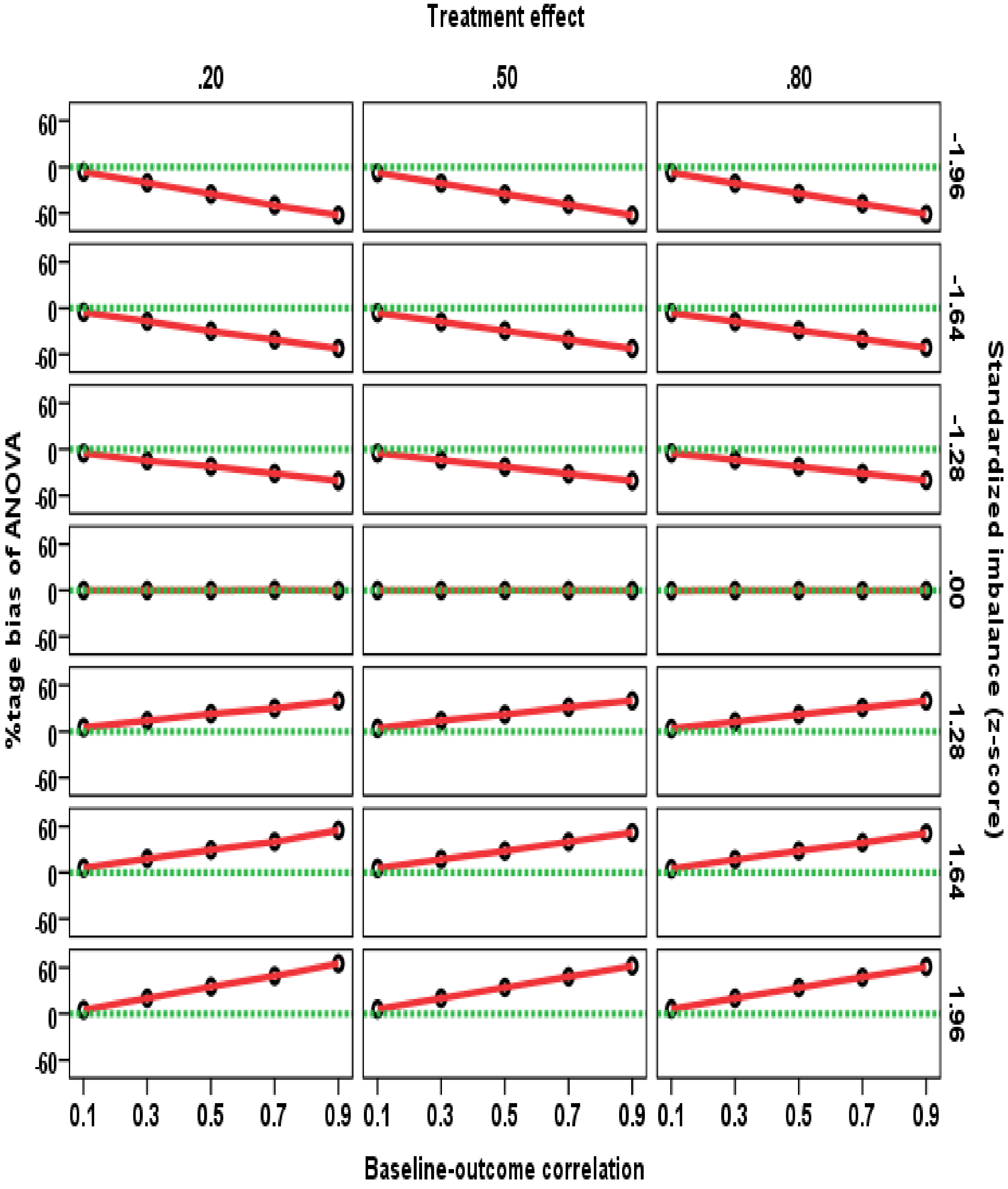
Here, results on the percentage bias associated with the estimates of effect using ANOVA and change score with ANCOVA as the reference analysis are presented. Table 4.11, presents the results on the percentage bias associated with the estimate of effect using ANOVA.

Table 4.11: Percentage bias in estimate of effect by ANOVA with reference to the unbiased estimator of effect by ANCOVA

Effect Z	**Levels of baseline-outcome correlations**				
0.2	**0.1**	**0.3**	**0.5**	**0.7**	**0.9**
–1.96	–6.50	–20.50	–35.50	–50.0	–62.50
–1.64	–5.50	–17.00	–30.00	–41.00	–54.50
–1.28	–5.00	–15.00	–22.25	–31.50	–41.00
0	0.00	0.00	0.00	0.50	0.00
1.28	5.00	14.00	23.00	30.00	40.00
1.64	6.00	18.00	29.50	40.00	55.00
1.96	5.00	20.00	35.00	49.00	65.00
0.5					
–1.96	–7.40	–21.12	–35.00	–48.80	–62.60
–1.64	–6.20	–17.80	–29.40	–41.00	–52.60
–1.28	–5.00	–14.00	–22.69	–32.00	–41.00
0	0.00	0.00	0.04	0.00	–0.20
1.28	4.00	14.00	22.00	31.60	40.00
1.64	5.60	17.20	28.00	40.00	52.00
1.96	6.00	20.00	34.00	48.00	62.00
0.8					
–1.96	–7.13	–21.25	–34.38	–47.88	–61.8
–1.64	–6.25	–17.50	–28.88	–40.25	–51.50
–1.28	–4.88	–13.75	–22.53	–31.38	–40.25
0	–0.40	0.00	0.00	–0.25	0.00
1.28	3.75	12.50	21.88	30.88	39.88
1.64	5.00	16.88	28.25	38.75	51.13
1.96	6.25	20.13	33.75	47.38	61.00

When Y = 0.2, n=788; Y=0.5, n=128, Y=0.8, n=52, 1000 iterations in each scenario

Figure 4.5: Pattern and percentage of directional bias of ANOVA in relation to the unbiased estimate of effect by ANCOVA at differing hypothetical trial scenarios

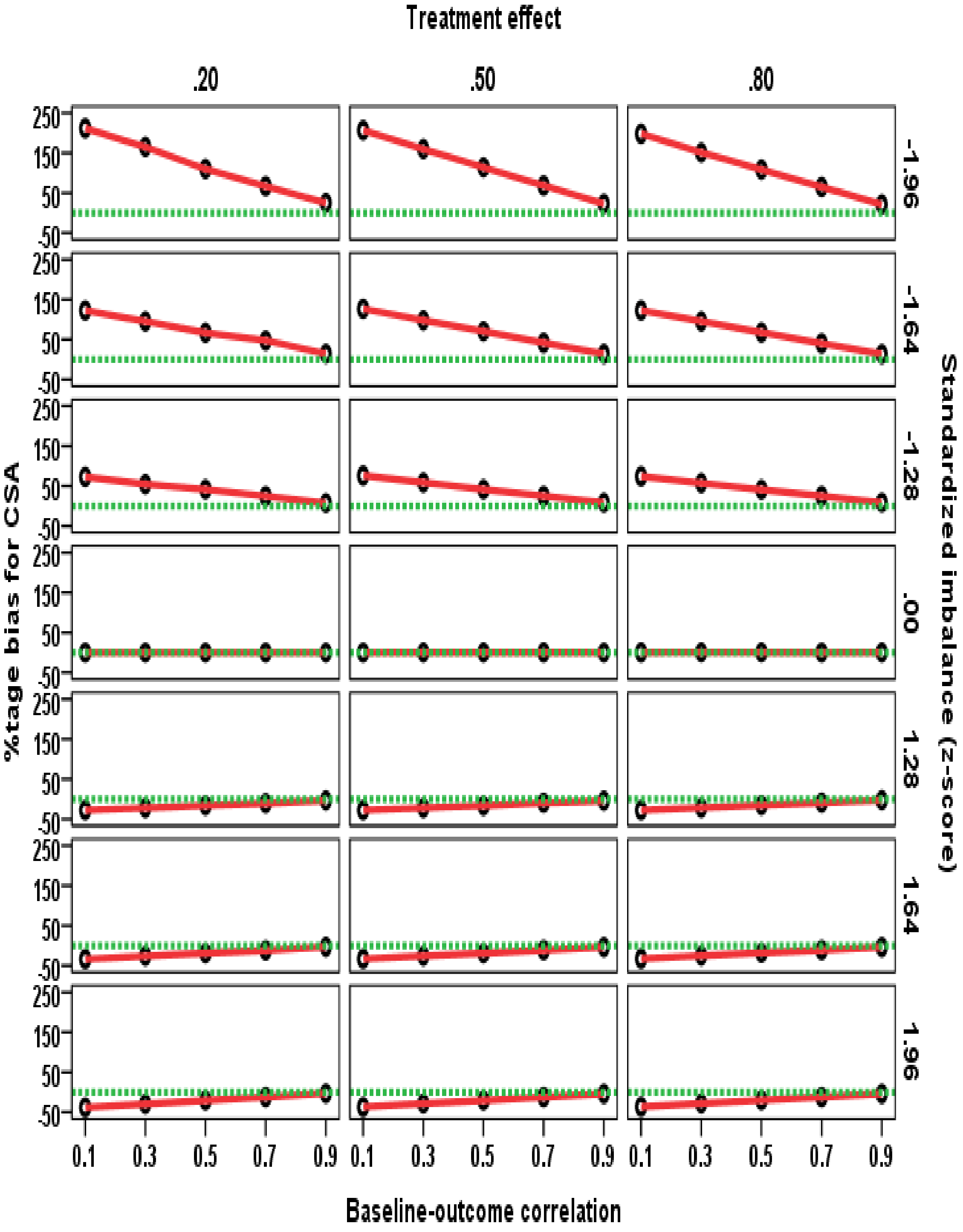


The green dotted line is the reference line for the unbiased estimate of effect

The percentage bias expresses the level in percentage of exaggeration or masking of treatment effect associated with the estimate of effect resulting from using ANOVA or change instead of ANCOVA.

When baseline imbalance is in the same direction as treatment effect, ordinarily, the unadjusted analysis exaggerates treatment effect. Accordingly, the exaggerated effect is reduced following the mechanism of adjustment by ANCOVA, making it smaller than the observed effect that results from using ANOVA. This explains why the percentage bias in the same direction presents with negative signs. The magnitude of bias in a given trial scenario is clearly being driven by the level of baseline-outcome correlation and the size of imbalance (see Figure 4.5). Also, the higher the baseline imbalance, the higher the percentage bias recorded. From Table 4.11, it is clear that treatment effect size (and hence size of trial) is not influential. Percentage biases by ANOVA in either direction are approximately equal in magnitude and opposite, thus, confirming the non-directionality of this bias. There is no bias in the estimate of treatment effect when treatment groups are balanced in baseline score; at all levels of baseline-outcome correlation the percentage bias is zero. This thus implies that in the context of bias, the level of baseline-outcome correlation is immaterial in the absence of baseline imbalance. A level of bias is jointly determined by the amount of baseline imbalance and the magnitude of the baseline-outcome correlation. It would appear that a large imbalance in a

Figure 4.6: Pattern and percentage of directional bias of CSA in relation to the unbiased estimate of ANCOVA at differing trial scenarios



The green dotted line represent the reference line for the unbiased estimate of effect variable with little or no prognostic relationship with the outcome does not cause as much bias as does a small imbalance in a strongly prognostic variable. For example, from Table 4.11, at an effect size of 0.5, small imbalance (z = 1.28) in a strongly prognostic variable (r ≥ 0.7) can have percentage bias in excess of 30%, whereas a large imbalance (z = 1.96) in a variable that has little prognostic relationship (r = 0.1) with the outcome presents with percentage bias that is just about 6%.

Table 4.12 and Figure 4.6 illustrate the percentage bias in the estimate of effect from a change score analysis relative to ANCOVA, given different levels of experimental factors. The results confirm the fact that when treatment groups are balanced at baseline, the estimate of effect using CSA is unbiased, as the percentage bias is approximately zero at each trial scenario. Higher percentage level of bias in the estimate of change is recorded for large imbalance in the same direction as effect. The absolute percentage bias is found to decrease as baseline-outcome correlation increases.

Although, percentage bias on estimate of effect for using CSA is scarcely related to the size of effect to be detected, both size and direction of baseline-imbalance determine bias to a very large extent. The percentage bias here is not symmetrical, considering the direction of imbalance. At a particular effect size, the percentage bias in a trial is significantly different from that obtained in a corresponding trial with imbalance in the opposite direction as treatment effect. For example, at effect size of 0.2 and baseline-outcome correlation of 0.5, the percentage bias for a trial that has a large imbalance in the same direction (–1.96) as effect is approximately 110% whereas the percentage bias in a corresponding trial with large baseline-imbalance in the opposite direction (1.96) as effect is approximately 20%. Generally, the associated bias in the estimate of effect using CSA to account for baseline imbalance in the same direction as the treatment is higher than that obtained when imbalance is in the opposite direction to the treatment.

Table 4.12: Percentage bias in the estimate of effect by change score analysis with reference to the unbiased estimator of effect by ANCOVA

Effect Z	**Levels of baseline-outcome correlations**				
0.2	**0.1**	**0.3**	**0.5**	**0.7**	**0.9**
–1.96	211.67	165.00	109.68	66.67	25.00
–1.64	122.35	95.29	66.67	47.50	14.46
–1.28	72.73	54.55	41.36	24.55	7.27
0	0.00	0.00	0.00	0.00	0.00
1.28	–27.59	–22.18	–16.04	–10.34	–3.45
1.64	–33.75	–26.02	–18.81	–11.95	–2.82
1.96	–38.24	–29.41	–20.59	–12.35	–2.94
0.5					
–1.96	206.62	159.47	113.82	68.42	23.03
–1.64	126.57	97.88	70.53	41.83	13.94
–1.28	75.93	58.67	41.54	25.00	8.46
0	0.00	0.40	0.20	0.40	0.00
1.28	–27.78	–21.92	–16.44	–9.24	–4.11
1.64	–33.08	–25.73	–18.88	–11.28	–3.80
1.96	–36.90	–28.57	–20.71	–12.43	–4.14
0.8					
–1.96	198.00	151.00	108.33	64.82	21.18
–1.64	122.55	95.27	67.50	40.18	13.45
–1.28	74.14	57.18	40.86	24.77	8.64
0	0.50	0.76	0.63	0.13	0.25
1.28	–27.70	–21.74	–15.23	–8.96	–2.70
1.64	–32.75	–25.20	–17.92	–11.20	–3.67
1.96	–36.42	–28.18	–20.09	–12.01	–3.88

When Y = 0.2, n=788; Y=0.5, n=128, Y=0.8, n=52, 1000 iterations in each scenario

A crude effect estimate, while baseline imbalance is in the same direction as treatment, will normally be smaller than the unbiased treatment effect from ANCOVA, and thus the crude estimate will have to be increased by a certain percentage to obtain the unbiased estimate. Similarly, when imbalance is in the opposite direction of the treatment, the biased effect from change score analysis will normally be greater than the unbiased estimate from ANCOVA, and thus the crude estimate will have to be decreased by a certain percentage to obtain the unbiased estimate. The size of treatment effect to be detected, and by implication the size of the trial, does not significantly influence the percentage bias of estimate using change score.

4.5 Discussion

4.5.1 Bias

The results of this simulation study demonstrate that when treatment groups are balanced at baseline, the estimate of effect by each of the methods for statistical analysis of RCTs is unbiased. When balanced at baseline, these methods yield the same estimate of treatment effect across all trial scenarios, except that CSA yields an estimate of effect that, though equal in magnitude to those of ANOVA and ANCOVA, has an opposite sign. This confirms Matthews' (2001) algebraic notations as mentioned in chapter 3, which indicates that using any of these methods yields an unbiased estimate if and only if the two treatment groups are exactly comparable at baseline. This finding also agrees with Senn (1991), who reports that, for groups that are perfectly balanced at baseline, the choice of CSA, ANOVA or ANCOVA makes no difference to the point estimate of treatment effect. In addition, Van Breukelen (2006) argues that if treatment groups are similar by randomization, both CSA and ANCOVA are unbiased. It is therefore a desirable feature of a trial to have the treatment groups comparable at baseline, since under these conditions both adjusted and unadjusted analyses will yield the same point estimate of treatment effect. Pocock et al (2002) believe that there is some credibility attached to demonstrating that covariate adjustment does not alter the conclusion derived from unadjusted analysis. This perhaps explains the huge efforts that researchers sometimes invest in the design of a clinical trial to ensure that treatment groups are the same in prognostic factors. Although views may differ as to their relative appropriateness, those design methods that have been variously used to attain balance in treatment groups in prognostic factors include; minimization, blocking usually combined with stratification and covariate adaptive response adjusted method - CARA (Hagino et al, 2004; Kernan & Makuch, 2001; Rosenberger & Sverdlov, 2008; Scott et al, 2002; Taves, 2004). However, as mentioned earlier in Chapter One, owing to one limitation or another, none of these design methods can ensure a total or complete balance in prognostic factors between the treatment groups.

In fact, the popular option is to include such prognostic factors that would require to be balanced at the design stage in the model for statistical adjustment, regardless of any of the design methods used to ensure balance (Altman &

Bland, 1999; Grimes & Schulz, 2002; Hagino et al, 2004; Minsoo et al, 2008; Roberts & Torgerson, 1998). Scott et al. (2002) and Senn (1995, 1989) contend that treatment groups will often differ with respect to some important prognostic covariate whose influence has proved impossible to control by design alone. Given the findings of this study that a small imbalance in a strongly prognostic baseline score is capable of inflicting a bias of over 30% on the estimate of effect, irrespective of the size of the trial, this is a matter of concern. This particularly strengthens the earlier view that advocates adjusting for imbalance in a prognostic variable despite minimization or stratification, as the case may be.

Moreover, this study has demonstrated that a large imbalance in a non-prognostic variable is not as important as a small imbalance in a strongly prognostic variable. This finding is consistent with Pocock et al (2002). Thus, it is suggested that if baseline-outcome correlation is weak, $\rho < 0.3$, baseline imbalance, irrespective of its level, is immaterial, whereas it is important to adjust for a strongly prognostic variable that manifests even a minimal baseline imbalance. With respect to bias of estimate, there should, however, be no concern whatsoever over a strongly related covariate that is balanced between treatment groups. Given this fact therefore adjusting for a balanced strongly prognostic covariate can be regarded as an extremity and should be avoided. However, since there is no known design method that promises complete balance in prognostic factors, particularly in pragmatic trial environments -and a small baseline imbalance in a strongly prognostic variable can cause major distortion to the estimate of effect, researchers are advised to adjust anyway.

Furthermore, one of the most striking findings of this study is the fact that, given the bias of the estimate of effect, covariate imbalance is just as much a problem for large studies as for small ones. This result corroborates earlier studies by Pocock et al (2002) and Senn (1989), but is in disagreement with Altman (1985) who asserted that, for large trials, any imbalance is likely to be very small and inconsequential. While trying to clarify this statement in a letter to the editor, (Altman & Caroline (1991) Altman appear to declare that his concern was with the bias of estimate of effect in the earlier paper and not any other trial attribute as might have been supposed. It is evidenced from this study that whether the trial is large or small, every little imbalance in a strongly prognostic variable poses the risk of a similar level of bias. There is no added advantage in terms of minimizing bias of the point estimate for studying large sample when in fact a small sample trial is also appropriate. Trial sizes are always in respect of the size of treatment effect to be detected. Large trials do not necessarily protect against biased estimates of effect. They are only necessary in trials that aim at detecting a small effect size, so as to have enough statistical power to do so. It is a matter of ethical concern to expose more patients than is necessary to a treatment whose efficacy has not yet been ascertained.

This study shows that, with respect to the bias in the treatment effect, ANOVA, CSA and ANCOVA all yield unbiased estimates only if the treatment groups are comparable at baseline. This result is consistent with what Matthews (2000) reports. However, when baseline imbalance is apparent, this study has shown that estimates from both ANOVA and CSA cannot be unbiased. Although the estimate of effect using ANOVA is close to that of ANCOVA when the baseline-outcome correlation is weak, ρ ≤ 0.3, as the correlation increases the level of bias in the estimate also increases especially with an increase in the level of imbalance.

Generally, when imbalance exists the level of bias of the point estimate of effect is much higher when using CSA instead of the crude and naïve ANOVA approach. This agrees with the argument by Senn (1991). Considering the level of bias associated with the estimate of effect using CSA, the method cannot be appropriate in dealing with imbalance. Senn (1994) maintains that when the covariate is the baseline measurement of the outcome variable, differences (changes) from baseline are often taken and mistakenly assumed to deal with imbalance. Depending on the level of imbalance and the prognostic relationship, the amount of bias associated with the estimate of effect (small, medium or large) using ANOVA can be very substantial, often in excess of 60%. Camilli & Shepard (1987) had earlier argued that ANOVA will fail to detect absurdly large amounts of bias, this is not surprising as there is no term in the ANOVA model that accounts for baseline imbalance when one exist. Thus it is incapable of detecting and accounting for bias. In this respect, CSA is worse than ANOVA as the level of bias associated with its estimate of treatment effect can be in excess of 15% when correlation (r) equals 0.3, a condition which is possible in empirical trial setting.

Although CSA has the potential to produce an estimate of treatment effect that approximates that of ANCOVA, this can only happen if the baseline- outcome correlation is close or equal to 1. This then shows that the bias in the estimate of effect by both ANOVA and CSA relates to baseline-outcome correlation differently. Bias in estimate of effect by ANOVA reduces as the correlation tends to 0, whereas bias in the estimate of effect by CSA reduces as the correlation tends to 1. In accordance with this finding, Matthews (2000) argues that theoretically, if baseline imbalance exists, CSA will yield a biased estimate of effect unless the correlation is 1. However, for a single variable to have a correlation of 1 with the outcome is highly unlikely in practice.

Lastly, this result also illustrates the directionality associated with the bias in estimate of effect using change scores. With CSA, the estimate is much more biased for trial scenarios that have the baseline imbalance in the same direction as the treatment effect. In practice, CSA cannot be a statistical method for controlling baseline imbalance since baseline - outcome correlation cannot be 1 and because of the issue of directionality associated with its estimate. The findings of this study thus suggest that, if the interest is in the unbiased estimates of effect, then it is erroneous to regard change scores as a method for statistical adjustment, given any level of baseline imbalance.

4.5.2 Precision

In this study, the precision of the estimate of effect for each of the methods varies widely with respect to the size of the trial. For each of the statistical methods, larger trials (relating to a small measured effect size, e.g. 0.2) present with smaller standard errors, and smaller trials are marked by larger standard errors (i.e. less precision), under same experimental conditions. All the methods, unadjusted or adjusted, have the potential to increase precision as the size of the sample increases. This result is consistent with the findings from previous authors (Pocock et al 2002; Altman, 1985; Rubberts & Russo, 2001). Standard error of treatment effect is a measure of the dispersion of the distribution of the estimate of effect in relation to the true effect size. As a result of the way in which it is computed, it depends upon the sample size being studied and the variability of scores in the sample. The larger the number of scores (i.e. the sample size), the smaller the variability within them, the smaller the standard error and the more likely the mean of the population will be more precisely measured.

Design strategies for treatment group allocation of patients in trials, such as stratification, minimization and blocking, are targeted to create uniformity both within and between treatment groups to minimize variability in scores. Thus, irrespective of the method for statistical analysis, there is always a reduction in standard error as sample size increases. This result is however limited to comparison within the same statistical method. Any of the three statistical methods will secure a higher precision of estimate with a larger sample size. Comparatively, this study has shown that the precision benefit of the adjusted analysis by ANCOVA is, however, independent of the sample size. For example, for trials in which baseline imbalance is ignored the ratio of the standard error of ANCOVA against ANOVA reduces markedly with increasing baseline – outcome correlation, irrespective of any change in the size of the trial. The percentage reduction, which is consistent over the various trial sizes, was observed to be as high as 57% for a correlation of 0.9. This means that, regardless of the size of the trial, there is a 57% reduction in the width of the confidence interval by using ANCOVA (as opposed to ANOVA), thus making the estimate of effect more plausible in relation to the true and unknown effect. In fact with this, there is 57% more confidence that the estimate is more accurate than that obtained from using ANOVA.

A reduction of 29% and 4% is observed in the width of confidence interval for correlations of 0.7 and 0.3 respectively. This result is consistent with Pocock et al 2002, who also record a reduction of 29% in the width of the confidence interval for a correlation of 0.7 and less than 10% (number not specified) when the correlation was 0.3.This again confirms the earlier submission, that in terms of the precision of estimate, the benefit of appropriate covariate adjustment is independent of the trial size. Rather, effort should be made to identify covariates that are strongly related with the outcome and adjust for such using ANCOVA even if the trial is balanced. Unless the correlation is greater than 0.3, there is no reason to consider statistical adjustment over the unadjusted analysis for a higher precision of the estimate of effect. The precision benefit of the adjusted analysis over the unadjusted has also been reported by Tsiatis et al (2007) and Wang and Hung (2005).

Furthermore, the results have demonstrated that statistical adjustment by ANCOVA of a strongly prognostic factor with large baseline imbalance yields the same results as adjustment for a strong prognostic factor with little or no imbalance. In this study, a ‘large’ trial comprises 788 hypothetical patients. Altman (1985) indicates a ‘large trial’ to mean n>500. In fact, following statistical adjustment, the difference between the precision of estimate of effect for large trials and small trials, irrespective of the size of baseline imbalance and baseline-outcome correlation is not beyond that which could have been explained by chance. Similarly, the difference in precision, following statistical adjustment by ANCOVA, of a trial with homogeneous treatment group at baseline and that with imbalance in a very strongly prognostic variable, does not also exceed that which could have been due to chance.

Thus, this simulation study has shown that with reference to the relative precision of estimate of effect, neither the size of the trial nor the size and direction of baseline imbalance in the treatment groups is important. This does not however seek to underestimate the usefulness of such strategies as blocking, stratification and minimization that are used for ensuring balance in treatment groups. They may seem not to add any extra value to the precision of estimate of effect so far the imbalance is accounted for by statistical adjustment, they do possibly have for other attributes as will be seen later. So, when any of such design methods as stratification or minimisation is used to make treatment groups similar in selected baseline covariates, the estimate of effect that results from using appropriate covariate adjusted method then is not more precise than the estimate which results from using the same appropriate covariate adjusted method following simple randomisation. The driver of the difference in precision of effect estimate between the ANCOVA adjusted and the unadjusted analysis is the degree of prognostic relationship between the baseline covariate and the outcome. The baseline-outcome correlation is also influential in relation to the contrast in precision between ANCOVA and CSA.

4.6 Conclusion

Generally, regarding precision, ANCOVA yields the most precise estimate of effect of the three methods and should always be used. When baseline imbalance exists in a strongly prognostic (r>0.3) variable, for example baseline values of the outcome variable, the estimate of effect from ANOVA presents with a less precise estimate compared to that of ANCOVA. However, the estimate from change score analysis is only more precise than that of ANOVA if the baseline-outcome correlation is greater than 0.5. The point estimate of effect from change score analysis is susceptible not only to the prognostic relationship between the baseline and outcome but also to both the size and the direction of baseline imbalance. The imbalance, and particularly the baseline-outcome correlation, is influential in distinguishing ANCOVA as the unbiased approach relative to unadjusted ANOVA.

Chapter 5: Statistical methods of analysis of RCTs with or without baseline imbalance: Implications on statistical power and trial sample size – efficiency

5.1 Introduction

In this chapter, results will be presented on the statistical power of methods for the analysis of randomised controlled trials at differing combinations of levels of treatment effect, baseline-outcome correlation and both levels and directions of imbalance. The effect of adjusted analysis on sample size requirements at various hypothetical trial scenarios is also presented. The results of the simulations demonstrate the effect of different directions and levels of baseline imbalance, levels of baseline-outcome correlation, and levels of treatment effect on the statistical power and sample size requirement of each of the methods of analysis of randomised controlled trials: ANOVA, ANCOVA and CSA (change score analysis). The first three sections are exclusively on the power of the statistical methods. The later sections provide results on relative sample size and efficiency of each of the methods at 80% power, with analysis of variance (ANOVA) as the reference unadjusted method.

The relative sample size of each of the methods for adjusted analysis was determined in relation to 80% nominal power. The chapter will present the difference in sample size required by each of the methods for statistical adjustment in relation to the original sample size. The original sample size represents the sample size for the unadjusted analysis, and percentage difference in sample size for each of the methods was computed and represented graphically. In this study, relative difference or percentage difference in sample size of each of the methods for statistical adjustment in relation to the original sample size are measures of the statistical efficiency of the methods in absolute and percentage terms, respectively. The method that requires the smallest sample size at 80% power is the most efficient and that which requires the largest sample size at the same 80% power to detect a specified effect size is the least efficient. Results are presented in tables and charts. In the tables, the hypothetical trial datasets for the information presented in the rows vary in the levels of baseline-outcome correlation, whereas, in the columns, the trial datasets vary in the levels of imbalance. Multiple scatter diagrams with interpolation lines fitted are used to represent statistical power and percentage efficiency of methods in different hypothetical trial scenarios.

As was mentioned in chapter 4, the primary concern in this thesis centres around outcome scales whereby treatment effect implies a decrease in baseline score of the outcome variable; for example, pain, depression and anxiety scales where low scores denote less health problems and high scores denote greater health problems. It is noted that at other times, however, it might mean an increase in baseline score of the outcome variable; examples of this are common generic quality of life scales and pain relief scales (though these are not the focus in this chapter). Although, this simulation study covers hypothetical trial scenarios in which treatment effect implies either decreasing or increasing baseline scores of outcome variables, the results in this section exclude those scenarios where increasing in baseline score of the outcome variable implies treatment effect.

Also, all results in this chapter are presented on experimental conditions at 80% nominal power; all other results in graphs and tables for 90% nominal power and when increasing in baseline scores implies treatment effect, are attached as appendices.

5.2 Statistical power of methods of RCTs at differing levels and direction of baseline imbalance

In the next three subsections, the effect of levels of groups' baseline heterogeneity at baseline on the statistical methods of ANOVA, CSA, and ANCOVA at different trial scenarios is investigated and results presented.

5.2.1 Statistical power of methods of analysis of RCT when groups are homogeneous

In this subsection, results on statistical power of the statistical methods being studied (ANOVA, ANCOVA and CSA) are presented. Table 5.1 below shows the statistical power of the methods at various effect sizes; small (0.2), medium (0.5) and large (0.8). The design in this section ignores the possibility of baseline imbalance on the basis that randomization produces similar treatment groups at baseline. Thus, the assumption is that there is no baseline difference, for example, treatment groups start with comparable or similar level of pain score. Here, baseline severity of pain does not influence treatment effect, as it is assumed to be balanced in the two groups.

From Table 5.1, level of treatment effect does not influence the power of the statistical test. Across levels of baseline-outcome correlation, the observed statistical power for ANOVA in the simulation approximates the nominal power of the study. However, especially as the level of effect increases, there appears to be a slight increase in the observed statistical power for ANOVA against the nominal power of 80%. As explained in the previous chapter, this slight deviation is likely to be a result of random fluctuation which is more pronounced for smaller sample sizes.

Table 5.14: Power (in percentage) of statistical methods for homogeneous trials at levels of baseline-outcome correlation (r) and differing levels of effect

Methods/ Effects	**Levels of baseline-outcome correlation**				
	0.1	**0.3**	**0.5**	**0.7**	**0.9**
ANOVA					
Small (0.2)	79.2	80.2	78.7	79.7	80.8
Medium(0.5)	80.3	79.3	80.2	80.1	80.1
Large(0.8)	81.0	81.7	81.9.	82.8	82.6
CSA					
Small(0.2)	57.2	68.6	80.7	95.0	>99.9
Medium(0.5)	56.7	66.3	79.9	95.2	>99.9
Large(0.8)	54.3	65.6	80.5	95.2	>99.9
ANCOVA					
Small(0.2)	79.8	82.2	89.0	96.8	>99.9
Medium(0.5)	80.0	83.4	90.1	97.5	>99.9
Large(0.8)	79.7	82.9	89.7	97.4	>99.9

When Y = 0.2, n=788; Y = 0.5, n=128; Y = 0.8, n=52, 1000 iterations in each scenario

As mentioned earlier, when the difference to detect is large (that is, Y=0.8), such that sample size to be studied is relatively small, it appears ANOVA slightly exceeds the nominal power that was set for the study. Again, this occurrence can be explained by random fluctuations in the simulations, the possibility of which is higher when the sample size is small, as is the case here. Small samples are more prone to higher variability in the sampling distribution of an identified population attribute. Another round of simulations in which a different seed (150) was used in the statistical program yielded a power for ANOVA that was closer to the 80% nominal power. This also presented a corresponding reduction in the observed power for both ANCOVA and CSA.

The power of the adjusted analysis increases with baseline-outcome correlation. For ANCOVA this increases upward from the nominal power (when the correlation is low) toward 100% (when the correlation is high). The power for CSA, though low at baseline-outcome correlation below 0.5, is similar to that of ANCOVA as covariate-outcome correlation exceeds 0.7. By and large, when covariate-outcome correlation is greater than or equal to 0.3, across all levels of experimental conditions, ANCOVA is most powerful. It is also evidenced from table 5.1, that large trials benefit as much as small trials from statistical adjustment when baseline-outcome correlation is sufficiently large, as a similar level of power is attained when appropriate statistical adjustment is applied. In most of the trial scenarios represented in this simulated study in which treatment groups are balanced in baseline score, ANCOVA is most powerful. Under the same trial conditions and given that groups are homogeneous, each of the methods yields approximately same level of statistical power across different levels of sample size, thus sample size or level of effect does not play any role in the statistical power of these methods. Although not shown in the table, further results in the appendix 2 confirm that when treatment groups are comparable at baseline, whether treatment effect implies an increase or a decrease in baseline score does not impact on the power of these statistical methods of analysis of RCTs.

5.2.2 Statistical power of methods of RCTs when groups are heterogeneous

When treatment groups are not comparable at baseline, the directions and levels of imbalance and the relative prognostic importance of the baseline difference play an important role in the statistical power of the various methods of analysis of a randomised controlled trial. The next subsections illustrate how statistical power of ANOVA, CSA and ANCOVA differs across levels and directions of baseline imbalance, baseline-outcome correlation and treatment effect.

5.2.2.1 Statistical power of methods of RCTs at levels of baseline imbalance in the same direction as treatment effect

The power of ANOVA, as observed from (Tables 5.2–5.4) does not respond to change in the level and direction of baseline imbalance and other experimental factors (baseline-outcome correlation and effect size). This is because there is no term in the ANOVA model that takes account of imbalance in baseline when one exists. Thus, the level and direction of the prognostic relationship of the covariate with the outcome is immaterial with ANOVA. The power of ANOVA essentially reflects the nominal power of the trial, which is unconditional on the disparity that may exist in baseline scores between treatment groups. Thus, irrespective of the direction of imbalance and the level of prognostic relationship (correlation) between the baseline and outcome, the power of ANOVA in the simulation approximates the nominal power of the study.

Table 5.15: Power (in percentage) of statistical methods at levels of baseline-imbalance (Z) in the same direction as effect and baseline-outcome correlation [Treatment effect size Y = 0.2; n=788]

	Levels of covariate-outcome correlations				
Methods (Z)	**0.1**	**0.3**	**0.5**	**0.7**	**0.9**
ANOVA					
1.28	79.2	80.2	78.7	79.7	80.8
1.64	79.2	80.2	78.7	79.7	80.8
1.96	79.2	80.2	78.7	79.7	80.8
CSA					
1.28	21.5	27.2	35.1	53.2	91.2
1.64	14.6	17.4	23.1	33.0	74.6
1.96	8.9	11.3	13.6	20.6	51.0
ANCOVA					
1.28	77.4	73.9	71.8	76.8	96.1
1.64	76.4	69.1	64.1	64.6	84.6
1.96	75.5	64.8	57.8	53.3	67.8

Y is the simulated effect size and Z is the standardized imbalance, 1000 iterations in each scenario

Table 5.16: Power (in percentage) of statistical methods at levels of baseline-imbalance (Z) in the same direction as effect and baseline-outcome correlation [Treatment effect size Y = 0.5; n=128]

	Levels of covariate-outcome correlations				
Methods (Z)	**0.1**	**0.3**	**0.5**	**0.7**	**0.9**
ANOVA					
1.28	80.3	79.3	80.2	80.1	80.1
1.64	80.3	79.3	80.2	80.1	80.1
1.96	80.3	79.3	80.2	80.1	80.1
CSA					
1.28	22.4	27.5	34.7	51.3	92.9
1.64	15.6	18.8	24.1	34.5	74.1
1.96	11.1	12.6	15.5	22.4	48.5
ANCOVA					
1.28	76.0	70.9	70.0	75.8	96.2
1.64	74.3	67.1	63.3	64.5	85.5
1.96	73.0	64.1	56.4	52.1	66.8

Y is the simulated effect size and Z is the standardized imbalance, 1000 iterations in each scenario

Table 5.17: Power (in percentage) of statistical methods at levels of baseline-imbalance (Z) in the same direction as effect and baseline-outcome correlation [Treatment effect size Y = 0.8; n=52]

	Levels of covariate-outcome correlations				
Methods (Z)	**0.1**	**0.3**	**0.5**	**0.7**	**0.9**
ANOVA					
1.28	81.0	81.7	81.9	82.8	82.6
1.64	81.0	81.7	81.9	82.8	82.6
1.96	81.0	81.7	81.9	82.8	82.6
CSA					
1.28	22.7	26.2	34.0	51.5	94.4
1.64	15.0	18.6	23.2	34.1	76.8
1.96	10.7	12.4	15.8	21.7	51.8
ANCOVA					
1.28	74.6	70.5	69.2	75.0	96.2
1.64	73.0	64.2	60.0	61.8	86.8
1.96	71.0	60.3	52.9	49.8	65.8

Y is the simulated effect size and Z is the standardized imbalance, 1000 iterations in each scenario

When there is baseline imbalance the statistical power of CSA and ANCOVA respond to this disparity in baseline scores. There is a shift from the nominal power of the study as it is now conditional on baseline imbalance; both the level and direction of baseline imbalance. In the two methods of statistical adjustment, the higher the imbalance the lower the conditional power recorded (for scenarios where the imbalance is in the same direction as the treatment effect, that is, the treated group has a better prognosis at baseline – a lower pain score than the control group). The result shows that with the imbalance in the same direction as treatment, the conditional power of CSA is very poor, especially when baseline-outcome correlation is low and the imbalance is very high (e.g. power can be as low as 10% when baseline imbalance is very high (z = 1.96) and baseline-outcome correlation is very low (r = 0.1). The higher the level of imbalance the lower the statistical power of CSA, and the greater the likelihood of false negatives (type II error) – the situation in which the test fails to detect the treatment effect when it exists. The level of treatment effect does not seem to have a marked impact on the power output of either CSA or ANCOVA.

When imbalance is in the same direction as effect, the pattern of conditional power of ANCOVA is somewhat different from that of CSA. The low power of ANCOVA at this instance of baseline imbalance being in the same direction as the treatment effect has to do with the way ANCOVA adjusts for imbalance; this was mentioned in chapter 4. In most trial conditions, across increasing levels of baseline-outcome correlation, the conditional power of ANCOVA decreases until a baseline-outcome correlation of 0.7, where it starts to increase. In this study, this attribute is observed when imbalance is either low or at medium level.

With imbalance in the same direction as treatment effect, the conditional power for ANCOVA decreases as imbalance increases, irrespective of the level of baseline-outcome correlation. This therefore leads to the possibility of an unadjusted analysis by ANOVA indicating a significant difference in effect whereas the adjusted analysis by ANCOVA fails to detect such a difference. The possibility of this occurring is high when there is a medium or large baseline imbalance in the same direction as effect in a strongly prognostic factor (in this case, the baseline value of the outcome variable). For example, in Table 5.3, when the baseline-outcome correlation is 0.7, for both medium and large imbalance, the statistical power to detect a treatment difference by ANCOVA are 64.5 and 52.1 respectively, whereas, that of ANOVA stood at 80.1. The statistical power for CSA then is 34.5 and 22.4 respectively.

5.2.2.2 Statistical power of methods of RCTs at levels of baseline imbalance in the opposite direction to effect

By contrast, when baseline imbalance is in the opposite direction to treatment effect – that is, the control group has a higher average score and better prognosis at baseline – the pattern of conditional power by the methods for statistical adjustment differs from what was observed in the previous section. Here, though there is little or no change in the unconditional power by ANOVA despite change in direction of an imbalance, there is an obvious increase in the statistical power of the adjusted analyses (CSA and ANCOVA). From tables 5.5–5.7, both ANCOVA and CSA give an increase in power from the nominal 80%. CSA demonstrates a higher probability to detect treatment effect compared to ANCOVA particularly when the baseline-outcome correlation is low (r = 0.1), though both methods have about 99% power or higher when the correlation is in excess of 0.5. Unlike the data for an imbalance in the same direction as the treatment effect, statistical power for CSA increases with an increase in level of imbalance and prognostic relationship of the baseline and outcome. ANCOVA is not dependent on the size of the baseline imbalance but is influenced by the size of the baseline-outcome correlation. Figure 5.1 illustrates the simulated power of ANOVA, CSA and ANCOVA across the different hypothetical trial scenarios.

Table 5.18: Power (in percentage) of statistical methods at levels of baseline-imbalance (Z) in the opposite direction of effect and baseline-outcome correlation [Treatment effect size Y= 0.2; n= 788]

	Levels of baseline-outcome correlations				
Methods (Z)	**0.1**	**0.3**	**0.5**	**0.7**	**0.9**
ANOVA					
1.28	79.2	80.2	78.7	79.7	80.8
1.64	79.2	80.2	78.7	79.7	80.8
1.96	79.2	80.2	78.7	79.7	80.8
CSA					
1.28	85.6	92.7	98.3	99.9	>99.9
1.64	90.5	96.0	99.3	>99.9	>99.9
1.96	93.8	97.9	99.7	>99.9	>99.9
ANCOVA					
1.28	82.8	90.6	97.3	99.7	>99.9
1.64	83.7	92.1	98.3	99.9	>99.9
1.96	84.2	93.8	99.0	>99.9	>99.9

Y is the simulated effect size and Z is the standardized imbalance, 1000 iterations in each scenario

Table 5.19: Power (in percentage) of statistical methods at levels of baseline-imbalance (Z) in the opposite direction of effect and baseline-outcome correlation [Treatment effect size Y= 0.5; n=128]

	Levels of baseline-outcome correlations				
Methods (Z)	**0.1**	**0.3**	**0.5**	**0.7**	**0.9**
ANOVA					
1.28	80.3	79.3	80.2	80.1	80.1
1.64	80.3	79.3	80.2	80.1	80.1
1.96	80.3	79.3	80.2	80.1	80.1
CSA					
1.28	85.1	92.6	98.7	>99.9	>99.9
1.64	90.0	96.0	99.2	>99.9	>99.9
1.96	97.5	98.2	99.8	>99.9	>99.9
ANCOVA					
1.28	83.4	91.3	97.5	>99.9	>99.9
1.64	84.1	93.1	98.6	>99.9	>99.9
1.96	84.4	94.1	98.8	>99.9	>99.9

Y is the simulated effect size and Z is the standardized imbalance, 1000 iterations in each scenario

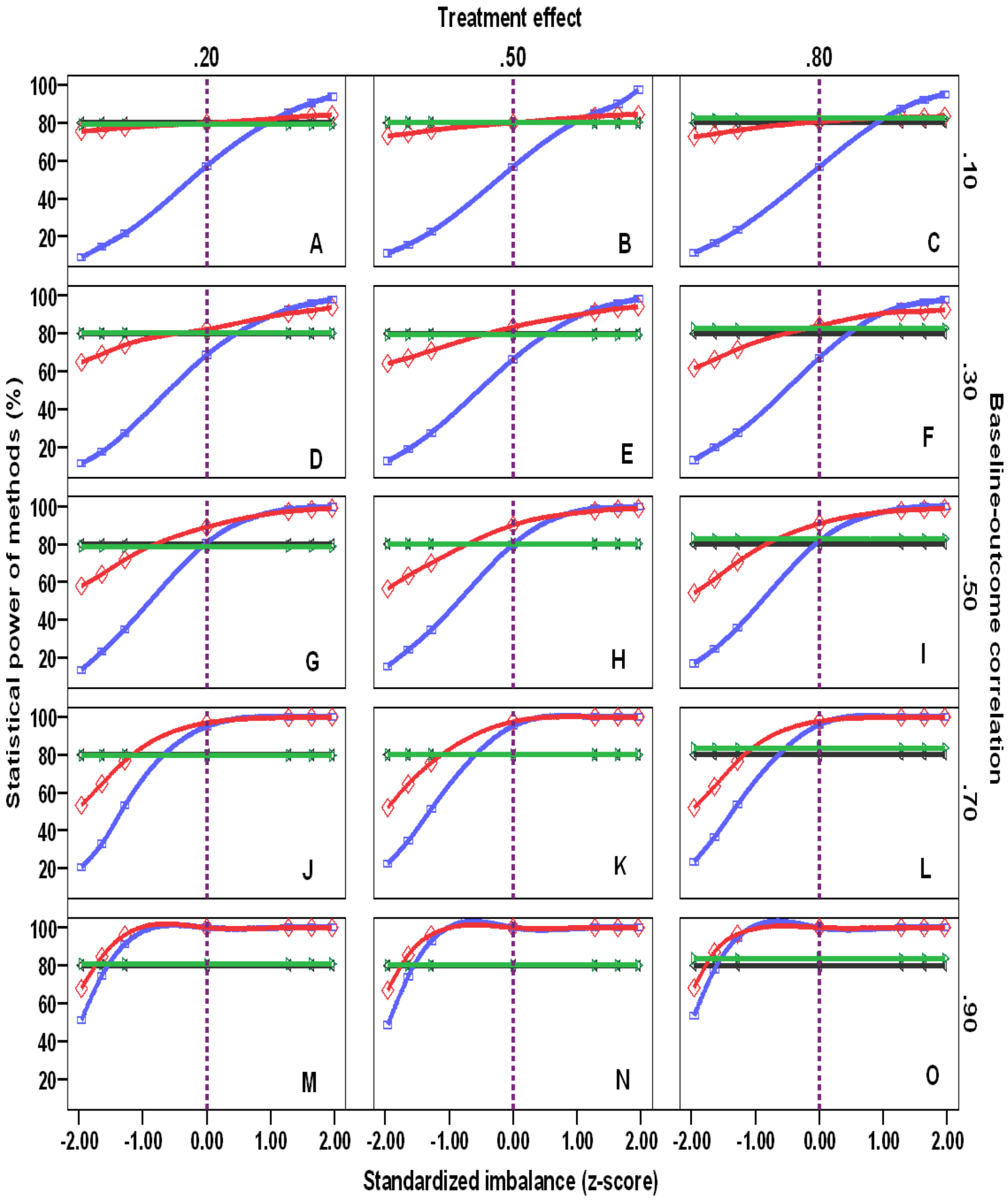
Table 5.20: Power (in percentage) of statistical methods at levels of baseline-imbalance (Z) in the opposite direction of effect and baseline-outcome correlation [Treatment effect size Y= 0.8; n=52]

	Levels of baseline-outcome correlations				
Methods (Z)	**0.1**	**0.3**	**0.5**	**0.7**	**0.9**
ANOVA					
1.28	81.0	81.7	81.9	82.8	82.6
1.64	81.0	81.7	81.9	82.8	82.6
1.96	81.0	81.7	81.9	82.8	82.6
CSA					
1.28	86.6	93.7	98.3	99.9	>99.9
1.64	91.7	96.2	99.8	99.9	>99.9
1.96	94.9	98.0	99.9	>99.9	>99.9
ANCOVA					
1.28	81.9	90.7	97.6	99.8	>99.9
1.64	81.7	91.4	98.1	99.9	>99.9
1.96	82.2	91.9	98.7	99.9	>99.9

Y is the simulated effect size and Z is the standardized imbalance, 1000 iterations in each scenario

Figure 5.1 illustrates how the statistical power of CSA and ANCOVA responds to levels of baseline-outcome correlation, levels of treatment effect, and directions and levels of baseline imbalance. In all cases, the simulated power for ANOVA approximately equals the nominal power, though the simulated power shows a very slight increase as effect size increases (due to wider random fluctuation dispersion for smaller trials). Statistical power for ANOVA (simulated) is non-responsive to change in level of baseline-outcome correlation and both direction and level of baseline-outcome imbalance.

Figure 5.8: Directional pattern of statistical power (%) of methods for analysis of RCT at levels of baseline imbalance, baseline-outcome correlation and differing level of treatment effect



KEY: ___ ANOVA nominal; ___ANOVA simulated; _____ CSA; ___ANCOVA
When Y = 0.2, n=788; Y = 0.5, n=128; Y = 0.8, n=52

Also, Figure 5.1 demonstrates how adjusted analyses are influenced by baseline imbalances and baseline-outcome correlations, but not the treatment effect size. Power is adversely affected for these approaches when the baseline-imbalance is in the same direction as the treatment effect, and is particularly problematic for CSA when the baseline-outcome correlation is low (e.g. $r \leq 0.3$). By contrast, the power of CSA is inflated when the baseline-imbalance is in the opposite direction to the treatment effect. Under such imbalances in the opposite direction to the effect CSA demonstrates greater power, and hence a higher potential to detect treatment effect than ANCOVA, when the baseline-outcome correlation is low ($r \leq 0.5$), whereas the two methods of analysis have roughly equal statistical power at higher level of baseline-outcome correlation ($r \geq 0.\ 5$).

The graph also shows the disparities in statistical power of these methods even when treatment groups are balanced in baseline outcome scores. As above, the disparity is independent of the size of treatment effect, and is therefore driven by the level of baseline-outcome correlation. This phenomenon parallels the way baseline-outcome correlation drives the precision of these statistical methods at different trial scenarios when groups are homogeneous. For randomised controlled trials in which groups are homogeneous, with nominal power at 80%, depending on the baseline-outcome correlation the conditional power for CSA ranges between about 54% and >99.9%, whereas, that of ANCOVA ranges between 80% and >99.9%. Thus, Figure 5.1 shows that if treatment groups are balanced the statistical power of ANOVA, though it may equal, cannot exceed that of ANCOVA. Whilst appreciating the need for statistical adjustment to overcome the issue of bias, the illustration highlights the need to be alert to the impact of the size and direction of imbalances and the magnitude of the baseline-outcome correlation on the conditional power of the statistical test (as it is the conditional power and not the nominal power that is the true basis of the test).

5.3 Efficiency or relative sample size of the statistical methods

The results on the ratios of the standard error (in chapter 4, tables 4.8–4.10) gave an insight into the number of patients that must be studied by these statistical methods at statistical power of 80%. This section is concerned with presenting the relative sample size for both ANCOVA and CSA in relation to the original sample size across different hypothetical trial scenarios. Efficiency, observable from the relative sample size for each trial scenario here, implies the level of reduction in the original sample size at the same level of power of 80%, which is brought about by using either of the methods for statistical adjustment instead of the unadjusted analysis (the reference). Also, results comparing the efficiency between methods of adjusted analysis (ANCOVA and CSA) are also presented (with ANCOVA the reference). The results of the simulations in the tables 4.8 to 4.10 had earlier shown that the ratio of the standard error of all the methods is exclusively dependent on the level of baseline-outcome correlation. These results are explored to determine their implications for the relative sample size requirement by each of these three methods.

Firstly, given that all three methods relate to the same standard deviation of original scores, at 80% power, it follows from section 4.2 that the ratio of the standard error of ANCOVA to ANOVA at baseline-outcome correlation of 0.7 in relation to equation (3.31) in chapter 3 for example, is given as;

$$\frac{\text{SE ANCOVA}}{\text{SE ANOVA}} = \frac{\text{SD}}{\sqrt{n'}} / \frac{\text{SD}}{\sqrt{n}}$$

$$= \frac{\text{SD}}{\sqrt{n'}} \text{X} \frac{\sqrt{n}}{\text{SD}} = 0.71$$

$$0.71\sqrt{n'} = \sqrt{n}$$

$$(0.71)^2 n' = n$$

$$0.5041 n' = n$$

where n' represents the sample size for ANCOVA and n the original sample size.

However, in this context, both n' and n are equal since the same sample size was originally used for all the three methods at each trial scenario. Therefore, the above mathematical expressions show that ANCOVA has a possibility of reducing the original sample by half, if the baseline-outcome correlation is 0.7. Following this reasoning, the results presented in subsequent tables in this section were derived, representing the relative reduction in the original sample size by taking the square of the ratio of the standard errors of ANCOVA and ANOVA at different hypothetical trial scenarios. These results of the simulated datasets illustrate the algebraic expression $(1-\rho^2)$labelled as equation (3.32) in chapter 3.

Table 5.21: Relative reduction in original sample size for using ANCOVA instead of ANOVA in different trial scenarios

Effect Z	**Levels of baseline-outcome correlations**				
0.2	**0.1**	**0.3**	**0.5**	**0.7**	**0.9**
-1.96	1.02	0.94	0.74	0.51	0.18
-1.64	1.03	0.94	0.76	0.53	0.20
-1.28	1.00	0.94	0.76	0.53	0.18
0	1.00	0.92	0.73	0.51	0.18
1.28	1.00	0.94	0.73	0.51	0.18
1.64	1.00	0.94	0.73	0.51	0.18
1.96	1.00	0.94	0.73	0.51	0.18
0.5					
-1.96	1.02	0.94	0.77	0.52	0.19
-1.64	1.01	0.93	0.77	0.52	0.19
-1.28	1.01	0.92	0.76	0.51	0.19
0	1.00	0.92	0.75	0.50	0.19
1.28	1.06	0.87	0.69	0.49	0.18
1.64	1.01	0.93	0.73	0.52	0.19
1.96	1.02	0.94	0.73	0.52	0.19
0.8					
-1.96	1.09	1.00	0.81	0.56	0.20
-1.64	1.04	0.98	0.81	0.55	0.21
-1.28	1.04	0.96	0.79	0.54	0.20
0	1.01	0.92	0.75	0.52	0.20
1.28	1.02	0.95	0.79	0.52	0.20
1.64	1.06	0.97	0.80	0.55	0.21
1.96	1.08	0.99	0.81	0.56	0.21

When Y = 0.2, n=788; Y=0.5, n=128, Y=0.8, n=52, 1000 iterations in each scenario

Table 5.8 above shows the reduction in the original sample size that needs to be studied if statistical adjustment by ANCOVA is preferred to ANOVA at different trial scenarios, maintaining 80% power. Remarkable reduction in the original sample size that is independent of both level and direction of imbalance was observed to change with levels of baseline-outcome correlation. For example, this simulation results demonstrate a reduction of 9% if baseline-outcome correlation is 0.3; whereas a 50% reduction in sample size was achieved with adjusting for a variable that has a baseline-outcome correlation of 0.7. Reduction in sample size exceeds 80% when the baseline-outcome correlation is 0.9 or higher.

Figure 5.9: Relative sample size for using ANCOVA instead of ANOVA at differing level of baseline-outcome correlation

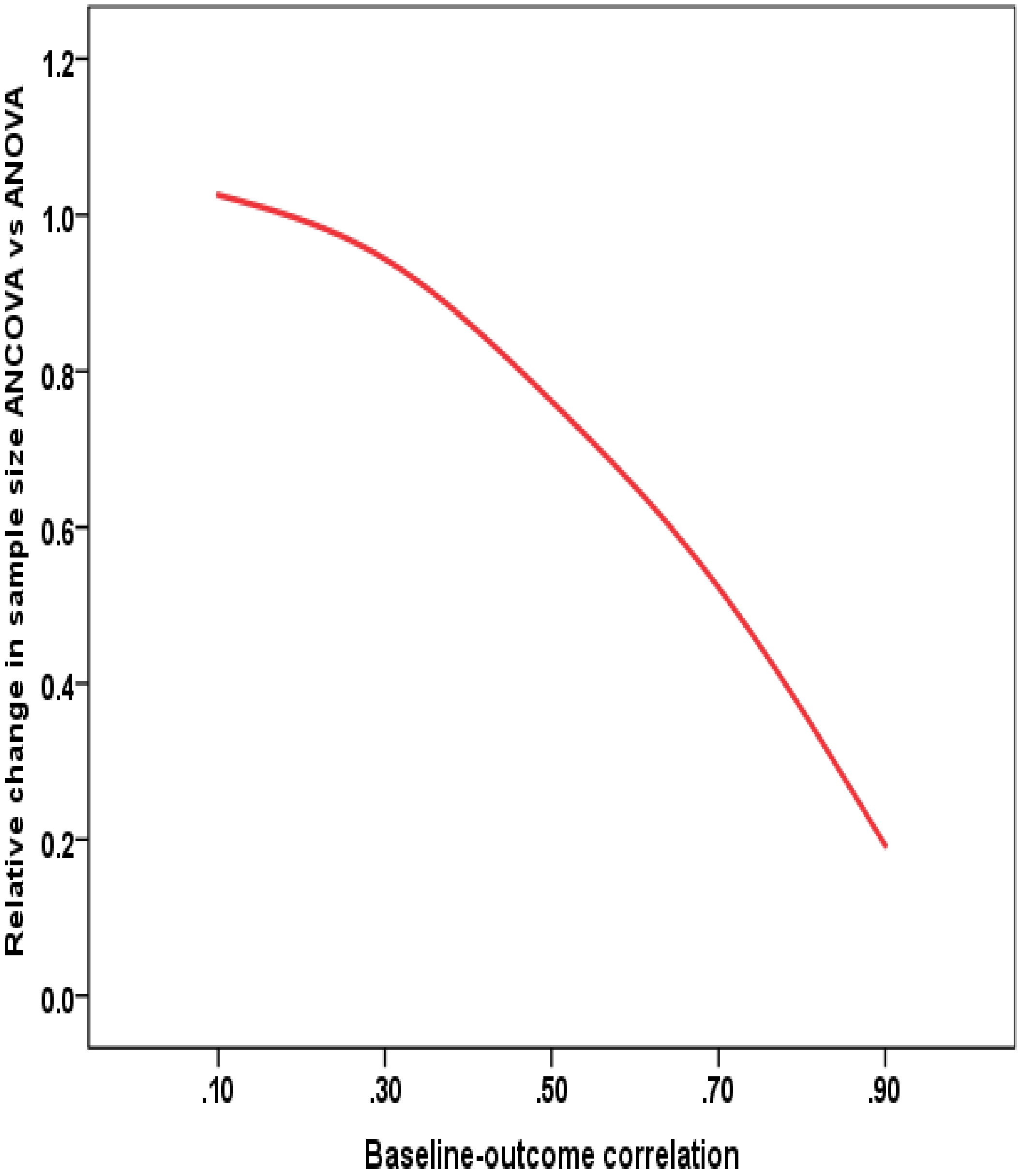
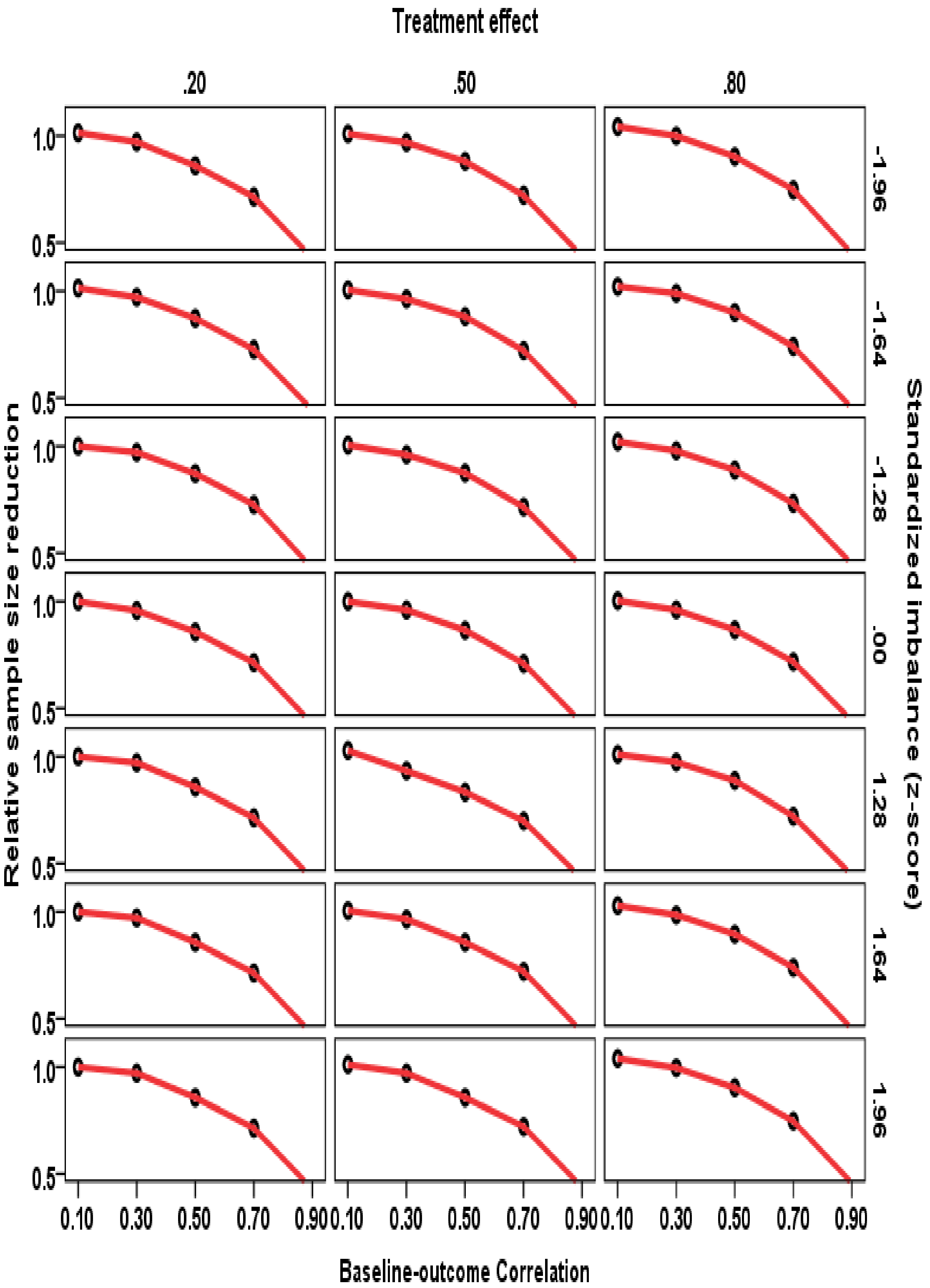


Figure 5.10: Relative reductions in sample size for using ANCOVA instead of ANOVA at differing trial scenarios



However, it may be impractical to find a variable that will have such level of prognostic relationship with the outcome. Clearly, the reduction in the required total sample

size for ANCOVA is compounded by the level of the baseline-outcome correlation as the algebraic expression includes a quadratic term for the correlation. Clearly, the size of the trial and level or direction of baseline imbalance does not matter as the efficiency is dictated by the baseline-outcome correlation only. Figure 5.2 illustrates the relationship between the reduction in sample size and the baseline-outcome correlation (without regard to the size of treatment effect and both levels and direction of imbalance – since these are not influential in this regard). Figure 5.3 however regards treatment effect and both size and direction of baseline imbalance.

The absolute size of the sample that needs to be studied in different scenarios using ANCOVA instead of ANOVA is presented in table 5.9. So, the sample size requirement of the trial (if carrying out ANCOVA analysis as opposed to ANOVA) is not greatly different when the baseline-outcome correlation is small, but a sizeable difference is noted when the baseline-outcome correlation is large. For example for detecting a treatment effect size of 0.5, when the correlation is 0.3 the required sample size for ANCOVA is 9% less (i.e. n=116 instead of 128), but when it is 0.7 the required sample size is 49% less (i.e. n=66 instead of 128), and for a correlation of 0.9 the required sample size is 81% less (i.e. n=24 instead of 128).

Table 5.22: Adjusted sample sizes for using ANCOVA instead of ANOVA in different trial scenarios

Effect Z	**Levels of baseline-outcome correlations**				
0.2	**0.1**	**0.3**	**0.5**	**0.7**	**0.9**
-1.96	811	744	582	403	145
-1.64	811	744	599	417	155
-1.28	788	744	599	417	145
0	788	722	579	402	145
1.28	788	744	579	402	145
1.64	788	744	579	402	145
1.96	788	744	579	402	145
0.5					
-1.96	131	121	100	67	25
-1.64	130	119	100	67	25
-1.28	130	118	98	66	25
0	128	118	96	65	24
1.28	136	112	89	63	24
1.64	130	120	94	67	25
1.96	131	121	94	67	25
0.8					
-1.96	57	52	43	29	11
-1.64	55	51	42	29	11
-1.28	55	50	41	28	11
0	53	50	39	27	11
1.28	54	50	41	27	11
1.64	55	51	42	29	11
1.96	57	52	43	29	11

Y=0.2, n=178; Y=0.5, n=128; Y=0.8, n=52, 1000 iterations in each scenario

Table 5.10 shows the proportion of the original sample size when using CSA instead of ANOVA. Here again, the only influential factor is the level of prognostic relationship between the baseline variable and the outcome. Levels of treatment effect or size of the trial and both levels and directions of imbalance have no impact on changes in sample size. Table 5.10 was derived by taking the ratio of the standard error of CSA versus ANOVA at different hypothetical trial scenarios in the simulated results. The changes are almost symmetrical at both sides of correlation of 0.5. For a baseline-correlation less than 0.5, the relative sample size for CSA is in excess of 1. For example, at a correlation of 0.3, the required sample size for CSA will be 31% in excess of the original sample size, whereas at correlation of 0.7, using CSA will save 38% of the original sample size.

Table 5.23: Relative sample size for using CSA instead of ANOVA at different trial scenarios

Effect Z	**Levels of baseline-outcome correlations**				
0.2	**0.1**	**0.3**	**0.5**	**0.7**	**0.9**
-1.96	1.84	1.31	1.00	0.62	0.18
-1.64	1.84	1.45	1.00	0.62	0.20
-1.28	1.84	1.31	1.00	0.62	0.18
0	1.84	1.31	1.00	0.62	0.18
1.28	1.84	1.31	1.02	0.73	0.20
1.64	1.87	1.44	1.00	0.62	0.18
1.96	1.84	1.31	1.00	0.62	0.18
0.5					
-1.96	1.80	1.40	1.00	0.60	0.20
-1.64	1.80	1.40	1.00	0.59	0.20
-1.28	1.80	1.40	1.00	0.60	0.20
0	1.80	1.41	1.00	0.60	0.20
1.28	1.80	1.36	1.00	0.56	0.19
1.64	1.80	1.40	1.00	0.59	0.20
1.96	1.80	1.40	1.00	0.59	0.20
0.8					
-1.96	1.80	1.40	0.99	0.60	0.19
-1.64	1.75	1.40	0.99	0.60	0.20
-1.28	1.80	1.40	1.00	0.57	0.20
0	1.80	1.40	0.99	0.60	0.20
1.28	1.80	1.39	0.99	0.57	0.20
1.64	1.79	1.39	0.99	0.57	0.20
1.96	1.78	1.39	0.99	0.57	0.20

Y=0.2, n=178; Y=0.5, n=128 ; Y=0.8, n=52, 1000 iterations in each scenario

When baseline-outcome correlation is 0.5, at all trial scenarios the sample size required for CSA equals that of the ANOVA. At the extremes of the levels of correlation (0.1or 0.9) there is 80% increase or decrease in the original required sample size respective of the trial size or effect size to be detected at nominal

80% power. Figures 5.4 and 5.5 illustrate the proportional changes in the original sample sizes for using CSA instead of ANOVA, given different trial scenarios. The two figures show the inverse relationship that exists between the proportion of change in the original trial sample units that have to be studied and baseline-outcome correlation for using CSA.

Figure 5.11: Relative changes in the original sample size for using CSA instead of ANOVA at differing level of baseline-outcome correlation

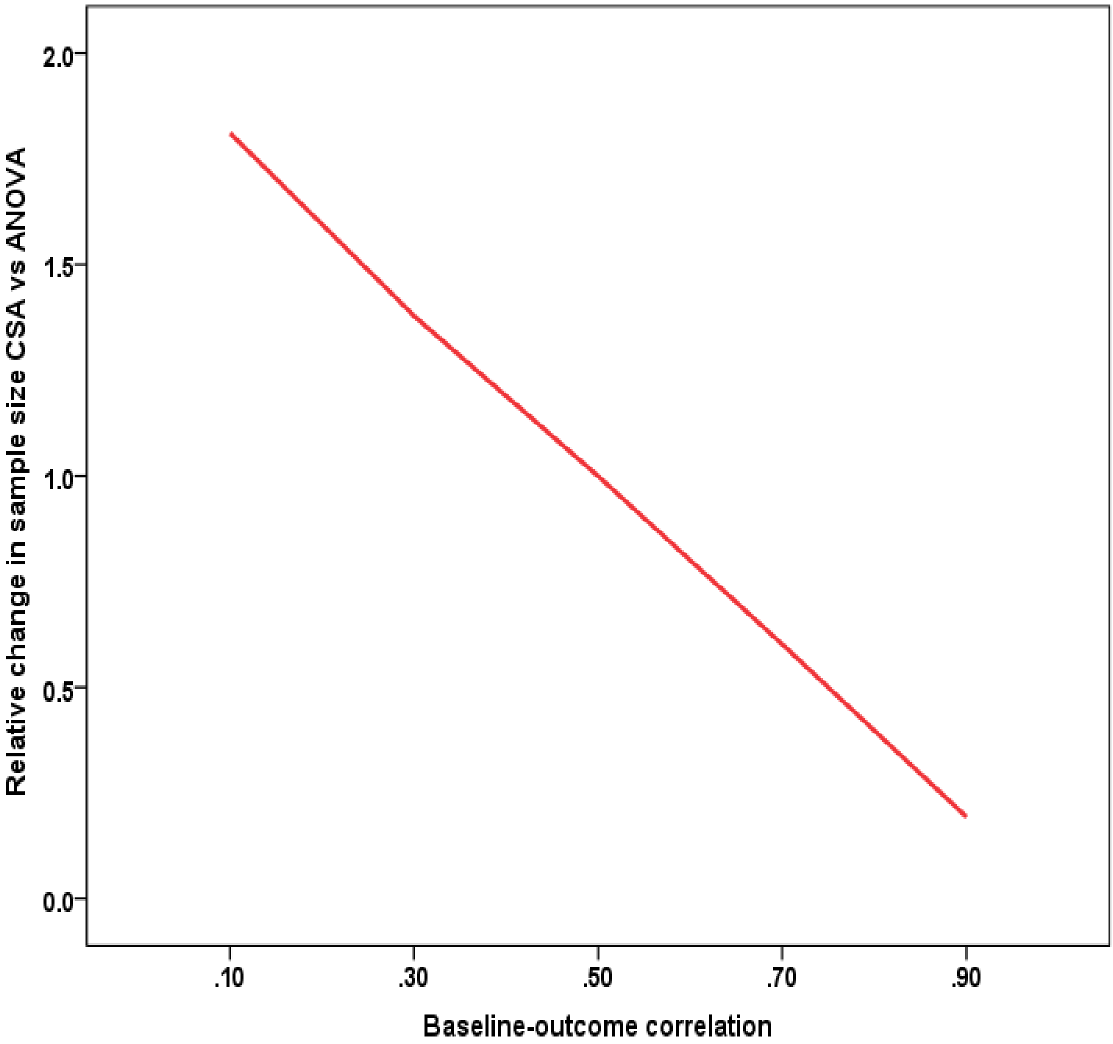


Figure 5.12: Relative changes in sample size for using CSA instead of ANOVA at differing trial scenarios

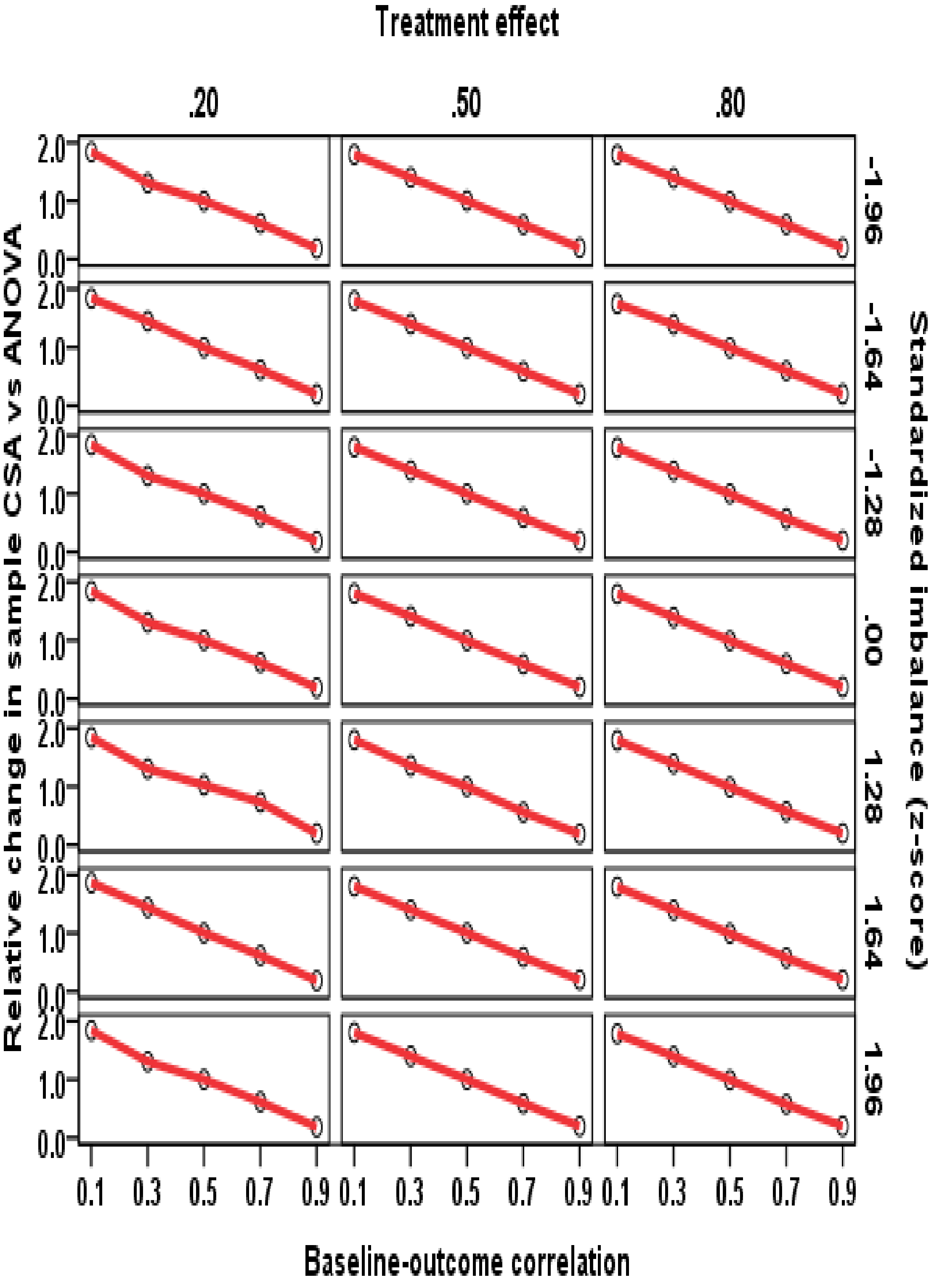


Table 5.11 presents the minimum absolute sample sizes that must be studied across different hypothetical trial scenarios with CSA as the primary analysis. At lower baseline-outcome correlation (r<0.5) ANOVA is more efficient than CSA, both are equally efficient at r=0.5, and CSA is more efficient at r>0.5.

Table 5.24: Adjusted sample size for using CSA instead of ANOVA

Effect Z	**Levels of baseline-outcome correlations**				
0.2	**0.1**	**0.3**	**0.5**	**0.7**	**0.9**
-1.96	1452	1030	788	487	145
-1.64	1452	1143	788	487	155
-1.28	1451	1030	788	487	145
0	1452	1030	788	487	145
1.28	1452	1030	811	579	155
1.64	1470	1135	788	487	145
1.96	1452	1030	788	487	145
0.5					
-1.96	230	180	129	77	26
-1.64	230	180	128	76	25
-1.28	231	179	128	76	26
0	231	182	128	76	26
1.28	231	175	128	72	24
1.64	231	179	128	76	25
1.96	231	179	128	76	26
0.8					
-1.96	94	73	52	31	10
-1.64	91	73	52	31	11
-1.28	94	73	52	30	11
0	94	73	52	31	11
1.28	94	73	52	30	11
1.64	94	73	52	30	11
1.96	93	73	52	30	11

Y=0.2, n=178;Y=0.5, n=128;Y=0.8, n=52, 1000 iterations in each scenario

Table 5.12 presents the relative sample size changes when using ANCOVA instead of CSA at different trial scenarios. The relative sample size formula follows the algebraic expression $(1+\rho)/2$ (from chapter 3, equation (3.34). Again, the formula is a function of the baseline-outcome correlation only (i.e. sample size reduction is not influenced by the directions and size of baseline imbalance and treatment effect). The results show that ANCOVA requires a smaller sample size compared to CSA given most trial scenarios, making it the more efficient of the two statistical methods. The benefit of sample size reduction by ANCOVA over CSA can reach nearly 50% when the correlation is close to 0.1. From table 5.12, the figures increase as correlation increases, meaning that the actual reduction in the sample size for using ANCOVA instead of CSA decreases as baseline-outcome correlation increases.

Table 5.25: Relative sample size reduction for using ANCOVA instead of CSA

Effect Z	**Levels of baseline-outcome correlations**				
0.2	**0.1**	**0.3**	**0.5**	**0.7**	**0.9**
-1.96	0.56	0.73	0.74	0.83	1.00
-1.64	0.56	0.65	0.76	0.86	1.00
-1.28	0.54	0.72	0.76	0.86	1.00
0	0.54	0.70	0.73	0.83	1.00
1.28	0.54	0.72	0.71	0.69	0.94
1.64	0.54	0.66	0.73	0.83	1.00
1.96	0.54	0.72	0.73	0.83	1.00
0.5					
-1.96	0.57	0.67	0.77	0.87	0.97
-1.64	0.56	0.66	0.77	0.89	0.99
-1.28	0.56	0.66	0.76	0.87	0.97
0	0.55	0.65	0.75	0.86	0.99
1.28	0.59	0.64	0.69	0.87	0.97
1.64	0.56	0.67	0.73	0.89	0.97
1.96	0.57	0.67	0.73	0.89	0.97
0.8					
-1.96	0.61	0.72	0.82	0.94	1.05
-1.64	0.60	0.70	0.82	0.92	1.03
-1.28	0.58	0.69	0.79	0.93	1.02
0	0.56	0.66	0.76	0.86	0.98
1.28	0.57	0.68	0.79	0.91	1.00
1.64	0.59	0.70	0.81	0.95	1.03
1.96	0.61	0.71	0.82	0.97	1.04

Y=0.2, n=178;Y=0.5, n=128;Y=0.8, n=52, 1000 iterations in each scenario

For example, there is at least a 44% reduction in the sample size if ANCOVA was used instead of CSA to adjust for a baseline variable that has a correlation of 0.1 with the outcome; whereas, with a correlation of 0.7, the reduction in sample size is just around 14%. Figure 5.6 shows a relationship between relative efficiency and baseline-outcome correlation. This simulation result is reasonably consistent with the algebraic expression earlier expressed for sample size reduction when using ANCOVA instead of CSA.

Figure 5.13: Relative changes in sample size for using ANCOVA instead of ANOVA at differing level of baseline-outcome correlation

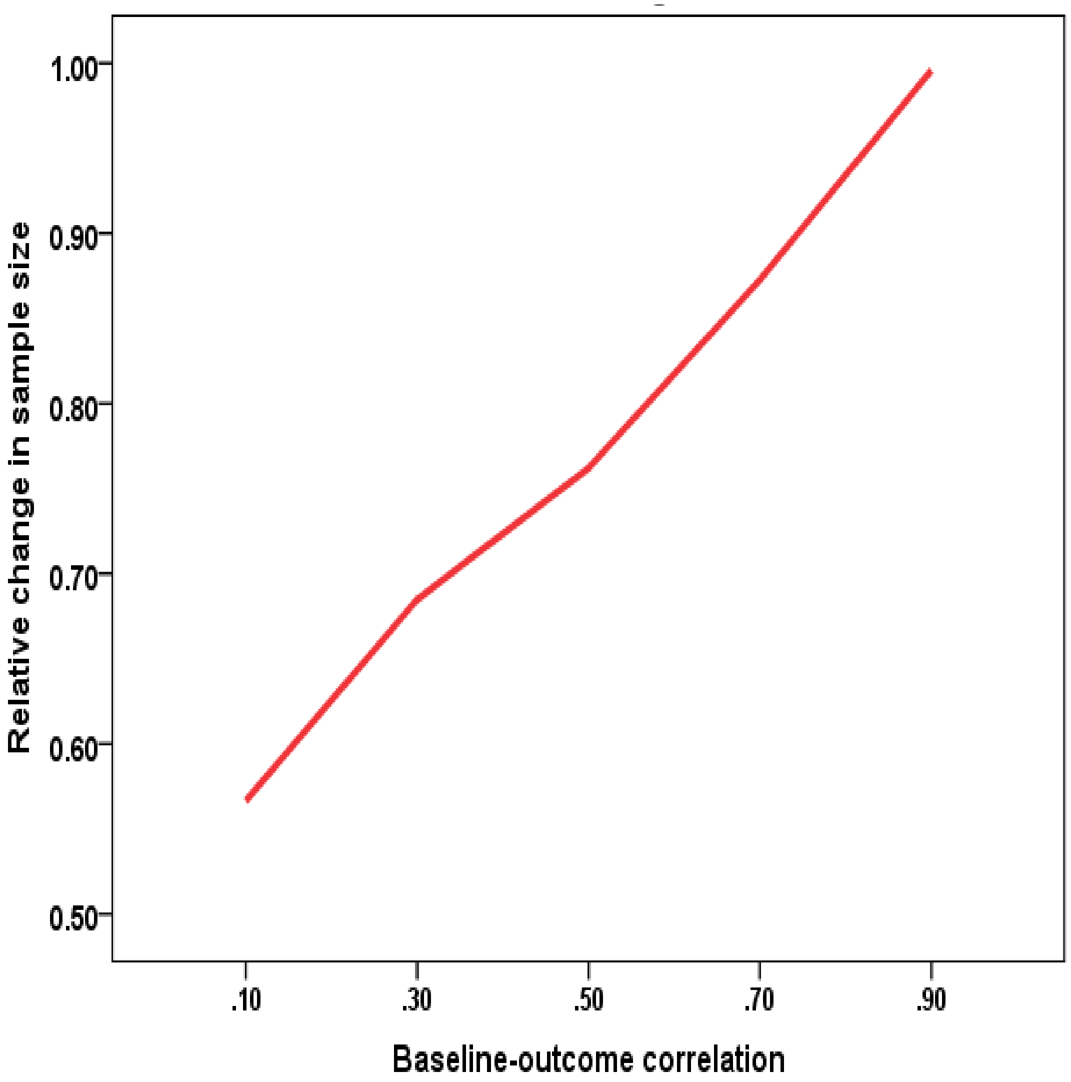
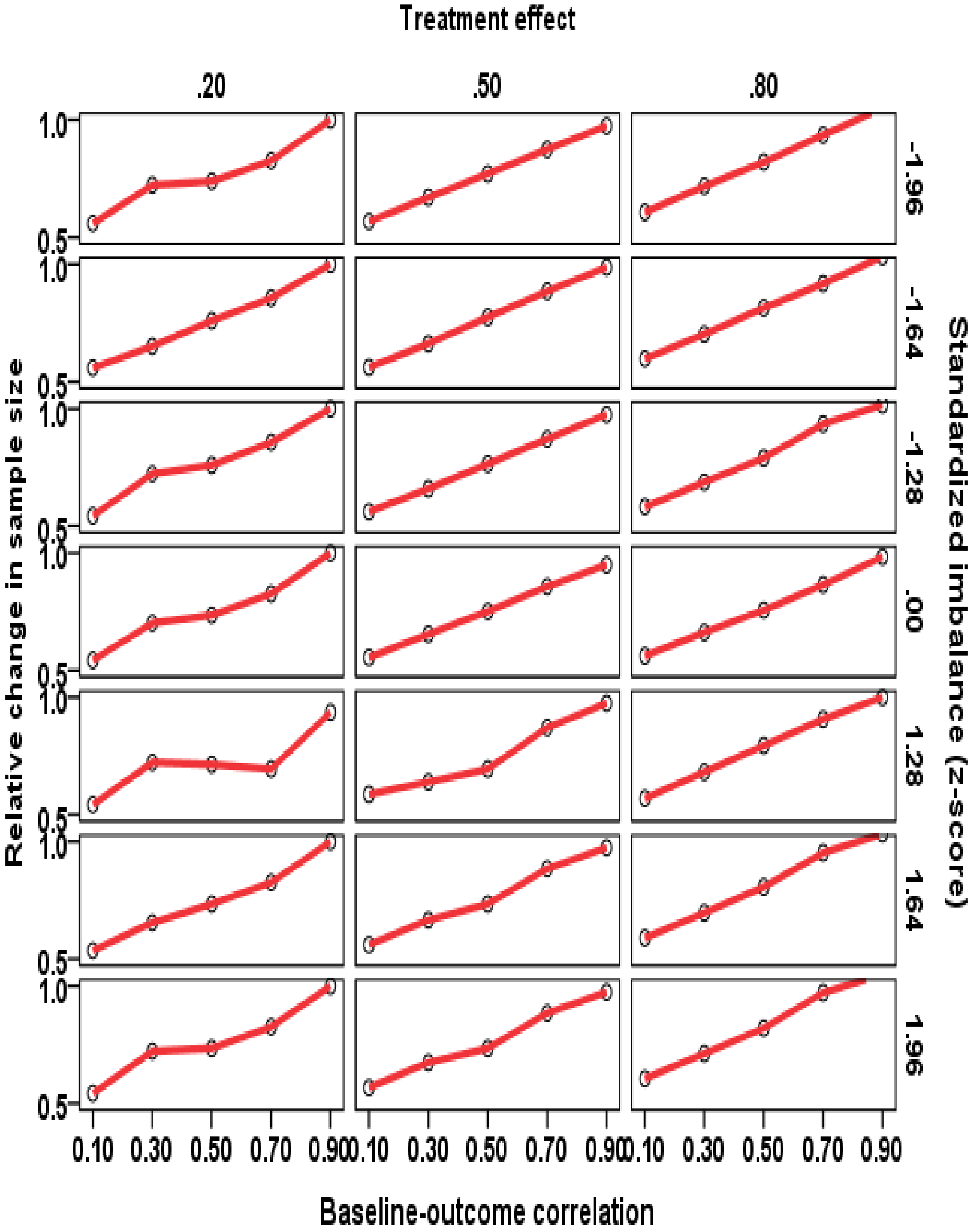


Figure 5.14: Relative change in sample size for using ANCOVA instead of CSA at different trial scenarios



Both ANCOVA and CSA are approximately equally efficient when adjusting for a baseline factor that has a correlation of 0.9 with the outcome. Accordingly, table 5.13 gives the number of additional patients that must be studied when using CSA instead of ANCOVA for statistical adjustment.

Table 5.26: Increase in the adjusted sample size for using CSA instead of ANCOVA

Effect Z	**Levels of baseline-outcome correlations**				
0.2	**0.1**	**0.3**	**0.5**	**0.7**	**0.9**
-1.96	641	286	207	85	0
-1.64	641	400	190	70	0
-1.28	664	286	190	70	0
0	664	308	209	85	0
1.28	664	286	232	177	10
1.64	682	392	209	84	0
1.96	664	286	209	85	0
0.5					
-1.96	100	60	30	10	1
-1.64	102	61	29	9	-1
-1.28	102	62	31	10	1
0	102	63	32	11	1
1.28	96	63	39	10	1
1.64	102	60	34	9	1
1.96	100	59	34	9	1
0.8					
-1.96	37	21	10	2	-1
-1.64	37	22	10	3	-1
-1.28	39	23	11	2	-1
0	41	25	13	4	-1
1.28	40	23	11	3	0
1.64	39	22	10	2	-1
1.96	37	21	10	1	-1

Y=0.2, n=178; Y=0.5, n=128;Y=0.8, n=52, 1000 iterations in each scenario

5.4 Discussion

5.4.1 Efficiency

The results of this simulation study demonstrate that, as a result of applying appropriate statistical strategies for adjustment, there can be an appreciable reduction in the sample units that have to be studied to detect a given effect. This reduction, as was shown in the results, is independent of both level and direction of imbalance, and it is wholly driven by the level of prognostic relationship between the baseline and the outcome. There is a lesser requirement for the number of patients in the study if statistical analysis takes strong prognostic factors into account, even if such factors are balanced between the groups at baseline. The sample size requirement is the same for trials in which baseline imbalance is zero and trials with significant level of imbalance as long as the baseline variable is appropriately adjusted. Thus, at any given level of baseline-correlation, relative efficiency is approximately the same across all levels of baseline imbalance for any pair of these statistical methods. This thus appears to suggest that in the context of relative efficiency in trials (reduced sample size) measures to balance the treatment arms (stratification, minimization etc) in prognostic factors are of little or no value. For example, the relative benefit in sample size reduction of ANCOVA against ANOVA is not a function of both direction and size of baseline imbalance. Kernan et al (1999) however, observe that stratification improves power and reduces sample size. These authors further argue that power losses for failure to stratify randomization could be made up by adding 6-14 patients to a given trial, their observations and comparison however, are limited to trials that do not use statistical methods of adjustment.

In addition, the results also show that in the design of a clinical trial, information on the amount or extent of possible baseline imbalance is not necessary to design a trial that minimizes the number of patient by using statistical method of adjustment. In the same vein, there is no need to be concerned with the direction of the possible imbalance in order to take advantage of the adjusted analysis for the purpose of minimising sample size. Once the treatment effect size to be detected is fixed, the efficiency of a statistical test depends on the precision with which treatment effects are estimated. From previous results (chapter 4), the precision of the treatment estimate in the context of a trial is dependent on baseline-outcome correlation and which statistical method is specified for analysis. Thus in terms of relative statistical efficiency, what matters is the information on the prognostic relationship between the relevant baseline variable and the outcome and then specify an appropriate method for statistical adjustment. Adjusting for a baseline covariate (by ANCOVA) that has the strongest relationship with the outcome guarantees the most efficient trial.

Depending on the prognostic relationship between the baseline and the outcome, these results demonstrate that up to 80% of the original sample size could be saved if ANCOVA or CSA is appropriately specified as the primary analysis, assuming baseline outcome variable has a correlation of 0.9 with the outcome. Taking a more realistic baseline-outcome correlation, about half the original sample size will be needed if a baseline variable that has correlation of

0.7 with the outcome is adjusted using ANCOVA. This result is consistent with Pocock et al (2002); these authors observed that with a baseline-outcome correlation of 0.7, the required sample size is roughly half the original sample size if the baseline variable is included in an ANCOVA model. Similarly, Porter & Raudenbush (1987) state that, at a baseline-outcome correlation of 0.8, only 33 patients per group will be needed when using ANCOVA instead of 84 per group if ANOVA is the statistical method to detect a large effect with 80% power. This of course has huge implications for: safety issues for the patients (as less patients will be randomised to receive treatment which in some cases may not be effective or presents with certain adverse reactions), trial cost, patient's recruitment and implementation time and other administrative issues. In this study, in the context of minimizing sample size – efficiency – CSA shows some potential to be more efficient than ANOVA, but only when correlation is greater than 0.5. On the other hand, in terms of efficiency ANCOVA shows itself to be the statistical method of choice across a wider range of experimental conditions typical of a clinical trial setting.

As a method for statistical adjustment, ANCOVA provides greater efficiency than ANOVA, especially where CSA performs poorly; for example, for a baseline-outcome correlation that is less than 0.5. The results show that there is a 25% reduction in the required sample size when using ANCOVA instead of CSA in a trial with a baseline-outcome correlation of 0.5. This finding corroborates the results of earlier studies (Egger eta al, 1985; Pocock, 2002; Van Breukelen 2006). Furthermore, whilst statistical adjustment of imbalance affects the statistical power in contrasting ways depending on the direction of imbalance, this is not so for efficiency. These results show that the potential for relative sample size reduction by ANCOVA does not depend on the original sample size or the size of treatment effect to be detected, nor does it depend on either the magnitude or the direction of imbalance.

This result agrees with Hernandez et al (2004), who argue that potential sample size reduction by the adjusted analysis does not depend on the size of treatment effect and sample size, although in the context of a dichotomous outcome variable in this instance. This implies that during the design of a trial, in order to minimise the sample size requirement, researchers need not worry about the possible effect of either size or level of chance imbalance if statistical adjustment is to be specified as the primary analysis. Efficiency – sample size reduction – of the statistical methods under consideration actually relates to the ability of each of the statistical methods to control the associated variability in the data and not necessarily the ability to reject the null hypothesis. A more efficient statistical method is expected to be more powerful than the other, however, statistical power responds to change in certain factors apart from baseline-outcome correlation, such as direction and level of baseline imbalance. For example, ANCOVA is the most efficient statistical method of the three, however, when imbalance is in the same direction as treatment then, with respect to the unbiased estimate of effect that it yields, the conditional power is lower than the nominal power of the trial at this instance.

The results further demonstrate that irrespective of the level of imbalance, CSA can be more efficient than the crude between-group comparison of treatment effect using ANOVA, but only if the baseline-outcome correlation is greater than 0.5. It also follows that while ANOVA can be more efficient than CSA at a correlation less than 0.5, both methods are equally efficient at correlation of 0.5. Thus, unless the prognostic relationship between the baseline and the outcome is represented by a correlation that is anticipated to be in excess of 0.5, CSA should not be specified as the primary analysis for the purpose of increased efficiency. This finding is again consistent with Senn's (1989, 1990), view that CSA is worse than simply using the unadjusted outcome values if the baseline-outcome correlation is less than 0.5. This result does not fully accord with Altman and Doré (1990), who present CSA as a reasonable statistical method for dealing with baseline imbalance without specifying the level of baseline-outcome correlation at which their observation is valid.

5.4.2 Statistical Power

The simulated statistical power for ANOVA within the range of study correlation, given a level of effect, is approximately equal, except for some random fluctuations that are more pronounced in small trials. Small trials are also known to be more susceptible to random or sampling error in the estimation of a population characteristic. In this study, for trials that ignore baseline imbalance as levels of effect increases, the statistical power for ANOVA shows capacity for slight increment. Although, on average, when correlation is low (r=0.1), ANCOVA has a statistical power that exceeds that of ANOVA, given the same level of correlation and a large effect size (y=0.8), ANOVA shows a slight increase in statistical power over ANCOVA as a result of random fluctuation. Random fluctuations of some sort are not unexpected in simulation studies, especially in small samples, as is the case when the effect size is large. Previously, Hernandez et al (2004) have also observed a reduction in the power of the adjusted analysis against the unadjusted under similar trial conditions. These results, in common with other previous studies, show that the benefit of statistical adjustment in terms of increased statistical power by ANCOVA is not noticeable until baseline-outcome correlation is greater than or equal to 0.3 (Altman, 1985; Senn, 1994).

Furthermore, the finding in this study that ANOVA is more powerful than CSA at baseline-outcome correlation below 0.5, irrespective of level of treatment effect to be detected, agrees with previous studies (Frison & Pocock, 1992; Vickers, 2001, Tu et al, 2005; Walters 2009). However, at a baseline correlation of 0.5 the pattern of statistical power of these methods (ANOVA & CSA), which shows slight variations across levels of treatment effect in this study, is similar to Tu etal's (2005) across a range of sample sizes studied. This therefore shows that there is approximately the same improvement in the statistical power of large trials as there is for small trials when a strong prognostic factor is adjusted for in the analysis. This finding is also consistent with previous studies by Pocock et al, (2002) and Senn (1989), who argue that appropriate covariate adjustment is also important for large trials. Various authors have also reported, similar to the findings in this study that increased statistical power increases with increase in baseline-outcome correlation (Porter & Raudenbush, 1987; Senn, 1989, Vickers, 2001;Pocock et al, 2002; Tu et al, 2005; Wang & Hung, 2005). This results show that adjusting for a strongly related baseline variable increases the statistical power of the test to detect the treatment effect between the groups.

The results generally demonstrate that for a randomised controlled trial that assumes that treatment groups are comparable in baseline outcome score, ANCOVA is the most powerful of the three methods of statistical analysis. The same findings have been reported by (Frison & Pocock 1992; Vickers 2001; Tu et al 2005; Van Breukelen 2006; Walters 2009). Overall, this study suggests that, a randomised controlled trial in which treatment groups are similar at baseline, and depending on covariate-outcome correlation, can benefit from adjusted analysis in terms of increased statistical power. This result is consistent with findings from other authors (Kent et al, 2009; Hernandez et al, 2003, Vanderlaan & Moore, 2007; Wang & Hung, 2005). For such trials, level of prognostic importance of the covariate with the outcome is important while deciding on whether or not to adjust and what method of statistical adjustment to use. The benefit of increased statistical power of the adjusted analysis is not significant unless baseline-outcome correlation is greater than or equal to 0.3.

Generally, CSA is no more powerful and no more efficient than ANOVA unless baseline-outcome correlation is greater than 0.5. In relation to CSA researchers should take advantage of the simplicity of ANOVA, especially for trials in which baseline-outcome correlation is low. The potential to increase statistical power with moderate baseline-outcome correlation ($r \geq 0.3$) by ANCOVA makes it a statistical analysis of choice for this design, and this quality should always be harnessed. Appropriate adjustment of baseline outcome will usually lead to statistical power that exceeds the nominal power by a margin that depends on how strongly correlated baseline and outcome scores are. Thus, if there is any treatment effect to be detected, the study is adequately prepared in terms of statistical power to detect such effect far more than the study nominal power.

The statistical model in these simulations is kept simple, with just a single covariate, the baseline score. However, one or two other covariates that is/are strongly correlated with the outcome predetermined or otherwise, can also be included in the model for adjustment. However, Pocock et al (2002) consider that other baseline covariates, either quantitative or binary, will usually have a weaker correlation with the outcome variable than that between the baseline and follow-up values of the outcome variable. In chapter seven of this study, the concept of covariate selection for statistical adjustment in relation to the prognostic strength of covariates was assessed using empirical trial datasets. The above results demonstrate the important effect imbalance in baseline prognostic factors can have, especially on the statistical power of the study. The issue with imbalance of a prognostic factor, as shown in this study, depends largely on the direction of such imbalance with respect to the treatment effect. Thus, the power of these statistical methods – ANOVA, CSA and ANCOVA – is not only driven by the size of baseline imbalance in prognostic variables, they are also dependent on the direction of the imbalance. Overall and Magee (1992) similarly assert that the effect of correlation on power is closely related to the direction of baseline imbalance.

Owing to the finding of this study that the direction of imbalance impacts on the statistical power of methods for adjustment, it is important to investigate the direction of imbalance to inform a suitable statistical method, given a particular trial scenario. This can be determined using summary statistics, but not necessarily tests of significance Schulz (1995), Senn (1997) and Altman (1985), to assess the direction of baseline imbalance in strong prognostic variables, especially the baseline values of the outcome variable. However, this is a post hoc activity, performed once data have been collected, which may mean that pre-specifying either the covariates to adjust for or the method of analysis to use is not appropriate at all time.

Pocock et al (2002) have argued that prior specification of the covariates to adjust for would be unrealistic in many instances. They also observed that direction of imbalance affects the two tails of a two sided test in contrasting ways. Similarly, according to Altman (1985), it should be remembered that if adjustment for prognostic factors affects the overall comparison, it is equally likely to do so in either direction. These two previous studies seem to corroborate the finding of this study especially of the effect of the direction of imbalance on the statistical power of the test and as a result the conclusion on the treatment effect. As earlier mentioned in this study (Chapter 4, Tables 4.2 - 4.7) the direction of baseline imbalance affect the size of the estimated treatment effect by each of these statistical methods, this thus has overall impact on the power with which the estimated effect is detected. There is a possibility of false positive and false negative errors associated with the estimate of effect by either ANOVA or CSA depending on the direction of imbalance.

This study has shown that, with respect to statistical power, it is not advisable to use CSA when imbalance is in the same direction as treatment effect. The reason is because, then, the method yields statistical power that might be too low to detect an effect that exists, thus giving rise to false negatives. Tu et al (2005) have also reported the possibility of false-negative results with CSA, although they did not mention the conditions under which this applies. The problem with CSA in this instance is associated with the way it adjusts for the imbalance. As earlier mentioned in chapter 3 (equation 3.15), CSA regards the regression coefficient for the covariate β_2 to be a value of 1 in its model. This leads to overcorrection of the imbalance and does not take the phenomenon of regression to the mean into account. Given the finding of this study that CSA can give a very different result from ANCOVA in respect of statistical power, even if the imbalance is low but in the same direction as treatment effect, the safest approach is to consider even a slight imbalance to be potentially important (Lewis, 1983) and use appropriate statistical adjustment – ANCOVA – to account for it. However, in terms of statistical power, the finding does not agree with Hewitt and Torgerson (2006) and Altman and Doré (1990) who had earlier reported that chance differences are generally of no consequence.

Although the conditional power for ANCOVA falls below the statistical power for ANOVA with imbalance in the same direction of effect, this should not be a problem as this only occurs when the unbiased treatment effect is smaller than the minimum effect the trial is designed to detect. The results show that when baseline imbalance is in the same direction as treatment effect the unadjusted analysis is most powerful. Consequently, a treatment effect may be indicated by the unadjusted analysis whereas the adjusted analysis by ANCOVA or CSA fails to detect any difference in effect. The issue here, however, is not necessarily because ANOVA is more sensitive than ANCOVA in identifying a given treatment effect, but rather because the two methods relate to different levels of effect. For ANOVA, the unadjusted large effect comprises the effect of the treatment and the influence of the imbalance that was not adjusted for. Conversely, ANCOVA or CSA operate with the adjusted effect – a smaller effect, because the effect of the imbalance (in the same direction as the treatment effect) is already accounted for. Camilli and Sheppard (1987) observe that ANOVA cannot but fail to detect even a markedly large amount of bias.

Since ANOVA fails to identify the effect of baseline imbalance, but rather regards it as treatment effect, this method cannot be free of false positives under the circumstance that imbalance is in the same direction of the treatment. This thus explains the higher level of statistical power ANOVA presents when baseline imbalance exists in the same direction of effect. These findings have implications for the design of clinical trials, specifically in relation to nominal power and sample size. When imbalance exists, not only in the same direction as effect, there will be problem with the interpretation of the results as was observed by Dougsheng et al, (2000). Researchers should therefore be careful of interpreting an effect size resulting from unadjusted analysis, especially when the appropriate adjusted method does not indicate one. As such, exaggerated effect may have arisen not only because the treatment is effective, but also because the treated group at baseline has a better prognosis – the baseline score of the outcome variable in the treated group is lower compare to the baseline score in the control group, which ANOVA fails to account for.

Thus, depending on the level and direction of imbalance and the size of the prognostic relationship between the baseline and outcome, one statistical test may have enough power to identify a significant difference in treatment effect whereas another may not, under the same experimental conditions. A classic example that illustrates this, cited by Altman (2005) and Assman et al (2002), is a trial by Christensen et al (1985) that concerned primary biliary cirrhosis. This study had a non-significant imbalance in a strong prognostic variable, serum bilirubin. Unadjusted and adjusted analyses yielded p=0.20 and p=0.02, respectively, for the treatment differences in survival. The only reason for this to have occurred in the context of the finding of this simulated study is if the baseline imbalance is in the opposite direction of the treatment. This implies that the treated group has a worse prognosis at baseline in the distribution of serum bilirubin. It is also possible to have a result that is further from the null and favouring the unadjusted analysis; this is expected if the baseline imbalance in the same direction as the effect, is large enough, and has an appropriate prognostic relationship with the outcome. This possibility is much higher, and will readily occur if CSA is the method for the adjusted analysis as CSA will usually presents with a very low statistical power much lower than that of ANCOVA, with imbalance in the same direction as effect.

When the treated group has a worse prognosis at baseline on the outcome variable, ANCOVA tends to account for the imbalance by adjusting the absolute value of the treatment effect in that group upward in relation to the overall score on imbalance – (average covariate score of the two groups). In contrast, CSA, by computing the difference between the baseline and outcome score, takes the treatment effect to be the sum of both baseline and treatment effect score. This explains the high overall effect size by this method for statistical adjustment of baseline imbalance in the opposite direction of effect. However, the high statistical power that results from using CSA can only lead to cases of false positives, in which the method suggests treatment effect when in fact there is none. Thus, when important baseline imbalance exists, and depending on its direction in relation to treatment effect, CSA is prone to statistical power that can be either inappropriately low or inappropriately high. This anomaly is as a result of the tendency to overcorrect baseline imbalance (Frison & Pocock, 1992; Assman et al 2000), due to the fact that CSA fails to take regression to the mean into account (Van Breukelen, 2006).

In a randomised trial, although post-test scores and baseline scores tend to be positively correlated, change and baseline scores tend to be negatively correlated (Senn, 1997, Vickers & Altman, 2001). With respect to pain and disability functions, patients with high baseline scores on the outcome variable tend to show more change than the average patient, and those with low baseline score tend to show less change than the average patient; in each case, regression to the mean is apparent.

5.5 Conclusion

Overall, in most conditions typical of clinical trial in which change from baseline score is a measure of improvement, ANCOVA is most powerful of the three statistical methods (ANCOVA, CSA and ANOVA) for analysis. However, when imbalance is in the same direction as treatment, ANCOVA has a reduced, lower power to detect an effect compare to ANOVA. This does not mean that ANOVA is better-off than ANCOVA, as the power of ANOVA at this time presents with false positive error just as that of CSA presents with false negative error. Thus when imbalance exists in the same direction as treatment, it poses serious implication for interpretation of treatment effect. The reduced statistical power by ANCOVA is in respect of the adjusted effect which is much lower compared to that on which the nominal power of the study is based. The reduced power by ANCOVA can also be viewed as a trade-off in respect of the unbiased estimate of effect that is yielded by this method.

Chapter 6: Baseline imbalance in randomised controlled trials: What does the literature say?

6.1 Introduction

In this chapter, an attempt is made to systematically review current practices on the subject of accounting for baseline imbalance in the clinical trial setting. Here, the review focuses on trials reported in five leading medical journals in which an increase in score from the baseline value of the outcome variable is a measure of treatment effect. The following journals provided articles that were reviewed in this study; Pain, Arthritis and Rheumatism, Rheumatology, Annals of the Rheumatic Diseases and Journal of Rheumatology. The impact factors of the journals and the geographical location were taking into consideration when making the choice of which journals should be considered in the review. All the journals have high impact factor and are also based in Europe, Canada and America. The systematic review exercise included all articles published in 2010 in these journals that met other inclusion criteria (as detailed in section 6.1) set for the purpose of this study.

The review seeks to cover current issues and practices related to statistical analysis involving baseline or covariate imbalance in clinical trial settings, as reported in these journals. However, since the primary endpoints covered in the review include both categorical and time to event, in addition to continuous primary endpoints, appropriate covariate adjusted methods in this context include; logistic regression, Cox regression, Mantel Haenszel test statistics and

ANCOVA. The more general focus of this PhD study is on those statistical methods that are used for the post-treatment assessment of a continuous outcome variable in RCTs, with or without baseline imbalance: the statistical methods of interest being ANOVA, CSA and ANCOVA

.

6.2 Methods

Following a careful selection of the journals in discussion with my supervisors, inclusion criteria were set out. The inclusion criteria for articles were: parallel-arm phase III clinical trial, published from 1 January–31 December 2010; overall sample size (n) greater than 50 – the minimum sample size (for large effect size) used in the simulation previously was 52 ; human subjects; results presented of a primary analysis/dataset; and published in the English language. All observational studies were excluded, as were cross-over trials, trials that analysed secondary datasets, phase I and II trials, and pilot trials. All issues regarding initial disagreement between the two reviewers on whether an article should be included or not were resolved by discussion.

The search was conducted in the Pub Med electronic database. The selected journals were entered as search terms. An abbreviation [TA] was however attached to each of the journals; this is to ensure that each search conducted returned only articles published in the specified journal. By making use of the 'limit' facility, the inclusion criteria were selected from the drop down menu. The title, abstract and method section of each article were searched to be sure that each article met the inclusion criteria. Each of the 40 eligible articles was assigned a study number as a means of identification. Once the ‘eligible’ trials had been identified, two reviewers (EE and ML) screened the journal articles and completed separate datasets for information (see section 6.2.1) relating to each trial. After completing this task, the two reviewers compared databases. Disagreements in relation to any of the recorded data between the matched data of the two completed datasets were to be resolved through discussion and consensus if possible; a third reviewer (‘arbitrator’ (JS)) was to be called on to resolve any disagreements.

6.2.1 Database

From each article, data were collected on the author’s name, centre status (single/multicentre) and, if multicentre, the number of centres. Data collected also included: allocation method, stratification/minimization factor if either stratification or minimization was used, sample size, disease category, statistical method for primary analysis (whether unadjusted or adjusted), primary outcome measure, type of endpoint (note, often a continuous primary outcome is categorized for the purposes of a main analysis), whether or not the primary outcome was based on ‘change’ from baseline (change in numerical terms (for numerical outcomes) or as a reference for categorical measures (e.g. ‘improved / better’ or ‘not improved / better’ compared to baseline .

In terms of the statistical method for primary analysis the emphasis was on whether adjusted analysis was used or not. Also of interest was: which statistical method for adjustment was used, the number of covariates adjusted for, whether correlation between baseline and posttest values on the outcome variable was determined, whether or not a test of significance of imbalance was performed, and whether or not imbalance was acknowledged. In addition, data were collected on whether or not centre status variable was adjusted; whether or not stratification or minimization factor(s) were adjusted as appropriate, and whether or not a subgroup analysis was done. The name of Journal was recorded, and reason(s) for exclusion if the article failed the inclusion criteria. Data were also collected on the treatment effect on the primary outcome and the conclusion pertaining to this.

Of these, 40 articles met the inclusion criteria for this study. Agreement between the two reviewers (EE and ML) was mostly achieved at the outset, and consensus was achieved through discussion without the need for a third reviewer/‘arbitrator’ (JS).

6.3 Results

6.3.1 Summary characteristics of included trials

The search yielded 33 articles in Pain, 41 in Arthritis and Rheumatism, 34 in Rheumatology, 133 in Annals of the Rheumatic Diseases and 36 in Journal of Rheumatology; thus a total of 277 full articles were accessed. Of the 277 articles that the search strategy yielded, (figure 6.1) only 40 (14.4%) articles met the inclusion criteria. More than half, 155(65.0%), of the total number of the excluded articles were not randomised controlled trials but observational studies (prospective/longitudinal cohort-, case control-, and cross sectional studies). Of the articles that were excluded 32 (13.5%) were ineligible owing to the sample size being less than 50; 26 (11.0%) articles were ineligible on the basis of reporting secondary data, 18 articles (7.6%) were non-phase III trials, and 6 (2.5%) articles were based on cross-over design. Of the 40 articles that were included in the study, 11(27.5%) were in each of Pain and Arthritis and Rheumatism journals respectively, 8 (20.0%) articles were in Annals of the Rheumatic diseases, 7 (17.5%) in the Journal of Rheumatology and 3 (7.5%) articles were in Rheumatology.

All the 40 trials included in this review randomised individuals into treatment arms; there were no cases of cluster randomisation. The trials were mostly (about two-thirds) multicentre trials, with the highest number of centres in a particular trial being 143. The authors of 6 of the 25 multicentre trials did not mention the number of trial sites or centres. The trial sample size was between 51 and 1025 inclusive and the median sample size was 160. Fifteen of the trials had a sample size of less than 100, the sample size for 18 trials was between 100 and 499 inclusive, and 7 trials had greater than or equal to 500 patients recruited. The sample size was classified into two categories (less than 200 and 200 and above) for the purposes of evaluation.

Figure 6.4: Flow chart illustrating the review procedures and outcomes

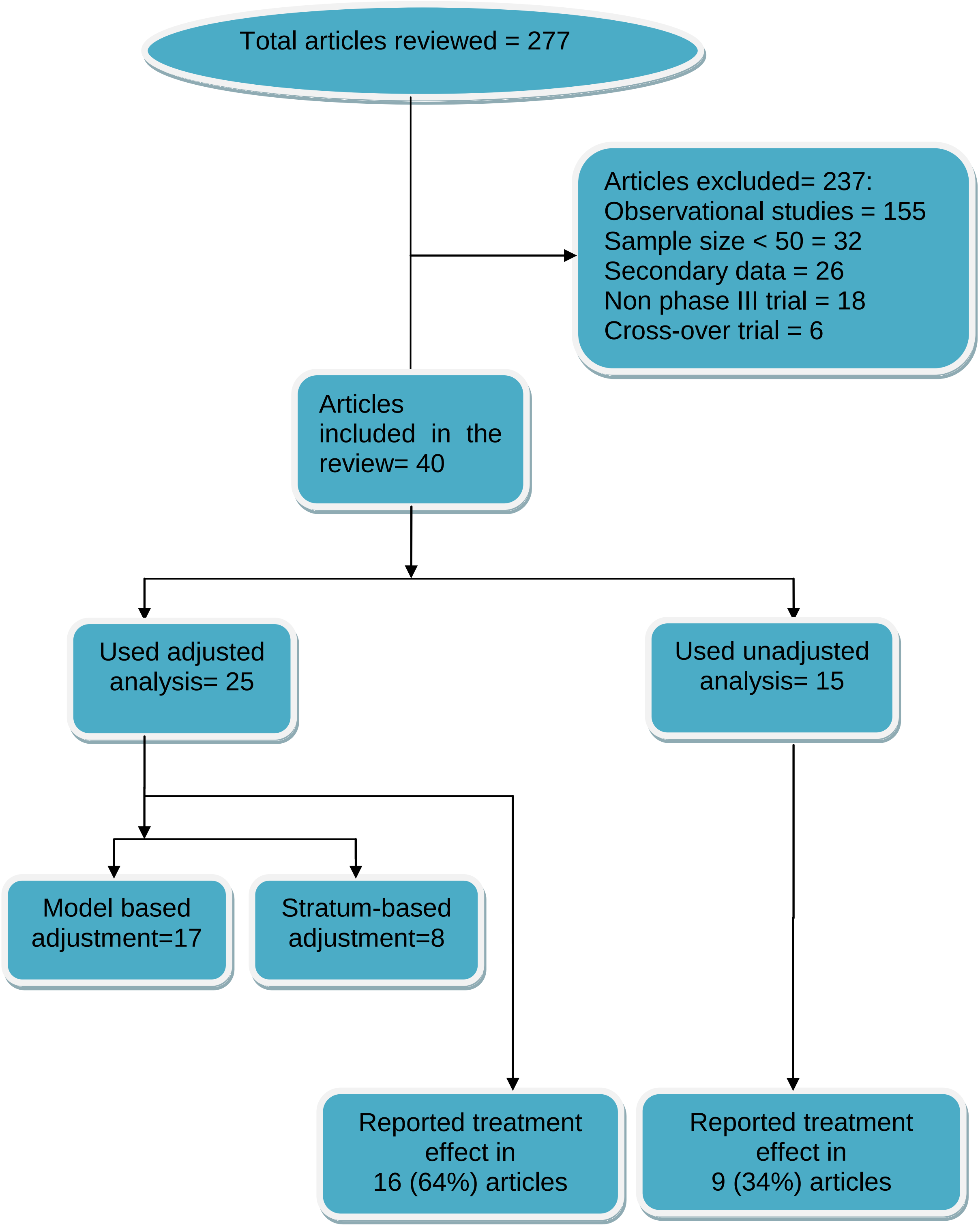

Total articles reviewed = 277
Articles excluded= 237:
Observational studies = 155
Sample size < 50 = 32
Secondary data = 26
Non phase III trial = 18
Cross-over trial = 6
Articles included in the review= 40
Used adjusted analysis= 25
Used unadjusted analysis= 15
Model based adjustment=17
Stratum-based adjustment=8
Reported treatment effect in 16 (64%) articles
Reported treatment effect in 9 (34%) articles

Table 6.4: Selected trial characteristics

Trial characteristics	**Number of trials (%)**
Trial centre	
Single	15 (37.5)
Multicentre	25 (62.5)
Methods of patient allocation	
Stratified blocking	17 (42.5)
Minimization	1 (2.5)
Simple random sampling	22 (55.0)
Primary outcome change score	
Yes	34 (85.0)
No	6 (15.0)
Primary outcome measure(s)	
Numerical	33 (82.5)
Categorical	4 (5.0)
Time-to-Event	3 (7.5)
Primary end point(s)	
Numerical	19 (47.5)
Categorical	18 (45.0)
Time to Event	3 (7.5)
Sample size	
< 100	15 (37.5)
100-499	18 (45.0)
≥500	7 (17.5)
Baseline test of significance	
Yes	12 (30.0)
No	28 (70.0)
Subgroup analysis done	
Yes	3 (7.5)
No	37 (92.5)

Also, regarding the treatment allocation method, in 22 trials the authors failed to specify the allocation technique beyond indicating that patients were randomised to treatment groups. In all of the 22 trials in which the authors did not specify the method of allocation, it was inferred that they had used a simple randomisation technique, on the basis that in none of these articles was a stratification factor mentioned. Thus, in all, stratification with blocking featured in 17 trials, 1 trial used minimization, and 22 trials adopted a simple random allocation technique.

Even though 33 trials had a numerical primary outcome measure (table 6.1), in 14 of these trials the continuous outcome variables were dichotomised; for example, one of the reviewed articles, Molsberger et al (2010) dichotomised the average pain level during the last seven days (VAS score) by using at least 50% reduction from baseline VAS following treatment as the yard stick for treatment effectiveness. Only three trials had primary outcome measures that were time-to-event measures. Thirty-four of the trials were explicit on the fact that change from baseline was the measure of treatment effect or improvement. Overall, 25 (62.5%) trials recorded significant improvement as a result of the treatment and two-third of all trials that had no record of statistically significant effect in primary outcome did not do so with a placebo.

In 12 (30.0%) of the trials the author used a test of significance to assess baseline comparability and only in one trial was baseline imbalance statistically significant. Statistically non-significant baseline comparisons were reported in 7 trials, and in 1 of the 12 trials the authors reported that baseline-imbalance occurred in certain variables and in the remaining 3 the authors declared no imbalance. In all, only in 17 trials did the authors report that there is no baseline imbalance and in another 10 trials, authors did not make a clear statement on the status of baseline-imbalance.

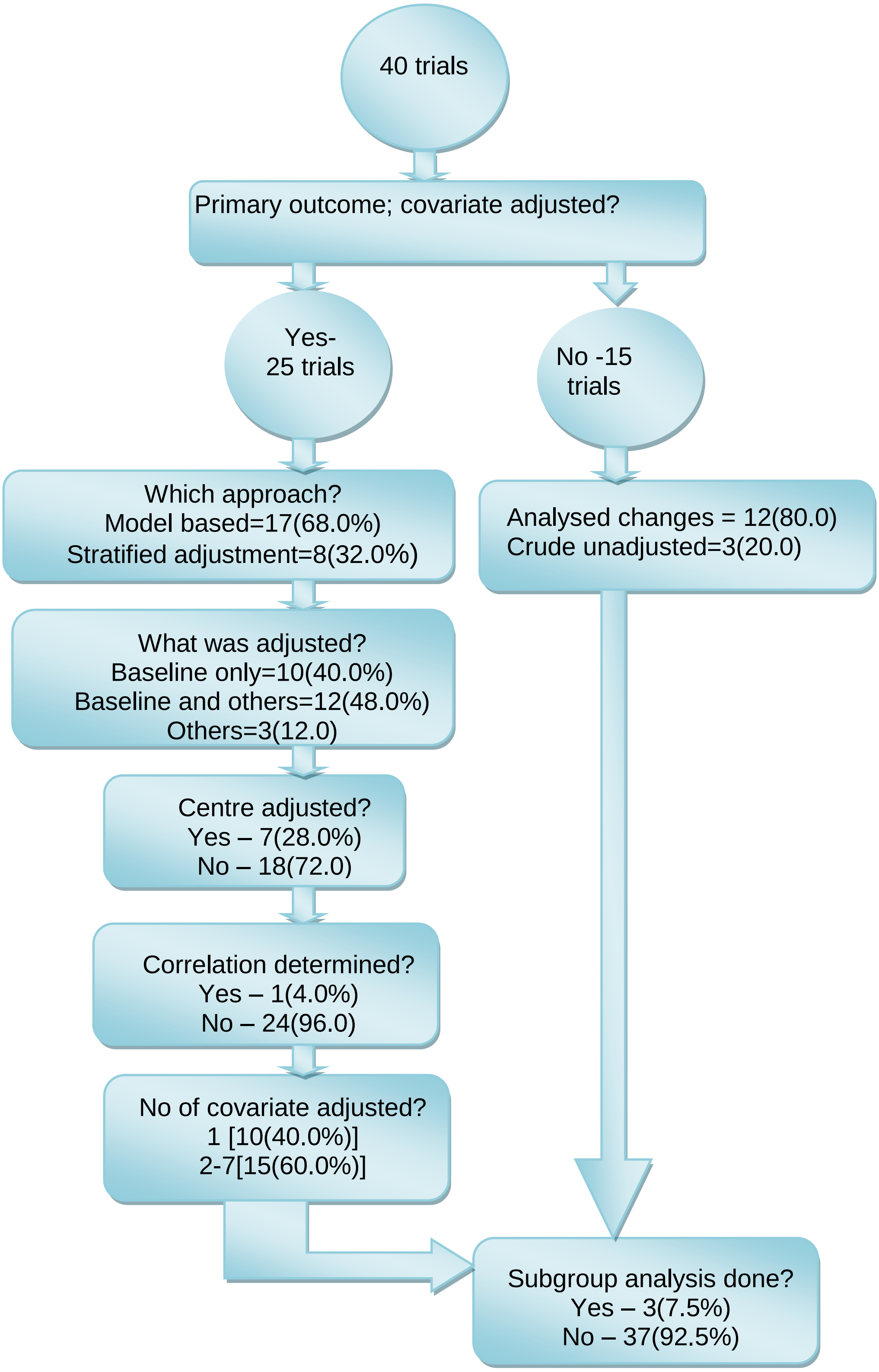
As shown in figure 6.2, 25 (62.5%) of the clinical trials reviewed statistically adjusted for baseline imbalance using regression or (pooled) stratification-based approaches. Model-based adjusted analyses, such as using ANCOVA for a continuous outcome variable, logistic regression for a binary outcome variable and Cox regression for time-to-event data, were more popular than stratum-based adjustment (using Mantel Haenszel and stratified Wilcoxon rank sum approaches), featuring in 17 trials compared to 8 respectively. Of the 25 trials that properly accounted for covariate imbalance, 10 (40%) adjusted for baseline-outcome only. In 7(28.0%) trials, the authors adjusted for centre or stratification factors, and only in 1 trial did the authors determine covariate-outcome correlation.

Trialists who used ‘change’ as the form of primary outcome often did so with statistical adjustment: of the 25 trials that did adjust using modelling- or a stratified- approach, 22 (88%) included ‘change’ as an outcome. Among the 15 studies that did not adjust, 12 (80%) included ‘change from baseline’ as the primary outcome. Thus, in three of the articles, the authors did not attempt to account for baseline imbalance. Most trials that used model-based adjustment reported adjustment for a single covariate (baseline values of the outcome variable); the largest number of covariates adjusted was 7. Subgroup analysis in which the baseline variable was stratified was reported in only 3 trials, one of which declared a subgroup treatment effect.

Table 6.5: Covariate adjustment practices and characteristics of the statistical adjustment

	Number of trials (%)
Primary outcome; covariate adjusted?**	
Yes	
N**o**	25 (62.5)
	15 (37.5)
If yes, what was adjusted?	
Baseline only	10 (40.0)
Baseline and others	12 (48.0)
Others	3 (12.0)
If yes, which approach?**	
Model based adjustment	17 (68.0)
Stratified adjustment	8 (32.0)
If no, which approach?**	
Analysis of ’ Change’(not necessarily CSA)	12 (80.0)
Unadjusted	3 (20.0)
If yes, Centre adjusted?**	
Yes	7 (28.0)
No	18 (72.0)
If yes, baseline-outcome correlation determined?**	
Yes	1 (4.0)
No	24 (96.0)
If yes, number of baseline covariates adjusted	
1	10 (40.0)
2–7	15 (60.0)
If stratified or minimization used, was the stratification/minimization factor adjusted for?	
Yes	7 (43.8)
No	11 (56.2)

Figure 6.5: Flow chart representing current statistical practices in RCTs



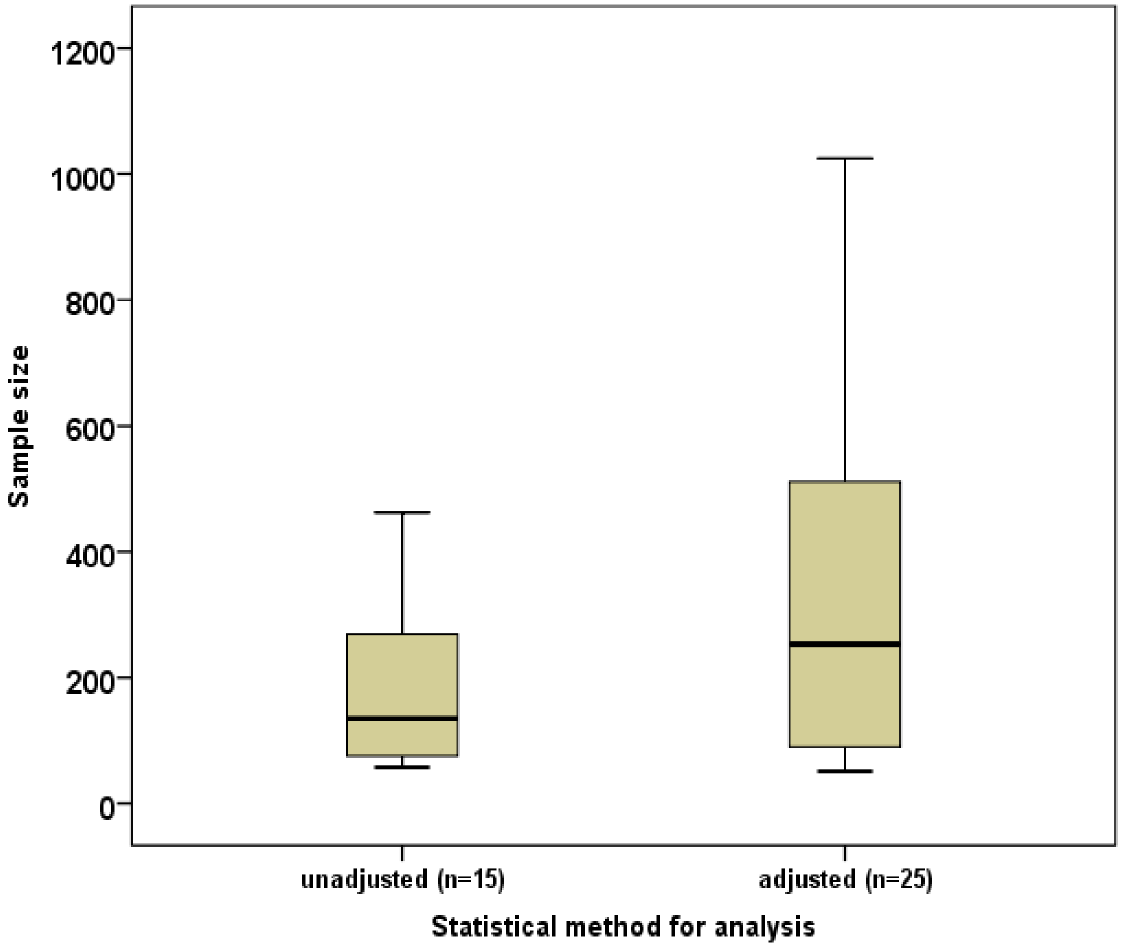
6.3.2 Association between study factors and statistical adjustment

Table 6.3, below, shows some association (albeit not statistically significant at the customary 5%-two tail level) between covariate adjustment and: baseline testing, trial status (single/multicentre), sample size, allocation technique and effect status.

Table 6.6: Association between statistical adjustment and selected trial attributes

Variables	**Used baseline significant test**		**Chi-square (X^2 ; df=1)**	**P-value**
	Yes (%)	**No (%)**		
Statistical adjustment				
Unadjusted (%)	6(50.0)	9(32.1)	1.14	0.311
Adjusted (%)	6(50.0)	19(67.9)		
Statistical adjustment	**Trial Centre**			
	Single	**Multicentre**	2.57	0.109
Unadjusted (%)	8(53.3)	7(28.0)		
Adjusted (%)	7 (46.7)	18(72.0)		
Statistical adjustment	**Sample Size**			
	< 200	**≥ 200**	1.93	0.165
Unadjusted (%)	10(47.0)	5(26.3)		
Adjusted (%)	11(52.4)	14(73.7)		
Statistical adjustment	**Allocation**			
	Simple R	**Stratified R**	1.30	0.251
Unadjusted (%)	10(45.5)	5(27.8)		
Adjusted (%)	12(54.5)	13(72.2)		
Statistical adjustment	**Treatment effect**			
	Yes (%)	**No(%)**	0.064	0.800
Unadjusted (%)	9(36.0)	6(40.0)		
Adjusted (%)	16(64.0)	9(60.0)		

Figure 6.6: Trial size and covariate adjustment



Furthermore, as seen in table 6.3, of the 25 trials that statistically adjusted for covariates 19(76.0%) did not use a test of significance compared with 9 (60.0%) of trials that did not adjust. Similarly, trials that adjusted for the primary outcome were more likely to be larger trials (see Figure 6.3), multi-centre trials and have stratification approaches to the design of patient allocation (as shown in

Table 6.3 is an assessment of the mean difference (by t test) in sample size between the trials that adopted a primary unadjusted analysis versus those that adopted a primary adjusted analysis yielded a statistically significant mean difference in sample size (t=2.06, 38df, p=0.047). Hence trialists were more likely to present a crude estimate of effect for small sample trials than they did in trials that had a large sample size.

Sixteen (64.0%) of trials that adjusted for covariates reported significant treatment effect and 9(36.0) of such trials had no significant effect. Similarly, 9(60.0%) of the trials that used unadjusted analysis recorded significant treatment effect and only 6(40.0%) of such trials had no significant effect.

6.4 Discussion

This review shows the current trend in statistical analysis of RCTs as one tending towards statistical adjustment of covariates. There is a considerable increase in the number of trials in which appropriate covariate adjusted analyses is specified as a primary method of analysis compared to what the practice was about ten years ago. For example in their review, Pocock et al (2002) recorded that in 12 (24%) of the 50 reviewed articles the authors specified covariate adjusted analyses as the primary method of analysis. This is low in comparison to the figure of 25 (62.5%) from 40 articles that were included in this review. This observed upward surge in the preference for the adjusted analysis could possibly be due to the various potential benefits that have been attributed to covariate adjusted analysis and increase in support for this statistical approach over the years. Since the review by Pocock et al (2002) for example**,** various authors have mentioned different benefits of covariate adjusted analysis over the unadjusted and these include: increase in statistical power (Kent et al (2009); Hernandez et al (2004) Moore and Vanderlan (2007); Wang & Hung, (2005)); improved type I error (Hagino et al (2004)); increased precision of estimates of treatment effect (Tsiatis et al, 2007; Wang & Hung, 2005), and reduced bias, giving more accurate estimates of the true value (Altman & Doré; 1990). The simplicity of the unadjusted analysis may no longer be a sufficient reason to continue to prefer this naïve method as the first line statistical approach in a clinical trial setting. There is evidence therefore of increasing in the usage and awareness of the merits of the 'adjusted' approach over the unadjusted approach. However, despite this increasing trend in the application of appropriate statistical adjustment, there still remain a substantial proportion of studies that do not properly adjust: in this review 37.5% of trials were unadjusted (crude comparison of effect or based solely on change from baseline).

Presently, the most popular statistical approach to the analysis of clinical trials is through 'change from baseline'. For numerical variables this equates to the CSA method presented in Chapters 4 and 5 (i.e. change from baseline score). For categorical variables, this equates to 'improvement' ascertained either directly through questioning such as 'how are your symptoms now compared to when you first presented at clinic: completely recovered, much better, better, no change, worse?', or indirectly by defining a threshold for change score improvement from a numerical measure whether this is in absolute (e.g. ≥ 2-point change on a pain-scale) or in relative terms (e.g. ≥30% improvement from baseline score). As established in earlier chapters, ‘change’ in itself is not a proper adjustment (and accordingly has not been merited as an ‘adjustment’ approach within the context of the results of this chapter). Overall, 3 (7.5%) of the trials in this review did not adopt any form of covariate adjustment in the primary analysis. This represents a marked reduction in the number of trials in which authors did not adopt any form of statistical adjustment, for example, Altman and Doré (1990) previously observed that in 39 (49%) of 50 reviewed articles the authors did not adjust at all. In this review, analysis based on ‘change from baseline’ (including CSA for numerical outcome) is the most popular, as 12 (30.0%) of all the articles reviewed are contented with ‘change’ as the primary analysis. This doubles the figure reported by Altman and Doré (1990) in which the authors reported a 15% utilisation. This finding confirms the fact that there is much more awareness for the need to adjust for covariate imbalance in the primary analysis of RCTs. However, the majority of authors settle for a basic adjustment in their trials.

Even though statistical methods such as ANCOVA that properly account for covariate imbalance especially the baseline of the outcome variable has been recommended by previous authors (Assman et al 2000; Senn 1989; Senn 1994; Tu et al 2005; Vickers 2001) the ease and convenience that accompany the use of CSA may well explain its use. Since CSA analyses differences from baseline score following intervention, it is sometimes believe to adjust for baseline imbalance (Altman & Doré 1991; Altman & Doré 1990). However, the mechanism of adjustment of imbalance by CSA that does not take the phenomenon of ‘regression to mean’ to cognisance also exposes the estimate of effect from this method to a degree of bias (as shown in Chapter 4). With respect to trial efficiency – relative sample size requirement and also the associated bias and precision of estimate of treatment effect, CSA alone is not better than the crude unadjusted analysis by ANOVA unless baseline-outcome correlation exceeds 0.5. Its statistical power is largely dependent on the direction of baseline imbalance; thus, CSA is prone to both false positive and false negative errors depending on the direction of imbalance.

In spite of the importance of the level of baseline-correlation in determining which covariate(s) to select for adjustment, the comparative benefit and appropriateness of the statistical methods for the analysis of RCTs, the practice by which the degree of prognostic relationship between covariate(s) and the outcome is empirically assessed to inform the choice of statistical method and or covariate selection is almost non-existant in this review. Most of the trials in this review lacked proper description of the allocation technique that was used in assigning patients to treatment groups. In this regard, there is an obvious deviation from the recommendation of the CONSORT document (Campbell et al, 2004; Schulz et al, 2010). CONSORT encourages authors to be explicit about the method used to generate and conceal random allocation sequence be described. It was observed that there is little or no improvement in the way authors report their trial allocation method compared to over ten years ago. This review found that over half of the authors did not properly document the allocation technique used in their trials, which is consistent with Assman et al

(2000), who also record that over half of the trials included in their review did not mention or describe the allocation method.

However the statistic in this review was an improvement compared to that reported by Altman and Doré (1990); in their study only one-fifth of the trials reviewed had stated the method of randomisation. Omission of such an important methodological issue represents a major flaw in the reporting but not necessarily the conduct of controlled clinical trials. A poorly reported trial will not allow the reader to understand how the study was conducted and how to assess the validity of the result.

Stratified random blocking was commonly used in the trials reviewed, and centre and baseline severity were the most specified stratification factors. The use of minimization as an allocation technique is still very limited; this might not be unconnected with the major drawbacks associated with this practice, such as: predictability of assignment, complex computation, and the fact that it is not completely a random process (Hewitt and Torgerson, 2006; Minsoo etal, 2008). However, Scott etal (2002) argue that minimization provides better balanced treatment groups when compared with unrestricted randomization and that it can incorporate more prognostic factors than stratified randomization methods such as permuted blocks within strata. The practice whereby most authors in this review did not adjust for the stratification or minimization factor(s) does not agree with the prevailing opinion on the issue (Scott et al, 2002; Hagino et al, 2004; Kent et al, 2009; Hernandez et al, 2003; Moore & Vanderlaan, 2007; Altman & Doré; 1991), who recommended that the stratification or minimization factor should always be included in the model for statistical adjustment. The expert view is that appropriate statistical adjustment is still necessary, despite efforts at the design stage to make treatment groups similar, as none of these methods is without one drawback or another. For example, in addition to the fact that stratification does not guarantee full and complete protection against imbalance, it becomes very complex to manage when there are several important prognostic factors to account for at the design stage (Rosenberger & Sverdlov, 2008).

In around one in three of the trials in this review, authors used a test of significance to assess baseline comparability. This represents a decrease in use compared to a similar review by Assman et al, (2000), where 50% of the authors used statistical tests of significance to assess baseline comparability of treatment groups. The common view of those that indulge in this practice is that once they are able to establish a non-statistically significant relationship or difference between groups on a specified baseline variable, adjusting for such variable is of no use. This is deemed to be an insufficient and improper practice as chance imbalance in a strongly prognostic variable has serious implication on the precision and estimate of effect. Thus, the fact that two-thirds of trials in this review did not engage in such a practise provides a reassurance that trialists are becoming less inclined to carry out hypothesis tests on a random phenomenon – which clearly does not make sense.

More often than not, in this review and in previous studies, when authors assess baseline imbalance between groups using a test of significance and such tests are non-significant, they tend to report that groups are comparable in baseline characteristics: which in itself should be a default position through a properly implemented randomisation procedure. The problem, as noted in earlier chapters, is that random differences can have a substantial effect – especially when it meets with a strong baseline-correlation. The issue is that by conducting such baseline testing with only about 1 in 20 tests being significant by the chance process the likelihood of overlooking important factors, notably ones that would be highly prognostic of outcome, is high.

In this review there was little association between trialists' use of adjustment in the hypothesis test of the primary outcome and observed significance in the treatment effect. As was shown in chapter 4, adjustment potentially results in increased power (particularly for a prognostic covariate) and thus an increased likelihood of attaining statistical significance for a true treatment effect. However, as was also shown in chapter 4, statistical significance after adjustment will also be dependent on the direction of covariate imbalance. When baseline imbalance is in the opposite direction of the treatment, treatment effect favours the covariate adjusted analysis; however, when imbalance is in the same direction as the treatment, higher estimate of treatment effect favours the crude unadjusted analysis. This observation may not hold for logistic regression and proportional hazard models as previous authors such as Robison & Jewell (1991) and Ford et al (1995) maintain that with covariate-adjusted estimates, odds ratios or hazard ratios become further from the null.

As explained in chapter 4, when a strong covariate that is in the same direction as treatment effect is statistically adjusted, the absolute value of the effect estimate is smaller compared to the effect estimate of the unadjusted analysis, and thus the F-ratio is closer to 1. When this happens, there is a higher likelihood for a treatment effect to be inferred with the unadjusted analysis than with the adjusted. Conversely, when the strong prognostic variable that is adjusted is in the opposite direction from the treatment effect, the absolute value of the effect estimate is larger than that of the unadjusted analysis – and the test yields a lower p-value.

Although the practice of using baseline significance testing to evaluate comparability is discouraged, its use however does not lead to significantly different results in this review. A higher likelihood of inferring treatment effect exists with non-usage of baseline significance tests; this may not be unconnected with the practice of covariate adjustment which is more popular in trials that do not use significance tests of baseline variables. From the review, trials in which baseline covariates are appropriately adjusted stand a higher chance of detecting a between-group difference when one exists. Again, this study suggests that those authors who use significance baseline tests will less likely adjust for covariate imbalance, as more often than not they will adjust only for imbalance that is statistically significant.

The practice of covariate-adjusted analysis in this review favours was more common among larger trials (i.e. those with greater sample size) than the smaller trials. However, since conditional benefit of covariate adjusted analysis is independent of sample size the size of a trial should not determine whether or not covariate adjusted analysis should be used. It has been previously found in this study that, in terms of precision, bias of estimate, statistical power and relative efficiency, both small sample trials and large sample trials benefit across different trial scenarios from appropriate statistical adjustment of covariate imbalance. Previous authors (Pocock et al 2002; Senn 1989) also maintain that in terms of bias, covariate imbalance is just as much a problem for large studies as for small ones.

6.5 Conclusion

The frequency of use of covariate adjusted analysis as the primary analysis in this review is reasonably high – compared to previous review studies. However, the number of trials that did not appropriately adjust for covariate by using CSA is still significant. Furthermore, trials that did adjust often excluded important covariates such as the corresponding baseline measure or severity indicator and/or the stratification/minimisation factors (applicable to block randomised trials) and/or Centre (applicable to multicentre trials). Lack of statistical adjustment has been shown to be more prominent amongst smaller (single-centre) trials. Adherence to standard guidelines on reporting of clinical trials is still an important issue.

Authors should endeavour to always provide concise information on randomisation procedure: type of randomisation, information on blocking and block size, method used to generate and implement the random allocation sequence, and how the sequence was concealed until assignment. Clear statements on which covariate was stratified on or minimized during the allocation process should be mentioned. Proponents of adaptive allocation techniques such as minimization still have the responsibility of creating awareness and promoting the usage of this alternative allocation method in the circles of clinical researchers in order to maximize the benefits and comparative advantages of using this procedure in the clinical trial setting.

Chapter 7: Comparative analysis of statistical models: the place of prognostic covariates in empirical datasets of clinical trials in musculoskeletal conditions

7.1 Introduction

The previous chapters, especially those on the results of the simulations, have shown that information on the level of correlation between pre and post-treatment scores on the outcome variable is crucial to precision, associated bias of estimate, statistical power and efficiency, with regard to the statistical method for the analysis. Depending on the level of correlation, there is considerable difference in these trial attributes with respect to the statistical methods of analysis: ANOVA, CSA or ANCOVA. For example, it has been shown (see table 4.8 in Chapter Four) that, for a given trial scenario and depending on the level of prognostic relationship between the baseline and the outcome, there can be up to 57% gain in precision by using ANCOVA instead of ANOVA. By contrast, the difference in precision between CSA and ANCOVA may not exceed 25% at any given trial scenario and this also is absolutely dependent on the level of baseline outcome correlation.

In addition, the associated bias of the effect estimate has varied considerably with levels of baseline-correlation in respect of the statistical methods used for analysis. However, depending on the statistical method, level and direction of imbalance also play an important role in the bias of estimate when using either CSA or ANOVA instead of ANCOVA. For example, in section 4.3 (Table 4.11) at an effect size of 0.2, with ANOVA as the method of analysis, for a large chance baseline imbalance, percentage bias on the estimate of effect varies from 6.5% to 62.5% depending on the level of pre and post-treatment outcome correlation. Direction of imbalance does not affect this bias. Also, with CSA as the method of analysis, in section 4.3 (Table 4.12) given that chance imbalance at baseline is large and in the same direction as treatment, depending on the level of pre and post-treatment outcome correlation, bias on the estimate of effect also varies from 25% to 211.7% whereas, at the same level of imbalance but in the opposite direction of treatment, bias varies from 7.3% to 72.7% depending on the level of pre and post-treatment outcome correlation.

Furthermore, in section 5.1.0, it is evidenced that the comparative advantage of the statistical methods under investigation with respect to gain in statistical power when treatment groups are balanced at baseline exclusively depends on pre and post-treatment outcome correlation. Although, when groups are heterogeneous (section 5.1.1, Figure 5.1) such that the chance imbalance is in the same direction as treatment, levels of chance imbalance and degree of baseline outcome correlation play an important role in determining the statistical power of both CSA and ANCOVA. Largely, the size of pre and post-treatment outcome correlation determines the power of CSA and ANCOVA when chance imbalance exists in the opposite direction from the treatment effect. The statistical power of ANOVA does not change with either level or direction of chance imbalance or changes in levels of other factors in the experiment.

Also, an important effect of pre and post-treatment outcome correlation is seen in the way it modifies the trial sample size requirement in respect of the statistical methods of ANOVA, CSA and ANCOVA. Usually, patients or other sample units are recruited into trials based on a specific figure calculated a priori. Such calculations are typically based on figures for the primary analysis that would be affiliated to crude unadjusted analysis (in this case, ANOVA) that does not make use of information on the covariance structure between the baseline and the post-treatment scores of the outcome. However, the simulation results in chapter 5 (Section 5.4.0) have shown that there is a remarkable difference in the sample size requirement between statistical methods of analysis that take pre and post-treatment outcome relationship into account and those that do not. This difference however, depends absolutely on the size of this relationship and is measured by the correlation between the pre and post-treatment scores. This difference which was earlier algebraically expressed (equation 3.32) by (Frison & Pocock 1992; Pocock et al, 2002) is independent of both size and direction of imbalance. With the information on the level of the pre and post-treatment outcome correlation, only a proportion of the original sample size will have to be studied if either of CSA or ANCOVA is specified as the method for the primary analysis. For example, specifying ANCOVA as the method for primary analysis instead of ANOVA (Table 5.4.3) in a trial scenario with pre and post-treatment outcome correlation (r) of 0.9 will lead to up to 80% reduction in the original sample size in any trial scenario, irrespective of the level of other experimental factors such as size of treatment effect to be determined and level and direction of baseline imbalance.

However, information on the level of correlation between the pre and post-treatment scores in musculoskeletal trials is not always available, making it almost impossible to harness the benefits that having such information offers in respect of design and methodological issues in the conduct of clinical trials in this setting. This chapter thus seeks to address this problem by exploring levels of correlation between baseline covariates and outcome, including baseline of the outcome variable, in empirical trial datasets in this Centre (Arthritis Research UK Primary Care Centre). Focus, specifically, is on trials involving back pain or low back pain. This information will not only dictate the path-way to future analysis of clinical trials in this setting, it also promises to inform a more efficient way of designing such trials. In subsequent sections in this chapter, an attempt is made to explore correlation between baseline variables and selected post treatment outcomes in three musculoskeletal trials. As have seen in chapters 4 and 5, when correlation between baseline variable and the post-treatment score reaches 0.3 and beyond, then adjusting or not adjusting for such prognostic covariates could have considerable effect on the performance of the statistical methods. Previous authors (Altman 1985; Cox & McCullough 1982; Senn 1994) have also specified a threshold of 0.3 as the minimum correlation a covariate should have with the outcome in order to include such covariate in the model for statistical adjustment. The overall effect of statistical adjustment of prognostic covariates on precision and bias in empirical trial settings is illustrated in this chapter. Exploration of levels of correlation in the covariates and the primary outcomes in these trials will inform the choice of potential covariates in the design and statistical analysis of future trials involving spine or low back-pain conditions.

7.2 Statistical methods for the empirical trial datasets

In this section, attempt is made to describe the statistical frame work necessary for the tasks set out in this chapter. Even though this study considers only continuous outcome variables, for example RMDQ and Northwick score, baseline covariates can either be continuous, binary or ordinal categorical. Thus, when inspecting the level of prognostic relationship between covariates and the outcome variables appropriate statistical methods depending on the scale of measurement of the covariates concerned were used. For example, whereas Pearson's correlation was used for assessing the degree of linear relationship between all numerical scales Spearman's correlation was used when the covariates being evaluated were ordinal categorical measures (since the scales were not linear) and in the event of binary/dichotomous covariates the point-biserial correlation was used to assess level of relationship between covariates and outcome. In each of the three trials, information on the level of correlation between the baseline covariates and the outcome were presented in tables. Any correlation that is greater than or equal to 0.3 is written in bold fonts for ease of identification.

In order to assess the stepwise selection of prognostic covariates in the model in terms of the amount of variability in the outcome that is explained by each progressively, all the covariates that met the criterion of having a minimum correlation of 0.3 with the outcomes at either of the follow-up times were entered into the regression model. In situations where a prognostic covariate is an ordinal categorical variable, appropriate numbers of dummy variables were manually entered alongside other covariates. In the regression mode, with outcome variables (for example RMDQ score at 4 months) as dependent variables a forward selection procedure was specified. However, the forward criterion: probability-of-F-to-enter which was <= 0.05 by default was preset to 0.98 so as to have all the entered covariates retained in the model outputs. Otherwise, as a result of the default 0.05 probability level for F, some covariates would not be retained and as such their explanation of the observed variability of the outcome would be missed.

Moreover two statistic were of interest at this time, they are: tolerance and R^2-change. Tolerance statistic assesses colinearity or multicolinearity as the case may be. It describes the percent of variance in the predictor that cannot be accounted for by the other predictors, hence very small values indicate that a predictor is redundant. Thus, a high tolerance value implies the covariates cannot explain each other and therefore merit further investigation. The minimum tolerance value allowed in this study for including a covariate into an existing model is 0.1. A tolerance value below 0.1 is not acceptable as it indicates unacceptable level of colinearity (on –line resource http://128.97.141.26/stat/spss/webbooks/reg/chapter2/spssreg2.htm). Colinearity is the term that describes the existence of two linearly related covariates in the model and multi-colinearity when more than two covariates are involved. In the absence of colinearity or multi-colinearity covariates that give desirable level of explanation of the variability in the outcome measured by R^2-change were selected and re-entered into a regression model this time together with the treatment group allocation variable.

In this study, a R^2-change value of 0.005 was carefully chosen (based on the observed change in the level of precision and bias associated with adding further covariate to the model) as the minimum amount of variability in the outcome a prognostic covariate should be capable of explaining independently to be considered for inclusion in the final model. In some cases, certain dummy variables of prognostic ordinal categorical variables are seen to be capable of independently explaining the variability in the outcome by at least the required amount of 0.005. Such prognostic categorical covariates include ‘Fear avoidance’, ‘Catastrophising’, and ‘Physical activity for age’. In such situations, summaries of estimates of the dummy variables were pooled together manually in each case. Thus, instead of reporting the estimates of the dummy variables individually the sum of each of their independent contributions was taken in each case. Lastly, in the final models, only those identified prognostic covariate(s) with a minimum correlation of 0.3 with the outcome and which are also capable of independently explaining at least a variability of 0.005 of the outcome are included together with the treatment allocation variable. Covariates according to their level of importance were added to the preceding model to show how they affect both estimate and precision of treatment effect.

Comparison of precision and estimate of treatment effect are later made with those obtained from using either ANOVA or CSA.

The focus of the evaluation are on pain and disability outcomes, which were the primary measures of treatment efficacy in these trials – and are the usual key clinical measures in musculoskeletal trials in primary care. Three empirical trials in the Centre provide the necessary datasets needed for the study objective set out in this chapter. The trials include the ‘StarTBack’ trial, the ‘Low back pain’ trial and the ‘PANTHER’ trial.

7.3 The StarTBack trial

7.3.1 Introduction to the trial including baseline distribution of selected variables[3]

This is a randomised clinical trial that has a primary objective of comparing the overall effectiveness of a ‘sub-grouping for targeted treatment’ approach with ‘best current care’ (non-targeted) physiotherapy practice, over a 12-month period, for low back-pain patients in the primary care setting. The trial used the StarTBack tool to classify patients into three groups for targeted treatment based on the presence of potentially modifiable risk factors. Participants, male and female, aged 18 years and above were recruited from 9 general practices within the Keele GP Research Partnership. Following completion of the baseline questionnaire, patients who consented to take part in the trial were randomly allocated to one of the two treatment arms: targeted treatment or best ‘usual care’ treatment. The allocation technique was block randomisation, stratified

[3] Adapted from the trial protocol and the main trial paper (Hay et al 2008; Hill et al, 2011)

according to centre and risk group. With a random allocation ratio of 2:1 (targeted treatment: best current care) and a block size of three, possible allocation blocks were: AAB; ABA; BAA.

Following information from the pilot study, which estimated that 25%, 50% and 25% of participants would be in the low, medium and high risk subgroups respectively, at 80% power and making a 20% allowance for lost to follow up, a total of 800 participants needed to be recruited to detect a 2.5-point difference in the Roland and Morris Disability Questionnaire (assuming a SD of 5). Overall, 851 patients were enrolled – the extra 51 patients were agreed with the Data Monitoring Committee to safeguard the power of the analysis based on a slightly higher than anticipated loss to follow up. The trial protocol specified ANCOVA for numerical outcomes and logistic regression for categorical outcomes on an intention to treat basis as the primary statistical analysis; a per protocol analysis would be performed as sensitivity analysis. Follow up was conducted at 4 months and 12 months post-randomisation. In this trial, outcome variables that were assessed at both 4 and 12 months included: RMDQ, 'intensity least painful back (last 2 weeks)' and average usual back pain (last 2 weeks) though the primary outcome variable was RMDQ.

Covariates	Low risk		Z_L-score	Medium risk		Z_M-score	High risk		Z_{H-} score
	Intervention	Control		Intervention	Control		Intervention	Control	
Age in years■	46.5 (14.3)	47.6(14.7)	-0.50	50.5(15.3)	49.3(13.5)	0.76	52.7(14.5)	50.1(15.3)	1.26
Sex, females (%)	82(55.4)	42(57.5)	-	160(60.8)	83(63.4)	-	88(56.1)	45(57.0)	-
Routine & manual occupation (%)	61(42.1)	26(27.1)	-	137(54.4)	65(51.2)	-	89(61.4)	58(76.3)	-
Currently in paid employment (%)	112(75.7)	50(68.5)	-	158(60.1)	83(63.4)	-	80(50.1)	41(51.9)	-
Time off work for back pain (%)	40(35.7)	12(24.0)	-	95(60.1)	51(61.4)	-	50(62.5)	27(65.9)	-
RMDQ disability score ■	4.6(3.5)	4.2(3.3)	0.87	9.9(4.5)	9.8(4.8)	0.19	14.4(4.6)	14.7(4.4)	-0.51
Back pain intensity■	3.4(1.6)	3.5(1.7)	-0.34	5.5(1.7)	5.3(1.8)	0.75	7.0(1.8)	6.8(2.0)	0.62
Average usual pain, last 2 weeks■	4.5(2.1)	4.8(2.3)	-1.20	7.0(2.0)	6.9(2.0)	0.60	7.9(1.9)	8.1(1.8)	-0.44
Referred leg pain (%)	61(41.2)	28(38.4)	-	176(66.9)	89(67.9)	-	115(73.2)	61(77.2)	-
Radiating pain below knee (%)	24(16.2)	10(13.7)	-	75(28.5)	47(35.9)	-	80(51.0)	36(45.6)	-
IPQR – Symptoms summary■	4.4(1.6)	4.2(1.6)	1.07	5.2(1.3)	5.6(1.4)	-2.57	6.01(1.1)	6.0(1.1)	0.28
EUROQOL – 5D Scores ■	0.7(0.2)	0.7(0.1)	-0.19	0.5(0.3)	0.6(0.3)	-1.04	0.3(0.3)	0.2(0.3)	1.71
PCS – catastrophizing score ■	8.6(5.8)	8.1(6.5)	0.57	14.6(8.1)	13.6(8.1)	1.15	26.4(10.6)	26.9(10.3)	-0.33
TSK – fear avoidance score ■	36.5(4.9)	36.5(5.8)	-0.02	39.2(5.0)	39.7(4.7)	-1.13	45.8(5.0)	46.0(5.7)	-0.20
HADS – anxiety subscale ■	5.2(2.9)	5.4(3.3)	-0.64	7.0(3.7)	7.4(3.7)	-0.83	10.1(4.2)	10.1(3.8)	0.04
HADS – depression subscale■	3.1(2.7)	3.0(2.5)	0.46	5.5(3.3)	6.0(3.8)	-1.35	8.9(4.3)	8.9(3.7)	-0.02
Widespread pain (%)	33(22.3)	16(21.9)	-	93(35.4)	56(42.7)	-	60(38.2)	32(40.5)	-
SF12 – Physical component ■	45.9(8.6)	46.0(9.1)	-0.08	35.7(9.5)	35.1(8.6)	0.49	30.8(7.6)	29.6(8.2)	1.17
SF12 – Mental component■	53.7(7.3)	53.0(8.1)	0.69	49.6(11.5)	48.6(11.3)	0.59	40.6(12.5)	41.5(12.3)	-0.55
Back pain – at the present■	3.1(2.1)	3.1(2.1)	0.09	5.0(2.4)	4.7(2.4)	1.04	6.7(2.2)	6.6(2.4)	0.32
Intensity of least painful ■	2.6(1.9)	2.5(1.8)	0.31	4.4(2.5)	4.2(2.6)	0.62	6.3(2.6)	5.8(3.0)	1.31

Table 7.1: Baseline distribution showing covariates Z-sores at levels of risk between the treatment groups
■ Numbers are mean and standard deviation in brackets

The trial results provided evidence that the stratified approach, by use of prognostic screening with matched pathways, was effective. An economic evaluation alongside the clinical effectiveness study provided further justification for the screening-and-targeted treatment approach. The findings are likely to have important implications for the future management of back pain in primary care.

With respect to the baseline distribution of the trial datasets in Table 7.1, baseline imbalances occurred similarly in the same and opposite directions as the estimated treatment effect.

7.3.2 Exploring levels of correlation between baseline and the post-treatment scores in StarTBack trial

Table 7.2a below shows the prognostic strength of the trial baseline variables and the pain/disability outcome variables at the two different follow-up periods (4 and 12 months). Over thirty baseline covariates and three outcome variables were evaluated. In this trial, the prognostic strength of the pre-treatment scores of the outcome variables exceeds that which exist between other covariates and the outcome variables, although, the prognostic strength of some of the other variables are somewhat close in size to that which exists between the pre and post treatment scores. For example, with Roland and Morris disability questionnaire (RMDQ) as the outcome variable, SF12-PCS score, Depression score, EUROQOL5D, ‘Expectation of back pain in 4 months’ and ‘having back pain at present time’ all have correlation coefficients that are close in value to that provided by the pre-post RMDQ correlation. Also, at 4 months follow-up, ‘having back-pain at present’ and ‘expectation of back-pain in 4 months’ have higher prognostic strength than baseline average usual pain; similarly, baseline RMDQ and EUROQOL5D are more strongly related with this outcome at 12 months follow-up period than its baseline score. ‘Expectation of back-pain in 4 months’ is also more strongly related with ‘average usual back pain (last 2 weeks)’ than the baseline score of this outcome.

In this trial, age and sex do not have a strong prognostic relationship with the outcome at either of the two time points for outcome assessment. In fact, they appear to have exhibited the least prognostic relationship with the outcomes across the different time points. In table 7.2a, figures in bold fonts are those that have a correlation coefficient of greater than or equal to 0.3 (r≥0.3) between the respective covariate and the outcome measure concerned.

When so many covariates have a relationship with the outcome variable that is high enough (r≥0.3) such that they are all included in the model for adjustment, there is a possibility for a linear relationship between two (collinearity) or more (multicollinearity) of the covariates. Thus, selected prognostic covariates have to be further investigated for the possibility of multicollinearity. For example, with respect to the primary outcome variable in this study, RMDQ, there are 14 potential covariates of the 34 that met the model inclusion criteria of (r≥0.3) and these may not be completely free of collinearity. The issue with multicollinearity is such that as it increases, the regression model estimates of the coefficients

Table 7.2a: Prognostic strength measured by correlation between the baseline and post-treatment score of the outcome variables – StarTBack [RMDQ: n=688 and 649 at 4 and 12 months follow-up periods respectively]

	RMDQ		**Intensity least painful back (last 2 weeks)**		**Average usual back pain (last 2 weeks)**	
Covariates	4	12	4	12	4	12
Age	0.11	0.20	0.13	0.22	0.05	0.13
Intensity least painful back (last 2 weeks)	**0.32**	**0.33**	**0.45**	**0.43**	0.25	0.28
Average usual back pain (last 2 weeks)	0.29	**0.31**	0.29	**0.33**	**0.30**	**0.34**
Back pain at the present Time	**0.38**	**0.37**	**0.41**	**0.40**	**0.34**	**0.34**
HADS-anxiety	0.26	0.30	0.21	0.28	0.19	0.24
HADS-depression	**0.42**	**0.41**	0.27	**0.33**	0.24	**0.30**
Pain severity	**0.39**	**0.40**	**0.46**	**0.46**	**0.35**	**0.38**
PCS score	**0.41**	**0.44**	**0.37**	**0.40**	0.30	**0.34**
RMDQ	**0.51**	**0.50**	**0.32**	**0.38**	0.28	**0.35**
SF12-MCS	**-0.31**	-0.29	-0.24	-0.25	-0.18	-0.20
SF12-PCS	**-0.45**	**-0.43**	-0.28	**-0.32**	-0.29	**-0.33**
TSK	0.30	**0.31**	0.27	0.28	0.16	0.23
EUROQOL5D	**-0.44**	**-0.43**	**-0.35**	**-0.38**	**-0.30**	**-0.36**
Expect back-pain in 4mth	**0.38**	**0.38**	**0.39**	**0.45**	**0.36**	**0.34**
Sex	0.14	0.01	-0.00	0.05	-0.11	-0.05
Currently employed	0.23	0.27	0.20	0.27	0.16	0.21
Social class	-0.04	0.15	0.17	0.19	0.10	0.12
Risk group◆	**0.32**	**0.34**	0.25	**0.31**	0.19	0.25
IPQR-symptom	**0.34**	**0.38**	0.27	**0.31**	0.27	**0.30**
IPQR personal control1◆	-0.15	-0.15	-0.15	-0.10	-0.14	-0.15
IPQR personal control2◆	-0.05	-0.04	-0.10	-0.11	-0.06	-0.10
IPQR treatment control◆	-0.10	-0.10	-0.11	-0.12	-0.10	-0.10
IPQR illness coherence◆	0.03	0.08	0.08	0.09	0.04	0.10
IPQR timeline cyclical◆	-0.17	-0.16	-0.13	-0.08	-0.13	-0.10
IPQR emotional◆	0.25	0.29	0.15	0.20	0.17	0.19
STarTBack tool◆	0.30	0.27	0.24	0.24	0.22	0.23
Bothersomeness	0.30	0.27	0.24	0.24	0.22	0.23
Pain spread down legs	-0.14	-0.12	-0.12	-0.15	-0.13	-0.17
Wide spread pain	0.13	0.17	0.18	0.17	0.20	0.17
SOC 2000	0.09	0.13	0.130	0.15	0.08	0.12
Currently employed	0.23	0.21	0.20	0.27	0.16	0.27
Time off work (back)	-0.10	-0.06	-0.05	-0.06	-0.05	-0.14
Time off work (other)	-0.03	-0.08	-0.16	-0.13	-0.1	-0.07
How long without pain	0.18	0.23	0.13	0.14	0.21	0.20

◆Spearman's correlation ‡Follow up time in months

become unstable and the standard errors of the coefficients can get wildly inflated. With this, the estimate of treatment effect will also present with more uncertainty and becomes less stable at the time.

7.3.3 Prognostic covariates rating of influence on the variability of the outcome variable (RMDQ)

The output of the regression analysis (following the procedure already described in section 7.1.0) showing the contributions of each covariates in terms of how much of the variability in outcome each of them independently explained as they are added to the model are presented - in an ordered fashion – in Table 7.2b and 7.2c. This order is completely a reflection of the importance of each of the covariates in terms of the variability in the outcome explained.

From tables 7.2b and 7.2c, 25.8% and 25.9% of the variability in the outcome were explained by baseline RMDQ only, at 4 and 12 months follow-up respectively. The last column in the tables gives information on the extra amount of variability in the outcome that was independently explained by adding the subsequent baseline variables to the preceding model. At both follow-up periods, the tables clearly show that baseline RMDQ explained most variability in the outcome and this is followed by ‘expectation of back-pain in 4 months’. There are thirteen individual models in each of the tables 7.2b and 7.2c. Each model consists of the covariate(s) at any point or number on the table and the (cumulative covariates) from the preceding model. For example, model 3 in table 7.2b has the following covariates: RMDQ, ‘Expectation of back pain at 4 months’

Table 7.2b: Model summaries and statistic at 4-month follow-up (RMDQ) [n=688]

Model	Covariates added	R	R^2	Adjusted R^2	R^2-change
1	RMDQ	.507	.258	.256	.258
2	Expect back pain 4 months*	.573	.328	.326	.071
3	HADS-Depression*	.589	.347	.345	.019
4	SF12PCS*	.601	.361	.357	.014
5	Risk*‡	.608	.369	.364	.008
6	SF12MCS*	.612	.373	.361	.004
7	Back pain at present*	.614	.375	.369	.002
8	IPQR-Symptoms summary*	.617	.377	.369	.002
9	EUROQOL-5D*	.618	.378	.369	.001
10	TSK score*	.618	.378	.369	.000
11	Pain severity*	.616	.378	.368	.000
12	PCS score*	.616	.378	.367	.000
13	Intensity of least painful back*	.616	.378	.366	.000

***Cumulative, ‡ added as 2 dummy variables (medium & high risk)**

Table 7.2c: Model summaries and statistic at 12-month follow-up (RMDQ) [n=649]

Model	Covariates added	R	R^2	Adjusted R^2	R^2-change
1	RMDQ	.509	.259	.258	.259
2	Expect back pain 4 months*	.577	.333	.331	.074
3	PCS score*	.590	.348	.345	.016
4	SF12PCS*	.599	.358	.354	.010
5	Risk*‡	.604	.365	.359	.007
6	HADS-Depression*	.608	.370	.363	.005
7	TSK score*	.611	.374	.366	.004
8	IPQR-Symptoms summary*	.615	.377	.368	.003
9	Ave. usual back pain (2wks)*	.616	.379	.368	.002
10	EUROQOL-5D*	.617	.380	.368	.001
11	Back pain at present*	.617	.380	.368	.000
12	Intensity of least painful back*	.616	.380	.367	.000
13	SF 12 MCS*	.616	.380	.366	.000

***Cumulative, ‡ added as 2 dummy variables (medium & high risk)**

and ‘HAD depression’. Each of the first five and six added covariates in the above model tables 7.2b (4 months follow-up) and 7.2c (12 months follow-up) respectively could explain the minimum variability in the outcome to be considered for inclusion in the final model.

In table 7.2b, most of the variability in the outcome was explained by the end of the fifth model; the adjusted R^2 indicates that 36.9% of the variability was explained by the time the fifth covariate was added into the model; this was close to the 37.8% variance accounted for by the full model that included all 13 baseline covariates. The result also indicates that by adding mental component score (SF12 - MCS) a further 0.004 (0.4%) of the variability in the outcome could still have been independently explained by this covariate; however, the value is not high enough to meet the criteria set for inclusion of covariate in the final model (according to the minimum criterion of 0.005 for R^2-change).

According to the result in table 7.2b, none of ‘TSK score’, ‘Pain severity’, ‘PCS score’ and ‘Intensity of least painful back pain’ could provide strong enough independent explanation of the variability of the outcome, despite the fact that they are related with the primary outcome through crude (univariate) association using the r≥0.3 criterion. This result suggests that a covariate can be prognostic of outcome without being able to independently explain the variability in that outcome when existing together with other covariates in the model. This brought into question the sufficiency of the criterion of r≥0.3 in determining which covariate is adjusted for during statistical analysis of RCTs. There is a slight change in the arrangement of the covariates in the model at both follow-up periods. Whereas at 12 months follow-up, ‘PCS score’ appears to be the third most important covariate regarding the amount of independent explanation of the variability in the outcome by these covariates, it is at the bottom of the table at the 4 months follow-up period. This suggests the need to repeat the check for which prognostic covariates to adjust for at different follow-up time points, which seems to contradict the practice by which covariates to be adjusted are determined a priori.

The results of diagnostic tests for collinearity show that collinearity is not an issue in having these covariates together in the model, as the associated tolerance value for adding each and every one of the covariates is considerably higher than 0.1 or 10% threshold for tolerance. However, if all the covariates that have correlation of greater than or equal to 0.3 with the post treatment RMDQ score are to be included in the model, the resultant full adjusted model will lead to redundancy of some covariates and also unduly increase the complexity of the model. Thus, a covariate being prognostic may not really be a sufficient reason for including it in the model for adjustment.

7.3.4 Comparative statistical models for the precision of the estimate and estimate of treatment effect at different follow-up periods - StarTBack

Tables 7.3a and 7.3b below, present results of the model-based covariate adjustment at five levels of adjustment each for the two follow up periods in comparison with the results of ANOVA and CSA. Results are presented on the estimates of treatment effect and precision for each of the methods. Standard error is highest for ANOVA models, though close to that which was obtained by using CSA; in fact the ratio of the two standard errors (CSA vs ANOVA) is 0.98 and 0.98 for 4 and 12 months follow-up, respectively. As observed in the simulation results, in chapter 4 of this study, the nearness to 1 of the ratios of standard error of both ANOVA and CSA at the two time points only reflects the

fact that the correlation between the baseline and post-treatment scores of the outcome variable is around 0.5 at each time. At baseline-outcome correlation of 0.5, both CSA and ANOVA are equally precise, so the ratio of their standard errors at the time is expected to be 1.

Table 7.3a: Models for the adjusted and unadjusted treatment effect comparison at 4months follow-up (StarTBack) [n=688]

		Treatment effect		
Models	**Covariates added**	**β_1**	**SE**	**95% CI**
1	RMDQ	1.46	0.40	0.68, 2.24
2	*Expect back pain	1.48	0.38	0.73, 2.22
3	*HADS-Depression	1.45	0.38	0.71, 2.19
4	*SF12PCS	1.42	0.37	0.69, 2.15
5	*Risk‡	1.41	0.37	0.68, 2.14
ANOVA	-	1.27	0.46	0.36, 2.19
CSA	-	-1.63	0.47	-2.53, -0.74

***Cumulative; ‡ 2 dummy variables added**

Table 7.3b: Models for the adjusted and unadjusted treatment comparison at 12months follow-up (StarTBack) [n=649]

		Treatment effect		
Models	**Covariates added**	**β_1**	**SE**	**95% CI**
1	RMDQ:	0.91	0.42	0.08, 1.75
2	*Expect back pain	0.98	0.40	0.19, 1.78
3	*PCS score:	1.02	0.40	0.23, 1.80
4	*SF12PCS	0.97	0.40	0.19, 1.75
5	*Risk‡	0.96	0.40	0.18, 1.74
ANOVA	-	0.72	0.49	-0.25, 1.68
CSA	-	-1.09	0.48	-2.04, -0.15

***Cumulative ‡ 2 dummy variables added**

These correlations were earlier observed, in table 7.2a, as 0.51 at 4 months and 0.50 at 12 months follow-up. The ratios of the standard errors suggest that CSA is yielding a slightly more precise estimate of effect; this is expected as

correlation between baseline and post-treatment RMDQ scores is slightly above 0.5. This phenomenon was also noted in chapter 4. Similarly, the ratios of the standard error of ANCOVA to ANOVA are 0.86 in table 7.3a and 0.86 in table 7.3b, representing the two follow-up time points. There is also a considerable gain in precision, especially when the second covariate ‘expectation of back-pain in 4 months’ was added to the models at the two time points. For example, at 4 and 12 month follow-up, adding ‘expectation of back-pain’ yield a decrease in standard error of approximately 0.02 (from 0.40 to 0.38 and from 0.42 to 0.40 respectively). Other levels of model-based adjustment involving adding other covariates did not show remarkable increase in precision, though each time the covariates were added in turn into the model they yielded a slightly more precise estimate of effect than the preceding model. Generally, estimates of effect at 4 months are more precise than the estimates at 12 months; this may not be unconnected with missing responses (as a function of sample size), which are higher at the 12 months follow-up period.

Similarly, the ratios of the standard error of ANCOVA and CSA were 0.88 and 0.88 at 4 and 12 months follow-up. The ratios of standard errors indicated here is in agreement with the level of correlation (0.51 & 0.50) in table 7.2a between the pre- and post-treatment scores of RMDQ. The results of the simulations in chapter 4 (tables 4.8 and 4.9) had earlier specified similar ratios of standard errors at a correlation of 0.5. At both time points, there is gain in precision following the addition of the third, fourth and fifth covariates to the model for statistical adjustment, however, whether the gain is worth the extra cost – in terms of the complexity of the model for including more prognostic covariates in the models is left to the discretion of the researcher.

There is a considerable difference in the estimate of effect from these statistical methods, ANOVA, CSA and ANCOVA, with CSA having the largest absolute effect size, followed by ANCOVA. This suggests that baseline imbalance in the RMDQ is in the opposite direction from the effect of the intervention; that is, the intervention group has the worse mean score in RMDQ at baseline (Figure 4.1). At 4 months follow-up, even though there is no difference in the conclusion by using any of ANCOVA, ANOVA and CSA – in all cases the null hypothesis is rejected indicating that the targeted treatment is superior. At 12 months however, both ANCOVA and CSA provide different conclusions to ANOVA; for ANCOVA and CSA the null hypothesis is rejected indicating superiority of the new treatment whereas for ANOVA the null hypothesis is accepted implying no evidence to reject the null hypothesis. This then implies that at 12 months follow-up which is actually the primary follow-up period being focused in this trial if the primary analysis had been based upon crude unadjusted analysis then the observed treatment effect attributed to the treatment being studied would have been missed out.

In table 7.3a, the percentage bias in the estimate of effect for ANOVA and CSA respectively with reference to effect estimate from ANCOVA is 14.5% and

10.6%; and in table 7.3b, the percentage bias in the estimate of effect for ANOVA and CSA respectively, is 27.1% and 61.24%. Also, the difference in the estimates of effect across the levels of the model-based adjustment at the two follow-up periods was minimal. However there is an obvious difference when estimates of effect from model-based adjustments (gold standard in this case) are compared to those by either ANOVA or CSA.

7.4 Low Back Pain Trial (LBPT)

7.4.1 Introduction to the trial including baseline distribution of selected variables[4]

This is a randomised controlled clinical trial in physiotherapy practice that compares the effectiveness of a brief pain-management programme with physiotherapy incorporating manual therapy for the reduction of disability at 12 months in patients consulting primary care with sub-acute low back pain (pain of no more than 12 weeks duration). In this trial, participants were recruited from 28 general practices in North Staffordshire, UK. All adults aged 18-64 years who consulted their general practitioners for the first or second time on account of non-specific low back pain (as defined by the UK Clinical Advisory Group) of less than 12 weeks' duration and who were able to give informed written consent were invited to participate. Patients were randomly allocated on a 1:1 basis (via simple randomisation) to either a brief pain management programme or a course of physiotherapy including manual therapy techniques. Treatment approaches

[4] Adapted from Hay et al (2005).

were agreed and standardised before the trial began. Interventions started within 10 days of randomisation and consisted of one 40-minute assessment and treatment session and up to six subsequent 20-minute treatment sessions.

Outcomes were measured at baseline, and at 3 and 12 months after randomisation. The primary outcome was change in disability related to the back pain measured at 12 months, rated on the self-completed (RMDQ) 24-item scale. The sample size calculation was based on change in RMDQ at 12 months after randomisation. To detect a clinically significant difference of 2 points between the two treatment groups with a significance level of 5% (2-tailed) and 90% power, 180 patients were needed in each group. A mean reduction in RMDQ score of 5.3 (SD 5.8) from a previous study was utilised. A total of 402 patients (including a 10% allowance for drop-out rate) were randomised into either of the two treatment arms. Analysis was by intention to treat (ITT). Primary and secondary analyses were by change score for numerical outcome and chi square for categorical variables. A sensitivity analysis used ANCOVA, which incorporated covariates based on the level of baseline random difference. There was little difference in clinical outcomes between the two active treatment groups. The trial concluded that brief pain management techniques delivered by appropriately trained clinicians offer an alternative to physiotherapy incorporating manual therapy and could provide a more efficient first-line approach for management of non-specific sub-acute low back pain. In this trial, the overall conclusion on treatment effect was not affected by the range of statistical methods used for the analysis.

Table 7.4: Baseline distribution showing Z-scores of selected covariates between the treatment arms.

Covariates	Brief pain management (n=201)	Manual Physiotherapy (n=201)	Z score
Age■	40.4(12.0)	40.9(11.6)	-0.41
Women (%)	100(50)	110(55)	-
Routine and manual occupation (%)	105(54)	130(66)	-
Currently in paid employment (%)	142(71)	152(76)	-
Time off for current episode (%)	97(48)	108(54)	-
RMDQ score■	13.8(4.8)	13.3(4.9)	1.00
Severity of pain today (VAS)■	55.8(23.3)	55.5(22.9)	0.14
Pain in past week (SF McGill VAS)■	68.3(21.2)	69.9(20.3)	-0.77
Radiating pain below the knee (%)	60(30)	67(33)	-
SF-McGill, present pain intensity■	2.13	2.07(0.704)	0.88
Using painkillers (%)	185(92)	189(94)	-
Duration of current episode of pain < 1 month (%)	150(75)	149(74)	-
Previous episode of low back pain (%)	157(78)	139(6)	-
MSPQ score■	5.6(4.3)	5.3(5.0)	0.56
Zung score■	24.9(7.9)	24.4(8.1)	0.78
CS-CSS score■	25.2(6.7)	25.1(6.2)	0.19
CS-PH score■	15.8(8.1)	15.1(7.7)	0.97
CS-CAT score■	8.4(6.7)	7.9(6.7)	0.73
CS-IBA■	22.4(6.3)	22.7(5.9)	-0.39
TSK■	40.7(6.2)	41.0(7.2)	-0.17
EUROQOL SCORE■	0.70(0.3)	0.70(0.3)	0.29
VAS main functional complaint■	60.9(20.0)	61.3(19.4)	-0.21
SF-McGill-sensory■	14.1(6.5)	14.1(6.2)	-0.05
SF-McGill-affective■	4.2(3.4)	4.2(3.3)	0.01
VAS main leisure complaint■	79.4(20.7)	76.7(21.2)	1.12
Wide spread pain (%)	25(12)	27(13)	-
Lateral flexion –right (start)■	67.9(4.5)	67.9(5.4)	-0.10
Lateral flexion – (finish)■	53.3(6.0)	54.2(6.5)	-1.40
Right flexion - distance■	13.9(6.1)	13.3(6.0)	1.098
Lateral flexion – left (start)■	67.9(4.6)	67.8(5.6)	0.19
Lateral flexion – left (finish)■	54.8(5.9)	54.9(8.3)	-0.15
Left flexion - distance■	12.5(5.7)	12.4(7.5)	0.20
Forward flexion (start)■	68.7(4.5)	68.9(5.6)	-0.26
Forward flexion (finish)■	55.9(12.4)	55.5(13.1)	0.28
Longstanding illness present (%)	64(32)	59(30)	-

■ Numbers are mean and standard deviation in brackets

Table 7.4 above shows the distribution of the baseline variables or covariates between the treatment groups in the low back trial. The Z-scores representing the standardized imbalance in baseline variables between treatment groups following random allocation of participants range from 0.01 to 1.40 in absolute terms. Treatment groups appear to be reasonably balanced in the between the two study groups.

7.4.2 Exploring levels of correlation between baseline and the post-treatment scores in the Low back-pain trial

Table 4a below, provides an assessment of the strength of the prognostic relationship of the baseline variables (covariates) with selected outcome variables at both 3 and 12 month follow-up periods.

Table 7.4a: Prognostic strength measured by correlation between the baseline and post-treatment score of the outcome variables [n=319 and 329 at 3 and 12 months follow-up periods respectively]

	RMDQ		VAS pain today		SF- McGill VAS average pain	
Covariates	**3**	**12**	**3**	**12**	**3**	**12**
Age	0.11	0.13	0.07	0.08	0.03	0.05
Gender	-0.06	0.05	-0.08	0.07	-0.09	0.05
Start of back pain	0.09	0.03	0.16	0.03	0.16	0.06
Widespread pain	0.14	0.12	0.10	0.07	0.12	0.10
VAS pain today	0.28	0.18	0.27	0.16	0.22	0.17
SF-McGill Vas average pain in last week	0.13	0.15	0.15	0.16	0.13	0.12
Referred pain to leg	0.08	0.08	0.09	0.08	0.11	0.08
SF-McGill-sensory	0.22	0.04	0.16	-0.04	0.15	0.01
SF-McGill effective	0.24	0.13	0.20	0.04	0.17	0.11
RMD-sum of all items	**0.32**	0.30	0.16	0.15	0.15	0.13
MSPQ-total score	0.23	0.21	0.18	0.17	0.16	0.19
ZUNG>90% items	0.28	0.25	0.24	0.14	0.16	0.16
CS-PH praying hop	0.18	0.13	0.14	0.12	0.11	0.13
CS-CAT catastrophising	0.28	0.24	0.21	0.12	0.20	0.14

CS-IBA inc. Activity	-0.22	-0.04	-0.01	0.03	-0.01	0.05
EUROQOL SCORES	-0.23	**-0.31**	-0.17	-0.22	-0.14	-0.22
EQ-VAS score	-0.13	-0.10	-0.13	-0.08	-0.11	-0.11
Employment status	-0.23	-0.16	-0.09	-0.10	-0.06	-0.08
Satisfaction with status	0.08	-0.00	0.13	0.03	0.10	0.01
Education after age 16	-0.17	-0.13	-0.10	-0.09	-0.08	-0.09
Education after age 18	-0.14	-0.12	-0.09	-0.09	-0.09	-0.12
Degree or equal qualification	-0.04	-0.04	0.02	-0.02	0.03	-0.05
Currently working	-0.13	-0.20	-0.07	-0.16	-0.04	-0.14
Time off employment	-0.14	-0.05	-0.05	-0.01	-0.06	-0.02
NS-SEC(Baseline)	-0.06	-0.08	0.02	-0.06	0.00	-0.08
Days of back pain	-0.08	-0.02	-0.05	-0.07	-0.07	-0.01
Something aggravate the pain	-0.07	-0.03	0.02	-0.02	-0.01	-0.00
VAS main functional complaint	0.17	0.11	0.14	0.12	0.11	0.07
TSK- total score	0.18	0.18	0.09	0.05	0.10	0.06
Low back pain before	0.19	0.10	0.14	0.04	0.13	0.02
Treatment for previous back pain	-0.03	0.01	-0.04	-0.00	-0.06	-0.00
Longstanding illness	0.20	0.09	0.16	0.07	0.16	0.06
Lateral flexion - right (start)	-0.02	0.01	-0.03	0.07	-0.07	0.05
Lateral flexion - right (finish)	0.09	0.04	0.03	0.08	-0.01	0.08
Right flexion – distance	-0.13	-0.04	-0.06	-0.05	-0.05	-0.06
Lateral flexion – left (start)	0.00	0.01	-0.01	0.06	-0.04	0.05
Lateral flexion – left (finish)	0.12	0.03	0.06	0.02	0.05	0.04
Forward flexion – distance	-0.06	-0.12	-0.02	-0.01	0.00	-0.08
Left flexion – distance	-0.14	-0.03	-0.07	0.01	-0.07	-0.02
Forward flexion (start)	-0.02	0.02	-0.02	0.07	-0.07	-0.02
Forward flexion (finish)	0.04	0.12	0.01	0.06	-0.04	0.08
Referred pain to leg◆	0.15	0.07	0.08	0.06	0.07	0.09
Satisfaction with status◆	0.11	0.05	0.15	0.06	0.13	0.05

Days off work◆	-0.01	0.13	-0.05	0.13	-0.03	0.11
Work physical load◆	-0.06	-0.01	-0.15	0.05	-0.09	0.01
Duration of back pain◆	0.06	-0.01	0.10	0.02	0.12	0.08
No of episode back pain more than 24hr◆	0.24	**0.31**	0.23	0.21	0.23	0.22
Sit to stand◆	0.16	0.11	0.12	0.11	0.07	0.11

◆Spearman's correlation

Baseline information was collected on both modifiable and non-modifiable risk factors. In all, close to seventy baseline variables measuring different attributes that were thought potentially to have influence on the prognosis of participants' condition were assessed. However, based on previous findings in the earlier chapters and recommendations of previous studies, a covariate may not have an important relationship with the outcome unless its level of relationship measured by its correlation with the outcome equals or exceeds 0.3.

Interestingly, as far as the primary outcome variable is concerned, RMDQ, only 3 of the assessed covariates managed to reach the threshold correlation of 0.3 with the post treatment scores at either of the follow-up periods. The three were: baseline (pre-treatment score) on the RMDQ, 'number of episodes of back pain more than 24 hours', and EUROQOL score. In this trial, the level of prognostic relationship between the baseline and the outcome variables in these sub-acute low back-pain patients was generally low. The highest correlation, 0.32, was observed between the baseline and the post treatment RMDQ scores at three months. This is rather weak for a relationship between the pre and post treatment scores of the outcome variable.

Theoretically, only these three covariates are qualified for statistical adjustment while estimating treatment effect. In fact, as a result of the weak relationship between the pre- and post-treatment RMDQ, the expected difference in the precision and the value of the estimate of the effect for ANOVA and ANCOVA was small.

7.4.3 Prognostic covariates rating of influence on the variability in the outcome variable (RMDQ) in the Low back pain trial

A careful study of these three covariates in the context of the amount of variability in the outcome that was independently explained by adding each of them to the model for adjustment shows a slight variation in the arrangement of the covariates at different follow up periods. Clearly, at the two follow-up times baseline score of the RMDQ is most important of the three covariates in terms of the explained variability in the outcome. At the 3 months follow-up period however, two of the ‘dummy’ levels of the covariate - ‘number of episode back pain more than 24 hours’ each independently explained at least 0.005 of the variability in the outcome. When these were added together, the aggregate of the variability explained by the covariate exceeds that which was explained by EUROQOL score. At the 12 months follow-up period however, there were three ‘dummy’ levels of ‘number of episode back pain more than 24 hours’ that independently explained variability in outcome of at least 0.005. When these were added however, the resultant aggregate was less than the amount of variability that was singly explained by EUROQOL score.

Table 7.4b: Model summaries and statistic at 3 months follow-up (RMDQ) [n=319]

Model	Covariates added	R	R^2	Adjusted R^2	R^2-change
1	RMDQ	0.277	0.076	0.072	0.076
2	*No of episodes of back pain‡	0.363	0.130	0.123	0.054
3	*EUROQOL scores	0.385	0.150	0.137	0.020

***Cumulative ‡ 2 dummy variables added**

Table 7. 4c: Model summaries and statistic at 12 months follow-up (RMDQ) [n=329]

Model	Covariates added	R	R^2	Adjusted R^2	R^2-change
1	RMDQ	0.321	0.103	0.099	0.103
2	*EUROQOL scores	0.408	0.166	0.159	0.063
3	*No of episodes of back pain lasting more than 24hr ‡‡	0.476	0.227	0.210	0.061

***Cumulative ‡‡ 3 dummy variables added**

The order of covariate appearance in tables 7.4b and 7.4c is a measure of the magnitude of the variability in the outcome that was explained by these covariates. The earlier a covariate appears in the model, the more the variability it explains.

7.4.4 Comparative statistical models for the precision and the value of the estimate of treatment effect at different follow-up periods – Low back pain trial

Although there is no difference in the conclusion regarding acceptance of the null hypothesis of no treatment effect in this trial as shown in tables 7.5a and 7.5b, by using any of the three different statistical methods of; ANOVA, CSA and ANCOVA at different levels of adjustment, CSA was observed to indicate treatment effect in the opposite direction of effect at 12 months (see table 7.5b).

A very considerable level of bias exists for not using ANCOVA especially with respect to the estimate of effect by CSA.

Table 7. 5a: Models for the adjusted and unadjusted treatment comparison at 3 months follow-up (LBP)[n=319]

		Treatment Allocation		
Models	**Covariates added**	**β_1**	**SE**	**95% CI**
1	RMDQ	-0.60	0.63	-1.83, 0.63
2	*No of episodes of back pain in 24hrs ‡	-0.16	0.75	-1.64, 1.31
3	*EUROQOL scores	-0.13	0.74	-1.59, 1.33
ANOVA	-	-0.84	0.66	-2.13, 0.46
CSA	-	0.22	0.71	-1.17, 1.62

***Cumulative ‡ 2 dummy variables were added**

Table 7. 5b: Models for the adjusted and unadjusted treatment comparison at 12 months follow-up (LBP) [n=288]

		Treatment Allocation		
Models	**Covariates added**	**β_1**	**SE**	**95% CI**
1	RMDQ	-0.50	0.60	-0.17, 0.67
2	*EUROQOL scores	-0.50	0.58	-1.64, 0.64
3	*No of episodes of back pain in 24hrs ‡‡	-0.47	0.65	-1.76, 0.82
ANOVA	-	-0.76	0.62	-1.98, 0.46
CSA	-	-0.01	0.69	-1.15, 2.34

***Cumulative ‡ ‡ 3 dummy variables added**

For example, in table 7.5a with ANCOVA as the reference gold standard statistical method, the percentage bias in the estimate of effect for using ANOVA or CSA is 28.2% or 172.4% respectively. In the same table, there is no marked change in the associated level of precision between ANOVA and ANCOVA. In fact, the ratios of the standard error between ANCOVA and both ANOVA and CSA are 0.95 and 0.88 respectively. This result is a reflection of the level of correlation 0.32 (in table 7.4a) that exists between the baseline and post-treatment RMDQ scores. In table 7.5b, the ratios of the standard errors between ANCOVA and both ANOVA and CSA are 0.960 and 0.866 respectively. In the simulation results in tables 4.8 and 4.9, the ratios of the standard error between ANCOVA versus either of ANOVA or CSA at correlation of 0.3 clearly resembles the ratios of the standard errors that was observed in this result. At the two different follow-up time points, the level of bias using CSA is higher compared to that from crude analysis of post treatment scores through ANOVA, so also, precision is better for ANOVA. Generally, in terms of precision and estimate of treatment effect this result is consistent with the result in the simulation chapter 4, however, in this particular trial, precision of ANOVA is sometimes better at higher level of adjustment possibly because of the involvement of the dummy variables of the prognostic covariate involved. This may need further investigating.

Thus, even though the primary objective of this trial focuses on the change from baseline regarding the primary outcome variable, since the level of the prognostic strength between the baseline and the post treatment scores is less than 0.5, it is not advisable to consider CSA as the primary analysis. As seen in chapter 4, while level of correlation is the sole driver of the difference between the precision of these statistical methods, it is also largely responsible for bias in the estimate of effect. Even though the disparity in the estimates of effect through using CSA and ANCOVA does not affect the conclusion about the treatment effect in this study, it no doubt constitutes a misleading basis for future studies that may rely on such biased information for sample size calculation.

7.5 PANTHER trial

7.5.1 Introduction to the trial including baseline distribution of selected variables[5]

This was a three arm, multicentre randomised controlled trial designed to determine whether manual therapy or pulsed shortwave diathermy, in addition to advice and exercise, provide better clinical outcome at 6 months than advice and exercise alone in primary care patients with nonspecific neck disorders. The trial involved 15 physical therapy outpatient facilities in the Midlands, UK. Eligible participants were aged 18 years or over with a clinical diagnosis of nonspecific neck pain. The randomisation procedure incorporated random-sized blocks stratified by physical therapy department and an allocation ratio of 1:1:1. All participants were randomised to 1 of 3 groups: 1) advice and exercise (A &E) with no further addition to treatment; 2) advice and exercise with the addition of manual therapy (MT); 3) advice and exercise with the addition of pulsed shortwave diathermy (PSWD).

The primary outcome measure was pain and disability measured with the Northwick Park Neck Questionnaire (9 items, scale 0-100). A sample size calculation was carried out to detect a mean difference in the 0-6 month change in Northwick Park scores of 5 points (±12.5 SD) at a 5% significance level (two tailed) at 80% power; a minimum of 99 patients were needed in each arm of the study. However, 350 patients were randomised to allow for loss to follow-up.

[5] Adapted from Dziedzic et al (2005).

Table 7.6: Baseline characteristics showing the Z-scores for selected covariates for PSWD and MT against the control treatment arm (A&E).

Covariates	Advice and Exercise (n=115)	Manual therapy (n=114)	PSWD (N=121)	Z-score (MT)	Z-score (PSWD)
Age in years ■	50.5(15.06)	52.8(13.24)	50.3(13.33)	1.24	-0.11
Sex, female, no.(%)	63 (55)	77(68)	81(67)	-	-
Routine and manual occupation, no.(%)	50 (51)	49(49)	47(47)	-	-
Currently employed, no.(%)	68 (59)	67(59)	68(56)	-	-
Time off work during last 3 months as a result of neck problem, no.(%)	20 (29)	22(33)	18(27)	-	-
Northwick Park score■	36.6 (13.54)	38.6(15.46)	36.9(13.56)	1.06	0.12
Pain severity last 3 days, mean (range)	4.6 (2.29)	5.4(2.26)	5.0(2.31)	2.46	1.29
Severity of main problem ■	5.0 (2.39)	5.6(2.25)	5.0(2.48)	1.98	0.25
EuroQol■	0.68(0.23)	0.62(0.27)	0.67(0.24)	-1.58	0.05
Painkillers in past 48 hours, no.(%)	45(39)	63(55)	60(50)	-	-
First episode, no.(%)	46(43)	37(34)	36(31)	-	-
Duration of pain>3 months, no.(%)	88(77)	96(84)	86(71)	-	-
Previous neck injury, no.(%)	24(26)	24(27)	33(33)	-	-
Previous physical therapy, no.(%)	30(27)	46(41)	47(39)	-	-
Widespread pain, no.(%)	18(16)	25(22)	24(20)	-	-
Chronic widespread pain, no.(%)	13(11)	24(21)	19(16)	-	-
SF-12 PCS score, mea■	41.3(10.4)	39.2(9.98)	41.9(9.96)	-1.49	0.43
SF-12 MCS score■	49.4(10.85)	48.9(9.82)	48.2(9.95)	-0.37	-0.88

■ Numbers are mean and standard deviation in brackets

Analysis was by intention to treat. In this trial, primary analysis was by CSA; ANCOVA and logistic regression were also used appropriately but as secondary analyses. No interim analyses were undertaken during the study period. There were no statistically significant differences in the primary outcome between the two active treatment arms and the control arm (advice and exercise alone). The study concluded that the addition of pulsed shortwave or manual therapy to advice and exercise did not provide any additional benefits in the physical therapy treatment of neck disorders.

With respect to the baseline distribution in the treatment arms in Table 7.6, participants in the PSWD and A&E groups are more comparable at baseline than those in MT. This observation is evidenced in the range of Z-scores; for example, whereas Z- scores range from 0.37 to 2.46 for MT versus A&E, the range of Z-scores for PSWD versus A&E is 0.11 to 1.29 in absolute terms.

7.5.2 Exploring levels of correlation between baseline and the post-treatment scores in PANTHER trial

Table 7.6a below presents the strength of the prognostic relationship between baseline variables (covariates) and selected outcome variables. Here prognostic covariates are observed to have a considerably high correlation with the outcome variables. The baseline scores of the outcome variables are particularly strongly correlated with post-treatment scores in each case. For example, the correlations between the baseline and the post-treatment primary outcome measure (Northwick Park) are 0.70 and 0.62 at the 6-week and 6-month

Table 7.6a: Prognostic strength measured by correlation between the baseline and post-treatment score of the outcome variables

[n=308 and 290 at 6 weeks and 6 months follow-up periods respectively]

	Northwick (total score)		Average pain past 3 days (NRS)		Most problem last 3 days (Ave. VAS)	
Covariates	6wks	6mts	6wks	6mts	6wks	6mts
Age	0.18	0.13	0.08	0.11	0.13	0.13
Gender	-0.10	-0.07	-0.06	0.11	-0.03	-0.14
Socio-economic class	0.26	0.26	0.21	0.24	0.18	0.26
Centre(stratification group)	0.03	0.05	0.02	0.07	0.03	0.07
Body chart (Baseline)	0.24	0.26	0.17	0.27	0.19	0.25
Body chart-widespread	0.24	0.27	0.15	0.26	0.16	0.24
Northwick(Base) total score	**0.70**	**0.62**	**0.50**	**0.47**	**0.51**	**0.50**
Medication use	0.26	0.24	0.16	0.17	0.21	0.18
Average pain past 3 days NRS)	**0.47**	**0.46**	**0.44**	**0.43**	**0.43**	**0.46**
Euroqol (Baseline) 5D score	**-0.57**	**-0.52**	**-0.40**	**0.42**	**-0.41**	**-0.42**
SF 12 PCS	**-0.54**	**-0.45**	**-0.33**	**-0.30**	**-0.39**	**-0.36**
SF 12 mental component	-0.30	**-0.31**	-0.20	-0.6	-0.17	-0.22
Average VAS problem)	**0.44**	**0.42**	**0.41**	**0.38**	**0.42**	**0.39**
Paid employment	-0.30	-0.20	-0.13	-0.13	-0.16	-0.11
Time off work	0.18	0.12	0.15	0.07	0.21	0.03
History of neck trauma	-0.04	0.06	0.03	-0.01	0.03	0.03
Patient's first episode	-0.04	-0.13	-0.01	-0.14	-0.03	-0.19
Fear avoidance◆	**0.40**	**0.34**	0.28	0.24	0.26	0.27
Catastrophising◆	**0.39**	**0.41**	**0.35**	**0.42**	**0.36**	**0.42**
Duration◆	0.22	0.24	0.16	0.24	0.20	0.26
Previous neck injury◆	0.02	0.09	-0.01	0.02	-0.01	0.04
Physio before◆	0.17	0.18	0.17	0.16	0.18	0.20
Patient expectation◆	0.12	0.14	0.07	0.15	0.06	0.13
Ability to influence work◆	0.02	0.07	0.03	0.10	0.03	0.08
Physical activity for age◆	**0.36**	**0.32**	0.25	0.23	0.24	0.26
Posture for age◆	0.11	0.09	0.04	0.06	0.02	0.04
Work satisfaction◆	0.02	0.11	-0.09	-0.00	-0.04	0.02
Duration episode◆	0.22	0.17	0.17	0.17	0.16	0.18

◆Spearman's correlation

Despite the high level of correlation that exists between both clinical and psychological covariates with the outcome variables, as in other trials, (Low back pain trial and StarTBack trial) socio-demographic characteristics (age, sex, socio-economic class) of patients do not have appreciable level of relationship with the outcome.

7.5.3 Prognostic covariates rating of influence on the variability in the outcome variable (PANTHER)

From tables 7.6b and 7.6c below, by the end of the fifth and sixth models, there was no prognostic covariate that could independently explain up to 0.005 or more of the variability in the outcome by including it in the model. To avoid unduly large tables, the covariates included beyond this point are not displayed (and have limited contribution to the variance of the NPQ at follow up).

Table 7. 6b: Model summaries and statistic at 6weeks follow-up (Northwick) [n=308]

Model	Variable added	R	R^2	Adjusted R^2	R^2-change
1	Northwick Park (Baseline) – Total	0.703	0.494	0.492	0.494
2	*EurQol (Baseline) 5D Score	0.717	0.514	0.510	0.020
3	*Fear avoidance‡‡	0.726	0.525	0.520	0.012
4	*SF12 (Baseline) physical component	0.726	0.528	0.521	0.007
5	*Physical activity for age (Baseline)‡	0.733	0.537	0.528	0.005

***Cumulative**
‡ ‡ 2 dummy variables added
‡ 1 dummy variable added

Table 7.6c: Model summaries and statistic at 6months follow-up (Northwick) [n=290]

Model	Variable added	R	R^2	Adjusted R^2	R^2-change
1	Northwick (Baseline) – Total	0.619	0.383	0.381	0.383
2	*EurQol (Baseline) 5D score	0.639	0.409	0.405	0.026
3	*Fear of avoidance‡‡‡	0.657	0.432	0.422	0.023
4	*Mental component score	0.670	0.448	0.436	0.016
5	*Catastrophising‡‡	0.677	0.458	0.443	0.010
6	*Physical activity for age‡	0.682	0.465	0.448	0.006

***Cumulative**
‡‡‡ 3 dummy variables added
‡‡ 2 dummy variables added
‡ 1 dummy variable added

With respect to Tables 7.6b and 7.6c, most of the variability in the outcome was explained by the baseline Northwick Park score. Again, multicollinearity is not a problem here as tolerance values were well above the threshold of 0.1 each time a covariate was added to the model.

7.5.4 Comparative statistical models for the precision and estimate of treatment effect at different follow-up periods - PANTHER

Tables 7.7a and 7.7b below present results of the models at both follow-up periods: 6 weeks and 6 months. Since there were three treatment arms/groups, one of the groups (Advice and exercise) was made a reference category. The trial was powered in such a way that separate comparisons of PSWD and MT against A&E are taking into consideration. There is a major difference between the unadjusted estimates of effect and the adjusted estimates either by CSA or ANCOVA. At both follow-up periods, the absolute size of crude treatment effect by ANOVA was much larger than that from either CSA or ANCOVA. This could have been due to the direction of baseline imbalance in Northwick Park Questionnaire score following randomisation. Often, as already seen in chapter 4, when the treated group has a better prognosis at baseline, the magnitude of treatment effect by the crude analysis is exaggerated. In this case, at both follow-up times, with respect to the estimate of effect using PSWD, there were 40.51 % and 45.52% associated biases of estimate for using ANOVA instead of ANCOVA at 6 weeks and 6 months respectively. The ratios of the standard error of ANCOVA vs ANOVA were 0.71 and 0.79 respectively. These results again confirm the previous results in table 4.8, that correlation between the pre and post-treatment scores is the sole driver of the precision of estimate of effect. The correlation between pre- and post-treatment Northwick scores in Table 7.6a is 0.70.

Table7.7a: Models comparison for the covariates adjusted and unadjusted analyses 6 weeks follow-up (Northwick) [n=308]

		Treatment Allocation		
Models	**Covariates**	**β_1**	**SE**	**95% CI**
1	Northwick (Baseline): Manual Therapy PSWD	 1.96 2.64	 1.64 1.62	 -1.27, 5.18 -0.54, 5.82
2	*EurQol (Baseline) 5D score: Manual Therapy PSWD	 1.67 2.48	 1.61 1.59	 -1.50, 4.83 -0.65, 5.60
3	*Fear avoidance:‡ Manual Therapy PSWD	 1.13 2.31	 1.61 1.58	 -2.03, 4.30 -0.80, 5.42
4	*SF 12 (Baseline) -Physical (PCS) : Manual Therapy PSWD	 1.06 2.50	 1.61 1.58	 -2.11, 4.22 -0.612, 5.61
6	*Physical activity for age:‡‡‡ Manual Therapy PSWD	 0.63 2.41	 1.62 1.58	 -2.56, 3.83 -0.70, 5.51
ANOVA	Manual Therapy PSWD	3.98 3.29	2.29 2.27	-0.53, 8.49 -1.17, 7.75
CSA	Manual Therapy PSWD	-1.47 -2.48	1.68 1.66	-4.77, 1.83 -5.75, 0.78

***Cumulative**
‡Compared to reference category (none of the time)
‡‡‡Compared to reference category (very good)
The reference category for the treatment groups is ‘Advice and exercise’

Table 7.7b: Models comparison for the covariates adjusted and unadjusted analyses 6 months follow-up (Northwick) [n=290]

		Treatment Allocation		
Models	**Variable added**	**β_1**	**SE**	**95% CI**
1	Northwick (Baseline):			
	Manual Therapy	1.69	2.10	-2.44, 5.83
	PSWD	1.49	2.07	-2.58, 5.56
2	*EurQol (Baseline) 5D score:			
	Manual Therapy	1.01	2.08	-3.08, 5.10
	PSWD	1.41	2.04	-2.60, 5.43
3	*Fear avoidance:‡			
	Manual Therapy	0.29	2.07	-3.77, 4.36
	PSWD	1.16	2.02	-2.82, 5.14
5	*SF 12 (Base) mental component:			
	Manual Therapy	0.21	2.07	-3.86, 4.28
	PSWD	0.74	2.04	-3.27, 4.75
	*Catastrophising ‡‡			
	Manual Therapy	-0.08	2.04	-4.09, 3.93
	PSWD	0.53	2.01	-3.42, 4.48
8	*Physical activity for age:			
	‘Very poor’‡‡‡			
	Manual Therapy	-0.25	2.03	-4.25, 3.75
	PSWD	0.39	2.00	-3.55, 4.32
ANOVA	Manual Therapy	3.60	2.70	-1.65, 8.85
	PSWD	2.74	2.63	-2.45, 7.92
CSA	Manual Therapy	-1.38	2.11	-5.53, 2.77
	PSWD	-1.29	2.08	-5.38, 2.81

***Cumulative**
‡ Compared to reference category (none of the time)
‡‡‡ Compared to reference category (very good)
The reference category for the treatment groups is ‘Advice and exercise’

Also, for using CSA instead of ANCOVA associated biases of estimate were 6.32% and 15.94% at 6 weeks and 6 months follow-up time respectively. In addition, the ratios of the standard error of ANCOVA to CSA at the two time points were 0.97 and 0.99; again these results corroborate the earlier finding in table 4.9. The results of the simulation in the earlier chapters had established that with similar levels of correlation (0.70) between pre- and post-treatment scores of the outcome variable such levels of the ratios of the standard error of estimate of effect would exist. The low biases associated with the estimate of effect from CSA in this instance are evidence of the fact that when the correlation between pre and post-treatment score is high, ≥ 0.7, then the estimates of effect by CSA and ANCOVA approximate each other. This idea was represented in figure 4.1 and table 4.12. Generally, in this study, the estimates of effect by CSA are more precise and less biased compare to the crude estimate by ANOVA owing to high level of correlation between pre- and post-treatment score of the outcome variable (Northwick Park).

7.6 Discussion

The relationship between baseline covariates and the outcome variables remarkably differs across the three empirical trials considered. For example, whereas baseline-outcome correlation reaches 0.70 for the PANTHER trial and 0.51 for the StarTBack trial it peaked at 0.32 for Low back pain trial. This difference in the pattern of prognostic relationship of covariates across the trials is a pointer to relative need with regards to statistical adjustment across these empirical trials. Surely there is more need for statistical adjustment of covariate imbalance in trials where such covariate has a high correlation with the outcome. In all the three trials, as expected, baseline of the outcome variables demonstrates the highest level of prognostic relationship with the outcome variables. Pocock et al, (2002) had observed that a correlation as high as 0.7 is quite plausible for the same variable measured at baseline and after treatment.

Some baseline covariates have at least correlation of 0.3 with the outcome; however, in some cases, they are seen to have a very similar prognostic strength to the baseline of the outcome variable. This emphasizes the need to consider some other prognostic covariates for adjustment apart from the baseline of the outcome variable – although this needs to be balanced in the context of 'independence'.

In all the three trials studied, socio-demographic characteristic such as: age, sex and social status do not have an appreciable level of correlation with the outcome variables. In fact their level of prognostic strength with any of the outcome variables falls below 0.3, which suggests a limited need to account for these factors. Even if significant imbalance exist in these covariates accounting for them does not in any way improve the status of estimate of treatment effect. Findings from the earlier chapters, 4 and 5, in this study agree with previous authors (Altman et al 2000) that it is not necessary to balance such covariates between groups. Again, unless a covariate has an established correlation of at least 0.3 with the outcome it is of little or no use including such a covariate in a model for statistical adjustment (Altman 1985; Cox & McCullough 1982; Senn 1994). This thus implies that the practice by which covariate adjustment is based on *a priori* selection of covariates which usually include age and sex needs re-visiting as only covariate that are prognostic enough may be included in the model. The review of current practices regarding covariate adjustment in clinical trial setting (in chapter 6) had shown that it is not a common practice to investigate the prognostic strength of a covariate before considering such for statistical adjustment. Sometimes, a covariate that has at least a correlation of 0.3 with the outcome is not capable of providing an independent explanation of the variability in the outcome, thus including such covariate in the model will only amount to redundancy and undue increase in the complexity of the model. Thus, the criterion of a minimum correlation of 0.3 with the outcome is only necessary but not sufficient in order to consider a covariate for statistical adjustment.

Generally, in all the three empirical trials, there is a tendency for reduction in the prognostic strength of the covariates at the latter follow-period. This may be connected with missing values as a result of loss to follow-up that is more apparent in the latter follow-up period (with correlation being greater among ‘responders’). Also, the appearance in the model of covariates which is a measure of the importance of the prognostic covariate varies with follow-up periods, this emphasizes the need to separately explore the covariates by the follow-up periods to determine which goes into the model for adjustment. The pattern of precision of the statistical method perfectly fits the earlier theoretical models in chapter 4 using simulated datasets. This thus lends credence and contributes to the plausibility of the statistical program that drives the simulation exercise in this study. Even though precision of ANOVA, CSA and ANCOVA differ across trials with the degree of difference which was dependent on the level of correlation between the covariate and the outcome, the conclusions around the null hypothesis is also the same for all the methods except in one instance. At the 12 months follow-up of the StarTBack trial, both adjusted methods (ANCOVA and CSA) provide different conclusions to ANOVA. For

ANCOVA and CSA the null hypothesis is rejected indicating superiority of the new treatment whereas for ANOVA the null hypothesis is accepted implying no evidence to reject the null hypothesis. This thus re-affirms the possibility of having different conclusions on the null hypothesis regarding treatment effect when both adjusted and unadjusted crude analyses are used; Christesen et al, (1985) had earlier reported this possibility. In both studies, imbalance could have only occurred in the opposite direction (Figure 4.1) of the treatment as treatment effect favours the adjusted analyses each time.

7.7 Conclusion

Overall, ANCOVA is better than the other two methods (CSA and ANOVA) in most trial scenarios. Further adjustment of prognostic covariates can enhance both precision and bias of estimates, researchers should however weigh the benefit and 'cost' of inclusion of more covariates in the model before-hand. However, when correlation is just around the threshold of 0.3, precision of estimate of effect may be influenced if further adjustment involve dummy variables of ordinal prognostic covariates. Findings of the results of this chapter alongside findings of previous chapters queried the practice of *a priori* specification of covariates for statistical adjustment. More often, covariates that are such specified are not prognostic enough to influence outcome, examples of covariate that always fall in this category are; age, sex and socio-class. Covariates that appear to be commonly prognostic and independent in effect across the three spinal pain trials include the corresponding baseline pain/disability variable; expectation of outcome and general health status (particular if scored through the EuroQoL EQ5D measure). By selecting covariates *a priori* researchers run the risk of missing some key independently prognostic factors that may have an important influence on the outcome – and hence, risk bias and reduced precision of the treatment effect. The inclusion of any covariate should be judged not only by its prognostic strength with the outcome variable, but also on the amount of variability in the outcome which it is capable of explaining *independently* amidst other covariates.

In general, the results of the empirical evaluation parallel those established from the simulation study (in chapters 4 and 5). It has been re-affirmed in this chapter that the correlation between the pre and post-treatment score plays important role in determining the relative bias and precision of the estimate of treatment effect. When this correlation is high, considerable amount of bias presents with crude estimate of treatment effect which is also less precise; adjusted analysis (CSA or ANCOVA) is better than ANOVA. However, when correlation between pre and post is low, especially below 0.3, ANOVA is reasonable and using CSA is not advisable.

Chapter 8: Conclusions, Recommendations and Limitations

8.1 Introduction

The RCT, although is the gold standard design for investigating treatment effectiveness, is not without its limitations. A major limitation of this design is the fact that it does not guarantee equality of treatment arms. Chance imbalance in prognostic covariates as a result of random allocation of patients if not accounted for, may have important effect on the estimate of treatment effect, causing bias. This study focuses on trials in which improvement is measured by change from baseline status or scores of a primary outcome variable following intervention. It represents a comprehensive assessment of three statistical methods that may be employed in this situation – ANOVA, CSA and ANCOVA – in respect of their precision, bias of effect estimates, statistical power and efficiency (relative sample size), which represents the overall aim of the study.

At the outset, a statistical program was developed to carry out the simulation aspect of the work. The program incorporates levels of certain experimental factors, such as: correlation between pre- and post-treatment outcome scores, magnitude of treatment effect, baseline chance imbalance, direction of imbalance and nominal power. The level of imbalance used in the simulation is proportional to the sample size or the size of treatment effect; large imbalance corresponds to small sample size or large effect size and small imbalance corresponds to large sample size or small effect size. At a given trial scenario, each level of these experimental factors was combined together and each of these three statistical methods was used simultaneously for comparison of results. In all, 210 trial scenarios in 210, 000 simulated datasets were involved. The same datasets were used to study all the four trial attributes: precision, bias, statistical power and efficiency. The statistical program also contained certain commands that tested three basic models assumptions; normality of residuals, linearity of the baseline covariates and outcome relationship and parallel regression lines assumption. A detailed account of the methods involved in the study has already been given in chapter 3.

The empirical datasets were from three clinical trials previously conducted by the Arthritis Research UK Primary Care Centre. The availability of empirical trial datasets played three major roles as far as this study is concerned: 1) it provided a means of validating the results of the simulation on real data; 2) It provided a platform to determine what level of prognostic relationship exists between baseline measures and primary outcomes in trials involving musculoskeletal conditions, thus making it possible to know what covariates are necessary to adjust for if adjustment is actually important (this information is important for the design and direction of statistical analysis of future trials related to musculoskeletal conditions); and 3) it made statistical model building and comparison possible.

8.2 Further Discussions

In respect of the overall aim of the study, the findings of this work agrees with previous authors on the subject of statistical adjustment of covariate imbalance in randomised controlled trial settings. For example, the simulated results in chapter 4 confirm what previous authors (Matthews, 2000; Pocock et al, 2002) have reported; that is, when treatment groups are equal at baseline, all the three methods for statistical analysis are unbiased as they yield the same estimate of treatment effect at that time. However, when there is baseline imbalance in treatment groups, with respect to the estimate of effect by ANCOVA which is theoretically unbiased (Mathews, 2000; Van Breukelen, 2006), this study is in agreement with previous authors (Camilli & Shepard 1987; Senn 1991, Van Breukelen 2006) in showing that the estimates of effect by CSA and ANOVA are biased. As was shown in previous chapters, major drivers of bias in the estimate of effect by CSA and ANOVA are the degree of prognostic relationship of the covariate with the outcome and both level and direction of baseline imbalance. The study has also shown the comparative benefits of the statistical methods in terms of precision at different trial scenarios. It pointed out the fact that the difference in precision of these statistical methods is solely dependent on the level of baseline-outcome correlation and that level of baseline imbalance does not really matter. This agrees with earlier findings by Fleiss (1986), Frison and Pocock (1992), Pocock et al (2002) and Walters (2009).

In addition, although previous authors (Vickers, 2001; Tu et al, 2005) have mentioned that ANCOVA has highest statistical powerof the three methods considered in this study (ANOVA, CSA and ANCOVA), their observations are limited to treatment scenarios in which treatment groups are assumed balanced at baseline. The result of this study agrees with these authors; however, if imbalance exists in baseline values of the outcome variable, depending on its direction, either ANOVA or CSA is more powerful than ANCOVA in most trial scenarios. For example, if imbalance exists in the same direction as treatment, ANOVA is more powerful than ANCOVA and if imbalance exists in the opposite direction of the treatment then, CSA is more powerful. However, this seemingly higher statistical power displayed by both ANOVA and CSA over ANCOVA is due to the fact that the estimate of effect that either of these methods yields is inherently marred, with false positive errors. Whereas, ANCOVA yields a smaller unbiased estimate of treatment effect, the estimate of treatment effect by either ANOVA or CSA at the time is unduly large and biased.

Furthermore, the results of the study show that, depending on the correlation between the baseline and the outcome variable, both ANCOVA and CSA can lead to a remarkable reduction in the original sample size if they are specified as methods for primary analysis. The higher the correlation the smaller the sample size required when using ANCOVA compared to ANOVA. However, the benefit of sample size reduction by CSA is only possible at a baseline-outcome correlation greater than 0.5. Irrespective of the level of baseline imbalance, ANCOVA secures a reduction of up to 50% in the original sample size when correlation between pre and post treatment score reaches 0.7. Previous authors (Porter & Raudenbush, 1987; Frison & Pocock, 1992; Pocock et al, 2002; Walters, 2009) have also mentioned the benefit of sample size reduction (efficiency) for using ANCOVA instead of ANOVA.

Similarly, the results of this study show that whereas the effect estimate and the precision of effect estimate for both ANOVA and ANCOVA are similar when baseline-outcome correlation is low even if baseline imbalance is large, they differ considerably when the baseline-correlation is high and baseline imbalance is low. Thus a large imbalance in a non-prognostic variable is not as important as a low imbalance that occurs in a highly prognostic covariate. This therefore argues against the use of baseline significance testing to determine which covariate(s) is/are selected for adjustment. This work perhaps presents the first graphic illustrations (Figure 4.1 & Figure 4.2) on the inappropriateness of selecting a covariate for statistical adjustment simply because of the random occurrence of large baseline imbalance in such a covariate, especially in combination with levels of other factors typical in the clinical trial setting. This finding is in agreement with previous authors (Altman, 1985; Begg, 1990; Schulz et al, 1994; Senn, 1994; Schulz, 1995; Senn, 1997; Fayer & King, 2009), who have variously criticised and condemned the practice.

In addition, this study observed a shift in the practice of covariate adjustment in statistical analysis of randomised controlled trials; more trials are specifying appropriate statistical adjustment as the primary statistical approach. For example, whereas, in 39 (49%) of 80 reviewed articles by Altman and Doré (1990) the authors did not adjust at all and only 12 (24%) of the 50 reviewed articles by Pocock et al (2002) specified a covariate adjusted approach as the primary analysis, in 25 (62.5%) of 40 articles that were included in the review chapter of this study the authors had specified and used appropriate statistical adjustment as primary analysis. This greater observed preference for the adjusted analysis could possibly be due to the various potential benefits that have been attributed to covariate adjusted analysis and an increase in support for this statistical approach over the years. Since the review by Pocock et al (2002) for example**,** various authors have mentioned different benefits of covariate adjusted analysis over the unadjusted and these include: increase in statistical power (Kent et al, 2009; Hernandez et al, 2004; Moore and Vanderlan, 2007; Wang & Hung, 2005); improved type I error rates (Hagino et al, 2004); increased precision of estimates of treatment effect (Tsiatis et al, 2007; Wang & Hung, 2005); and reduced bias, giving more accurate estimates of the true value (Altman & Doré, 1990). The simplicity of the unadjusted analysis may no longer be a sufficient reason to continue to prefer this naïve method as the first line statistical approach in a clinical trial setting. There is evidence therefore of increasing usage and awareness of the merits of the 'adjusted' approach over the unadjusted approach. However, despite this increasing trend in the application of appropriate statistical adjustment, there still remain a substantial proportion of studies that do not properly adjust; in this review 15 (37.5%) of 40 trials were unadjusted (crude comparison of effect or that based on change from baseline). In fact, this study observed a 100% increase in the use of analysis based on 'change' compared with what was recorded by Altman & Doré (1990); these authors recorded a 15% utilisation compared to 30% observed in this study.

With respect to covariate selection in empirical trial datasets, this study questioned the idea of a prior specification of covariates to include in the model for adjustment. None of the variables (age, sex, socio-economic status) which are usually treated as such had meaningful prognostic relationship with the outcome. In all three real trials examined in this study, they all had a correlation of less than 0.3 with the outcome. Previous authors (Cox & McCullough, 1982; Senn, 1994) have recommended a minimum covariate-outcome correlation of 0.3 for covariate selection in statistical adjustment. Thus, the practice of a priori specification negates the possibility of gathering maximum evidence for or against the effectiveness of treatment under investigation based on the current dataset. It was however observed in this study that the condition of having a minimum correlation of 0.3 with the outcome would not be sufficient to select a covariate for adjustment, as some of the covariates that met this selection criterion did not have reasonable amount of independent explanation of the variability of the outcome variable. Thus, by inspecting the estimate of treatment effect and the associated precision for adding each of the covariates that met the inclusion criterion of a minimum of 0.3 in the empirical datasets, this study suggests that for a covariate to be worthy of model inclusion it should be capable of independently contributing an R^2- change of at least 0.005.

This study has investigated the subject of covariate adjustment in a randomised controlled trial involving a single assessment of a continuous outcome variable in a manner that no previous study had done. For example, no previous study has used the same datasets to investigate trial or treatment attributes of bias, precision, statistical power and efficiency in relation to the three statistical methods (ANOVA, CSA and ANCOVA) as was done in this study.

A major contribution of this work is that it attempts to settle the age-long divided views and opinions on the benefit of using of CSA as a method for statistical adjustment in an RCT (Altman, 1985; Senn, 1989; Altman, 1991; Senn, 1991). The study has made clear statements on the comparative advantages of each of the methods for statistical analysis in different trial scenarios; it is hoped that these will go a long way to facilitate informed decisions on when to use and when not to use particular methods. Trial situations in which model-based adjustment (ANCOVA), basic adjustment by CSA and unadjusted analysis will possibly yield different estimates of effect have been highlighted to provide guidance on future analysis of a randomised controlled trial with a continuous outcome variable. The three methods have been observed to yield different conclusions on the estimate of treatment effect in empirical trial settings (Christensen et al, 1985; Piantadosi, 1997).

This work study, in a way that has not previously been attempted investigate (in relation to three statistical methods ANOVA, CSA and ANCOVA) the combinations of various levels of experimental factors obtainable in empirical trial situations that can have serious influence on the selected trial attributes . For example, the effect of direction and size of covariate imbalance in combination with various levels of covariate-outcome correlation at different levels of anticipated treatment effect – small, medium and high (Cohen, 1982) – have been investigated. Previous authors (Vickers, 2001; Tu et al, 2005) dealing with specific attributes such as statistical power have assumed that baseline scores in outcome variable are equal. The same datasets were used to study these trial attributes.

8.3 Summary of findings

In the subsequent subsections the findings of this study are briefly summarised.

8.3.1 Bias

In agreement with previous findings (Matthews 2000) there is no difference in the estimate of treatment effect for all the statistical methods (ANOVA, CSA and ANCOVA) when treatment groups are identical at baseline – they are all unbiased. This situation rarely occurs in the real life clinical trial setting: due to random variation (randomisation rules out systematic variation). However, when groups are not balanced, only ANCOVA is unbiased and the level of bias associated with estimate of effect by ANOVA and CSA is dependent on the level of baseline-outcome correlation and baseline imbalance. When baseline imbalance is small following randomisation, the estimate of effect that results by using ANOVA is less biased compared with the estimate of effect of an unadjusted analysis when there is a larger baseline imbalance. This therefore establishes the benefit of design methods (such as stratification and minimisation) i.e. to ensure that baseline imbalances are small. Unadjusted analyses based on trials with a stratified-design will therefore yield estimates of effect that are less biased; although adjustment is preferred as it will ensure unbiased estimates and certainly in relation to improved precision and power (see following subsections). There is an indication that if appropriate statistical adjustment follows, the benefit of stratification and or minimisation in reducing bias in the estimate of effect is the same for both small and large trials.

ANOVA yields an equal degree of bias but in opposite directions, depending on the direction of baseline imbalance. A negative bias occurs when baseline imbalance in a prognostic variable favours the treated group (treated group is better at baseline) and the estimate of effect is positively biased when the baseline imbalance favours the control group. When the level of baseline-outcome correlation is low ($r<0.3$), the bias in the effect estimate by ANOVA is minimal. Also with ANOVA, whereas large imbalance does not matter unduly if correlation is low, a small imbalance in strong prognostic variable causes considerable bias in the estimate of effect. Given a level of prognostic relationship of covariate with the outcome variable, the degree of bias by ANOVA is directly related to the level of chance imbalance. In addition, size of treatment effect to be determined, or, by implication, trial sample size, does not affect bias in effect estimate. Appropriate statistical adjustment is as much of benefit to small trials as it is to large trials. Even though a small absolute imbalance results when a large sample is randomly allocated to treatment groups and large absolute imbalance when the randomly allocated sample size is small, at the same level of other factors, the associated bias is approximately the same in both cases if the imbalance were standardized) when compared with the estimate of effect by ANCOVA.

In the case of CSA, the estimate of effect is positively biased when imbalance is in the same direction as the treatment or intervention and negatively biased if the control group has a better prognosis at baseline. The problem with CSA is that it does not take regression to the mean into consideration. The amount of bias associated with CSA varies hugely with the direction of imbalance. The bias is minimal in relation to a strongly prognostic covariate. Unless there is an established strong relationship between the baseline and the outcome variable, CSA should not be used for the primary analysis. By and large, bias in the estimate of effect by ANOVA reduces as baseline-outcome correlation tends to 0 and bias in the estimate of effect by CSA reduces as correlation tends to 1.

Thus unless baseline-outcome correlation is large (as close to1 as possible) do not use CSA as a primary method of analysis and unless baseline-outcome correlation is reasonably low (< 0.3) do not use ANOVA as the primary analysis.

8.3.2 Precision

Within each of the statistical methods, there is evidence of improved precision of estimate when sample size increases or the effect size to be determined reduces. The precision of estimate of effect by ANOVA does not respond to baseline-outcome correlation and not to the size and direction of imbalance. The reason is that there is no term in the ANOVA model that relates to covariate imbalance. The key component in distinguishing the three statistical methods in relation to their precision is the baseline-outcome correlation. Even when treatment groups are balanced, there is considerable difference in the precision of these statistical methods, and selecting the appropriate method of statistical adjustment can lead to markedly higher precision in the estimate of effect (Altman 1985).

When baseline-outcome correlation is less than 0.3, there is little benefit in terms of precision of using ANCOVA instead of ANOVA; however, at a higher level of correlation, ANCOVA presents a more precise estimate than ANOVA. The precision of estimates of effect from CSA is not superior to that from ANOVA unless the baseline-outcome correlation is greater than 0.5; when the correlation is below 0.5 ANOVA yields a more precise estimate than CSA. Both CSA and ANCOVA have very similar standard error estimates, and hence similar precision of estimates, at a baseline-outcome correlation of 0.9. The higher the correlation, the more closely the precision of CSA aligns with that of ANCOVA. Theoretically, both methods are equal in the precision of estimate at a correlation of 1.0. Unless baseline-outcome correlation is high, it is not advisable to use CSA. Relative precision benefits between statistical methods are the same across various sample sizes and are also independent of level and direction of imbalance.

8.3.3 Statistical Power

When treatment groups are comparable at baseline, appropriate statistical adjustment by ANCOVA increases the statistical power of a randomised controlled trial. At such time, improvement in statistical power is not obvious until baseline-outcome correlation is greater than or equal to 0.3.Comparatively, CSA and ANOVA are equally powerful at correlation of 0.5; however, CSA is less powerful to ANOVA at a baseline-outcome correlation of less than 0.5 and more powerful at baseline-outcome correlation of greater than 0.5. When treatment groups are not balanced, both sizes and direction of imbalance, as well as the degree of baseline-outcome correlation, play important roles in the power of the statistical methods. Depending on the direction of imbalance, CSA is prone to both false positive and false negative errors. When baseline imbalance is in the opposite direction of treatment effect, overestimation of treatment effect as a result of the way CSA corrects the imbalance leads to false positive error, thus indicating treatment effect when there is none. Similarly, when baseline imbalance is in the same direction as the treatment effect, there is underestimation of treatment effect by CSA, also as a result of the way it handles baseline imbalance. Here, the test based on CSA fails to identify an effect when one exists.

When the effect estimate is conditional on a baseline imbalance that occurs in the same direction as the treatment effect, then the nominal power of the trial which equals the statistical power of the unadjusted analysis is exaggerated by the effect of the imbalance that was not accounted for by the unadjusted analysis – resulting in false positive error. Also, when effect estimate is conditional on baseline imbalance that exists in the opposite direction from the treatment effect, then the nominal power or the power of the unadjusted analysis by ANOVA is underestimated by the effect of the imbalance which the test cannot account for, thus resulting in false negative error. When imbalance exists, ANCOVA presents a mechanism of handling such imbalance in such way that the effect of the imbalance is removed, and thus does not affect the conclusion as to the treatment effect.

Depending on the levels of other factors, when imbalance is in the same direction as treatment, there is a risk that ANOVA will indicate a treatment effect which the adjusted analyses (CSA and ANCOVA) fail to identify; CSA has the least propensity of detecting a treatment effect in this circumstance. Similarly, when imbalance is in the opposite direction of a treatment effect, the adjusted analyses are better placed to indicate a treatment effect since they are more powerful than the unadjusted analysis. Thus, in this circumstance, there is a chance of indicating treatment effect by either ANCOVA or CSA which ANOVA may fail to detect. The risk of ANCOVA and CSA making different inferences on treatment effect is higher with baseline imbalance occurring in the same direction as treatment than with imbalance in the opposite direction – the possibility is for ANCOVA to indicate a treatment effect which CSA may fail to identify here.

In addition, with imbalance in the same direction as treatment, there is a possibility for all the statistical methods (ANOVA, CSA and ANCOVA) to be inconclusive on the estimate of treatment effect. The effect of imbalance in the same direction as treatment that is not accounted for by ANOVA makes the method prone to false positives – i.e. this method will sometimes indicate a treatment effect when none exists. Also, CSA is known to be susceptible to false negatives and as a result will be unable to identify a treatment effect when one exists. On the other hand, ANCOVA in this situation also ends up having a conditional power that is less than the nominal power with which the study is designed to detect a given magnitude of effect.

However, since ANCOVA yields an unbiased estimate of effect, it thus implies that the conditional power by ANCOVA is actually correct but in relation to the adjusted effect that it produces. Thus, the reduced power ‘suffered’ by ANCOVA when baseline imbalance is in the same direction of treatment effect though proportional to its estimate of effect can be viewed as a trade-off for using this method which of course guarantees unbiased estimate of effect. This is no problem as it only occurs when the treatment effect so detected is smaller than the minimum clinically important difference (MCID). The reduced power by ANCOVA here is not sufficient to compromise appropriate statistical adjustment using this method for the crude estimate of effect by ANOVA.

8.3.4 Efficiency

The choice of statistical method for the analysis of a randomised controlled trial has an important influence on sample size requirements. Appropriate statistical adjustment has the benefit of a considerable reduction in the original sample size that would have been required if the crude unadjusted analysis by ANOVA were to be specified as the primary analysis. The level of the baseline-outcome correlation is the sole determinant of the comparative benefit in terms of sample size reduction between the unadjusted and the adjusted analysis. This observation is parallel to the findings for the precision of the estimate of effect; for a given trial scenario. Both attributes – efficiency (sample size reduction) and precision of estimate of effect – relate to the ability of the statistical methods to control the variability observed in the data.

With ANCOVA as the primary analysis, reduction in the original sample size when groups are comparable is approximately the same compared to when there is maximum or large imbalance in the treatment groups at baseline. Thus, if trial efficiency is in terms of relative reduction in the sample size requirement by these statistical methods, design methods (stratification, minimisation) that are aimed at making treatment groups similar have little or no benefit. Both size and direction of imbalance do not affect how much reduction in the original sample size will be recorded if ANCOVA is used instead of ANOVA. Realistically, this reduction in sample size by ANCOVA can reach half of the original sample size, and more in some trial scenarios.

Unless the correlation of the covariate with the outcome exceeds 0.5, CSA is not more efficient than ANOVA. Both ANOVA and ANCOVA are more efficient or require a lesser number of patients at a given statistical power than CSA when correlation is less than 0.5. At such time also (when $r<0.5$), since CSA would require more patients to be studied on top of the original sample size, given a level of statistical power therefore, the estimate of effect by CSA cannot be valid.

8.3.5 Review on current practices around baseline imbalance in RCTs

There is increased awareness and preference for statistical adjustment in RCT compared to unadjusted or crude comparison of treatment effect between groups. However, a very considerable proportion of RCTs are still being analysed with baseline imbalanced not properly accounted for, ‘change from baseline’ being the commonest single statistical approach for the assessment of a continuous outcome variable. Presently, exploring baseline-outcome correlation to inform which covariate(s) is included in the model for adjustment is far from being practiced. There has not been any improvement whatsoever in the way authors report the allocation techniques adopted in their trials compare to the practice 11 years ago despite various CONSORT statements that recommend clear description of the allocation procedure. Most authors that utilized stratified- blocking did not adjust for stratification factor and this is not consistent with the popular opinion that recommends that stratification or minimisation factors should be adjusted.

Although, compared to existing information, this study records a reduction in the level of practice regarding the use of test of significance to assess baseline comparability with a view to select covariates for statistical adjustment, an unacceptable level of authors of RCTs are still guilty. This practice lacks theoretical support and should be ignored in statistical analysis in clinical trial setting. A large imbalance in a non-prognostic covariate does not warrant statistical adjustment, whereas little or no baseline imbalance in a strong prognostic covariate should be appropriately accounted for to maximise precision of estimate of effect.

8.3.6 Handling covariate(s) in empirical trial datasets

Although all the trials are within the same health domain, the level of baseline-outcome relationship differ from trial to trial; this observation has implication on what is statistically adjusted in each case. In real-life clinical trial settings, covariates specified a priori and entered in the model may be less prognostic than some others which are not specified and thus not accounted for. For example socio-demographic variables such as: age, sex, and socio-economic status which often fall into that category are among the baseline variables with the least level of prognostic relationship with the outcome variable in each of the empirical trials. This raises doubt on the correctness of prior specification of variables for adjustment. Prior specification of covariates to adjust may be necessary but there should still be a cross-checking on such pre-specified covariates by assessing the baseline-outcome correlation to determine the actual prognostic importance of such pre-specified covariates. Information on prognostic strength of a covariate with an outcome is a post hoc issue and should be treated as such.

Allowance should be made in the protocol to include any covariates that are considered as strongly prognostic with the outcome in the model for adjustment, even if such are not pre-specified. No other covariates may be included in the model apart from the baseline-outcome correlation unless it is considered prognostic (r>=0.3). A baseline-correlation of 0.3 has been recommended by previous authors though few in numbers (Cox & McCullough 1982; Senn 1994) as the least correlation a covariate should have with the outcome to be considered necessary for adjustment. This condition as was detailed in chapter 7 is though necessary (results from the simulation in chapters 4 & 5 agree to this) but not really sufficient for covariate selection for adjustment. A covariate may have a correlation of at least 0.3 with the outcome and yet not be able to independently explain the variability in the outcome.

Researchers should avoid making their model unnecessarily complex, thus should include only few important covariate(s) that have important contributions to the variability in the outcome variable. Adjusted analysis that accounts for strong prognostic baseline factor(s) should be specified as primary analysis and should be emphasized in case the effect estimate that results is noticeably different from that obtained from the unadjusted analysis. In trial scenarios where there are many covariates that meet this criterion, adjusting for them all will not only increase the complexity of the statistical model for adjustment, it will also lead to the redundancy of some covariates in the model. A necessary and sufficient condition for inclusion of covariates in the model for adjustment includes subjecting such prognostic covariates with correlation of greater than or equal to 0.3 with the outcome variable to test of collinearity and then establishing the amount of independently explained variance in the outcome variable for each of them. This study proposes that a covariate should be capable of explaining at least 0.005 of the variability in the outcome independently before including it in the final model. The reason for this as evidenced from the analysis of the empirical datasets is that there is little or no implication on both the precision and size of estimate of effect for including a covariate with lesser R^2-change in the final model.

This proposition for covariate selection ensures the best estimate of effect with as parsimonious an overall model as possible. As a result of the importance of the information on the prognostic strength of covariates in the statistical analysis of clinical trials, authors should endeavour to report on this by specifying baseline-outcome correlations in their trials especially those that are greater than or equal to 0.3.

8.4 Recommendations

Following the findings of this study, the following are recommended in view of the future statistical analysis of randomised controlled trials.

- Appropriate statistical adjustment of prognostic covariates is important in all real life trial scenarios irrespective of whether measures are taken to balance baseline imbalance or not.
- Using test of significance to assess baseline imbalance lacks theoretical support and should be ignored in practice.
- In any given trial, prognostic strength of covariates should be assessed by using appropriate correlation coefficients; such covariates with correlation of at least 0.3 with the outcome should be marked and reported. This will inform what information to collect in future trials. Some pre-specified covariates are often less prognostic than some other covariates that already exist in the trial data; they are thus less useful. Pre-specification of covariate to be adjusted will fail to make use of existing information that can lead to aless biased estimate of effect. Pre-specification is therefore not efficient and should be avoided in practice.
- Not all covariates that have at least correlation of 0.3 with the outcome are fit for inclusion in the model for statistical adjustment. Many variables have overlapping effect; it is those variables that have major independent

effects on the outcome variable that statisticians need to more readily target as covariates. Covariates with correlation of at least 0.3 which are also capable of independently explaining a considerable variability in the outcome should be entered into the model. In this study, I propose a prognostic covariate should be capable of independently explaining at least 0.005 of the variability in the outcome to be considered for inclusion in the model.

- The statistical analysis section of any trial protocol should be clear on the path-way to statistical analysis and be transparent on the approach to statistical adjustment.
- Change score analysis does not perform better than ANOVA or crude estimate of effect unless the correlation between the pre and post treatment score exceeds 0.5. Thus, CSA should not be used unless there is an established correlation of at least 0.5 between the pre and post treatment score.
- Even then, neither ANOVA nor CSA should be specified as the method for the primary analysis as they are prone to false positives and false negative errors depending on the direction of imbalance.
- ANCOVA should be specified as the primary analysis. Adjusting for variables that are independently prognostic will yield a markedly better estimate of effect. However, there are usually no considerable changes in both the estimate of effect and the precision of estimate after the third or fourth covariate is added to the model.

- Simplicity of models is very important and should be preserved as much as possible.
- ANCOVA is most efficient and specifying it as the method for the primary analysis leads to considerable reduction in the original sample size irrespective of the direction and size of imbalance. This benefit should be appropriated at the design stage of the trial. It minimises trial costs and time of result delivery. From an analysis viewpoint, it ensures that a reduced number of patients are exposed to adverse effects associated with drug or trial of any intervention whilst maintain the power of the study to address the hypothesis is sets out to test .

8.5 Limitations

This study concerns randomised controlled trials in which there is a single post treatment analysis of a continuous outcome variable; thus, the primary outcome variable is assumed to be normally distributed. The findings of the study therefore are limited in interpretation to RCTs with normally distributed outcomes and they should not be generalised to RCTs in which the outcome variable is non –normal; such as that which involve a binary outcome or that in which time-to-event is the primary outcome. Another limitation of the study is that it does not give considerations to missing data, which are almost a natural phenomenon in follow-up studies. Although there are formal procedures and techniques by which missing data are handled, this topic was deemed to be outside the scope of this study.

8.6 Future research areas

In view of the foregoing, future research could usefully investigate the trial attributes already considered in this study but for outcome variables on different scales of measurement, for example, binary outcomes or time-to-event outcomes for trials involving survival analysis. Another possible investigation in this domain would be to assess the effect of repeated measures on the trial attributes (bias, precision, statistical power and efficiency) already studied in relation to particular statistical methods appropriate for such trials. It might be of interest to incorporate the idea of missingness into the hypothetical trials datasets to determine how this might affect the results. A further possible area of research interest is to investigate what happens if any of the underlining assumptions for ANCOVA are violated; for example, the assumption of linear relationship between the covariate and outcome variable.

8.7 Conclusion

If CSA or ANOVA must be used as a primary method, the level of baseline-outcome correlation must be checked to ensure that it is high or low enough respectively for the purpose. Statistical adjustment that properly adjusts for covariate imbalance must be specified as primary method of analysis; for example, model-based adjustment methods; ANCOVA or linear regression for numerical endpoints, logistic and Cox regression for the binary and time to event outcomes respectively. It is no longer sufficient to include covariates that have been specified a priori in the model for adjustment, researchers should assess the level of relationship covariates have with the outcome. Only when a covariate has a correlation of at least 0.3 with the outcome variable and capable of explaining appreciable amount of the variability in the outcome say 0.005 that such covariate should be included in the model for adjustment. Even though conclusion on treatment effect across the statistical methods may not differ, the need for an unbiased estimate of effect which is close to the true value should make authors always specify appropriate covariates adjusted method as the first line statistical methods in RCTs.

References

Abdeljaber MH, Monto AS, Tilden RL, Schork MA, Tarwotjo I. 1991. The impact of vitamin A supplementation on morbidity: a randomized community intervention trial. *Am J Public Health*, 81 (12), 1654–1656.

Abdulla G, Critchlow T, Arrighi W. 2004. Simulation data as data streams. *SIGMOD Record*, 33 (1), 89–94.

Alford L, 2006. On differences between explanatory and pragmatic clinical trials. *N Z J Physiother*, 35(1), 12–16.

Altman DG.1985. Comparability of randomised groups. *Statistician*, 34, 125–136.

Altman DG. Practical Statistics for Medical Research. Chapman and Hall, London, 1991.

Altman DG, Doré CJ. 1990. Randomisation and baseline comparisons in clinical trials. *Lancet*, 335 (8682), 149–153.

Altman DG, Doré CJ. 1991. Baseline comparisons in randomized clinical trials. *Stat Med* , 10 (5), 797–799.

Altman DG, Bland JM. 1999. How to randomise. *BMJ*, 319 (7211), 703–704.

Altaman DG, Schultz KF. 2001.Concealing treatment allocation in a randomised trials.BMJ23;446-447.

Altman DG, Bland JM.2005.Treatment allocation by minimisation. BMJ 330;843-

Andersson GB, Lucente T, Davis AM, Kappler RE, Lipton JA, Leurgans S. 1999. *N Engl J Med* 4, 341 (19),1426–31.

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

Austin PC. 2008. The performance of different propensity-score methods for estimating relative risks. *J Clin Epidemiol*, 61 (6), 537–545.

Bamgboye EA. A Companion of Medical Statistics; FOLBAM Publishers, Ibadan, 2006.

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

Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, Pitkin R, Rennie D, Schulz KF, Simel D, Stroup DF. 1996. Improving the quality of reporting of randomized controlled trials. The CONSORT statement. *JAMA*, 276 (8), 637–639.

Begg CB. 1990. Suspended judgment. Significance tests of covariate imbalance in clinical trials. *Control Clin Trials*, 11, (4) 223–225.

Burgess DC, Gebski VJ, & Keech AC. 2003. Baseline data in clinical trials. *Med J Aust*, 179, (2) 105–107.

Cai H, Xia J, Xu D, Gao D, Yan Y. 2006. A generic minimization random allocation and blinding system on web. *J Biomed Inform*, 39 (6), 706–719.

Callas PW. 2008. Searching the biomedical literature: research study designs and critical appraisal. *Clin Lab Sci*, 21, (1) 42–48.

Campbell MK, Elbourne DR, Altman DG. 2004. CONSORT statement: extension to cluster randomised trials. *BMJ*, 328 (7441), 702–708.

Camilli G, Shepard L.A. 1987. The inadequacy of ANOVA for detecting test bias. *J Educ Stat*, 12 (1), 87–99.

Canner PL.1991. Covariate adjustment of treatment effects in clinical trials. *Control Clin Trials*, 12 (3), 359–366.

Chalmers I.1999. Why transition from alternation to randomisation in clinical trials was made. *BMJ*, 319: 1372.

Chan AW, Altman DG. 2005. Epidemiology and reporting of randomised trials published in PubMed journals. *Lancet*, 365 (9465), 1159–1162.

Chan S, Jonsson A, Bhandari M. 2007. Planning a clinical research study. *Indian J Orthop*, 41 (1), 16–22.

Chow SC, Liu JP. Design and Analysis of Clinical Trials: Concepts and Methodologies. John Wiley, New York ,1998.

Christensen E, Neuberger J, Crowe J, Altman DG, Popper H, Portmann, B, Doniach D, Ranek L, Tygstrup N, Williams R. 1985. Beneficial effect of azathioprine and prediction of prognosis in primary biliary cirrhosis. Final results of an international trial. *Gastroenterology*, 89 (5), 1084–1091.

Christie J, O'Halloran P, Stevenson M. 2009. Planning a cluster randomized controlled trial: methodological issues. *Nurs Res,* 58 (2), 128–134.

Cohen J. Statistical Power Analysis for the Behavioral Sciences, 2nd edition. Lawrence Erlbaum, Hillsdale, 1988.

Clinical Standards Advisory Group on Back Pain. Back pain. HMSO Stationery Office, London,1994.

Cook TH, DeMets DL. Introduction to statistical methods for clinical trials; Chapman & Hall/CRC, Madison, 2008.

Cotton PB. 2000. Randomization is not the (only) answer: a plea for structured objective evaluation of endoscopic therapy*. Endoscopy***,** 32 (5),402–405.

Cox DR, McCullough P. 1982. Some aspects of covariance. *Biometrics*, 38, 541–561.

Curtis LM. Clinical trials design, conduct and analysis. Oxford University Press, New York, 1986.

Dziedzic K, Hill J, Lewis M, Sim J, Daniel J, Hay EM. 2005. Effectiveness of manual therapy or pulsed shortwave diathermy in addition to advice and exercise for neck disorders: a pragmatic randomized controlled trial in physical therapy clinics*. Arthritis Rheum*, 53 (2), 214–222.

Di MC, Martinez M, Menard JF, Petit M, Thibaut F. 2001. Evidence of a cohort effect for age at onset of schizophrenia. *Am J Psychiatry*, 158 (3), 489–492.

Doig GS, Simpson F. 2005. Randomization and allocation concealment: a practical guide for researchers. *J Crit Care*, 20 (2), 187–191.

Doncaster CP, Davey AJH. Analysis of variance and covariance: how to choose and construct model for the life sciences. Cambridge University press, Cambridge, 2007.

Eccles M, Grimshaw J, Campbell M, Ramsay C. 2003. Research designs for studies evaluating the effectiveness of change and improvement strategies. *Qual Saf Health Care*, 12 (1), 47–52.

Egger MJ, Coleman ML, Ward JR, Reading JC, Williams HJ. 1985.Uses and abuses of analysis of covariance in clinical trials. *Control Clin Trials*, 6:12–24.

Ekouevi DK, Azondekon A, Dicko F, Malateste K, Toure P, Eboua FT, Kouadio K., Renner L, Peterson K, Dabis F, Signate-Sy H, Leroy V, Pwada IP. 2011. 12-month mortality and loss-to-program in antiretroviral-treated children: The IEDEA Pediatric West African Database to evaluate AIDS (PWADA) 2000-2008. *BMC Public Health*, 11 (1), 519.

Everitt BS, Pickles A. Statistical Aspects of the Design and Analysis of Clinical Trials – revised edition. Imperial College Press, London, 1999.

Fayers PM, King MT. 2009. In reply to Berger "don't test for baseline imbalances unless they are known to be present?" *Qual Life Res,* 18(4),401–402

Fedorov V, Jones B. 2005. The design of multicentre trials. *Stat Methods Med Res,* 14 (3), 205–248.

Fleiss JL. The Design and Analysis of Clinical Experiments. John Wiley, New York,1986.

Frison L, Pocock SJ.1992. Repeated measures in clinical trials: analysis using mean summary statistics and its implication for design. *Stat Med,* 11 (13),1685–1704.

Ford I, Norrie J, Ahamadi S. 1995. Model inconsistency, illustrated by Cox proportional hazards model. *Stat Med,* 14 (8),735–746.

Funai EF, Rosenbush EJ, Lee MJ, Del PG. 2001. Distribution of study designs in four major US journals of obstetrics and gynecology. *Gynecol Obstet Invest*, 51 (1), 8–11.

Godwin M, Ruhland L, Casson I, MacDonald S, Delva D, Birtwhistle R, Lam M, Seguin R. 2003. Pragmatic controlled clinical trials in primary care: the struggle between external and internal validity. *BMC Med Res Methodol,* 3, 28.

Grimes DA, Schulz KF. 2002a. An overview of clinical research: the lay of the land. *Lancet*, 359 (9300), 57–61.

Grimes DA, Schulz, KF. 2002b. Descriptive studies: what they can and cannot do. *Lancet*, 359 (9301), 145–149.

Grizzle JE. 1982. A note on stratifying versus complete random assignment in clinical trials. *Control Clin Trials,* 6, 146–155*.*

Hagino A, Hamada C, Yoshimura I, Ohashi Y, Sakamoto J, Nakazato H. 2004. Statistical comparison of random allocation methods in cancer clinical trials. *Control Clin Trials*, 25 (6), 572–584.

Hall NS. 2007. R.A. Fisher and his advocacy of randomisation. *J Hist Biol,* 40 (2),295–325.

Hay EM, Mullis R, Lewis M, Vohora K, Main CJ, Watson P, Dziedzic KS, Sim J, Lowe CM, Croft P. 2005. Comparison of physical treatments versus a brief pain-management programme for back in primary care: a randomised clinical trial in physiotherapy practice. *Lancet,* 365, 2024–2030.

Hallstrom A, Davis K. 1988. Imbalance in treatment assignments in stratified blocked randomization. *Control Clin.Trials*, 9 (4), 375–382.

Hedman C, Andersen AR, Olesen J. 1987. Multi-centre versus single-centre trials in migraine. *Neuroepidemiology*, 6 (4), 190–197.

Helms, P.J. 2002. 'Real world' pragmatic clinical trials: what are they and what do they tell us? *Pediat Allergy Immunol.* 13 (1), 4–9.

Hernandez AV, Steyerberg EW, Habbema JD. 2004. Covariate adjustment in randomized controlled trials with dichotomous outcomes increases statistical power and reduces sample size requirements. *J Clin Epidemiol*, 57 (5), 454–460.

Hennekens CH, Buring JE. Epidemiology in Medicine. Little, Brown and company, Boston ,1987.

Herbert R, Jamtvedt G, Mead J, Hagen KB. Practical Evidence-Based Physiotherapy. Butterworth Heinemann, London, 2005.

Hewitt CE, Torgerson DJ. 2006. Is restricted randomisation necessary? *BMJ*, 332 (7556), 1506–1508.

Hill JC, Whitehurst DG, Lewis M, Bryan S, Dunn KM, Foster NE, Konstantinou K, Main CJ, Mason E, Somerville S, Sowden G, Vohora K, Hay EM. 2011. Comparison of stratified primary care management for low back pain with current best practice (STarT Back): a randomised controlled trial. *Lancet,* 378 (9802),1560–1571.

Hollis S, Campbell F. 1999. What is meant by intention to treat analysis? Survey of published randomised controlled trials. *BMJ*, 319 (7211) 670–674.

Horwitz RI, Viscoli CM, Berkman L, Donaldson, RM, Horwitz, S.M., Murray, CJ, Ransohoff DF, Sindelar J. 1990. Treatment adherence and risk of death after a myocardial infarction. *Lancet*, 336 (8714), 542–545.

Hotopf M (2002): The pragmatic randomized controlled trial. *Adv Psychiatr Treat,* 8, 326–333.

Hughes GR Currie CSM, Corbett EL. Modelling Tuberculosis in areas of high HIV prevalence. Proceedings of the 2006 winter simulation conference; 459–465.

Huwiler-Muntener K, Juni P, Junker C, Egger M. 2002. Quality of reporting of randomized trials as a measure of methodologic quality. *JAMA*, 287 (21), 2801–2804.

Jadad A. Randomised Controlled Trials. BMJ, London, 1998.

Jafar TH, Islam M, Hatcher J, Hashmi S, Bux R, Khan A, Poulter N, Badruddin S, Chaturvedi. N. 2010. Community based lifestyle intervention for blood pressure reduction in children and young adults in developing country: cluster randomised controlled trial. *BMJ*, 340, c2641.

Jamieson J. 1999. Dealing with baseline differences: two principles and two dilemmas. *Int J Psychophysiol*, 31 (2),155–161.

Juni P, Altman DG, & Egger M. 2001. Systematic reviews in health care: Assessing the quality of controlled clinical trials. *BMJ*, 323 (7303), 42–46.

Kang M, Ragan BG, Park JH. 2008. Issues in outcomes research: an overview of randomization techniques for clinical trials. *J Athl Train.*, 43, (2) 215–221.

Kent DM, Trikalinos TA, Hill MD. 2009. Are unadjusted analyses of clinical trials inappropriately biased toward the null? *Stroke*, 40 (3), 672–673.

Kernan WN, Makuch RM.2001. Letter*. J Clin Epidemiol, 54: 104–107*.

Kernan WN, Viscoli CM, Makuch RW, Brass LM, & Horwitz RI. 1999. Stratified randomization for clinical trials. *J Clin Epidemiol*, 52 (1), 19–26.

Kranzer, K, Ford, N. 2011. Unstructured treatment interruption of antiretroviral therapy in clinical practice: a systematic review. *Trop Med Int Health,* doi: 10.1111/j.1365-3156.2011.02828.x.

Lachin, J.M. 1988. Statistical properties of randomisation in clinical trials. *Control Clin Trials*, 9, (4) 289–311.

Lachin JM. 2000. Statistical considerations in the intent-to-treat principle. *Control Clin Trials*, 21(3),167–189.

Lewis JA.1983. Clinical trials: statistical developments of practical benefit to the pharmaceutical industry. *J R Stat Soc A*, 146: 362–393.

Liu J, Kjaergard LL, Gluud C. 2002. Misuse of randomisation: a review of Chinese randomized trials of herbal medicines for chronic hepatitis B. *Am J Chin Med*, 30 (1), 173–176.

Luepker RV, Perry CL, McKinlay SM, Nader PR, Parcel GS, Stone EJ, Webber LS, Elder JP, Feldman HA, Johnson CC . 1996. Outcomes of a field trial to improve children's dietary patterns and physical activity. The child and adolescent trial for cardiovascular Health. CATCH collaborative group. *JAMA*, 275 (10), 768–776.

Machin D, Campbell M, Fayers P, Pinol A. Sample Size Tables for Clinical Studies. Blackwell Science, Oxford, 1997

Macpherson, H. 2004. Pragmatic clinical trials. *Complement Ther Med*, 12, (2), 136–140.

Mark RG. Computational models of cardiovascular function for simulation, data integration and clinical decision support.

http://www.nsbri.org/research/projects/viewsummary. [Accessed – August 21, 2009]

Markham IS, Rakes TR.1998.The effect of sample size and variability of data on the comparative performance of artificial neural networks and regression. *Computer Ops Res,* 25(4), 251–263.

Matthews J.N.S. An Introduction to andomised Controlled Clinical Trials. Arnold, London, 2000*.*

McLeod, R.S., Wright, J.G., Solomon, M.J., Hu, X., Walters, B.C., Lossing, A. 1996. Randomised controlled trials in surgery: Issues and problems. *Surgery*, 119 (5), 483–486.

Meinert CL. Clinical trials: Design, Conduct, and Analysis. Oxford University Press, New York,1986.

Meinert CL, Tonascia S, Higgins K. 1984. Content of reports on clinical trials: a critical review. *Control Clin Trials*, 5 (4), 328–347.

Mellor R Georgina, Christine SM Currie, Elizabeth L. Corbett, Russell, CH Cheng. Targeted strategies for tuberculosis in areas of high prevalence: A simulation study. Proceedings of the 2007 Winter Simulation Conference.

Miguel R, Miguel AM.2000 Intention to treat analysis is related to methodological quality*. BMJ*, 320:1007.

Millard SP and Neerchal NK. Environmental statistics with S-Plus; Chapman & Hall CRC, Florida, 2001.

Moher D, Schulz KF, Altman D. 2001. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. *JAMA*, 285 (15), 1987–1991.

Molsberger AF, Schneider T, Gotthardt H, Drabik A. 2010. German randomised acupuncture trial for chronic shoulder pain (GRASP) – A pragmatic, controlled, patient-blinded,multi-centre trial in an outpatient care environment. *Pain*, 151; 146–154.

Montori VM, Guyatt GH. 2001. Intention-to-treat principle. *CMAJ,* 165 (10), 1339–1341.

Moore K.L, Vanderlaan MJ. Covariates adjustment in randomised trials with binary outcomes: targeted maximum likelihood estimation. [online resource: http://www.bepress.com/ucbbiostat/paper215/].

Orrell C, Kaplan R, Wood R, Bekker LG. 2011. Virological breakthrough: a risk factor for loss to followup in a large community-based cohort on antiretroviral therapy. *AIDS Res Treat*, 2011, 1–6.

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

Overall, JE, Doyle, S.R. 1994. Implications of chance baseline differences in repeated measurement designs. *J Biopharm Stat,* 4 (2), 199–216.

Parker AB , Naylor CD. 2000. Subgroups, treatment effects, and baseline risks: some lessons from major cardiovascular trials. *Am Heart J*, 139 (6), 952–961.

Peto R, Pike MC., Armitage P, Breslow NE, Cox, DR, Howard SV, Mantel N, McPherson K, Peto J, Smith PG. 1976. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. *Br J.Cancer*, 34 (6), 585–612.

Peto R, Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV, Mantel N, McPherson K, Peto J, Smith PG. 1977. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. II. analysis and examples. *Br J Cancer*, 35 (1), 1–39.

Piantadosi S. Clinical Trials: A methodological Perspective, Wiley, New York, 1997.

Pocock S. 1982. Statistical aspect of clinical trial design. *Statistician*. 31:1–18.

Pocock SJ. Clinical Trials: A practical Approach. John Wiley, New York, 1983.

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

Porter A.C, Raudenbush H.1987. Analysis of covariance:Its model and use in psychological research. *J Couns Psychol*, 34 (4), 383–392.

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

Rangaka MX, Gideon HP, Wilkinson KA, Pai M, Mwansa-Kambafilwe J, Maartens G, Glynn JR, Boulle A, Fielding K, Goliath R, Titus R, Mathee S, Wilkinson RJ. 2011. No discriminatory value of interferon release added to smear negative HIV-tuberculosis algorithms. *Eur Respir J* (in press).

Roberts C, Torgerson D. 1998. Understanding controlled trials: Randomisation methods in controlled trials. *BMJ,* 317:1301.

Roberts, MJ, Russo, R. A student's Guide to Analysis of Variance. Routledge, New York, 1999.

Robinson LD, Jewell NP. 1991. Some surprising results about covariates adjustment in logistic regression models*. Int Stat Rev,* 58:227–40.

Roland M, Torgerson, DJ. 1998. What are pragmatic trials? *BMJ*, 316, (7127) 285.

Rosenberger WF, Sverdlov O. 2008. Handling covariates in the design of clinical trials. *Stat Sci* 23(3), 404-419.

Rothwell, PM. 2000. Responsiveness of outcome measures in randomised controlled trials in neurology. *J Neurol Neurosurg Psychiatry*, 68, (3) 274–275.

Rothwell, PM. 2005. Treating individuals 2. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *Lancet*, 365, (9454) 176–186.

Rothwell PM. Bhatia M. 2007. Reporting of observational studies. *BMJ*, 335 (7624), 783–784.

Rutherford A. Introducing ANOVA and ANCOVA: a GLM Approach. Sage Publications, London, 2001.

Ruiz-Canela M, Martinez-Gonzalez MA., de Irala-Estevez J. 2000. Intention to treat analysis is related to methodological quality. *BMJ*, 320 (7240), 1007–1008.

Sacks FM, Bray GA, Carey VJ, Smith SR, Ryan DH, Anton SD, McManus K, Champagne CM, Bishop LM, Laranjo N, Leboff MS, Rood JC, de JL, Greenway FL, Loria CM, Obarzanek E, Williamson DA 2009. Comparison of weight-loss diets with different compositions of fat, protein, and carbohydrates. *N Engl J Med*, 360 (9), 859–873.

Schulz KF, Chalmers I, Grimes DA, Altman DG. 1994. Assessing the quality of randomization from reports of controlled trials published in obstetrics and gynecology journals. *JAMA*, 272 (2), 125–128.

Schulz KF. 1995. Subverting randomisation in controlled trials. *JAMA*, 274 (18), 1456–1458.

Schulz KF, Grimes DA, Altman DG, Hayes RJ. 1996. Blinding and exclusions after allocation in randomised controlled trials: survey of published parallel group trials in obstetrics and gynaecology. *BMJ*, 312 (7033), 742–744.

Schulz KF, Chalmers I, Altman DG. 2002. The landscape and lexicon of blinding in randomized trials. *Ann Intern Med*, 136 (3), 254–259.

Schulz KF, Altman DG, Moher D. 2010. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. *BMJ*,340:c332

Scott NW, McPherson GC, Ramsay CR, Campbell MK. 2002. The method of minimization for allocation to clinical trials. a review. *Control Clin Trials*, 23 (6), 662–674.

Senn, SJ. 1989a. The use of baselines in clinical trials of bronchodilators. *Stat Med,* 8 (11), 1339–1350.

Senn SJ. 1989b. Covariate imbalance and random allocation in clinical trials. *Stat Med*, 8 (4), 467–475

Senn SJ. 1991. Baseline Comparisons in randomised clinical trials*. Stat Med,10,1157-1160.*

Senn S, 1990. Statistics in clinical trials. *Lancet,* 335, 514

Senn.S 1994. Testing for baseline balance in clinical trials. *Stat Med,13,1715–1726.*

Senn SJ. Statistical Issues in Drug Development. John Wiley and sons, Chichester, 1997.

Sim J, Wright C. Research in Health Care. Concepts, Designs & Method. Cheltenham, Nelson Thornes, 2000.

Sim J. 2003 Baseline Imbalance in Randomised Controlled Trials and the Suitability of different analytical strategies: a simulation study (unpublished)

Taves DR. 2004. Faulty assumptions in Atkinson's criteria for clinical trial design. *J R Stat Soc* 167 (1), 179–181.

The European Agency for the evaluation of medicinal products; Evaluation of medicine for human use (2003); Committee for proprietary medicinal products CPMW/EWP.

The National Emergency Medical Services to Children Data Analysis Resource Centre [online resource:http://www.nedarc.org. Accessed - 04/04/09

Tolerance. [on–line resource, http://128.97.141.26/stat/spss/webbooks/reg/chapter2/spssreg2.htm. Accessed - 06/07/11].

Torgerson DJ, Roberts C.1999. Understanding controlled trials. Randomisation methods: concealment. *BMJ*, 319 (7206), 375–376

Tsiatis AA, Davidian M, Zhang M, Lu X. 2007. Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: a principled yet flexible approach. *Stat Med*, 27 (23) 4658–4677.

Tu D, Shalay K, Pater J.2000. Adjustment of treatment effect for covariates in clinical trials: Statistical and Regulatory issues; *Drug Inf J,* 34, 511–523.

Tu Y K, Blance A, Clerehugh V, Gilthorpe MS. 2005. Statistical power for analyses of changes in randomized controlled trials. *J Dent Res,* 84 (3), 283–287.

Twisk J, Proper K. 2005. Is analysis of covariance the most appropriate way to analyse changes in randomised controlled trials? *J Clin Epidemiol*, 58 (2), 211–212.

Van Breukelen GJP. 2006. ANCOVA versus change from baseline had more power in randomized studies and more bias in nonrandomized studies*. J Clin Epidemiol*, 59, 920–925.

Vickers AJ. 2001. The use of percentage change from baseline as an outcome in a controlled trial is statistically inefficient: a simulation study. *BMC.Med.Res.Methodol,* 1, 6.

Vickers AJ, Altman DG. 2001. Analysing controlled trials with baseline and follow up measurements. *BMJ,* 323: 1124.

Von EE, Altman DG, Egger M, Pocock SJ, Gotzsche PC, Vandenbroucke JP. 2007. Strengthening the Reporting of Observational Studies in Epidemiology

(STROBE) statement: guidelines for reporting observational studies. *BMJ*, 335 (7624), 806–808.

Wang D, Bakhai A. Clinical Trials: A Practical Guide to Design, Analysis and Reporting. Remedica Medical Education and Publishing, London, 2006.

Walters SJ. Quality of life outcomes in clinical trials and health care evaluation: a practical guide to analysis and interpretation. Chichester: Wiley 2009.

Wang SJ, Hung HM. 2005. Adaptive covariate adjustment in clinical trials. *J Biopharm Stat,* 15 (4), 605–611.

Wei LJ, Lachin JM. 1988. Properties of the urn randomization in clinical trials. *Control Clin Trials*, 9 (4), 345–364.

Wieder HA, Beer AJ, Lordick F, Ott K, Fischer M, Rummeny EJ, Ziegler S, Siewer JR, Schwaiger M, Weber WA. 2005. Comparison of changes in tumor metabolic activity and tumor size during chemotherapy of adenocarcinomas of the esophagogastric junction. *J Nucl Med.* 46 (12), 2029–2034.

Wingo PA, Higgins JE, Rubin GL, Zahniser SC, Eds. An epidemiologic approach to reproductive health. World Health Organization Bulletin, Geneva: WHO, 1994.

Wright CC, Sim J. 2003. Intention-to-treat approach to data from randomized controlled trials: a sensitivity analysis. *J Clin Epidemiol* 56 (9), 833–842.

Appendices

Appendix 1

```
set obs n
   g g=mod(_n,2)
   g z=invnorm(uniform())*1
   g k=invnorm(uniform())*1
   g r=0.3
   g y= z*r+k*(1-r^2)^.5
   replace z=z-g*z’
   replace y=y-g*y
   g c=z-y
   regress y g
   regress c g
   regress y g z
```

Appendix 2

Figure 1: Directional pattern of precision of statistical methods for the analysis of RCTs at levels of baseline imbalance, baseline-outcome correlation and differing effect sizes – 90% power

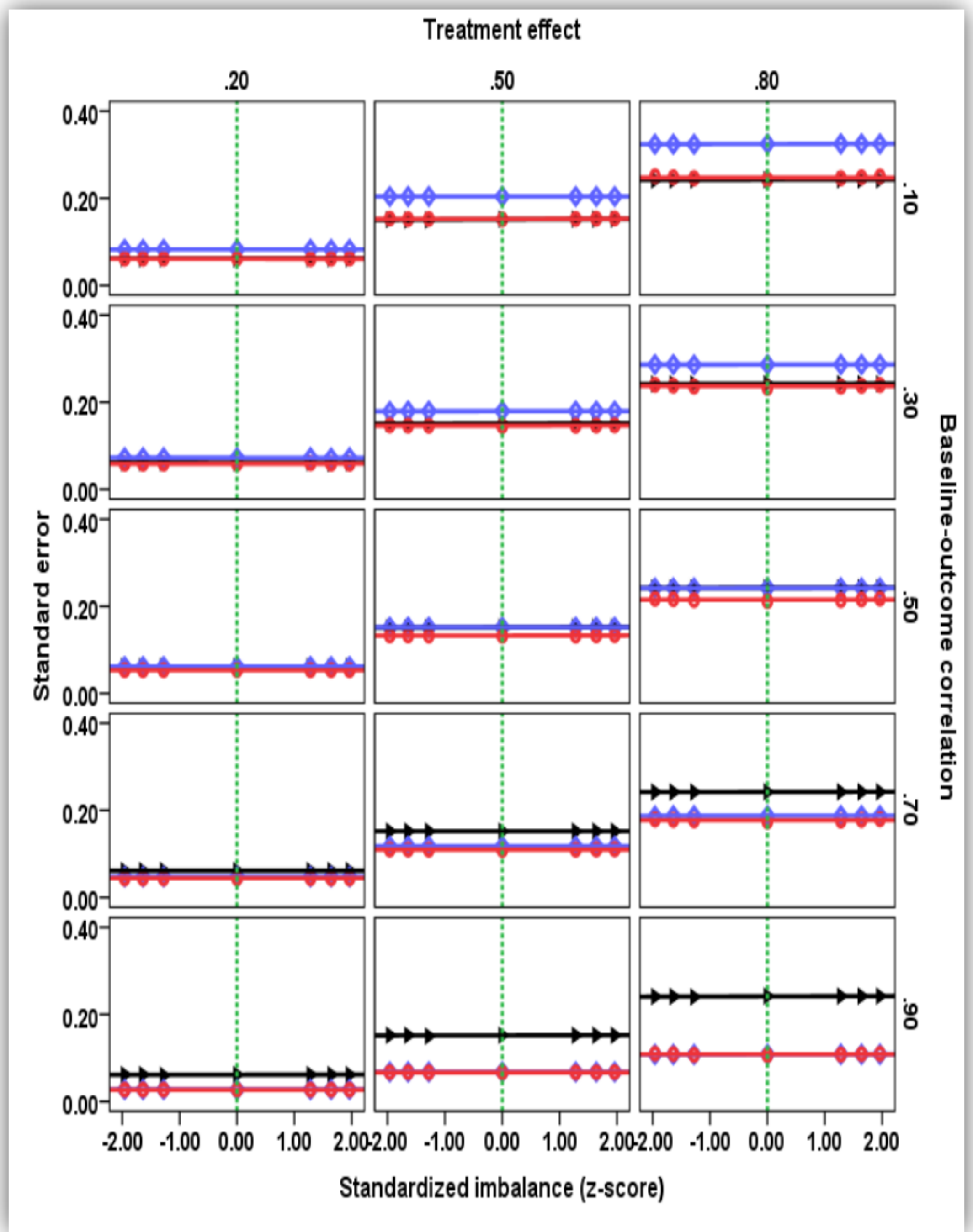


Figure 2: Directional pattern of bias of statistical methods for the analysis of RCTs at levels of baseline imbalance, baseline-outcome correlation and differing effect sizes – 90% power

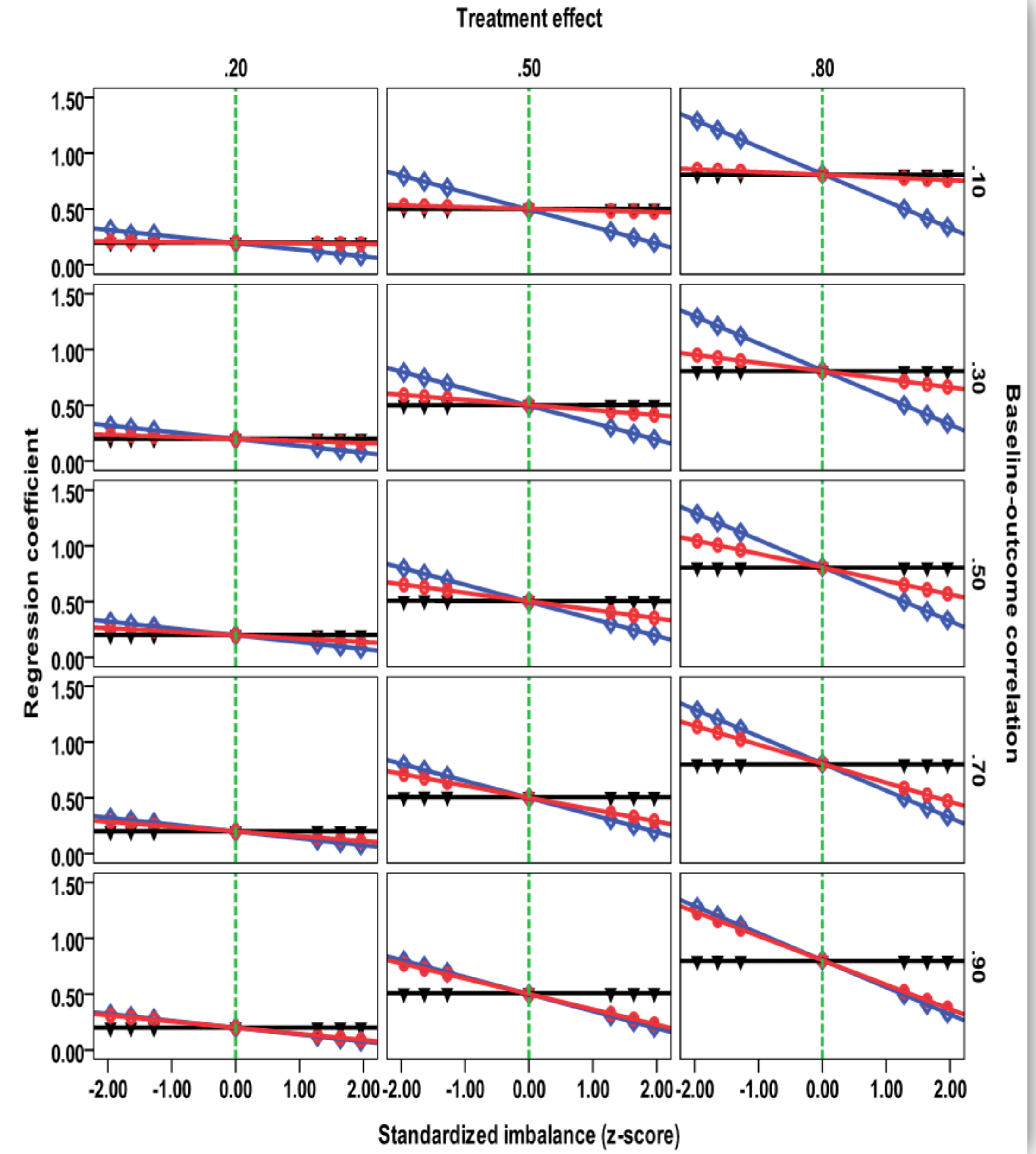


Figure 3: Ratio of the standard error (SE) of the adjusted analysis (ANCOVA) to unadjusted analysis (ANOVA) at different hypothetical trial scenarios - 90% power

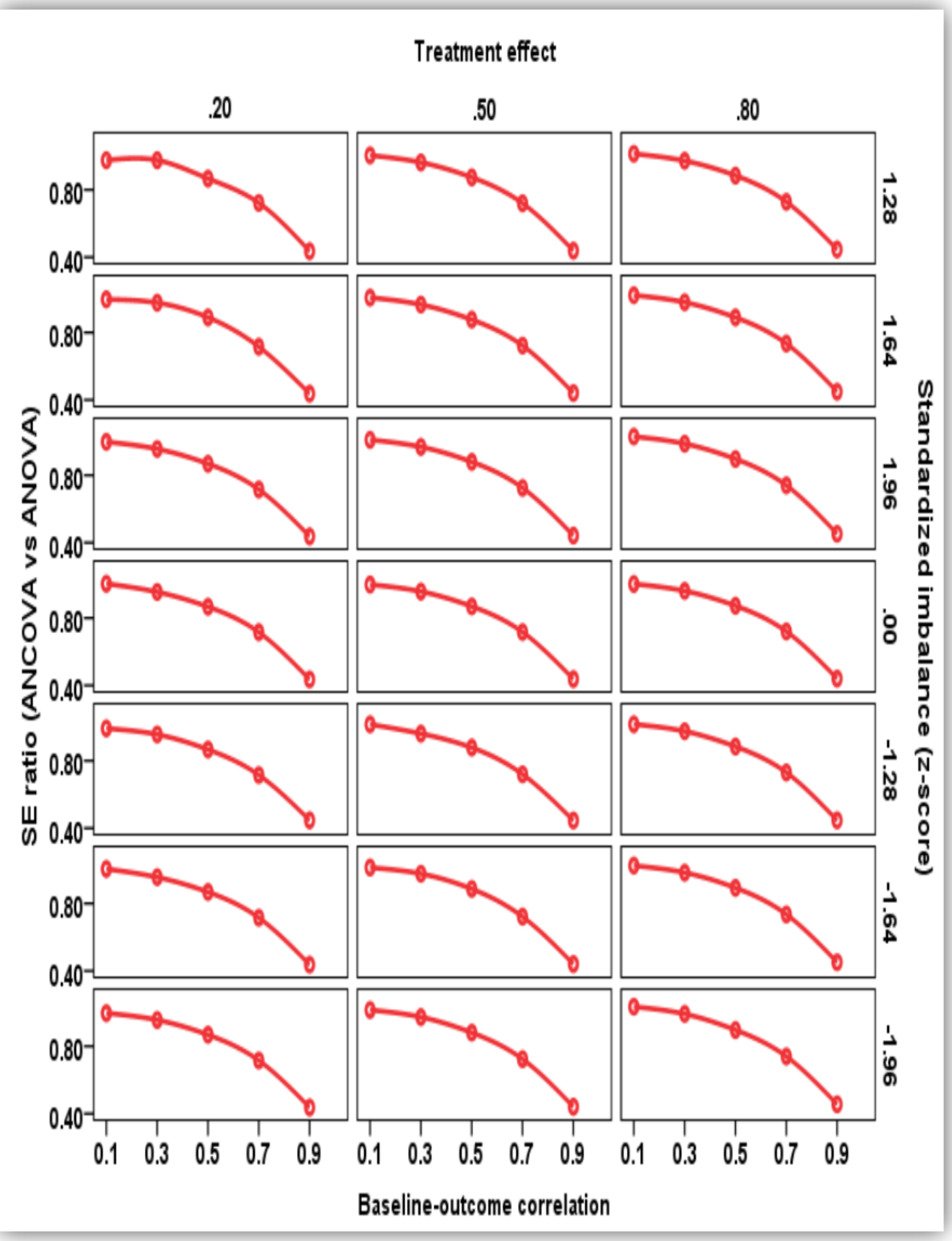


Figure 4: Ratio of the standard error (SE) of the two adjusted analysis (ANCOVA and CSA) at differing hypothetical trial scenarios – 90% power

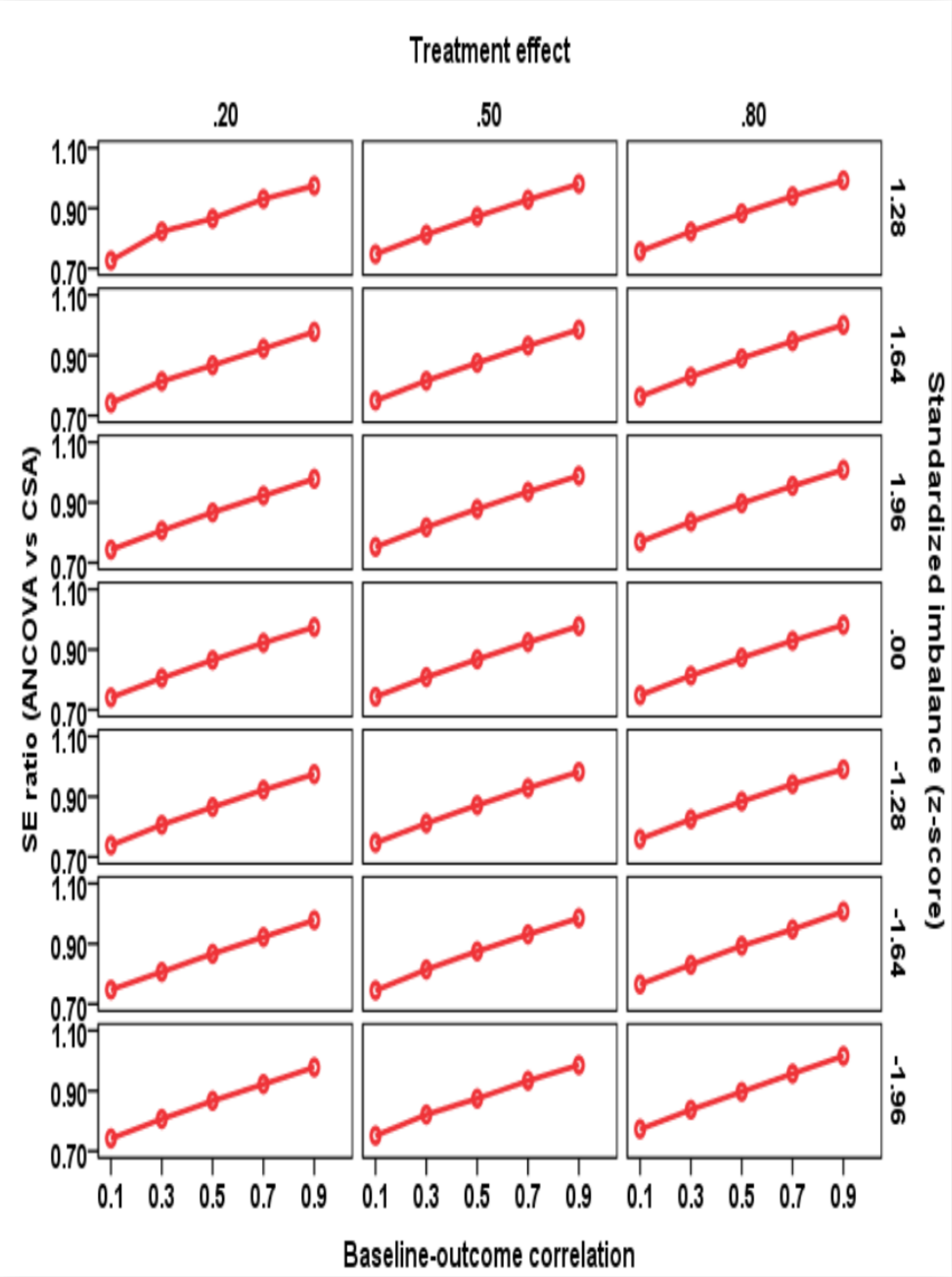


Figure 5: Ratio of the standard error (SE) of CSA and ANOVA at differing hypothetical trial scenarios – 90% power

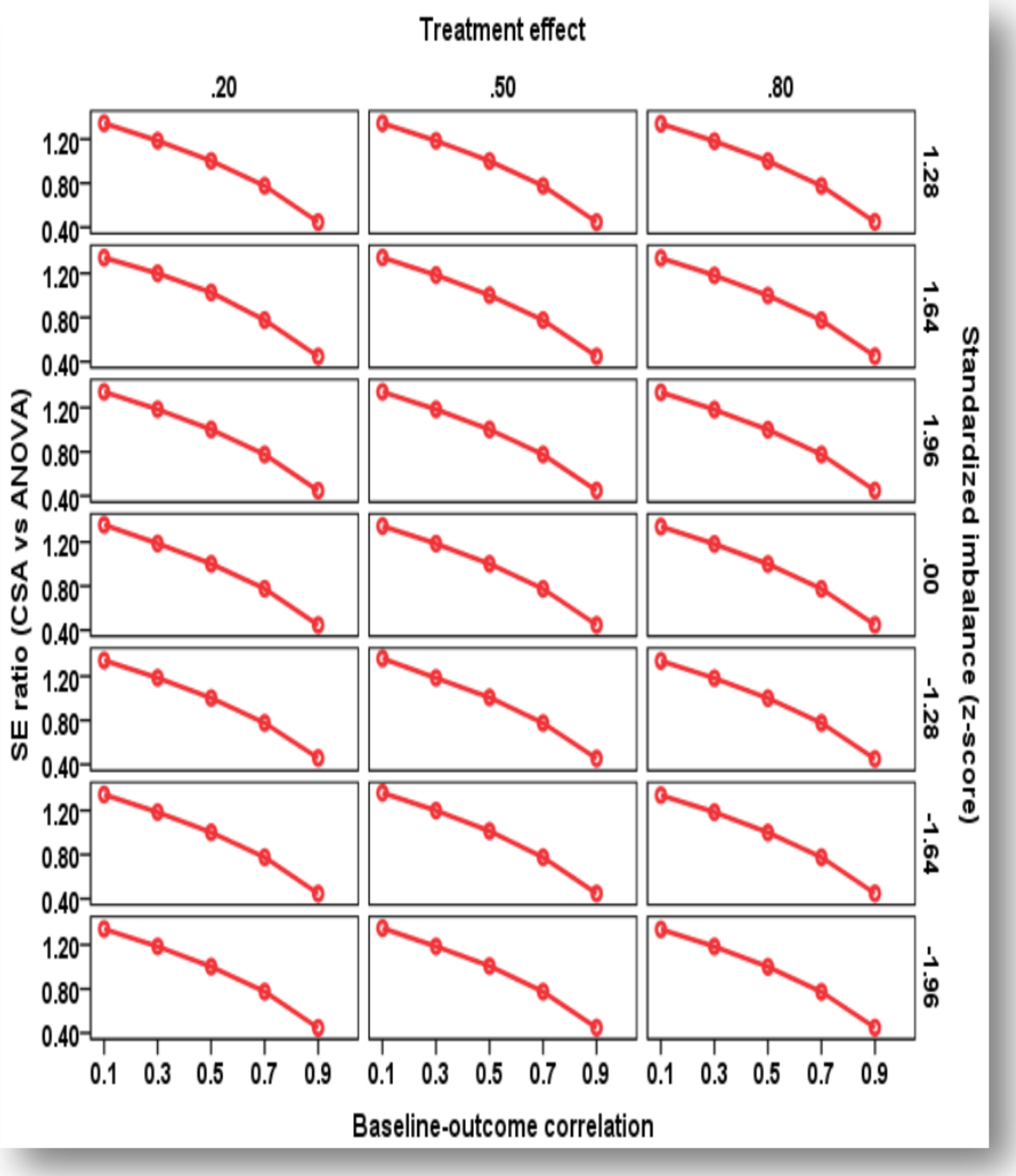


Table 1: Power (in percentage) of statistical methods at levels of baseline-imbalance (Z) in the opposite and same direction of effect and baseline-outcome correlation [Treatment effect size Y= 0.2; n= 1054]

Methods Z	**Levels of Correlation**				
ANOVA	**0.1**	**0.3**	**0.5**	**0.7**	**0.9**
-1.96	90.4	90.1	89.5	90.1	89.2
-1.64	90.4	90.1	89.5	90.1	89.2
-1.28	90.4	90.1	89.5	90.1	89.2
0	90.4	90.1	89.5	90.1	89.2
1.28	90.2	90.1	89.5	90.1	89.2
1.64	90.4	90.1	89.5	90.1	89.2
1.96	90.4	90.1	89.5	90.1	89.2
CSA					
-1.96	97.5	99.3	99.9	99.9	99.9
-1.64	91.4	98.9	99.9	99.9	99.9
-1.28	91.4	97.2	99.4	99.9	99.9
0	65.7	76.5	89.4	98.9	99.9
1.28	39.7	27.0	48.5	71.2	99.0
1.64	22.4	26.3	34.6	52.3	95.3
1.96	16.1	19.7	25.0	36.9	82.6
ANCOVA					
-1.96	93.5	98.2	99.7	99.9	99.9
-1.64	92.7	97.7	99.6	99.6	99.9
-1.28	92.7	97.0	99.2	99.9	99.9
0	90.5	92.7	96.8	99.4	99.9
1.28	72.7	85.5	85.6	90.9	99.7
1.64	86.9	83.6	81.3	84.7	98.3
1.96	86.6	80.7	74.9	75.2	92.3

Table 2: Power (in percentage) of statistical methods at levels of baseline-imbalance (Z) in the opposite and same direction of effect and baseline-outcome correlation [Treatment effect size Y= 0.5; n= 172]

Methods Z	**Levels of Correlation**				
ANOVA	**0.1**	**0.3**	**0.5**	**0.7**	**0.9**
-1.96	91.5	91.6	91.4	91.2	90.7
-1.64	91.5	91.6	91.4	91.2	90.7
-1.28	91.5	91.6	91.4	91.2	90.7
0	91.5	91.6	91.4	91.2	90.7
1.28	91.5	91.6	91.4	91.2	90.7
1.64	91.5	91.6	91.4	91.2	90.7
1.96	91.5	91.6	91.4	91.2	90.7
CSA					
-1.96	96.7	99.1	99.9	99.9	99.9
-1.64	94.3	98.1	99.7	99.9	99.9
-1.28	91.7	96.5	99.4	99.9	99.9
0	68.0	77.5	89.6	98.7	99.9
1.28	31.5	39.0	51.5	73.1	99.4
1.64	23.9	27.9	36.5	56.6	95.8
1.96	17.9	21.2	26.4	39.1	83.6
ANCOVA					
-1.96	93.7	98.2	99.9	99.9	99.9
-1.64	93.5	97.8	99.7	99.9	99.9
-1.28	93.3	97.1	99.3	99.9	99.9
0	91.8	93.3	97.0	99.5	99.9
1.28	88.9	86.2	86.4	92.1	99.9
1.64	88.2	83.2	81.5	84.8	98.5
1.96	87.5	80.7	77.0	77.4	93.5

Table 3: Power (in percentage) of statistical methods at levels of baseline-imbalance (Z) in the opposite and same direction of effect and baseline-outcome correlation [Treatment effect size Y= 0.8; n= 68]

Methods Z	Levels of Correlation				
ANOVA	**0.1**	**0.3**	**0.5**	**0.7**	**0.9**
-1.96	91.5	90.7	90.6	90.2	89.8
-1.64	91.5	90.7	90.6	90.2	89.8
-1.28	91.5	90.7	90.6	90.2	89.8
0	92.5	91.4	90.8	91.0	90.7
1.28	91.5	90.7	90.6	90.2	89.8
1.64	91.5	90.7	90.6	90.2	89.8
1.96	91.5	90.7	90.6	90.2	89.8
CSA					
-1.96	98.2	99.5	99.9	99.9	99.9
-1.64	96.3	99.0	99.8	99.9	99.9
-1.28	92.3	97.9	99.6	99.9	99.9
0	71.1	81.3	91.3	99.4	99.9
1.28	31.6	39.2	50.8	74.3	99.5
1.64	23.9	27.9	37.1	55.5	96.7
1.96	17.3	20.8	26.6	39.7	83.9
ANCOVA					
-1.96	93.7	97.6	99.6	95.1	99.9
-1.64	93.8	97.4	99.6	99.9	99.9
-1.28	93.5	97.0	99.3	99.9	99.9
0	92.7	94.3	97.7	99.6	99.9
1.28	86.7	83.3	83.5	90.8	99.8
1.64	85.2	80.0	77.6	81.8	98.5
1.96	84.1	76.4	72.2	72.2	93.3

Figure 6: Relative sample sizes for using ANOCVA instead of ANOVA at differing level of baseline-outcome correlation

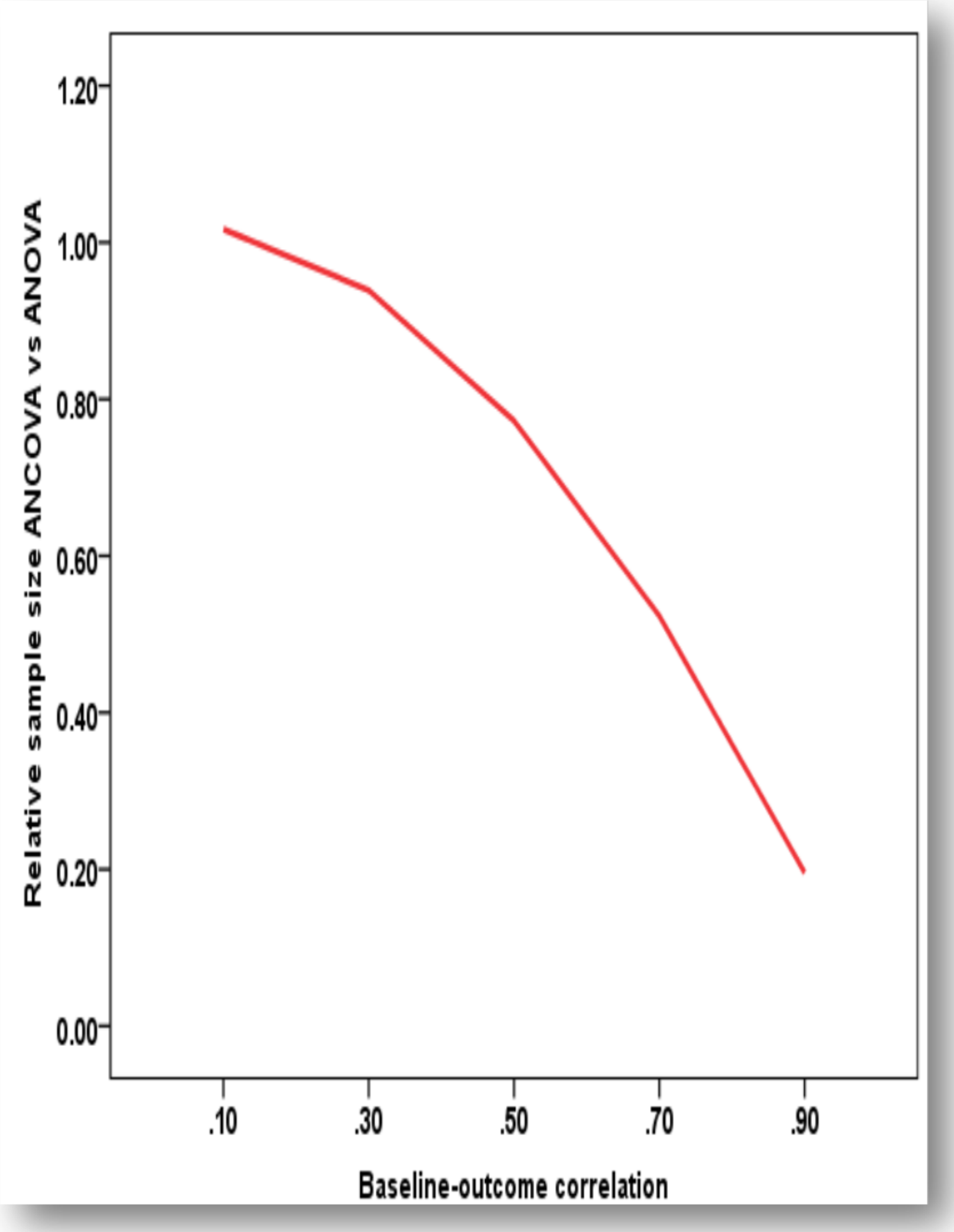


Figure 7: Relative sample sizes for using ANCOVA instead of ANOVA at differing trial scenarios

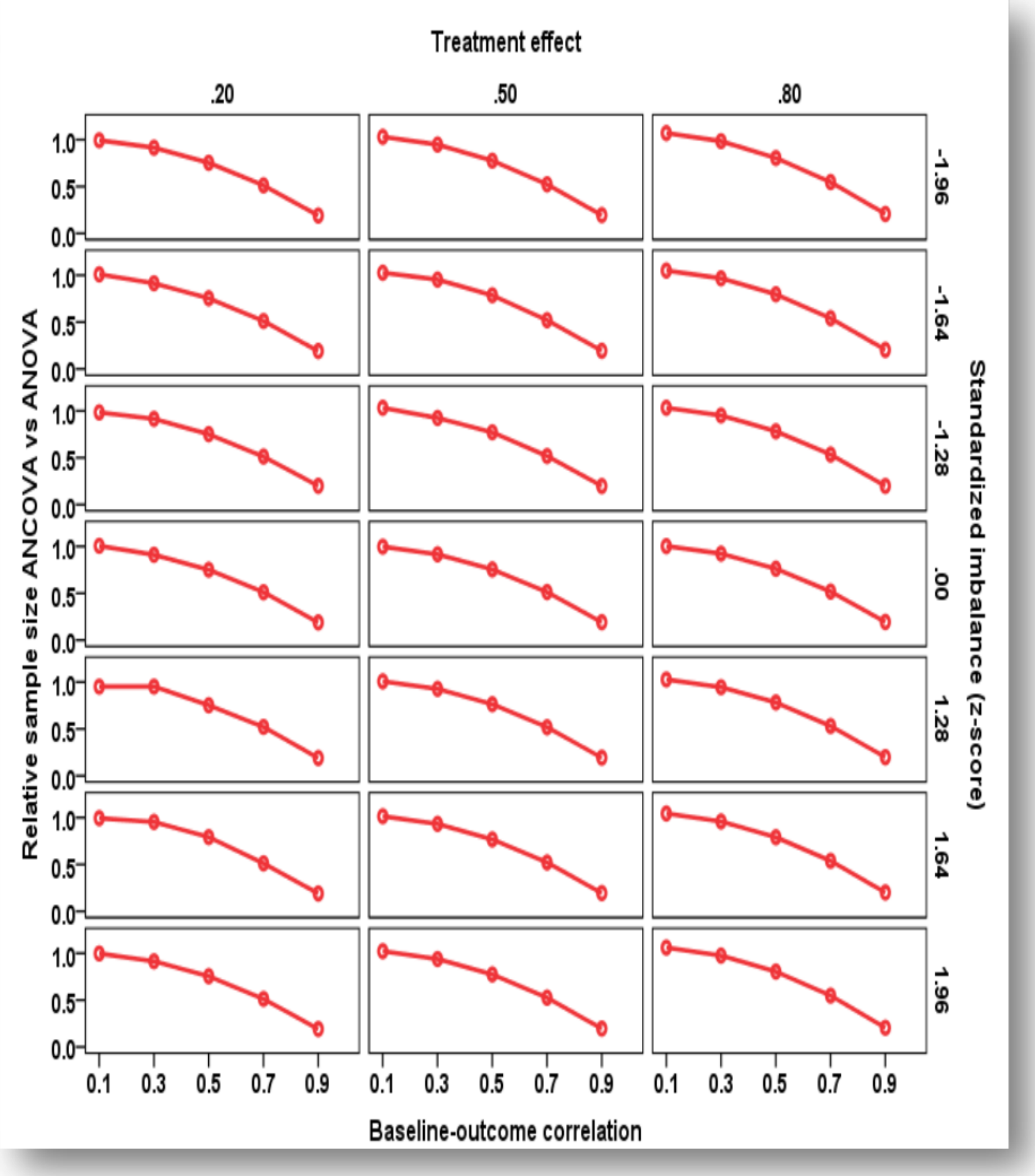


Figure 8: Relative sample sizes for using ANCOVA instead of CSA at differing level of baseline-outcome correlation

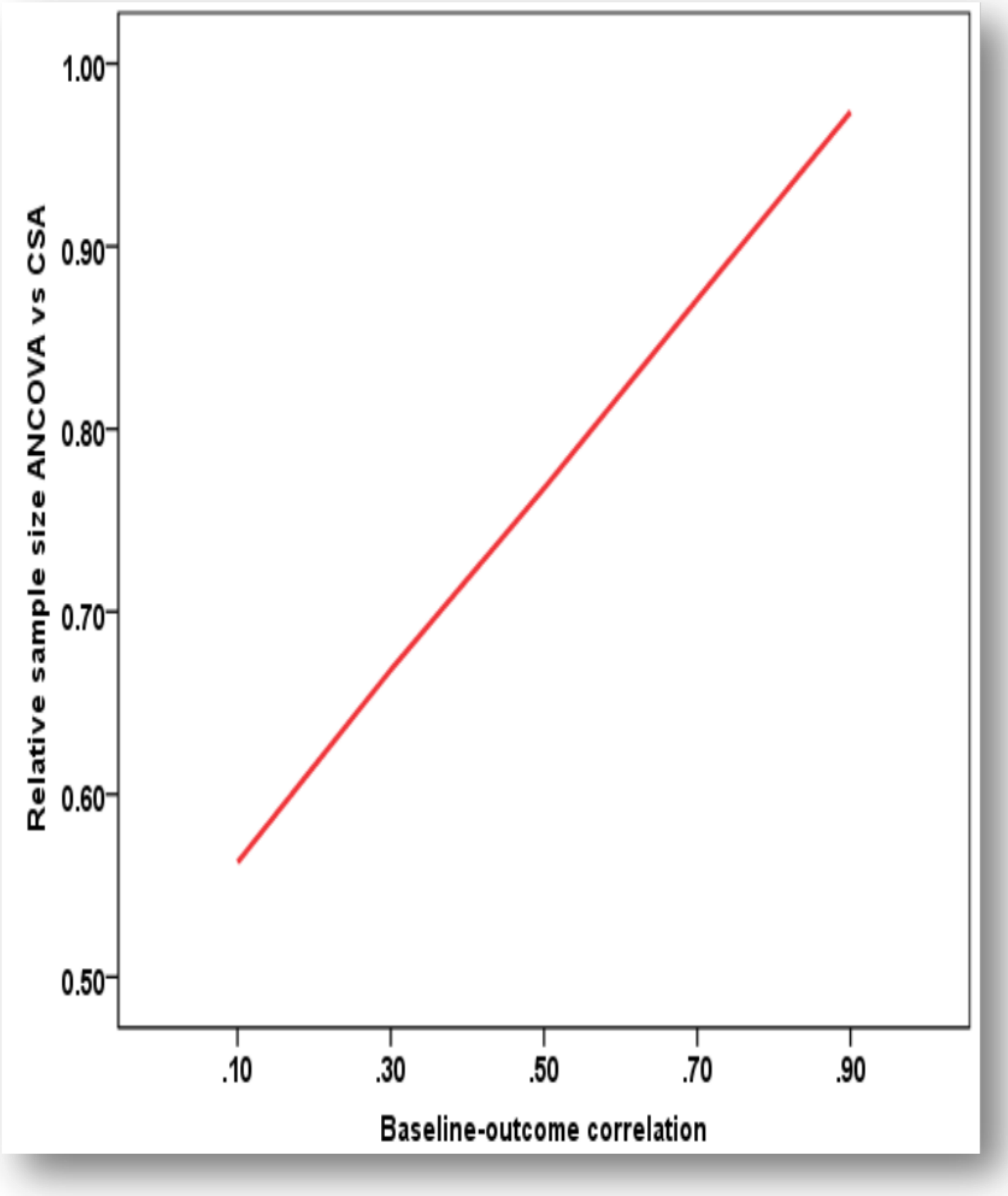


Figure 9: Relative sample sizes for using ANCOVA instead of CSA at different trial scenarios

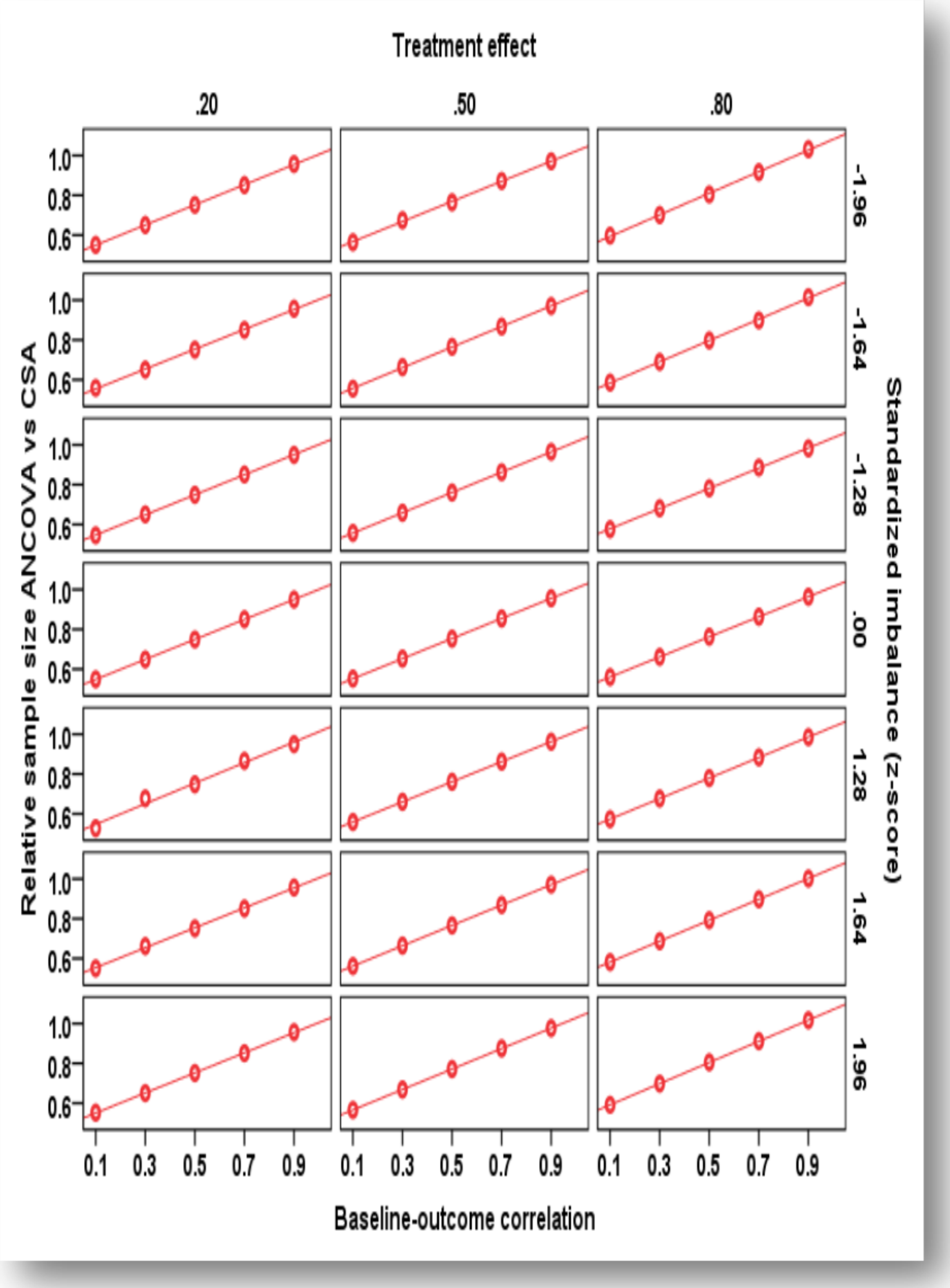


Figure 10: Relative sample sizes for using CSA instead of ANOVA at differing level of baseline-outcome correlation

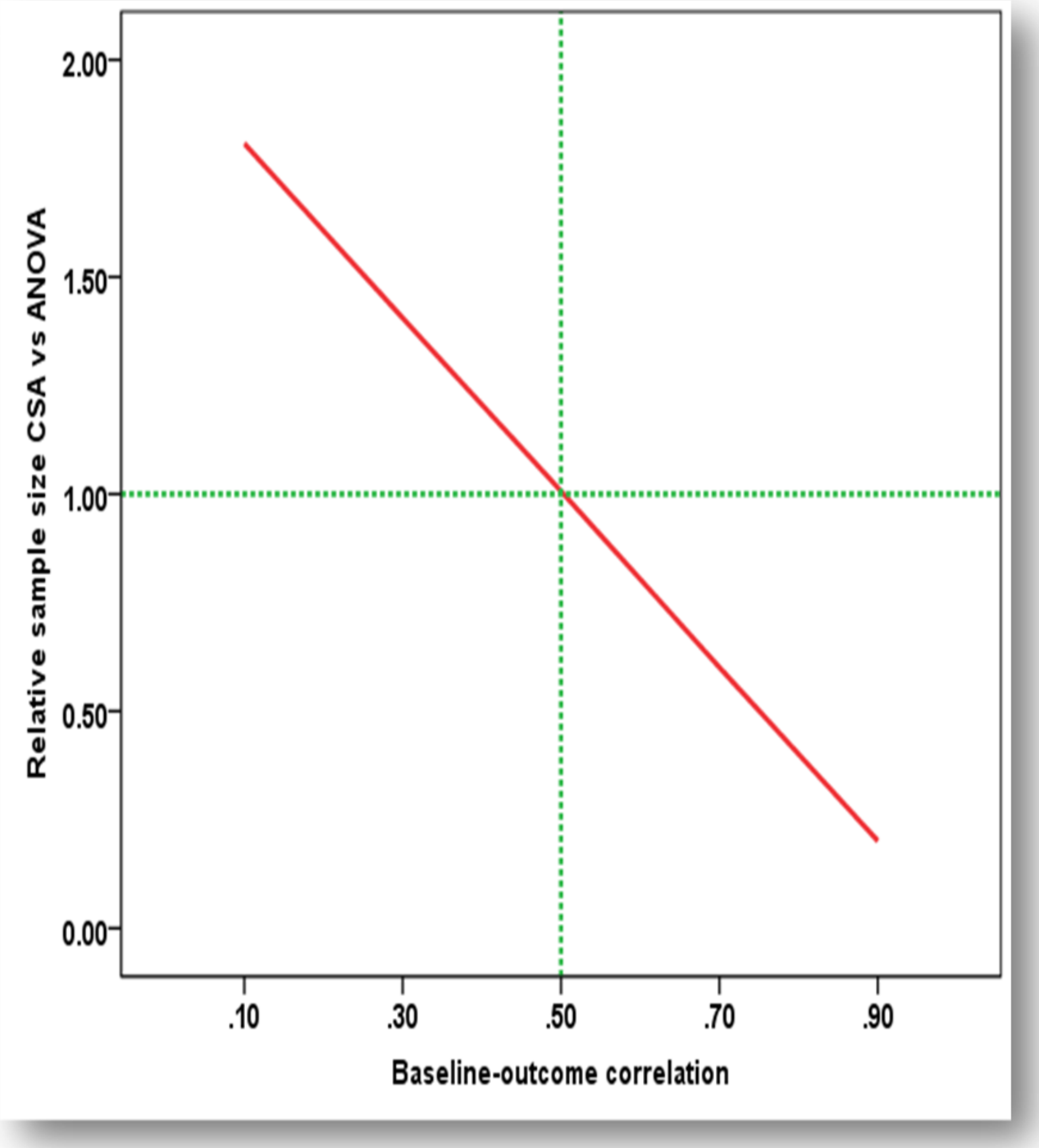


Figure 11: Relative sample sizes for using CSA instead of ANOVA at differing trial scenarios

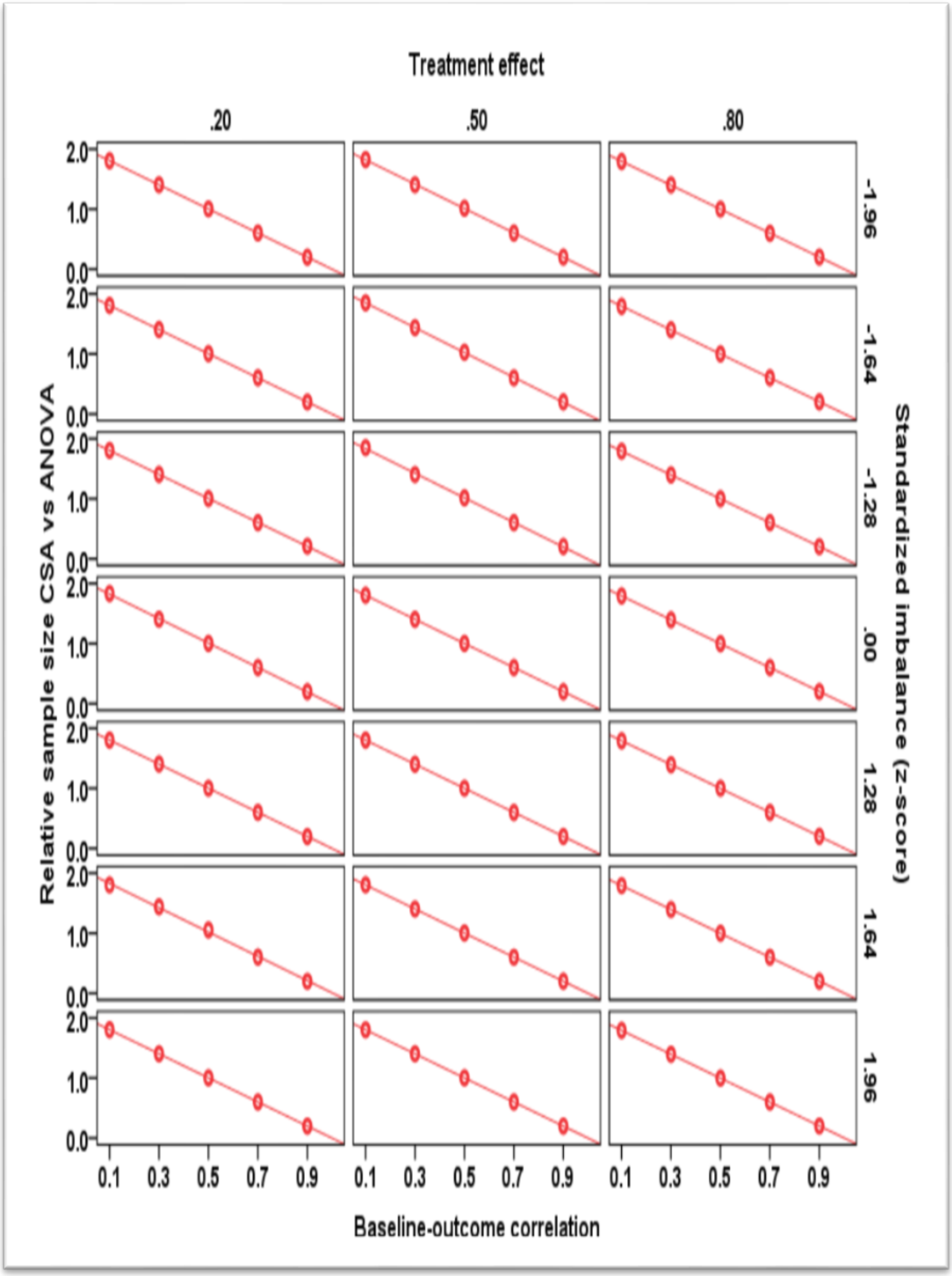
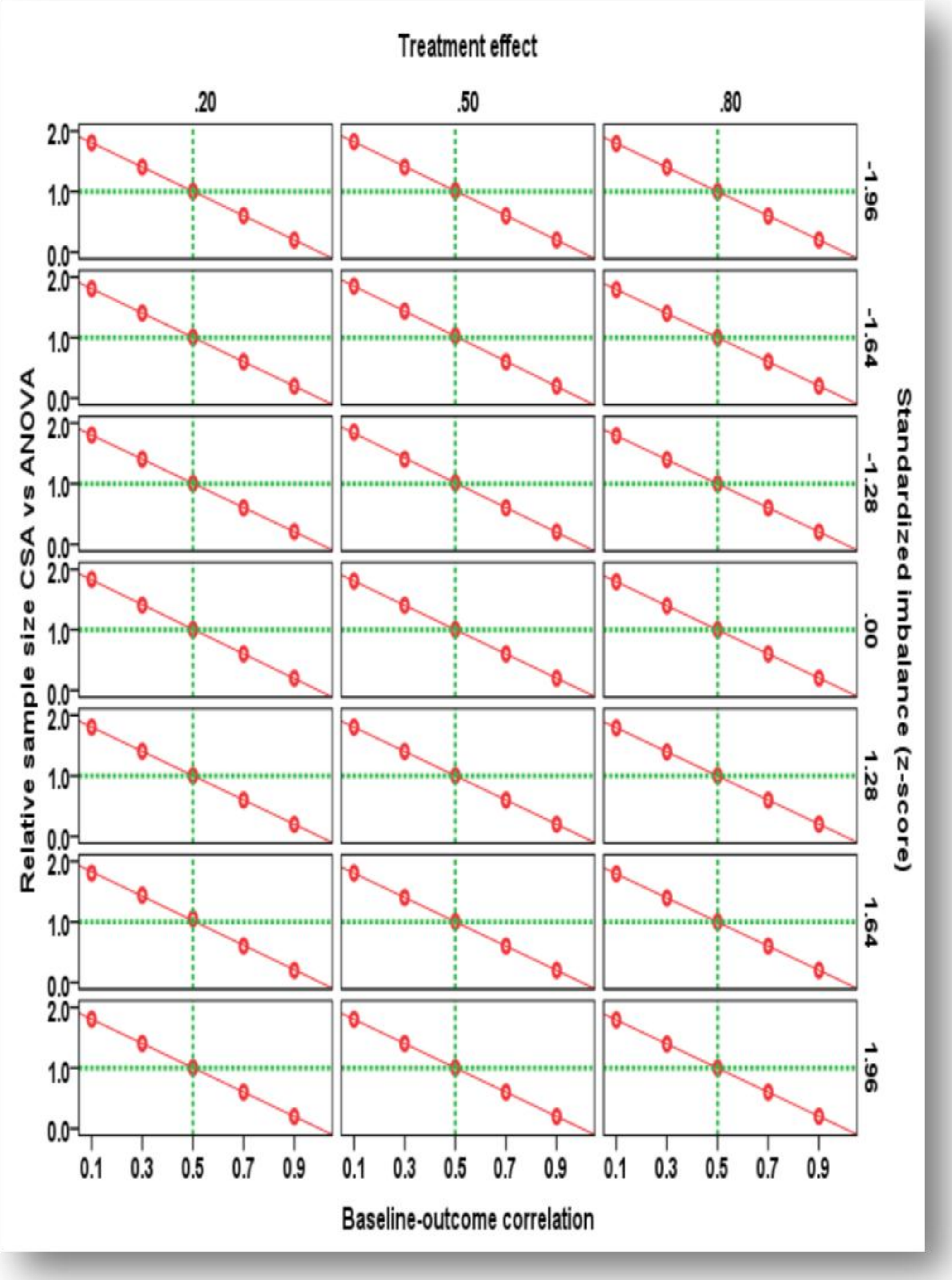


Figure 11b: Relative sample sizes for using CSA instead of ANOVA at differing trial scenarios



Appendix 3

Up-to-date thesis Plan

The whole thesis is an eight chapters (8) document with references and appendices

Chapter 1: This chapter provides background information to the study. It makes statements on existing issues and concerns around covariate imbalance in clinical trial setting. It also touches on what the study seeks to do and how it would be done.

Chapter 2: This chapter is more or less the general chapter of the study; it brings together in single chapter existing information on selected concepts on the design and analysis of clinical trials.

Chapter 3: This is the chapter on the methodology employed in the study. It gives; information on the procedure and the statistical programme that drives the simulation, information on the level of factors involved in the trial scenarios and theory of the statistical methods involved.

Chapter 4: This is the first chapter on the result of the simulation. It presents result on the relative precision and bias of the statistical methods of ANOVA, change score analysis (CSA) and ANCOVA.

Chapter 5: This is the second chapter on the results based on simulation. It presents result on the statistical power and efficiency of these methods and as in chapter 4, it uses both tables and graphs for illustration.

Chapter 6: This chapter is a systematic review on the current practices around baseline imbalance in RCT setting.

Chapter 7: This chapter seeks to inform the analysis of empirical trial datasets in the centre by the result of the simulation. It seeks information on the levels of baseline-outcome correlations that exist and what covariates to statistical adjust in trials involving musculoskeletal conditions

Chapter 8: This is the chapter on the final conclusion and recommendations of the study.

Appendix 4

Hypothetical trial scenarios at differing levels of treatment effect, levels of standardized imbalance and sample sizes at 80 and 90% nominal powers.

Outcome					Baseline covariate mean imbalance		
Effect size		Power	n (per group)	n (total)	SE	Z score	Absolute imbalance
1	0.2	0.8	394	788	0.07124705	1.281551566	0.091306768
2.	0.2	0.8	394	788	0.07124705	1.644853627	0.117190969
3.	0.2	0.8	394	788	0.07124705	1.959963985	0.139641652
4.	0.2	0.9	527	1054	0.06160411	1.281551566	0.078948844
5.	0.2	0.9	527	1054	0.06160411	1.644853627	0.101329744
6.	0.2	0.9	527	1054	0.06160411	1.959963985	0.120741838
7.	0.5	0.8	64	128	0.176776695	1.281551566	0.226548451
8.	0.5	0.8	64	128	0.176776695	1.644853627	0.290771788
9.	0.5	0.8	64	128	0.176776695	1.959963985	0.346475956
10.	0.5	0.9	86	172	0.15249857	1.281551566	0.195434782
11.	0.5	0.9	86	172	0.15249857	1.644853627	0.250837827
12.	0.5	0.9	86	172	0.15249857	1.959963985	0.298891706
13.	0.8	0.8	26	52	0.277350098	1.281551566	0.355438452
14.	0.8	0.8	26	52	0.277350098	1.644853627	0.456200315
15.	0.8	0.8	26	52	0.277350098	1.959963985	0.543596203
16.	0.8	0.9	34	68	0.242535625	1.281551566	0.31082191
17.	0.8	0.9	34	68	0.242535625	1.644853627	0.398935603
18.	0.8	0.9	34	68	0.242535625	1.959963985	0.47536109
19.	0.2	0.8	394	788	0.07124705	-1.28155157	-0.091306768
20.	0.2	0.8	394	788	0.07124705	-1.64485363	-0.117190969
21.	0.2	0.8	394	788	0.07124705	-1.95996398	-0.139641652
22.	0.2	0.9	527	1054	0.06160411	-1.28155157	-0.078948844
23.	0.2	0.9	527	1054	0.06160411	-1.64485363	-0.101329744
24.	0.2	0.9	527	1054	0.06160411	-1.95996398	-0.120741838
25.	0.5	0.8	64	128	0.176776695	-1.28155157	-0.226548451
26.	0.5	0.8	64	128	0.176776695	-1.64485363	-0.290771788
27.	0.5	0.8	64	128	0.176776695	-1.95996398	-0.346475956
28	0.5	0.9	86	172	0.15249857	-1.28155157	-0.195434782
29.	0.5	0.9	86	172	0.15249857	-1.64485363	-0.250837827
30.	0.5	0.9	86	172	0.15249857	-1.95996398	-0.298891706
31.	0.8	0.8	26	52	0.277350098	-1.28155157	-0.355438452
32.	0.8	0.8	26	52	0.277350098	-1.64485363	-0.456200315
33.	0.8	0.8	26	52	0.277350098	-1.95996398	-0.543596203
34.	0.8	0.9	34	68	0.242535625	-1.28155157	-0.31082191
35.	0.8	0.9	34	68	0.242535625	-1.64485363	-0.398935603
36.	0.8	0.9	34	68	0.242535625	-1.95996398	-0.47536109

37.	0.2	0.8	394	788	0.07124705	0.000000000	0.000000000
38.	0.5	0.8	394	128	0.07124705	0.000000000	0.000000000
39.	0.8	0.8	394	52	0.07124705	0.000000000	0.000000000
40.	0.2	0.9	527	1054	0.06160411	0.000000000	0.000000000
41.	0.5	0.9	527	172	0.06160411	0.000000000	0.000000000
42.	0.8	0.9	527	68	0.06160411	0.000000000	0.000000000

Appendix 5

List of Abbreviations

ANCOVA	Analysis of covariance
ANOVA	Analysis of variance
A &E	Advice and exercise
Back pain	Pain intensity: 0-10 numerical rating scales of least and average pain in last 2 weeks and current pain
CARA	Covariate-adjusted response adaptive
CONSORT	Consolidated standard of reporting trials
CPMP	Committee for proprietary medicinal products
CSA	Change score analysis
CS-CAT	Catastrophising
CS-CSS	Coping self-statement
CS-IBA	Increasing activity level
CS-PH	Praying or Hoping
EQ-5D	Preference – based health utility
HADS_ANX	Anxiety subscale score
HADS_DEP	Depression subscale score
IPQR	Illness perception questionnaire revised
ITT	Intent-to-treat analysis
MT	Manual therapy
MCS	Mental component Score
MSE	Mean square error
NPQ	Northwick park total score

PCS	Physical component score
PSWD	Pulsed short wave diathermy
RCTs	Randomised controlled/Clinical trials
RMDQ	Roland and Morris Disability Questionnaire (Back pain Disability)
SE	Standard error
SF12	Short Form 12 (Health related quality of life)
SF-Mc Gill VAS	Average pain in last week
STROBE	Strengthening the reporting of observational studies in Epidemiology
TSK	Tampa Scale Kinesiophobia