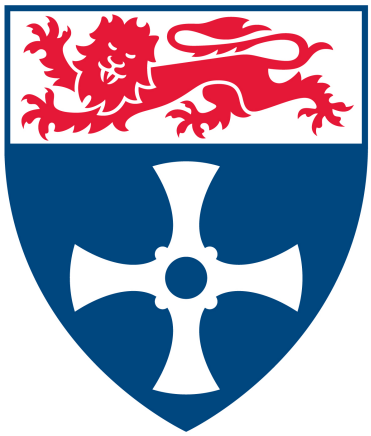


Newcastle University

MAS8391: MMathStat Project

School of Mathematics & Statistics

Sample Size Re-estimation in Clinical Trials

Author:
Isabella Hatfield

Supervisor:
Professor. J.N.S Matthews

April 30, 2015

ISABELLA HATFIELD

Abstract. In Randomised Clinical Trials, the initial variance estimate required to calculate a sample size is often misspecified using fixed sample size methods. Sample size re-estimation trials make use of emerging information to modify the sample size of a trial, with an aim to save the power of the trial compared to conventional fixed-sample methods. The aim of the present dissertation was to simulate and analyse sample size re-estimation trials, where the initial variance was smaller than the true variance, comparing the results and conclusions with those of the original fixed sample trials. It was found that sample size re-estimation is an effective method of achieving an appropriate sample size. However, this method is not without limitations and these should be weighed up against the possible benefits when designing a clinical trial.

Contents

1. Introduction

1.1. **Randomised Clinical Trials.**

Clinical trials have become a central component in the assessment of new therapies. Hundreds of new treatments are approved every year for development and manufacturing around the world, to treat a variety of different diseases. Typically in Phase 2 and Phase 3 studies, Randomised Clinical Trials (RCTs) are used to reduce bias and ensure the work is robust and of maximum benefit to the patient.

1.2. **History.**

An RCT is an unbiased and scientific way of testing treatments, in which randomization is used to allocate treatments to patients. One of the first published RCTs was in 1948 by Bradford Hill, (Bradford Hill, 1948). It involved the use of streptomycin as a treatment for pulmonary tuberculosis. RCTs became standard procedure in medical experiments using human subjects, and strict regulations were created, such as the Declaration of Helsinki (1964) (Association, 2013), to make them an ethical method for analysing prospective treatments. Many rules are now in place regarding RCT design to ensure effective conduct, including rules for derivation of sample size.

1.3. **Sample Size.**

To put the aim of this report into context it is important to understand the function of sample size, with particular reference to its interdependence with other factors, in determining the value of RCTs.

1.3.1. *Definition.*

Sample size means the number of participants we should intend to participate in an RCT. Determining sample size is one of the first and most important steps in designing a successful study. The ICH E9 (1998) (Phillips and Haudiquet, 2003) guideline states that: *'The number of subjects in a clinical trial should always be large enough to provide a reliable answer to the questions addressed. This number is usually determined by the primary objective of the trial.'* For the purpose of this dissertation, sample size, n, is defined as the number of patients in each group, where the groups are equally sized. It is important to note that if the intended sample size is incorrect it can have serious consequences, ultimately affecting the 'power' of the trial.

The power of a trial means its ability to reveal a difference between treatments when such a difference exists. Power is a function of the difference in treatment means and it should be large enough when

the true treatment difference is important. Without sufficient power, it is possible that the treatment difference sought will not be identified in the trial. Power is determined by study design, of which sample size is a critical part. In fact, the power of a trial and sample size used within it are interdependent; we choose sample size to determine how powerful we want a trial to be, and for a trial to have a reasonable chance of answering the research question it addresses, the sample size must be large enough.So the power of a trial is maintained by choosing an appropriate sample size.

The determination of sample size is often referred to as 'power analysis' (Shein-Chung Chow, 2008). If the number of patients recruited to the trial is too small, this leads to the trial being underpowered, which means that it may be impossible to detect important differences between the treatments. Conversely, if too many patients are recruited, the trial is overpowered, which may have ethical implications if the study's outcomes are met before the end of the trial and some participants must still complete treatment for what might have been revealed as an inferior treatment.

1.3.2. *Power Analysis.*

Power, $1-\beta$, is a function of the Type II error, β. A Type II error occurs when, for example, a treatment is claimed to be ineffective when it is, in fact effective. Conversely, a Type I error, α, occurs when the treatment is claimed to be effective when it is not. One way to derive the power of a trial is by the testing of hypothesis of new treatment difference. Where the null hypothesis is accepted when no important difference between the means is identified and the alternative hypothesis is accepted when an important difference between the means is identified. So the Type II error occurs under the alternative hypothesis and the Type I error occurs under the null hypothesis.

Type II errors are a function of sample size and, consequently, we set sample size in RCT design based on what we consider to be acceptable levels of such errors. It is important to set a sample size which specifies the Type II error when the true difference is the minimal clinical difference.

Figure 1 illustrates this for a Normal distribution with variance $\sigma^2 = 1$ and with $\triangle$ =the minimal clinical difference = 0.1, it shows the effect different sample sizes will have on the power of a trial. So a sample size of 1,600 patients was required to obtain 80% power. You can see that power decreases with a decrease in sample size however, dropping to just 20% when the sample size is 250. This demonstrates the importance of calculating an appropriate sample size. In common practice, the choice of power is either 80% or 90%.

Normal Outcomes.

So far the methods explained are general and can be applied to all types of outcome- normal, binary,

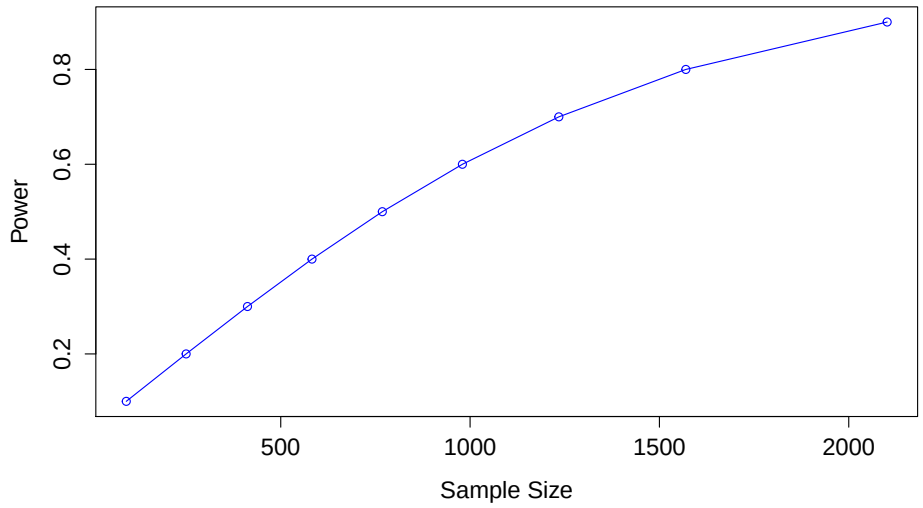


FIGURE 1. The effect of sample size on power, where $\triangle = 0.1$ and $\sigma^2 = 1$

continuous, etc. From this point forwards, the dissertation will only focus on normally distributed outcomes.

1.3.3. *Sample Size Equation.*

In order to determine an adequate sample size, the following values must be specified:

- a two sided Type I error, α, that the investigator can tolerate. Generally this is set at 5%.
- a clinically important treatment difference, $\triangle$. This is the difference between treatment arms that the trial seeks to identify in order to conclude that it is effective. It is seen as acceptable to miss any difference which is less than $\triangle$.The larger the treatment difference sought, the smaller the sample size required to identify it, and vice versa.
- a desired power, $1 - \beta$. This reflects the chance of correctly detecting a difference, when a difference of $\triangle$ does exist.

A sample size equation is used to provide an estimate for the required sample of a trial. Different equations are used in different circumstances. This report will investigate the simplest case for normal outcomes, using the assumption that the groups are equally sized. For these circumstances, also requied is:

- an estimate of the standard deviation of the treatment difference, σ^2. When this is small, detection of a treatment difference will require a smaller sample size.

This sample size equation is:

$$n = \frac{2\sigma^2(z_{\frac{\alpha}{2}} + z_\beta)^2}{\triangle^2} \tag{1}$$

where n is the sample size for each group $\triangle$ is treatment difference, $z_{\alpha/2}$ is such that $Pr(Z > z_{\alpha/2}) = \alpha/2$, (1.96 for a two-sided α of 5%), and $Pr(Z > z_\beta) = \beta$ (0.84 for 80% power) where $Z \sim N(0, 1)$ and σ^2 the common variance in both treatment groups.

1.3.4. *Estimating* σ^2.

To use the sample size equation, information on some of the parameters of the trial is required. We specify $\triangle$ when discussing the principal aim of the trial. However, we also need a value for an estimate of the variance of the outcome, σ^2. This can be difficult to determine, especially when there is a lack of information regarding the outcome variable. Which is, in turn, problematic as misspecification of variance can have a sizeable impact on the power of the trial.

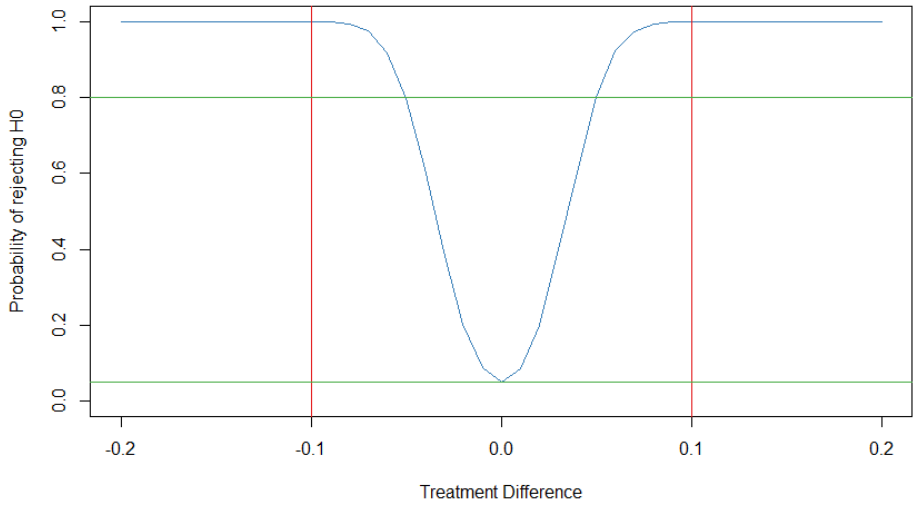


FIGURE 2. Effect of overestimation of Variance, with $\triangle = 0.1$, $\sigma^2 = 1$ and initial estimation of $\sigma^2 = 2$. At 80% power; less than the minimum clincal difference between treatment means is identified

If the variance is over estimated, the trial is over powered. In Figure 2, the target power of 0.8 is reached prematurely, suggesting that the number of patients recruited would have been unnecessarily large, which has ethical implications.

Conversely, in Figure 3, the target power is not reached because of the underestimation of variance, so not enough patients would be recruited to determine the treatment effect with adequate variance.

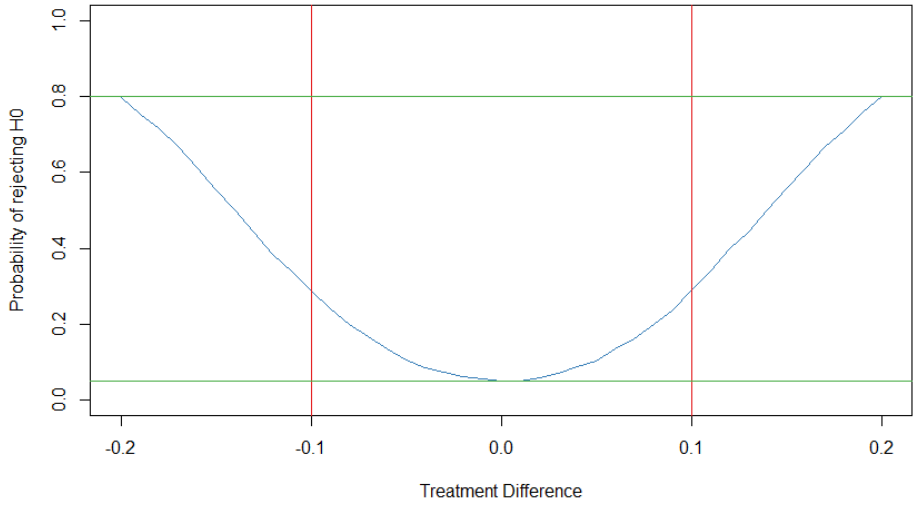


FIGURE 3. Effect of underestimation of Variance, , when $\triangle = 0.1$, $\sigma^2 = 1$ and initial estimation of $\sigma^2 = 0.5$. At 80% power; the minimum clinical difference between treatment means is not identified

These are hurdles that most investigators face when planning trials. Variance specification has a sizeable influence on trial integrity and it is vital that values are as accurate as possible

1.3.5. *Example.*

There have been a number of instances of trials failing due to variance estimates being too low. When this happens, trials fail to identify a meaningful difference between treatments. A trial run by Julious (2004) found a misspecification of variance led to a potentially significant treatment being deemed ineffective. One of the key problems identified was the unexpectedly high variance of 47% observed, which hugely exceeded the estimated 30%. If the variance had been estimated more accurately a larger sample size would have been specified, increasing the power of the study and the likelihood of an important trial result being identified.

1.4. **Aim of the Report.**

Misestimation of variance is one of the major pitfalls of fixed sample size design. So, in a fixed sample size design, an estimate of sample size is derived using the sample size equation and data is collected until this is achieved. The sample is labeled 'fixed' because once the initial estimate is made, this does not change. Alternative methods, including sample size re-estimation, have been created as a solution to these pitfalls.

This paper aims to compare some of the relative merits and limitations of fixed sample size and sample size re-estimation methods.

It will achieve this by:

- Performing an audit of fixed sample size trials
- Exploring the statistical properties of the fixed sample size design
- Conducting a literature review, to provide an overview of the various methods of sample size re-estimation
- Conducting simulations of both sample size re-estimation and fixed sample designs, to provide a direct comparison.

2. Investigation into estimation of Variance: An Audit

2.1. **Aims of Investigation.**

This Audit set out to examine the characteristics of sample size calculation of completed and published trials which used fixed, sample size design.

More specifically, the aims of the audit were as follows:

- To identify the methods used in such trials for producing an initial estimate of variance for the sample size calculation
- To determine how successfully these methods worked
- To examine characteristics of studies which used different ways to obtain the initial variances
- To discuss the transparency of the different methods

2.2. **Methods.**

2.2.1. *Data Collection.*

The archives of four separate journals were used to identify trials that had been completed and published, and which had normally distributed outcomes. The journals used were:

(1) Heart *http://heart.bmj.com/*

(2) Thorax, *http://thorax.bmj.com/*

(3) Emergency Medicine, *http://emj.bmj.com/*

(4) Journal of emergency medicine, *ttp://www.journals.elsevier.com/the-journal-of-emergency-medicine/*

The data was last accessed, 23 November, 2014. These journals were chosen because they were likely to contain a large number of trials which had normally distributed outcomes. Journals focusing on cancer trials were omitted, because their focus on reporting 'time to' survival outcomes was considered less helpful to the present study. The search was conducted on 17th November 2014, using key words 'CLINICAL TRIALS' in the title or research summary.

The search results were then sorted by their eligibility for further analysis, according to the criteria set out below:

- They were randomised controlled trials

- They were completed and published
- They had normally distributed primary outcomes
- They were of parallel design; as other designs used different sample size equations
- They were classified as interventional; no bioequivalence trials
- The participants were patients not healthy volunteers
- The trial was of classic design; no adaptive designs

Only trials that had been completed and published were included in the analysis, in order to get the most valid impression of the sample sizes being used in informative trials, presented for scientific readership. Trials conducted on healthy volunteers were not included, as these are not usually efficacy studies.

Adaptive designs were not included, as these might include sample size re-estimation, which was not the focus of the audit.

Data from trials in the preliminary literature search were exported to Excel, where relevant data was captured. After the trials had been assessed against the inclusion criteria, the eligible examples were imported to RStudio for analysis.

2.2.2. *Data Extraction.*

Data on the target sample size and any trial components that might influence it - such as funding body, number of treatment arms and disease area - were collected. Information was extracted from research articles on the journals' databases, when this was available. To complement the published articles search, an Internet search was also undertaken to locate any wider reporting of the trials and to seek information which appeared to be missing from the published articles.

2.3. **Results.**

Applying the search term 'CLINICAL TRIALS' to the archives of the four selected journals yielded 4,579 studies over the time since the archives were created. The feasibility of filtering so many trials was unrealistic, so it was decided that a subset of 40 trials from each archive would be collected. These subsets were identified by sorting the trials according to citation -high to low - and the first 40 trials from each re-sorted archive were selected. After eliminating duplicates, removing studies that did not meet the inclusion criteria and removing studies with no available data, a total of 47

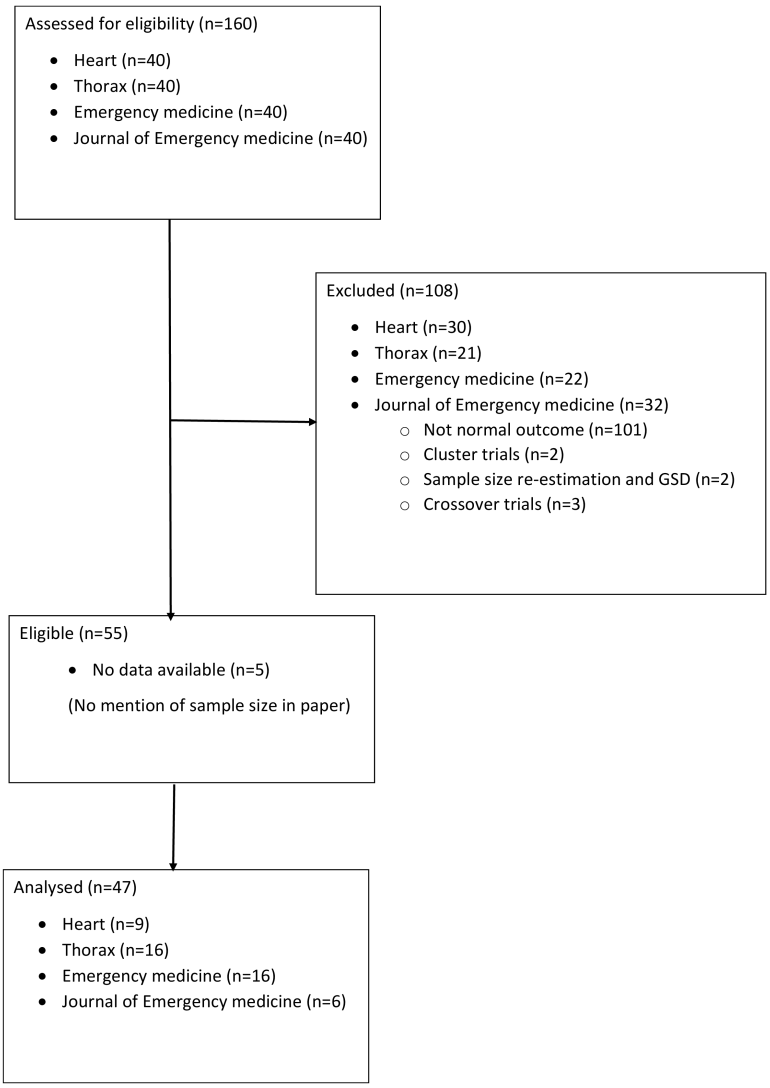


FIGURE 4. Flow diagram; the extraction and eligibility of data from each source

trials progressed to be analysed. 'No available data' refers to trials published with no information regarding trial sample size, and neither was this available from other sources. Figure 4 illustrates the flow of trials through the subset review.

2.3.1. *Trial Characteristics.*

Table 1 summarises the characteristics of the trials that met the inclusion criteria. The majority of trials (n=41, 87.2%) consisted of two arms, one control treatment and one experimental treatment.

Most of the trials (n =38 , 80.9%) were publicly funded, with the remaining trials being funded by industry (n = 9, 19.1%).

		Journal									
		Heart		Thorax		E M		J E M		All	
		n	%	n	%	n	%	n	%	n	%
Standard deviation, σ^2	Provided	8	88.9	10	62.5	12	75.0	4	66.7	34	72.3
	Not provided	1	11.1	6	37.5	4	25.0	2	33.3	13	27.7
Treatment difference, δ	Provided	9	100.0	15	93.8	16	100.0	6	100.0	46	97.9
	Not Provided	0	0.0	1	6.7	0	0.0	0	0.0	1	2.1
How estimate was made	Provided	7	77.8	10	62.5	12	75.0	3	50.0	32	68.1
	Not Provided	2	22.2	6	37.5	4	25.0	3	50.0	15	31.9
Funder	Public	9	100.0	10	62.5	14	87.5	5	83.3	38	80.9
	Private	0	0.0	6	37.5	2	12.5	1	16.7	9	19.1
Arms	2	9	100.0	15	93.8	12	75.0	5	83.3	41	87.2
	3	0	0.0	0	0.0	4	25.0	1	16.7	5	10.6
	4	0	0.0	1	6.3	0	0.0	0	0.0	1	2.1

TABLE 1. Trial characteristics of the 47 studies included in the final analysis

The majority of trials (n=34, 72.3%) stated the initial standard deviation and nearly all (n=46, 97.9%) stated the important treatment difference, $\triangle$. How the sample size was calulated was stated less frequently (n=32, 68.1%).

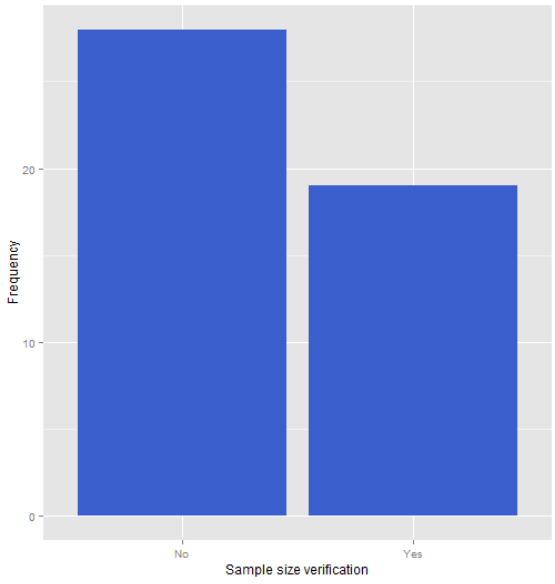


FIGURE 5. The number of trials that produced the same sample size as the equation used in this dissertation

2.3.2. *Assessing the Sample Size Equation.*

In the case of trials stating the necessary details, a standard sample size equation was calculated, using a function in R, to see if this verified the actual sample size used in each case. The majority of actual sample sizes (n=29, 61.7%) were not verified by the sample size equation, as shown in Figure 5.

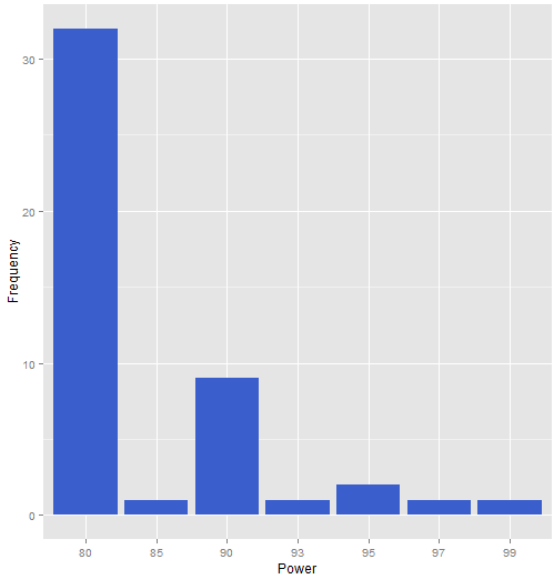


FIGURE 6. Designed power of the 47 trials included in the analysis

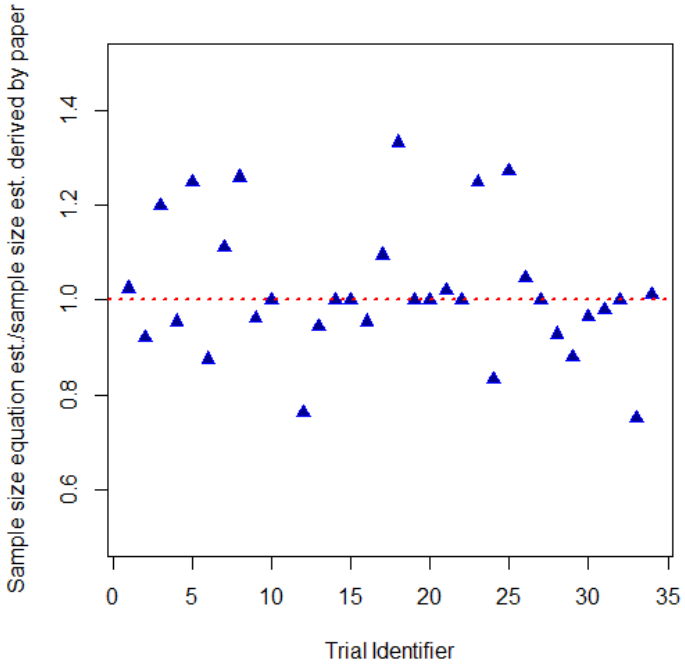


FIGURE 7. Comparison of sample size equation to derived sample size of the 47 trial in the final analysis

Sample sizes calculated specifically using Sample Size Equation 1 (see Section 1.3.3) were then compared to the sample sizes derived by the trials. We can see from data presented in Figure 7, that results were similar in the majority of trials. The sample size equation derived a, roughly, 75% smaller sample size than that estimated in two of the papers and a 30% bigger sample size in another. Although not extreme, these discrepencies could have an important effect on the power.

2.3.3. *Power used in Studies.*

The power of the equation was stated in every trial analysed. Most trials (n=32, 68.1%) used 80% power. However, a range from 80% to 99% was observed all the trials. Figure 6 shows the power used in the trials.

2.3.4. *Use of Sample size.*

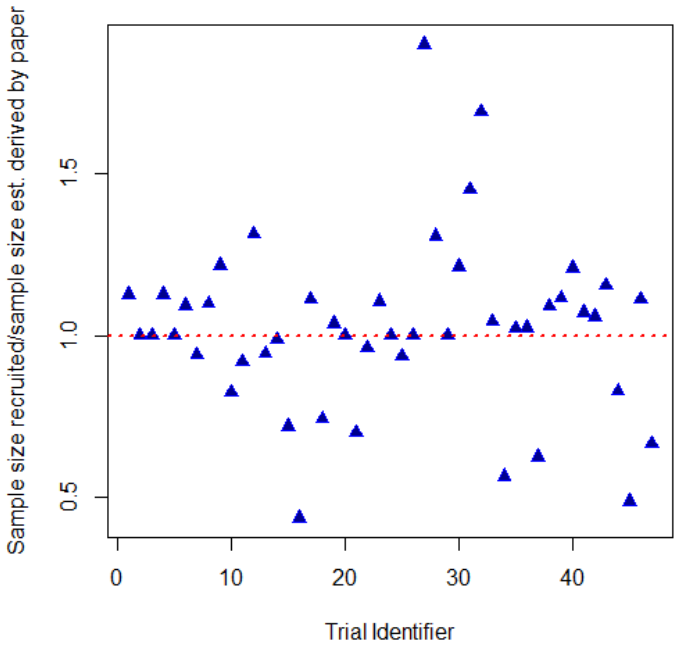


FIGURE 8. Ratio of sample size estimated over used sample size for the 50 trials used in the final analysis

Frequently, trials were seen to estimate a sample size, but not go on to recruit the sample size stated. Conversely, some studies over recruit to compensate for dropout. It is common practice in RCTs to recruit up to 10% more subjects than required (Watson and Torgerson, 2006), because of the probability that not all subjects will follow the trial through to analysis.

Other studies were found to have a smaller sample size than estimated. On further inspection, the principal reason for this was problems with recruitment. In studies where there is a small population to select from, such as rare diseases or when consent is difficult to obtain, it can often be difficult to recruit the full sample size required, leading to loss of power in the trial. In Figure 8 we can see that more studies were overpowered than underpowered in the audit, however, with one study using nearly twice the estimated sample size. Some of the underpowered trials had a very small sample

size, in comparison to the estimated size, causing these studies to lose validity. It was discovered that the study with the least power in Figure 8 terminated early, due to poor recruitment.

2.3.5. *Methods of Variance estimation.*

Table 2 displays the source of information for initial variance of the trials observed in the audit. Out of the 32 trials that gave information on how their initial values were obtained, the majority (n=20, 62.5%) cited previous trials, with the remaining estimates coming from pilot studies (n=5, 15.6%), literature (n=5, 15.6%) and audits (n=2, 6.3%). To see how effective these methods were, the final variance of each trial was compared to the estimated variance.

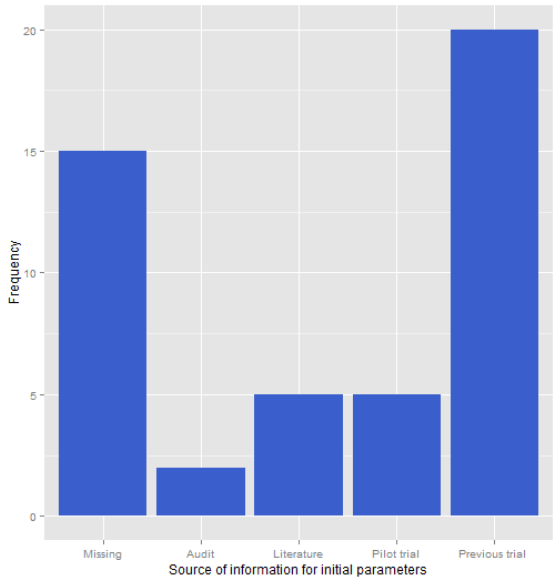


FIGURE 9. The source of information of initial variance

	Journal				
	Heart	Thorax	EM	JEM	All
Missing	2	6	4	3	15
Audit	0	0	2	0	2
Previous trial	5	7	6	2	20
Literature	1	1	2	1	5
Pilot trial	1	2	2	0	5

TABLE 2. Where the parameters for sample size equation were derived from

In Figure 10 the estimated variances are mostly less than the real variances at the end of the trials. It is hard to determine which source of variance estimation works best, as there were only a limited number of trials in the audit. The 'previous trial' method produced a spread, where some were accurate and others not, whilst all trials that used 'literature' to determine estimated variance performed well. 'Pilot studies' had scattered results, with two performing satisfactorily and two not. Thus pilot studies were not as accurate as expected. Rather than give a best source of information our audit has given an idea of how these estimates are derived.

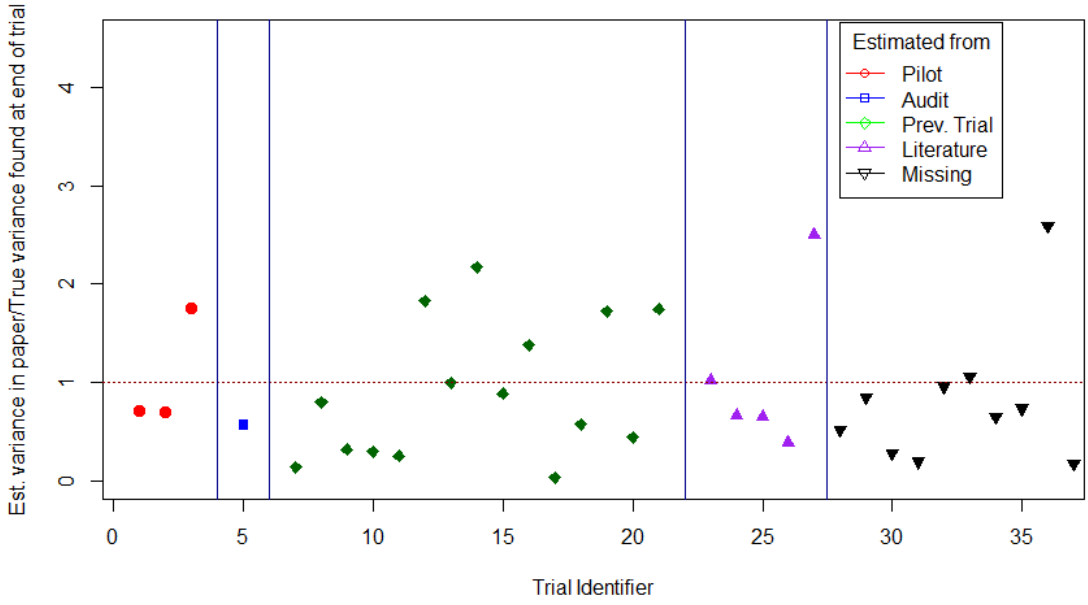


FIGURE 10. Comparison of estimated variance and real variance

2.4. **Discussion.**

2.4.1. *Under reporting.*

Under reporting of the parameters used to derive sample size was an issue identified while reviewing the trials. The standard deviation is frequently not reported, and so it is impossible to know how the estimated sample size was derived. The power and Type I error were reported in every trial, and treatment difference in all but one. However, why specific treatment differences were chosen is not always stated. The sample size calculation used is not often stated, as most journal editors do not make it a mandatory criterion for publication.

Many of the trials did not state what method of sample size derivation was used. There is an important issue concerning the under reporting of statistical details in medical journals. It is often not clear what statistics have been applied in published trials, and the missing information on sample size calculation identified in this review is just one of numerous examples (Chan, 2008).

2.4.2. *Sources of Information for initial parameters.*

Literature and medical knowledge were some of the methods used to gain preliminary information for design of the audited trials, with sources being medical journals and articles. Audits were also conducted to estimate the treatment differences used in trials, particularly with reference to pain scores.

Information from previous trials was the most common method used to derive the parameters required for sample size estimation and, where the trial was novel, pilot studies were often included to aid the estimation.

In some cases, multiple previous trials were available, and so an average of the treatment difference and standard deviation could be used as estimators. Alternatively, parameters from the largest trial available were used. The number of participants in pilot trials was small (ranging from 5-15 people) in comparison to the number participating in full trials. It has been stated by Billingham *et al* (2013) , from a previous audit review, that the median sample size for pilot and feasibility studies is 36 participants, which is relatively high compared to the numbers observed in this audit.

2.4.3. *Insufficient pilot sample sizes.*

On observation, the pilot trials did not perform as well as expected, seen in Figure 10. A pilot study should, in theory, perform well as its design is the same as the trial being conducted, just with a smaller sample size. However it can be proved that if the number of the patients in the pilot trial is small, then the pilot trial will not provide a good estimator for the variance of the main trial. The number of subjects in the pilot trials were 6,10,12 and 26.

The effect of using such small sample sizes can be assessed as follows.

We first note that the sample variance, s^2 is distributed as,

$$s^2 = \sigma^2 \frac{\chi^2_\nu}{\nu}$$

where ν is the degrees of freedom and σ^2 is the true unknown variance.

We can determine a 95% confidence interval from noting

$$0.95 = Pr(\frac{\chi^2_{\nu:L}}{\nu} < \frac{s^2}{\sigma^2} < \frac{\chi^2_{\nu:U}}{\nu}) \tag{2}$$

where $\chi^2_{\nu:L}$ and $\chi^2_{\nu:U}$ are the lower and upper limits of the interval respectively.

If we aim to have $\frac{s^2}{\sigma^2}$ in the interval

$$[\frac{1}{1.1}, 1.1]$$

then the degrees of freedom required is ≈ 40 using the χ^2 tables, so at least 40 patients are required to make a pilot trial adequate. This is much bigger than the sample size of the studies observed in the audit.

2.5. **Limitations.**

The limitations of this study include the fact that only 47 trials were analysed, which is a small sample of data and may reduce the validity of the results. The search was carried out by just one reviewer and was not repeated to check for accuracy, owing to time constraints.

2.6. **Conclusion.**

Most trials examined within the audit have stated treatment difference, power and Type I error. However, standard deviation needs to be included more regularly. Derivation of these estimators should also be explicitly stated, for clarity and for justification of sample size.

In this study it was found that most trials obtain these estimators from previous trials. The accuracy of the sample size estimation is often not as good as statisticians would like, and it is suggested that new methodology could be used, when inference around the variance is uncertain.

3. Methods of Sample Size Re estimation

3.1. **Internal or External Pilot?**

It is clear from the results of the Audit in Section 2, that the external pilots in these studies did not provide an accurate estimate of variance. George Cochran once said that 'pilot studies would often lead to regret' (Wittes and Brittain, 1990). You could argue that, if an external pilot study is significant, what is the point of a larger study? And if not significant, a larger study is a wasted investment anyway. To follow this line of thinking, however, is to risk losing the wider insights that might emerge from full size studies. Either way, many are of the opinion that external pilots are a questionable use of time and resources.

Sample size re-estimation methods, however, use an internal pilot. By incorporating the pilot study within a two stage trial design, data collected from it is not wasted. Instead, it informs the second stage of the trial and allows for correction of inaccurate initial values.

3.2. **Motivation for Sample Size Re estimation.**

The Audit in Section 2 highlighted the need for sample size re-estimation methods, in trials where the parameters are uncertain. Classical designs however, with fixed sample sizes, are still the most common method used in clinical trials. These were designed by pioneering statisticians, including Sir Ronald Fisher. His methods were based on agricultural research, where a large amount of time is needed to see the outcome of experiments. Armitage (1993, as cited in (Jennison and Turnbull, 1999)) argued that statistical theory has evolved to require different methodology, for use in fields like medical and industrial research, where outcomes are observed more quickly. Adaptive designs, such as sample size re-estimation, have been developed to achieve more efficient methods, leading to clinical gain. The need to improve the efficacy and effectiveness of the clinical development process has also been recognized by regulatory bodies such as the Food and Drug Administration (FDA) in America, and the European Medicines Agency (EMEA). They believe the design, conduct and analysis of clinical trials is a key area for improvement. Note that the motive for investigating sample size re-estimation in this report is to explore a way of avoiding the hazards of inaccurate estimation of parameters; it is not about feasibility (where an external pilot might be more useful).

3.3. **Basic Framework.**

Sample size re-estimation is based on the following four steps:

FIGURE 11. Illustration of internal pilot method

(1) The first step echoes the classical design described in Chapter 1. Estimates for the important treatment difference and standard deviation are required, and Type I error and power need to be set to get an initial estimate of the sample size. These parameters are then put into the sample size equation:

$$n_0 = \frac{2S_0^2(z_{\frac{\alpha}{2}} + z_\beta)^2}{\triangle^2}$$

where S_0^2 is the initial variance.

(2) A proportion π is chosen and πn_0 patients are recruited to each arm of the trial. This constitutes the 'internal pilot study'.

(3) Once there are πn_0 patients in each arm, the trial is paused for sample size re-estimation. The internal pilot data is analysed (in various different ways depending on method) to produce a new estimate for the variance of the outcome, called S_1^2. This estimate is informed by the internal pilot data, and so should be more accurate than the estimate used initially. The new variance estimate is then put into the sample size equation:

$$\hat{n_1} = \frac{2S_1^2(z_{\frac{\alpha}{2}} + z_\beta)^2}{\triangle^2}$$

to get a new estimate $\hat{n_1}$ of the sample size in each treatment group.

(4) This estimate is not necessarily the number of patients recruited as this depends on what restrictions, if any, are used:

- the 'restricted': $n_1 = max(n_0, \hat{n_1})$; proposed by Wittes and Brittain, 1990. The study will always recruit at least n_0 patients in each treatment arm. This was imposed to safeguard the method further against Type I error inflation. See later in the Chapter.
- the 'unrestricted': $n_1 = max(\pi n_0, \hat{n_1})$; proposed by Birkett and Day, 1994. This allows the study to terminate, if the internal pilot, πn_0 is the same or larger than the new

sample size estimate, $\hat{n_1}$. However this design is not popular, as a minimum number of patients are often required in a trial, in order to meet the treatment's safety profile. Throughout this dissertation, the 'restricted' approach will be used.

(5) The trial continues until n_1 patients outcomes are collected in each treatment arm. The full data will consist of the internal pilot data, collected in the first stage, combined with the additional data recruited from the new sample size estimate using S_1^2.

(6) At the end of the trial, the data is then assessed by a t test statistic applied to the full data.

3.4. **Implications of the Sample size reestimation method.**

Two potentially problematic implications of using the sample size re-estimation method are the unblinding of the treatment groups at the end of the first stage in the trial and inflation of the Type I error. The ICH (1999) E9 (Phillips and Haudiquet, 2003) guideline states that: *'The steps taken to preserve blindness and consequences, if any, for the type I error* [...] *should be explained'.* Many authors have proposed different methods to overcome these two problems.

3.5. **Unblinded Methods.**

Stein (1945) and Wittes and Brittain (1999) proposed methods of sample size re-estimation using the internal pilot, where the estimate of variance uses pooled variance of the data. These methods estimate pooled variance of each treatment group and so require that the patients allocation to treatment is revealed and hence unblinded. To ensure this unblinding is performed in an unbiased manner, the estimate of variance could be made through a third party, which would require DMEC approval. Alternative technical solutions have also been proposed. The Wittes and Brittain method follows the basic framework using the 'restricted' design. It shall be called the unblinded method throughout this dissertation.

3.6. **Blinded Methods.**

To avoid having to unblind the data to determine a revised value of S_1^2 for the variance estimate at the end of the internal pilot, several methods have been proposed.

3.6.1. *The 'lumped variance' method.*

A method where the total variance of all the treatments is used as an estimator for standard deviation, S_1^2, can be used to avoid unblinding the treatment allocations of the data. This is coined by Zucker

et al (1999) as the 'lumped variance', and is defined as:

$$S^2_{1,total} = \frac{1}{\pi n_0 - 1} \sum_{j=1}^{2} \sum_{k=1}^{\pi n_0} (X_{1jk} - \bar{X}_1)^2$$

where $\bar{X}_1$ is the mean of the total data for both treatment groups, $2\pi n_0$, j denotes that we sum over two treatment arms and k is the number of patients in each treatment group; which, in all cases throughout this dissertation, are equal.

The lumped variance is most desirable when the true treatment effect is not too large, as the total variance will be similar to the 'within group' variance.

3.6.2. *The 'Adjusted variance' method.*

Zucker *et al* (1999) modified the 'lumped variance' method. The authors proposed the estimator:

$$S^2_{1,adj} = S^2_{1,total} - \frac{\pi n_0}{4(\pi n_0 - 1)} \triangle^2$$

Their method adjusts the total variance by subtracting a value that is a function of the clinically important treatment difference $\triangle$ and the size of the internal pilot πn_0. This estimator is most desirable when the true effect of the treatment is close to the value we designed the trial to detect, i.e $\mu_T - \mu_C = \triangle$.

Gould and Shih (1992) proposed an EM algorithm based procedure for sample size recalculation which produces a maximum likelihood estimate of the within-group variance while preserving the blind. However, others have claimed that the method can produce non-unique and/or severe underestimates of the true within-group standard deviation. Friede and Kieser (2002) noted that this approach was flawed and it was not suggested for practical use. The authors recommended that the 'adjusted variance' method, seen above, is instead used to maintain blinding of treatment allocations.

Throughout this dissertation the lumped variance approach shall be referred to as the blinded method and the adjusted total variance approach as the adjusted blinded method.

3.7. **Inflation of the Type I error; use of the *t* statistic.**

It was seen in Figure 1, with the fixed sample design, use of an incorrect initial variance, S^2_0 did not effect the Type I error. This is because when conducting a independent two sample t test, we make

certain assumptions such as sample size is unaffected by statistics that are informative about the treatment outcome means and variance.

In the literature review on methods of sample size re-estimation, an inflation of the Type I error has been recognized as one of the main drawbacks of its use. Using the internal pilot data, πn_0 to inform the new sample size, n_1, has infringed the assumption of independence regarding sample size, which is made when using a t test.

3.7.1. *The 'naïve' t statistic.*

Wittes and Brittain advocate use of the following test statistic, coined the 'naive t statistic' by Kieser and Friede (2004):

$$\frac{\left|\bar{X}_T - \bar{X}_C\right|}{\sqrt{(S^2/n_1)}}$$

where $S^2 = \frac{1}{2}(S_T^2 + S_C^2)$, the pooled variance of both treatment arms, T and C.

This method uses information from Stage 1 and Stage 2 in the final test statistic. It has the advantage of not sacrificing any data.

E(N)	α	Power	S^2
86.0	0.050	0.996	1.0
86.5	0.050	0.97	1.5
93.2	0.050	0.93	2.0
128.4	0.051	0.89	3.0
170.0	0.052	0.90	4.0

TABLE 3. Adapted from Wittes and Brittain, 1990, where the S^2 is the initial variance value and true $\sigma^2 = 1$ and $\triangle = \sqrt{2}$

The issue of Type I error inflation seen in the literature is illustrated in Table 3. Using simulation, Wittes and Brittain were able to prove that the α exceeds the Type I error. Stein (1945) suggested use of a t test that only uses the internal pilot variance, S_0^2. However, this is often regarded as a waste of information; as the inflation obtained when using all the data is minimal. So, Wittes and Brittain (1999) recommend the use of the 'naive' t statistic, nonetheless. Their paper states that the gain in power (by using all the data's variance) is preferable at the expense of a small bias in the Type I error. Similar levels of Type I inflation have been identified in other papers, such as Friede and Kieser (1992), where there was a slight inflation of Type I error using the blinded methods, which is to be expected as the independence assumption is broken.

4. Statistical Properties of Normal Random Variables

It has been stated that the sample size re-estimation method leads to problems such as inflated Type I error. To grasp why this occurs we first must fully understand some properties of the t statistic. These properties include why a t statistic has a t distribution and what the components of this procedure require. We therefore begin with a careful derivation of the t test.

4.1. **Independence of the sample mean and sample variance.**

One important property of random normal variables, is that the sample mean is independent of the sample variance. When the unpaired t test is performed, it is required that the difference in means is independent to the pooled variance. To see this we state the following:

Suppose $y \in \mathbb{R}^N$ and $y = (y_1, ..., y_n)^T$ is an iid sample from a $N(\mu, \sigma^2)$

Now, define a matrix O which is orthogonal, i.e. $OO^T = O^TO = I$.

Choose $\tilde{O}$ such that $x = Oy$ and the first row of $\tilde{O}$ is constant,i.e. i.e $x = \begin{pmatrix} \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} \\ & \tilde{O} & \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$.

$\mathsf{var}(y) = \sigma^2 I$ and

$$\mathsf{var}(x) = \mathsf{var}(Oy) = O\mathsf{var}(y)O^T = \sigma^2 I$$

So, $\mathsf{var}(y) = \mathsf{var}(x)$, the xs are constant too and so are independent. Also the ys are Normal, so

$$x_1 = \frac{y_1 + \cdots + y_n}{\sqrt{n}} = \bar{y}\sqrt{n}$$

where $\bar{y}$ is the mean of the ys

Also,

$$x^Tx = (Oy)^TOy = y^TO^TOy = y^TO^TOy = y^Ty$$

Therefore,

$$y_1^2 + y_2^2 + \cdots + y_n^2 = x_1^2 + x_2^2 + \cdots + x_n^2$$

which allows us to write,

$$y_1^2 + y_2^2 + \cdots + y_n^2 - n\bar{y}^2 = x_2^2 + \cdots + x_n^2$$

Now, the sample variance of the ys is S^2 where,

$$(n-1)S^2 = \sum (y_i - \bar{y})^2 = \sum {y_i}^2 - n\bar{y}^2$$

The sample mean, $\bar{y} = \frac{x_1}{\sqrt{n}} = f(x_1)$ and $(n-1)S^2$ is $g(x_2^2 + \cdots + x_n^2)$, so is independent of x_1 and hence of $\bar{y}$

4.2. **The sample distribution of S^2 is proportional to a χ^2 distribution.**

The result that the sample distribution is proportional to a χ^2 distribution can also be seen from the following calculations. Note that

$$E(y) = \mu 1_n$$

where 1_n is an n-dimensional vector of ones and

$$E(x_k) = E(O_k^T y) = \mu O_k^T 1_n = 0, k > 1$$

when O_k is the k^{th} row of O and hence is orthogonal to the first row of O.

$$(n-1)S^2 = \sum_{i=2}^{n} x_i^2$$

has the same distribution as

$$\sigma^2 \sum_{i=2}^{n} Z_i^2$$

where $z_i \sim N(0,1)$ so $S^2 \sim \sigma^2 \chi_\nu^2/\nu$ where $\nu = n-1$

It follows that the variance also has a distribution proportional to χ_ν^2. For the two sample case, the pooled variance is $\chi_{\nu_1+\nu_2-2}^2$.

One way to check that the sample variance is proportional to a χ^2 is to check the skewness of the distribution.

4.3. **The skewness of a χ^2 distribution.**

Skewness is defined as the measure of asymmetry of a distribution around its mean. Positive skewness indicates a distribution with an asymmetric tail extending towards more positive values and negative skewness indicates a distribution with an asymmetric tail extending towards more negative ones. It is often used as measure of the departure from the normal distribution, which is about symmetrical, so is not skewed, but small variation of the skewness can occur by chance alone. The χ^2 is asymmetrical, but as $n \to \infty$ it becomes more symmetric, as it tends to the Normal

distribution because of the Central Limit Theorem. The χ^2 distribution is always positively skewed, shown by the following proof.

Skewness is mathematically defined as

$$\frac{E(X-\mu)^3}{\sigma^3} \tag{3}$$

where X is χ^2 and its mean is $\mu = E[\chi^2_\nu] = \nu$ and $\mathsf{var}(\chi^2) = 2\nu$. This is now scale free and so is also the skewness of S^2.

$$E(X-\nu)^3 == E[(\sum Y_i)^3]$$

where $Y_i = Z_i^2 - 1$ and $E[Y_i] = 0$. Now

$$E[\sum (Y_i)^3] = nE[Z^2-1]^3 = nE[Y_i^3]$$

because $(Y_1 + Y_2 + \cdots + Y_n)^3 = \sum_{i,j,k} Y_iY_jY_k = 0$ and $E(Y_iY_jY_k) = 0$ unless all the subscripts coincide.

Therefore, we need to evaluate $E[Y_1^3] = E[(Z_1^2-1)^3]$

$$Skewness = \frac{nE(Z^2-1)^3}{(2n)^{\frac{3}{2}}} = \frac{1}{\sqrt{n}}\frac{E(Z^2-1)^3}{2^{\frac{3}{2}}}$$

So,

$$E(Z^2-1)^3 = E(Z^6 - 2Z^4 + 3Z^2 - 1) = E(Z^6) - 3E(Z^4) + 3E(Z^2)$$

To determine these expectations we use the moment generating functions,

$$E(e^{tZ}) = exp(1/2t^2)$$

$$E(1 + tZ + \frac{t^2Z^2}{2!} + \frac{t^3Z^3}{3!} + \frac{t^4Z^4}{4!}\cdots) = 1 + \frac{t^2}{2} + \frac{t^4}{8} + \frac{t^6}{48}$$

Equating coefficients,

$$E(Z^2) = 1$$

$$E(Z^4) = \frac{4!}{8} = 3$$

$$E(Z^6) = \frac{6!}{48} = 15$$

Putting this back into the equation for skewness,

$$E(Z^6 - 3Z^4 + 3Z^2 - 1) = E(Z^6) - 3E(Z^4) + 3E(Z^2) - 1 = 15 - (3 \times 3) + (3 \times 1) - 1 = 8$$

Therefore,

$$Skewness = \frac{8n}{(2n)^{\frac{3}{2}}} = \sqrt{\frac{8}{n}} \tag{4}$$

Note that the skewness must be a positive value. A sample skewness, seen in Joanes and Gill (1998), can be defined as

$$g_1 = \frac{m_3}{m_2^{3/2}}$$

where

$$m_r = \frac{\sum_i (x_i - \bar{x})^r}{N}$$

is the sample moments of order r where N is the sample size.

4.4. **The error of the skewness.**

It can be proven that the standard error of the skewness is

$$\sqrt{\frac{6}{N}} \tag{5}$$

where N is the sample size, seen also in Joanes and Gill (1998).

These theoretical properties of random normal variables will be used as a comparison to the results of a simulation study, which will be discussed later in this paper.

5. The Two Stage Design: A simulation study

5.1. **Setting and Notation.**

Consider a study comparing two groups, treatment (T) and control (C), assumed for simplicity to be equally sized. Let the outcome, X be normally distributed with μ_T as the mean of group T and μ_C the mean of group C, both of which have common unknown variance σ^2. The hypothesis tested is:

$$H_0 : \mu_T - \mu_C = 0 \text{ versus. } H_1 : \mu_T - \mu_C \neq 0$$

The aim of the trial is to detect a clinically important difference of $\triangle = 0.175$, with a two sided Type I error rate of 0.05 and a power of 80%. This will require $n_0 = 513$ per group, if the variance is $\sigma^2 = 1$, the value used for the true variance in the study simulations.

To test the implications of the variance σ^2 being underestimated, as this is a common problem which can lead to trials that are too small, a range of initial estimates, S_0^2, were proposed as follows:

- $S_0^2 = 1.0$; the initial estimate is correct
- $S_0^2 = 0.7$; the initial estimate is slightly underestimated
- $S_0^2 = 0.3$; the initial estimate is severely underestimated

5.1.1. *The original fixed-sample trial.*

For direct comparison with the results of the sample size re-estimation trials, results from the original study were generated using traditional fixed sample size methods. Information was collected regarding the sample size, skewness of pooled variance, resulting Type I error and power.

S_0^2	n_0	α	Power
1.0	513	0.0489	0.800
0.7	252	0.0492	0.497
0.3	47	0.0496	0.131

Table 4. Results of the fixed sample size method: Estimated sample size, Type I error and power for varying $S_0^2 = 1.0, 0.7, 1.0$, where α collected when $\mu_T - \mu_C = 0$ and power collected when $\mu_T = \mu_C = \triangle$. The trials were simulated 100,000 times.

Table 4 presents results of the fixed sample trial, including calculated final sample size, the Type I error (how many trials out of 100,000 were significant when there was no treatment difference) and

power (how many trials were significant when there was a treatment difference) for each estimate of the variance at 0.3, 0.7 and 1.0.

5.1.2. *Sample size re-estimation.*

Next, the simulations ran trials using sample size re-estimation where:

- πn_0 patients are collected, where π named the 'information fraction' is:
  - $\pi = 0.25$ or
  - $\pi = 0.5$
- the new sample size estimate has a downward restriction such that:
  - $n_1 = \max(n_0, \hat{n_1})$; proposed by Wittes and Brittain

A study was then implemented to assess and compare the properties of different sample size re-estimation methods. Three types of method were used, with further details on each presented in Section 3). These are:

(1) The unblinded method: proposed by Wittes and Brittain (1990)

(2) The unadjusted blinded method;

(3) The adjusted blinded method; proposed by Zucker et al. (1999)

As for the fixed sample size trial, the parameters collated were final sample size, power, Type I error and bias and skewness of the pooled variance.

The sample size re-estimation study was performed using the software package R, and the simulations were performed 100,000 times.

5.2. **Results.**

5.2.1. *Mean and standard deviation of final sample size.*

The means and standard deviations of the final sample size were computed for each sample size re-estimation method, on each estimate of variance, S_0^2.

As shown in Tables 5 and 4, had the hypothesized standard deviation been correct all methods, fixed and sample size re-estimation, on average, returned similar results. However, when the standard deviation estimate is underestimated, the sample size re-estimation methods , on average, maintain the same sample size, but the sample size of the fixed sample method decreases relative to the

S_0^2	π	$E[N](SD(N))$ Unblinded	Blinded	Adjusted Blinded
1.0	0.25	512.1(45.3)	512.1 (45.3)	508.1 (45.3)
	0.5	512.0 (31.9)	512.0 (31.9)	508.1 (31.9)
0.7	0.25	511.9 (65.2)	511.9 (64.8)	507.9 (64.9)
	0.5	512.0 (45.8)	512.0 (45.7)	508.1 (45.7)
0.3	0.25	512.1 (154.7)	512.1 (151.2)	508.0 (151.2)
	0.5	512.1 (106.4)	512.1 (105.2)	508.1 (105.3)

TABLE 5. Mean final sample size of 100,000 studies with their standard deviations; seen in brackets, for each S_0^2; 1.0, 0.7, 0.3 and π; 0.25 or 0.5.

proportion of underestimation of the standard deviation. However, the variable nature of the sample size returned by sample size re-estimation methods, show that this mean sample size returned, which is suitable, is not guaranteed. This is shown by the standard deviations in Table 5.

Figure 12 shows the difference in certainty about sample size, as the information fraction is lowered and as the underestimation of standard deviation becomes more severe. When the information fraction $\pi = 0.25$ and the initial standard deviation estimate $S_0^2 = 0.3$, as shown in Figure 12, there is a very large standard deviation, with the implication that the sample size could be half or double that required.

5.2.2. *Power.*

The power of the trials was collected from simulations conducted when $\mu_T - \mu_C = \triangle$. This probability was found by dividing the number of significant trials by the number of times the simulation was run.

S_0^2	Fixed Sample Est. Power	π	Sample Size Re-estimation Est. Power Unblinded	Blinded	Adjusted Blinded
1.0	0.800	0.25	0.800	0.800	0.797
		0.5	0.799	0.799	0.797
0.7	0.497	0.25	0.794	0.795	0.792
		0.5	0.798	0.799	0.796
0.3	0.132	0.25	0.772	0.774	0.770
		0.5	0.786	0.786	0.783

TABLE 6. Estimated power of fixed sample method when S_0^2=1.0, 0.7, 0.3 and estimated power for sample size re-estimation methods when S_0^2=1.0, 0.3, 0.7 and π= 0.25, 0.5

To provide a clear comparison, Table 6 provides the power of the fixed design and the sample size re-estimation methods. Figure 13 displays the comparison between the fixed and sample size re-estimation methods and also shows a clearer comparison of the different versions of the sample size

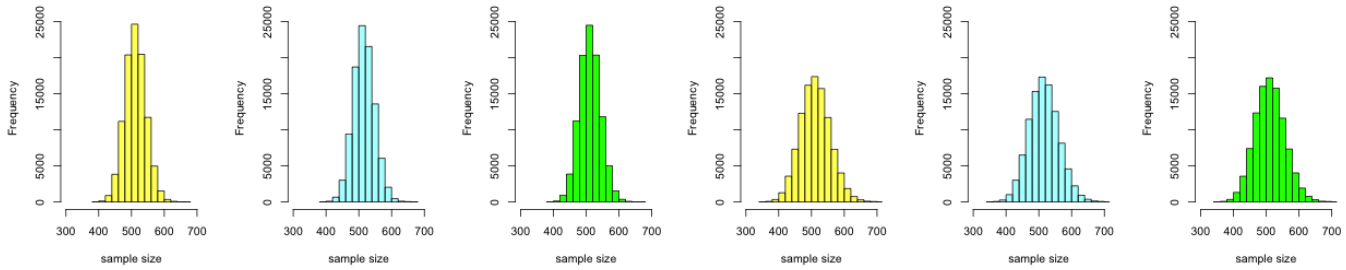


Figure 12.A: $\pi = 0.5$ and S_0^2=1

Figure 12.C: $\pi = 0.25$ and S_0^2=1

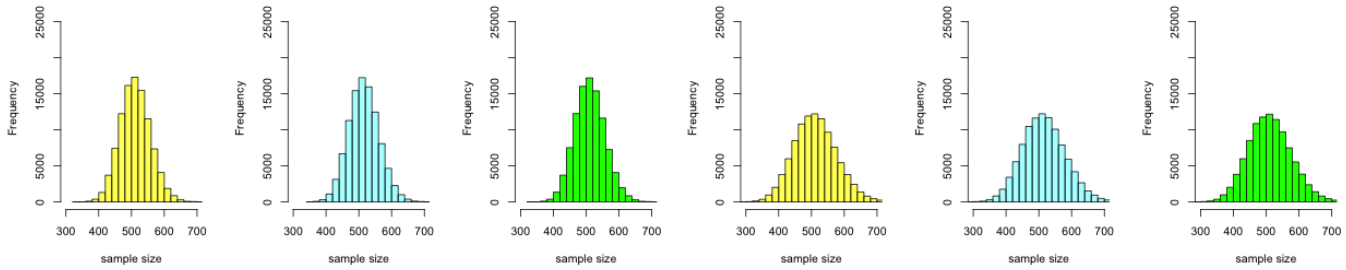


Figure 12.B: $\pi = 0.5$ and S_0^2=0.7

Figure 12.D: $\pi = 0.25$ and S_0^2=0.7

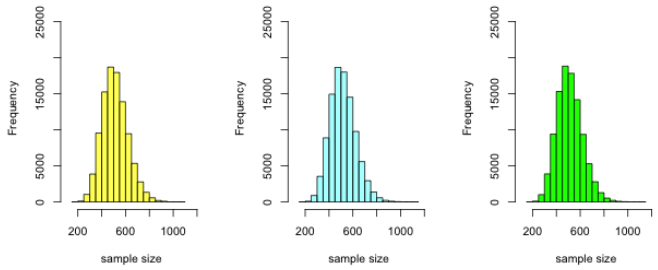


Figure 12.E: $\pi = 0.5$ and S_0^2=0.3

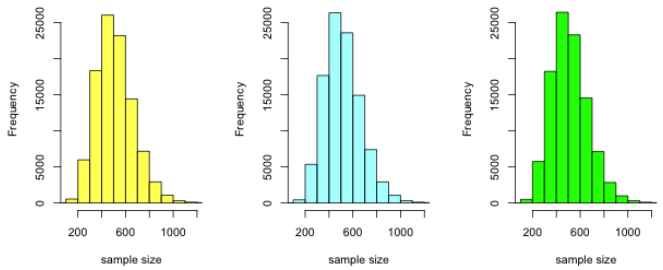


Figure 12.F: $\pi = 0.25$ and S_0^2=0.3

FIGURE 12. Histograms of the Sample Size for each sample size re-estimation method where S_0^2=1.0, 0.7, 0.3 and π=0.25, 0.5

re-estimation methods for varying variance estimates. On both plots on Figure 13 there is a line when the power is 0.8, the power the trial was designed initially to achieve.

According to the simulations performed in this dissertation, the average power achieved did not fall below 77% under the alternative hypothesis for the sample size re-estimation methods. If the fixed

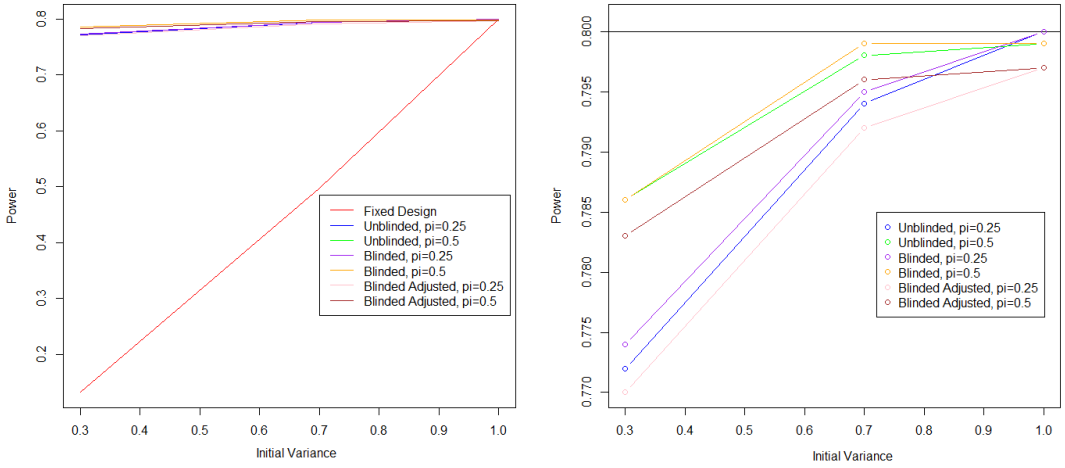


FIGURE 13. Left:Estimated power against S_0^2 for fixed sample design and sample size re-estimation methods; Right: Magnified graph to distinguish difference between sample size re-estimation methods

sample design was used, the power of the trial decreased to as little as 13%, when the trial was severely underestimated, when S_0^2=0.3.

S_0^2	π	10%	25%	50%	75%	90%
1.0	0.25	0.75	0.77	0.80	0.82	0.84
	0.5	0.77	0.79	0.80	0.82	0.84
0.7	0.25	0.73	0.77	0.80	0.83	0.86
	0.5	0.75	0.78	0.80	0.82	0.84
0.3	0.25	0.62	0.70	0.79	0.86	0.91
	0.5	0.68	0.70	0.80	0.85	0.89

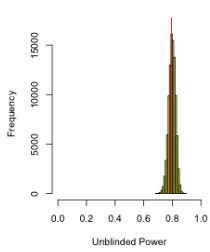
TABLE 7. Quantiles of estimated power using the unblinded sample size re-estimation method where S_0^2=1.0, 0.7, 0.3 and π=0.25, 0.5

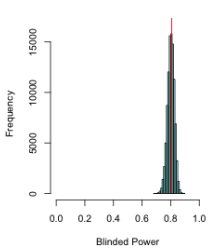
These powers were found by replicating a trial 100,000 times and taking an average. So, looking at the whole picture, the sample size re-estimation methods work well. However, how about on an individual basis? This is of most concern to practicing statisticians- how well will this method work in any specific trial? The power of each trial was captured, using a re-arranged version of the initial sample size equation.

$$z_\beta = \frac{\triangle}{\sigma\lambda} - z_{\frac{\alpha}{2}}$$

where $\lambda = \sqrt{(2/\pi n_0)}$; as there are equal number of patients in each treatment group, β is found and hence $1-\beta$ =power.

Figure 14 represents the power of each individual trial. It is seen that when π and S_0^2 are small, there is much more variability in power, with some trials being extremely underpowered. The quantiles seen in Table 7 shows that 10% of trials achieve only 62% power when the initial value for variance,





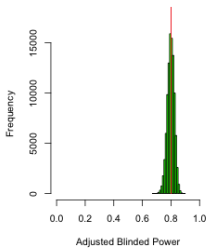
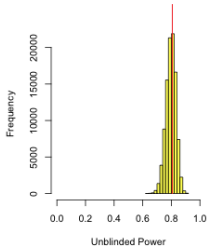
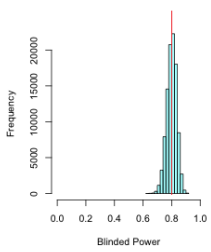


Figure 14.A: $\pi = 0.5$ and S_0^2=1.0





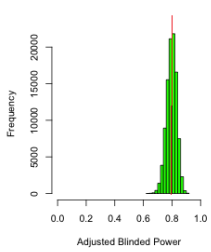
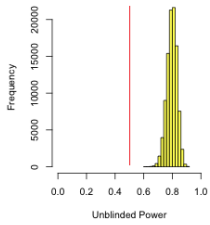
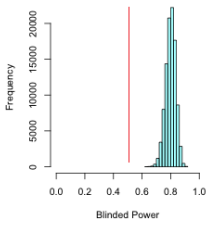


Figure 14.C: $\pi = 0.25$ and S_0^2=1.0





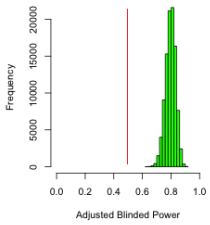
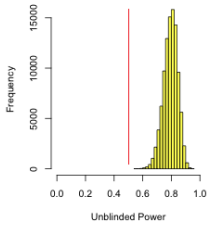
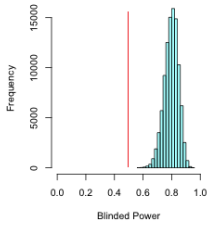


Figure 14.B: $\pi = 0.5$ andS_0^2=0.7





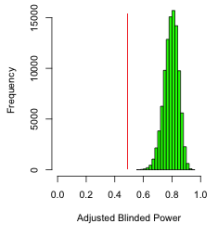
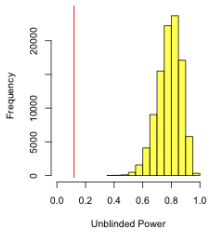
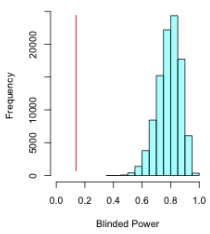


Figure 14.D: $\pi = 0.25$ andS_0^2=0.7





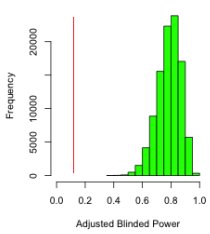
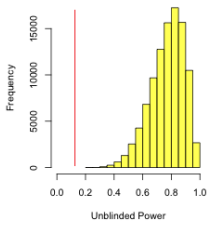
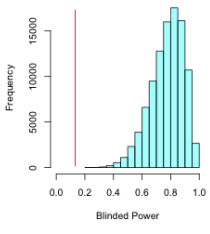


Figure 14.E: $\pi = 0.5$ and S_0^2=0.3





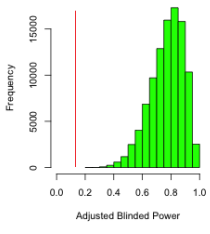


Figure 14.F: $\pi = 0.25$ and S_0^2=0.3

FIGURE 14. Histograms of the power for the sample size re-estimation methods where S_0^2=1.0, 0.7, 0.3 and π=0.25, 0.5, red line displays fixed power of the fixed sample design

S_0^2, is small; however, this is still a much larger power than the fixed sample design which only obtained 13.2%.

5.2.3. *The Type I error.*

The Type I error was collected from simulations under the null hypothesis. It was derived similarly to the power. Table 8 displays the errors for both types of design, fixed sample and sample size re-estimation.

$\frac{\sigma_{est}}{\sigma_{real}}$	Fixed Sample, α	π	n_0	Type I Error, α Unblinded	Blinded	Adjusted Blinded
1.0	0.049	0.25	513	0.050	0.050	0.050
		0.5	513	0.050	0.050	0.050
0.7	0.049	0.25	252	0.051	0.051	0.051
		0.5	252	0.051	0.050	0.051
0.3	0.050	0.25	47	0.051	0.050	0.051
		0.5	47	0.050	0.049	0.049

TABLE 8. The Type I error of the fixed sample method where S_0^2=1.0, 0.7, 0.3 and the sample size re-estimation methods where S_0^2= 1.0, 0.7, 0.3 and π = 0.25, 0.5

5.2.4. *Bias of* σ^2.

Table 9 displays the mean pooled variance of the sample size re-estimation trials compared to the pooled variance of the fixed sample design. The pooled variance for the sample size re-estimation methods is always smaller than the true variance of 1. The fixed sample design, however, is correct to 3 decimal places for all S_0^2. Wittes, et al (1999) discussed how the unblinded method could lead to a downward bias, and the results of the simulations in Table 9 verify this theory.

S_0^2	Fixed Design $E[\sigma^2]$	π	$E[\sigma^2]$ Unblinded	Blinded	Adjusted Blinded
1.0	1.0000	0.25	0.9983	0.9983	0.9982
		0.5	0.9981	0.9981	0.9981
0.7	1.0000	0.25	0.9980	0.9980	0.9980
		0.5	0.9981	0.9980	0.9980
0.3	1.0000	0.25	0.9977	0.9978	0.9977
		0.5	0.9979	0.9980	0.9979

TABLE 9. The bias of σ^2 for the fixed sample method where S_0^2=0.3, 0.7, 1.0 and sample size re-estimation methods where S_0^2=0.3, 0.7, 1.0 and π=0.25, 0.5

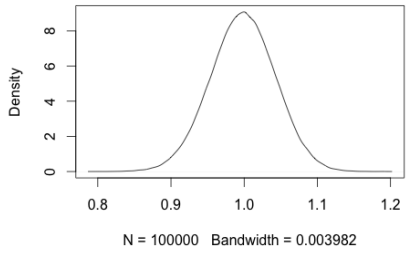


FIGURE 15. π = 0.5 and S_0^2=1

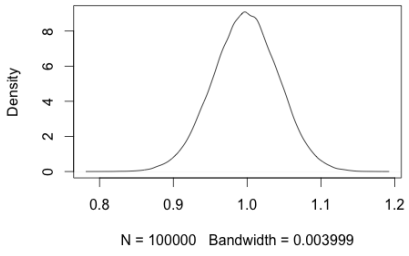


FIGURE 16. π = 0.5 andS_0^2=0.7

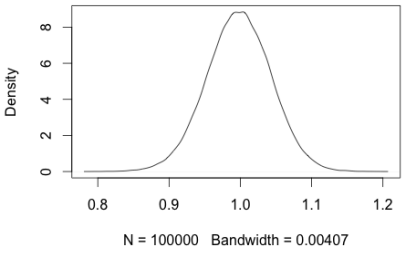


FIGURE 17. π = 0.5 andS_0^2=0.3

FIGURE 18. Histograms of the σ^2 for the sample size re-estimation methods where S_0^2=1.0, 0.7, 0.3 and π=0.5 using the unblinded method

This bias and the distribution of σ^2 are related. It is seen in Figure 18 that the distribution is not quite the bell shaped curve we would expect of the Normal distribution. There is evidence of a slight negative skewness.

5.2.5. *Skewness.*

To see the implications the interim look at the data can have on the statistical properties of the t test, the skewness of the pooled variance used in the final analysis for each of the sample size re-estimations methods was derived from the simulations.

S_0^2	Fixed Sample Design Skew	Theoretical Skew	π	Unblinded Skew	Blinded Skew	Adjusted Blinded Skew	Theoretical Skew
1.0	0.08, [0.07, 0.10]	0.09	0.25	-0.07, [-0.08, -0.05]	-0.07, [-0.08, -0.05]	-0.07, [-0.08, -0.05]	0.09
			0.5	-0.05, [-0.06, -0.03]	-0.05, [-0.06, -0.03]	-0.05, [-0.06, -0.03]	0.09
0.7	0.12, [0.11, 0.14]	0.13	0.25	-0.05, [-0.07, -0.04]	-0.05, [-0.07, -0.04]	-0.05, [-0.07, -0.04]	0.09
			0.5	-0.05, [-0.06, -0.03]	-0.05, [-0.06, -0.03]	-0.05, [-0.06, -0.03]	0.09
0.3	0.29, [0.28, 0.31]	0.30	0.25	-0.07, [-0.09, -0.06]	-0.07, [-0.09, -0.06]	-0.07, [-0.09, -0.06]	0.09
			0.5	-0.05, [-0.07, -0.04]	-0.05, [-0.07, -0.04]	-0.05, [-0.07, -0.04]	0.09

TABLE 10. Table comparing the skewness of fixed sample method with 95% confidence intervals with theoretical skewness for S_0^2=1.0, 0.7, 0.3 and comparing the skewness of the sample size re-estimation methods with 95% confidence intervels and their theoretical skewness for S_0^2=1.0, 0.7, 0.3 and π=0.25, 0.5

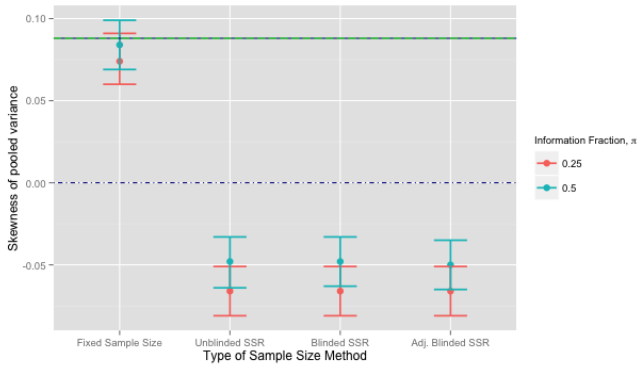


FIGURE 19. S_0^2=1

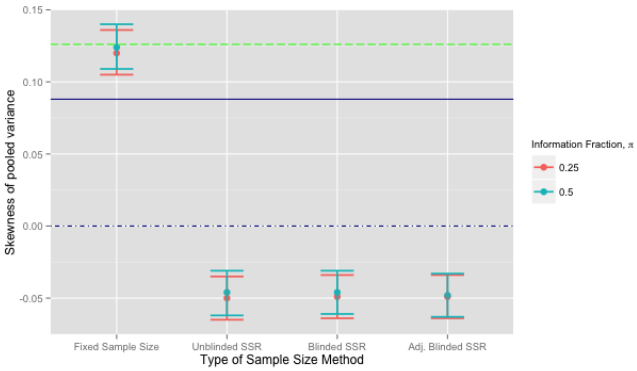


FIGURE 20. S_0^2=0.7

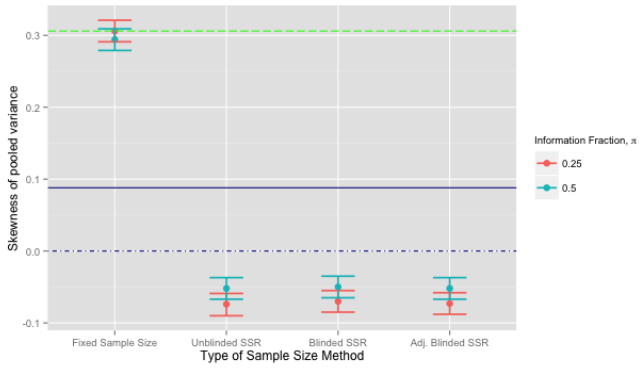


FIGURE 21. S_0^2=0.3

FIGURE 22. Skewness of the fixed sample size method and the sample size re-estimation methods with 95% confidence intervals. Line 0.09 represents the theoretical skew when σ^2=1.

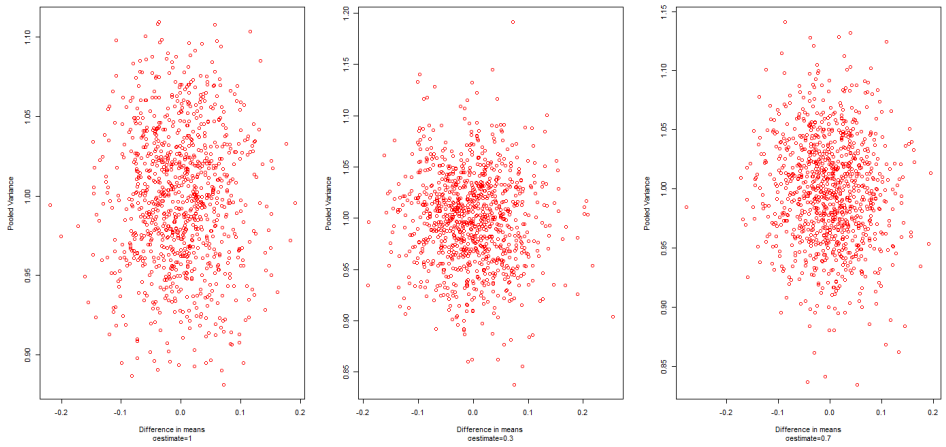


FIGURE 23. Correlation graphs between the estimated treatment difference the estimated pooled variance, S^2, for the unblinded sample size re-estimation method

Figure 22 shows the skewness for each estimation of initial standard deviation S_0^2 for the sample size re-estimation methods. The standard error for each were calculated using Equation 5 which are compared to the theoretical value derived from Equation 4. It is clear, that the pooled variance of the internal pilots have different properties to the fixed design which correspond to their theoretical values.

S_0^2	Correlation
1.0	-0.01
0.7	-0.06
0.3	-0.07

TABLE 11. Correlation between estimated treatment difference and pooled variance estimate for unblinded method

One of the theoretical assumptions made is the independence between $\bar{X}_T - \bar{X}_C$, the estimated treatment difference, and the pooled variance estimate of the final sample. Figure 23 shows the relationship between these two variables for each estimation of the standard deviation, S_0^2. At inspection, it can be concluded that there is no distinguishable pattern to identify any specific relationship between these two variables, for the unblinded method (seen in Figure 23) or the blinded methods (results not shown). All correlations are slightly negative, seen in Table 11, but these are not large enough values to prove dependence.

6. Discussion

6.1. **Comparison of original and internal pilot methods.**

6.1.1. *Results and Conclusions.*

The sample size re-estimation method was successful in providing a mean estimate of sample size close to the required number of patients. When the original estimation of standard deviation matched the true values, the fixed sample size and sample size re-estimation methods, on average, returned similar results. That is, for each method and information fraction of the sample size re-estimation, the mean sample sizes returned were similar and sufficient. When the initial variance was underestimated, with the sample size re-estimation method the final mean sample size remained the same as when the initial variance was correctly estimated. This was not the case with the classic, fixed sample size method however, where the sample size generated was much smaller following underestimation of the initial variance. This was most severe when the standard deviation was believed to be only 30% of the true value. Therefore the sample size re-estimation method was much more efficient in generating an appropriate sample size when substantial initial misspecification of the variance occurred. Perhaps the biggest advantage of the sample size re-estimation method is the saving in power. When initial estimates of variance used in the fixed sample size trial were too small, the predicted power was reduced significantly. As a result, many of the fixed sample size trials were would fail to detect important treatment differences. However, the sample size re-estimation approach enabled the study to achieve power at a level close to the one desired (80%) by readjusting the estimate for variance at the interim review. The integrity of the trial was saved and the clinically important difference between the means of the two treatments was identified.

6.2. **Limitations of Sample Size Recalculation.**

6.2.1. *Why does the Type I error change?*

The use of sample size re-estimation is not without controversy. One of the main criticisms is the potential for Type I error inflation, but this has been proved to be very small in most cases (see Methods, Section 3.5 for qualification). In the sample size re-estimation simulations presented in this paper, the Type I error found was only marginally bigger than the 0.05 value required, and so would not disturb the integrity of the trial. In the literature review, Wittes and Brittain (1999) discussed high Type I error inflation in small internal pilots, πn_0 with sample sizes of 10 or less. In the

current simulations however, much smaller inflation was observed, although it is worth noting that the smallest sample size, containing 12 subjects, was slightly larger than Wittes and Brittain's.

$n_0\pi$	N	Type I Error, α
10	30	0.03394
20	40	0.0294
30	52	0.02794
50	76	0.02678

TABLE 12. Adapted from Kieser and Friede (2004), Type I error rates, α for the nominal level $\alpha = 0.025$ and various internal pilot sizes, $n_0\pi$, where N is the true but unknown sample size required

Kieser and Friede (2004) conducted a review of unblinded and blinded sample size re-estimation methods. The authors found some levels of high inflation, again particularly when the size of the pilot was small. Their results are shown in Table 12. Results from the current simulations, however, do not reflect such extreme values of inflation. It should be noted that Kieser and Friede designed their trial using a one-sided t-test, as opposed to the two sided t-test used in this dissertation.

However, with sample size re-estimation, the dependence of the second stage data collected on the parameters of estimation complicates the distribution. The negative skewness of the pooled variance, an output of the simulations presented in the Results section in this dissertation, proves this statement. This is because Equation 4, seen in Section 4.3 proved that the skewness of a χ^2 must be positive and so the pooled variance is therefore cannot follow a χ^2 distribution. So the final test statistic does not follow a t distribution, due to its dependence on the pooled variance. This complicates the sample size re-estimation process further still, as the final test statistic distribution is now not known. Ignoring this discrepancy may inflate the Type I error, which is observed in the simulations in Table 8. However, the χ^2 is nearly symmetrical for trials of any reasonable size. The skewness of S^2 is also very close to zero, which implies that the usual test statistic will be close to a t distribution and so inflation of the Type I error will be small.

The bias of σ^2.

It was seen from the present simulations that, in comparison to the fixed design, there was a downward bias of the estimated standard deviation, when using the sample size re-estimation methods. This bias was first identified by Wittes and Brittain (1999) and further explored by Coffey and Muller (2000).

This bias holds importance, as the test size inflation varies directly with it. This bias can also help explain the skew of the pooled distribution. The negative bias means that more trials fall below the expected value and so a negative skewness follows, due to the heavy negative tails.

Some Type I error inflation can be explained by this downward bias of the variance estimate used (Kairalla, 2007). Coffey and Muller (2000) stated that this bias is minimized in the final t test when the stage I data collected in the internal pilot is much smaller than the stage II data.

6.2.2. *Implications of using a small internal pilot.*

Sample size re-estimations have the potential to save trials from the potentially negative impact of variance misspecification which, as we have seen, is a drawback of conventional, fixed-sample trials.

This is seen in the results of the present study, while the mean sample size was sufficient, the variability of sample size was high. This is because using sample size re-estimation causes the final sample size to be a random variable, due to its dependence on characteristics of the internal pilot data. Variability is greater with lower estimates of standard deviation, as this leads to a smaller internal pilot on which new estimates are made, decreasing the certainty of the new standard deviation estimate.

When it comes to deciding what proportion of the sample size should be used to form the internal pilot, it is a question of achieving the right balance between power and efficacy. It would be preferable to have a small internal pilot to avoid over recruiting to give an early indication when the trial size has to change. However, if the internal pilot is too small, the new estimate for the standard deviation, S_1^2, varies more widely around the actual value. Wittes and Brittain (1990) chose to use 50% of the initial sample size calculated in their simulations, which may reflect this thinking.

The results of the study in this dissertation showed that the mean sample size was the same for both information fractions, π=0.25 and π=0.5. However, the standard deviation of the final sample size when π=0.25 was twice as variable as when π=0.5.

The present findings reflect the work of Galli and Mariani (2014), who also evaluated two stage re-estimation data using simulations. The authors simulated a two-arm trial aimed at comparing two means of normally distributed data, finding an unwanted increase in sample size due to the variability of the re-evaluated standard deviation, the result of a small internal pilot. The authors declared that a sufficiently high information fraction, π, of $50\% - 70\%$ is required for sample size

re-estimation to have good estimated properties. However, in large trials such a high proportion would be hugely inefficient.

One suggestion might be to propose a fixed sample size for the internal pilot, rather than using the information fraction, π, to extract a proportion from the original estimated sample size. It was seen in Section 2.4.3 that a minimum of 40 patients to prove a 95% confidence interval within 10% of the true value. The fixed internal pilot method may be preferable for large studies, where use of an information fraction might result in unnecessarily large internal pilots.

6.2.3. *Comparison of blinded and Unblinded methods.*

Some critics of sample size re-estimation have argued that information gained throughout the course of a study introduces the potential for bias in blinded studies. Gould and Shih (1992) highlighted several potential causes of such bias.

Although many statisticians see this as an issue, in practice today there are many ways of avoiding bias without the need to keep the studies blinded. A third party could be enlisted, for example, to unblind the treatment allocations at the interim stage, so that those assessing the trial are not aware of such allocations.

Concerns over the consequences of bias lead researchers to establish blinded methods, to avoid information about a treatment difference being revealed at the interim stage. The simulations in this dissertation showed that blinding the allocation of treatments did not change the outcome of the sample size re-estimation compared to unblinded allocations.

The adjusted blinded method used in the present study produced results comparable to the other sample size re-estimation methods used. However, the mean final sample size observed was marginally less than that observed with the unadjusted blinded and unblinded methods, and so was consequently less powerful than the other methods. The power achieved was still of a satisfactory level, however, especially in comparison to fixed sample size trial design.

6.3. **Potential barriers to the adoption of Sample size recalculation.**

Bauer and Einfalt (2006) conducted a review of the use of sample size re-estimation between 1989 and 2004, concluding that there has been an increase in use of these methods. However, this novel process is still not commonplace in medical research. Although very effective, there are still a number of issues that investigators need to consider before using sample size re-estimation:

- The ability to re-scope or extend a clinical trial earlier than would happen conventionally could lead to funding issues. If a request to extend is based on a small internal pilot, or the standard deviation is much larger than originally estimated, then funding agencies may recommend termination of the trial.

- There are many possible methods for re-estimating the sample size of an ongoing trial which, whilst providing potential for more efficient trials, may discourage researchers who are confused by the multitude of options available and their own lack of familiarity with required technique. A study conducted by Scott and Baker (2007) found that regulatory uncertainty, the lack of popularity of a novel approach and logistical problems all contribute to the lack of implementation of sample size re-estimation in clinical trials.

6.4. **Recommendations for future work.**

6.4.1. *Overestimation of* σ^2.

In this dissertation the focus has been on trials where the variance is initially underestimated. This is because, in practice, this is a common issue. Statisticians will hope for a small variance to make the trial feasible in terms of recruitment. However, there will be cases when the standard deviation is overestimated, leading to an overpowered trial. Along with subjecting patients to an inferior treatment, this could also lead to an issue recruiting sufficient numbers. This difficulty would be eased if fewer patients were needed.

When a trial is struggling in recruitment, it is common practice for a trial steering committee (TSC) to be put together, to see if any modifications could be made to the recruiting process or the trial itself. Often, if a trial is very lacking in the required patients, the trial is terminated as it is seen as no longer feasible. One solution, to avoid the trial being shut down, might be to construct an internal pilot based on the data already collected. Provided the number of data collected already is a realistic number (i.e. at least 10), the variance could be recalculated. If this value is found to be less than the estimate used in the initial sample calculation, the new final sample size required will be much smaller than the original, saving the trial. An extension of the current work would be to perform simulations based on $S_0^2 > 1$, when σ^2=1. However, due to project constraints, it was not possible to review both underestimation and overestimation of the variance.

References

Association, W. M.

2013. World medical association declaration of helsinki: Ethical principles for medical research involving human subjects. *JAMA*, 310(20):2191–2194.

Bauer, P. and J. Einfalt

2006. Application of adaptive designs - a review. *Biometrical Journal*, 48:493–506.

Benjamin, A.

2008. Audit: how to do it in practice. *BMJ*, 336(7655):1241–1245.

Billingham, S., A. Whitehead, and S. Julious

2013. An audit of sample sizes for pilot and feasibility trials being undertaken in the united kingdom registered in the united kingdom clinical research network database. *BMC Medical Research Methodology*, 13(1):104.

Birkett, M. and S. Day

1994. Internal pilot studies for estimating sample size. *Statistics in medicine*, 13(23-24):2455–63.

Bradford Hill, A.

1948. Streptomycin treatment of pulmonary tuberculosis. *BMJ*, 2(4582):769–782.

Chan, A.-W.

2008. Bias, spin, and misreporting: Time for full access to trial protocols and results. *PLoS Med*, 5(11):e230.

Chang, M.

2008. Adaptive design theory and implementation using sas and r. *London: Chapman & Hall.*

Coffey CS, M. K.

2000. Some distributions and their implications for an internal pilot study with an univariate linear model. *Communications in statistics: theory and methods*, 29.

Denne, J. and C. Jennison

2000. A group sequential t-test with updating of sample size. *Biometrika*, 87(1):125–134.

Friede, T. and M. Kieser

2006. Sample size recalculation in internal pilot study designs: A review. *Biometrical Journal*, 48(4):537–555.

Gould, A. L. and W. J. Shih

1992. Sample size re-estimation without unblinding for normally distributed outcomes with unknown variance. *Communications in Statistics - Theory and Methods*, 21(10):2833–2853.

Jennison, C. and B. Turnbull

1999. *Group Sequential Methods with Applications to Clinical Trials*, Chapman & Hall/CRC Interdisciplinary Statistics. CRC Press.

Joanes, D. N. and C. A. Gill

1998. Comparing measures of sample skewness and kurtosis. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 47(1):pp. 183–189.

Julious, S. A.

2004. Designing clinical trials with uncertain estimates of variability. *Pharmaceutical Statistics*, 3(4):261–268.

Kairalla, J.

2007. *An Internal Pilot Study with Interim Analysis for Gaussian Linear Models*. University of North Carolina at Chapel Hill.

Matthews, J.

2006. *Introduction to Randomized Controlled Clinical Trials, Second Edition*, Chapman & Hall/CRC Texts in Statistical Science. CRC Press.

Phillips, A. and V. Haudiquet

2003. Ich e9 guideline 'statistical principles for clinical trials': a case study. *Statistics in Medicine*, 22(1):1–11.

Scott, C. and M. Baker

2007. Overhauling clinical trials. *Nature Biotechnology*, 25:287–292.

Shein-Chung Chow, Jun Shao, H. W.

2008. Shein-chung chow, jun shao, hansheng wang (2008): Sample size calculations in clinical research, 2nd edition. *Statistical Papers*, 52(1):243–244.

Watson, J. and D. Torgerson

2006. Increasing recruitment to randomised trials: a review of randomised controlled trials. *BMC Medical Research Methodology*, 6(1):34.

Wittes, J. and E. Brittain

1990. The role of internal pilot studies in increasing the efficiency of clinical trials. *Statistics in medicine*, 9(1-2):65–72.

Zucker, D., J. Wittes, O. Schabenberger, and E. Brittain

1999. Internal pilot studies ii: comparison of various procedures. *Statistics in medicine*, 18:3453–509.