Adjusting for Bias in the Mean for Primary and Secondary Outcomes when Trials are in Sequence

Joanne C Rothwell[1,2*], Steven A Julious[2] and Cindy L Cooper[3]

[1] Parexel International, Sheffield, UK

[2] Design Trials and Statistics, School of Health and Related Research (ScHARR), University of Sheffield, UK

[3] Sheffield Clinical Trials Unit, ScHARR, University of Sheffield, UK

*Corresponding Author:

Joanne C Rothwell, Parexel, Navigation House, 1 South Quay Drive, Sheffield, S2 5SU

Email: joanne.rothwell@parexel.com

Abstract

When designing a clinical trial, one key aspect of the design is the sample size calculation. The sample size calculation tends to rely on a target or expected difference. The expected difference can be based on the observed data from previous studies, which results in bias. It has been reported that large treatment effects observed in trials are often not replicated in subsequent trials. If these values are used to design subsequent studies, the sample sizes may be biased which results in an unethical study. Regression to the mean (RTM) is one explanation for this. If only health technologies which meet a particular continuation criterion (such as $p < 0.05$ in the first study) are progressed to a second confirmatory trial, it is highly likely that the observed effect in the second trial will be lower than that observed in the first trial.

It will be shown how when moving from one trial to the next, a truncated Normal distribution is inherently imposed on the first study. This results in a lower observed effect size in the second trial.

A simple adjustment method is proposed based on the mathematical properties of the truncated Normal distribution. This adjustment method was confirmed using simulations in R and compared with other previous adjustments. The method can be applied to the observed effect in a trial which is being used in the design of a second confirmatory trial, resulting in a more stable estimate for the "true" treatment effect. The adjustment accounts for the bias in the primary and secondary endpoints in the first trial with the bias being affected by the power of that study. Tables of results have been provided to aid implementation, along with a worked example.

In summary, there is a bias introduced when the point estimate from one trial is used to assist the design of a second trial. It is recommended that any observed point estimates be used with caution and the adjustment method developed in this paper be implemented to significantly reduce this bias.

Key words: regression to the mean, target effect size, effect size, trials in sequence

Introduction

During the design stage of a randomised controlled trial (RCT), the most sensitive parameter in the standard sample size calculation is the target difference, *d*. The most common method used to inform the quantification of *d* is to use an observed difference from a previously conducted study[1].

When designing trials in sequence, such that the second trial will only begin if the first trial is "successful", one must be wary of a bias which is inherently introduced when using the results from one trial to design the next. This bias is known as regression to the mean (RTM) and is introduced by the second trial depending on the result of the first.

The definition of success for the first trial could mean it failing to reach a specified decision point for the outcome of interest, whether this is reaching statistical significance or having an observed effect size within the bounds of a 95% confidence interval, the follow-up trial would be unlikely to be undertaken.

This decision point in the first trial introduces bias into this trial which we will now discuss in the paper. The methods in this paper assume that the true treatment difference is known, and we will consider the implications of this assumption in the discussion.

Background

It has been reported that large treatment effects observed in trials are often not replicated in future trials[2]. For example, when replicating a small trial with a trial with a larger sample size the effect seen in the second trial may decrease in size[3]. Clinical programs where small or early trials report observing very large treatment effects often progress to larger trials or later stage trials. These trials are then considered as "failing"[4] when the main trial does not demonstrate an effect size close to that estimated in the smaller, early phase trial.

There are a number of plausible reasons for this. Regression to the mean has been discussed as being one possible reason[4-6]. Regression to the mean is defined by McCall et al as[7]

"the likelihood that an outcome variable will show a significant change depending upon how much baseline values depart from the mean".

If only the 'promising' trials are taken forward to Phase III from Phase II, the average of the Phase III results will be less than the average of the 'promising' Phase II trials, due to an expected truncated distribution. This is caused by trials in Phase II having to exceed a pre-specified criterion to move to Phase III. The left-truncated Normal distribution results in a higher mean difference for that trial. The average mean difference of the subsequent trial is less than that of the first and the distribution would not be truncated.

With regards to phase II or pilot studies, careful design of the early study could reduce the impact of regression to the mean, however it is unlikely to eradicate the issue completely. The occurrence of regression to the mean in the context of trial design and moving from one trial to the next needs careful management.

There are two circumstances where regression to the mean could occur; those are when multiple measurements are being taken on the same patients (this can be defined as within-study) and when there are similar trials being conducted (this can be defined as between-study).

Regression to the mean occurs, according to Morton and Torgerson, when[8]

"an extreme group is selected from a population based on the measurement of a particular variable."

They comment that if another measurement of the same variable is taken from this same group, the mean of the second measurement will be "closer to the population mean than the first measurement." This definition could be thought of as similar to what happens when moving from a Phase II to Phase III trial, or when using results from a Cochrane review or meta-analysis to design a large publicly funded trial. The first result must be "encouraging" the definition of which will be discussed later, if the second trial is to commence. So, in this situation, the extreme group contains all the trials which are "encouraging" though the results observed will likely decrease towards the population mean at the next trial. This highlights the importance of investigating current methods for adjusting for regression to the mean which could be applied to the case of moving from one trial to the next.

The issue of analyses being done in sequence and the impact on inference also impacts studies for adaptive designs. For a group sequential trial if there is to be an interim analysis the data will be left truncated (if there is an assessment of futility) and right truncated (if there is a stop for efficacy). To obtain an unbiased estimate there is a need to allow for the interim analysis[9]. It has also been recently highlighted how the sequential nature of studies can bias results when undertaking studies in sequence[10].

This paper will highlight when two studies are done in sequence how the statistical power of the first study will impact on the bias in the estimates of effect from that study. It will also extend the work to show how, even when the first study is powered on a different primary endpoint to the second study, and the primary endpoint of the second study is assessed in the first study as a secondary, then it will also be biased. Simple methods are proposed to allow for the adjustment and tables are provided to assist in their application.

Methods of Adjustment for Trials in Sequence

According to Zhang et al., proposed methods of adjusting for overestimation of the treatment effect are not regularly implemented[11]. A search of the literature identified two simple comparable methods which could be used for adjustment for trials in sequence[12,13]. There are more advanced methods of adjustment which are briefly mentioned in the discussion.

Wang et al.[12] proposes a method of adaptation for the sample size calculation when using data from Phase II trials. The context of this adaptation is industry-based (Phase II to Phase III, or early-phase to late-phase) and considers the use of surrogate endpoints in Phase II trials to be one of the causes for a high failure rate of Phase III trials. This could be, for example, the use of tumour shrinkage as the endpoint for the Phase II trial when the primary outcome is survival in the Phase III trial. Another example could be using a 1-month outcome in the early phase trial when the main trial requires a 6-12-month outcome.

Later in the paper we will highlight how designing a study around a surrogate endpoint could mean the results for secondary outcomes – which may be used as the primary endpoint for the next trial – could be biased.

Wang et al. focused on the calculation of the sample size for Phase III trials based on either the point estimate from the Phase II trial or the lower confidence limit. It recommends having a bias adjustment of

$$\hat{\Delta} - 1 \times s.e(\hat{\Delta})$$

where s.e is the standard error and $\hat{\Delta}$ is the point estimate from the Phase II trial. However, this result can lead to very small estimated effect sizes and therefore not many Phase III trials being started. Kirby et al. developed an adjustment method which was tested on the scenarios used by Wang. This method is a multiplicative adjustment[12] which is based on the concept of assurance. The general adjustment is

$$\hat{\Delta} \times 0.9$$

Where $\hat{\Delta}$ is the point estimate from the Phase II trial.

These methods of adjustment can be applied to the context of previous research to inform a new study, such as Phase II to Phase III, or pilot study to confirmatory trial.

Mathematical Development

Methods

As mentioned earlier in this paper, it could be assumed that the initial trial will follow a truncated Normal distribution. The truncation occurs because there is a specific criterion above which the trial will be progressed to the second trial.

Two types of trial designs will be considered in this paper. The first is two trials in sequence for example a Phase 2 to Phase 3 trial. The Phase 2 trial will be deemed successful if $p \geq 0.05$. The second will be a pilot study to main trial sequence. For the pilot study to be deemed successful the criteria will be that $d_{observed} > 0$.

The Truncated Normal Distribution

The truncated Normal distribution is the general Normal distribution bounded by a random variable from either above, below or both. This could occur if there was a floor- or ceiling-effect with the data, for example if trial eligibility criteria included a pre-specified threshold for blood-pressure values, there could be a truncation at that threshold value.

Suppose $X \sim N(\mu, \sigma^2)$ and let Y be a truncated Normal, denoted $TN(\mu, \sigma^2, a, b)$ where (a, b) are restrictions on the domain of X $(-\infty \leq a < b \leq \infty)$[14]. These results are for a two-sided truncated Normal distribution. The probability of X lying within the internal of (a, b) is given by

$$\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right) \tag{1}$$

Where $\Phi(...)$ is the cumulative density function of the standard Normal distribution.

The probability density function of a left-truncated Normal distribution is

$$f\big(y\big|(a,b)\big) = \frac{\frac{1}{\sigma}\phi(\frac{y-\mu}{\sigma})}{1-\Phi\left(\frac{a-\mu}{\sigma}\right)} \tag{2}$$

for $a \leq y$ and $f(y) = 0$ otherwise.

From Equation (1) the probability of X lying within the interval $(a, \ \infty)$ is given by

$$1 - \Phi\left(\frac{a-\mu}{\sigma}\right) \tag{3}$$

Where $\Phi(..)$ is the cumulative distribution function, since $b \rightarrow \infty$.

The expectation for the truncated Normal distribution is given by the following, if we let $E(Y) = \mu^*$ where μ^* is the expectation of the truncated Normal distribution

$$E(Y) = \mu^* = \mu + \sigma\left[\frac{\phi\left(\frac{a-\mu}{\sigma}\right)}{1-\Phi\left(\frac{a-\mu}{\sigma}\right)}\right] \tag{4}$$

As previously mentioned, if only statistically significant studies were taken forward to the second trial, the results would form a left-truncated Normal distribution. Therefore, this will remain the focus for this research.

From (4), if we simplify this such that

$$A = \frac{a-\mu}{\sigma}$$

then this becomes

$$\mu^{*} = \mu + \sigma\left[\frac{\phi(A)}{1-\Phi(A)}\right] \tag{5}$$

Where μ is the expectation or mean of the underlying Normal distribution (the untruncated Normal distribution), and σ is the population standard deviation. It can be observed that $\mu^{*} > \mu$ since

$$\sigma\left[\frac{\phi(A)}{1-\Phi(A)}\right] > 0,$$

so when the distribution is left-truncated, the mean is higher than the standard Normal expectation. This method is similar to the Maximum Likelihood method discussed by Kirby et al.[10]

If one is able to find the truncation point, then it is possible to calculate the mean for the distribution if it were not truncated. This can then be used to assess the bias and determine an adjustment by which one can estimate an unbiased mean for the first trial.

The probability of X lying in the area greater than a becomes

$$P[X > a] = 1 - \Phi\left(\frac{a-\mu}{\sigma}\right) \tag{6}$$

Since $b \to \infty$ this result looks similar to that for the power of a trial $(1-\beta)$. Here, β is the probability of making a Type II error, whilst the power of a trial is the probability of observing a difference if there is truly a difference to be observed (i.e. if the alternative hypothesis is true).

Note it is often the t-distribution which is used for the test statistics and to estimate the power. However, for large sample sizes, the t-distribution tends to the standard Normal distribution.

Results

To derive an adjustment that can be applied to reduce the bias when moving from initial study results on to designing a later phase study, there are a number of steps. Further details of these can be found in Appendix 1.

It has been highlighted that when moving from one study to the next, a truncated Normal distribution is observed for the results of the first study. The truncation point of this distribution is associated with the power of the first study. Thus, the power is linked to the bias observed when estimating the effect size for the second study. This will be expanded upon in the next section.

It can be shown that the distribution of the effect sizes in the first trial is

$$N\left(ES\sqrt{\frac{n}{2}},1\right).$$

Let $E(Y)$ be denoted μ^*, which is the mean of the truncated Normal distribution. Since the truncation point, a, can be calculated using $t_{2n-2,1-\alpha/2}$, and the truncated mean μ^* is known, we can rearrange the equation (5) in terms of the true mean to be

$$\mu = \mu^* - \sigma\frac{\Phi(A)}{1-\Phi(A)}$$

where $A = \frac{a-\mu}{\sigma}$.

Power-based Truncation Point

The purpose of this section is to highlight the association between the truncation point and the power of a study, whereas the previous section highlighted the association between the truncation point and the bias. Therefore, the bias and the power are both directly linked to the truncation point.

The results from the previous section can be used to investigate the truncation point using the Minimum Detectable Difference (MDD). The MDD is the smallest difference which can be statistically detected in a particular study[16].

One intuitive method to calculate the MDD is to set the power to 50% for a sample size calculation, this gives the minimum value that the 95% confidence internal around the point estimate will exclude the null value. Using a standard sample size calculation,

$$n = \frac{2\sigma^2\left(Z_{1-\beta}+Z_{1-\frac{\alpha}{2}}\right)^2}{d^2}$$

With the power set to 50%, the $Z_{1-\beta}$ term becomes equal to 0. If the resulting equation is re-arranged in terms of d, it becomes

$$d_{det} = \sqrt{\frac{2\sigma^2 Z^2_{1-\alpha/2}}{d^2}}$$

If the P-value observed in the trial is less than the (two-sided) statistical significance of 0.05, and the sample size is achieved, then the ratio of the detectable difference to the target difference can be calculated for various powers by

$$\frac{d_{det}}{d}$$

This will provide an adjustment value for the target difference, d, showing the minimum detectable difference. These results can be seen in Table 1.

[Insert Table 1 here]

If a trial has a power of 80% and a P-value of 0.05, the detectable difference would be $0.7d$, where d is the target difference. These values are closely associated with the truncation point, since the truncation point is the value at which the p-values become significant, and the detectable difference is the proportion of d which will observe a P-value of 0.05.

Results of *a priori* truncation point calculation

This section compares the results of a series of simulations, where the bias is quantified and observed, with the results following the implementation of the adjustment derived in Table 1. The adjustment outlined in Table 1 can be used in the design of subsequent trials based on current or previously observed results.

Each variable in the results Table 2 and Table 3 are described here. The values for μ are those of the true effect sizes and μ^* show the biased mean estimate on the standardised scale.

- Sample size per arm (n) is set to pre-specified values shown in the tables.
- μ is the mean difference based on the non-central t-distribution. Thus $\mu = ES$ where ES is the effect size
- a_{det} is calculated by $x.\mu$, where x is the associated value from Table 1.
- a is the mathematical truncation point calculated as $a = t_{2n-2,1-\alpha/2}$

The difference between μ^* and μ quantifies the level of bias which occurs from the truncation of the underlying distribution, with μ^* being the mean of the truncated distribution (the observed mean) and μ being the mean of the non-truncated distribution (the 'true' mean).

A series of simulations were performed, based on the concept of trials in sequence. There were 10,000 pairs of trials run in sequence, following a Normal distribution with varying "true" mean and a fixed standard deviation. The sample sizes were calculated then the results of the simulations represented the individual patient observed values. The average was taken to be μ^*.

[Insert Table 2 here]

[Insert Table 3 here]

Pilot study to Main Trial

This section presents an extension of the previously described adjustment method to the case of a pilot study to a main trial. The adjustment has been altered slightly to assume the "success" of a pilot study being that there is a positive treatment effect.

For a pilot study to main trial scenario, we are considering only pilot studies where the observed treatment effect is positive. The same adjustment as above would be developed by setting the truncation point, a, equal to zero. The results are given in Table 4.

[Insert Table 4 here]

It can be observed that the ratio of mean differences observed from the simulations are similar to the ratio of mathematical mean differences developed here.

Some assumptions have been made for the context of pilot study to main trial, the first being that the sample size calculation for the pilot study, which is capped at a minimum of 10 participants per arm. The number of subjects has been derived from the results presented by Whitehead et al.[17]. The second assumption is that we are considering only the context of having a positive $(d > 0)$ result in the pilot study resulting in a confirmatory study.

However, these results for the mathematical solution depend on the conditions set moving forward with the main trial. Since these results are all based on only observing a positive effect size in the pilot study, these results can be extended to any pilot trial which has a decision point before starting the next trial.

Development of Adjustment Method

Up to this point, we have highlighted the bias in the first trial when two trials are done in sequence. By being able to quantify this bias from a known cut-point, it is possible to adjust for it by extending the results from Table 2 to Table 4 by using the bias expected for the decision point as an adjustment.

Adjustments that can be applied to the results of trial 1 (T1) is presented in Table 5 (for powered studies) and Table 6 (for pilot studies).

The values presented in Table 5 are the adjustment values which the observed trial 1 effect size should be multiplied by to get the adjusted effect size for trial 1. This gives a more unbiased estimate of the 'true' effect size. It can be seen that the adjustment for constant power is relatively stable irrespective of varying observed effect size.

[Insert Table 5 here]

The adjustment values developed for pilot study to main trial scenarios are presented in Table 6. These are dependent on the continuation criteria imposed on the pilot study, for example this could be that the observed effect size was positive $(d_{pilot} > 0)$ or that the 95% confidence interval contains a pre-specified minimum clinically important difference.

[Insert Table 6 here]

Powering on a Surrogate or Secondary Endpoint

So far in the paper, the scenario under discussion has been when two studies being done in sequence are having decisions made based on the same endpoint. Often, however, the primary endpoint for the first study will be different to that in the second study. Thus, when undertaking a preliminary study, it will not be powered on the primary outcome for the definitive study but a surrogate or other outcome such as the primary outcome but at an early timepoint. The outcome for the definitive study may also be still assessed in the trial but is a secondary outcome.

For a primary outcome, let μ_1 be the effect from the underlying Normal distribution and μ_1^* be the effect from the truncated Normal distribution. The estimate of the mean effect for the secondary outcome would be obtained from[16]

$$\mu_2 = \mu_2^* - \rho\frac{\sigma_2}{\sigma_1}(\mu_1^* - \mu_1)$$

Where σ_1 and σ_2 are the standard deviations for the primary and secondary outcomes and ρ is the pooled correlation coefficient between the primary and secondary outcome. Thus, if there is a bias in the primary outcome there will also be bias in the secondary outcome – which in the context of this paper will be the primary outcome for the study being planned.

Table 7 extends the results in Table 5 and Table 6 by giving the ratio of effects for the secondary outcomes between studies run in sequence assuming that $\sigma_1 = \sigma_2$. This latter assumption may be reasonable if the primary outcome in the pilot trial is an early time point for the outcome in the definitive trial.

[Insert Table 7 here]

The work in Table 7 could also be extended to pilot trials. If the correlation between the primary and secondary outcomes are known, and the bias in the primary outcome is known then this approach can be used to adjust to get an unbiased estimate for the primary outcome.

In practice, the correlation between the primary and secondary endpoints and the bias are not always known, however these could be estimated from results from other studies with similar target populations. Alternatively, a sensitivity assessment could be performed so a range of feasible correlation assumptions are evaluated and the impact on the adjusted effect size can be considered.

Comparison with other methods

In order to assess the developed adjustment, it will be compared with the Wang and Kirby adjustments discussed earlier in the paper. The adjustments are compared and presented in Table 8.

The adjustments have been applied to the simulation results, which were outlined in an earlier section, and which were produced to confirm the mathematical methodology. A series of simulated parallel, two-arm trials were produced with varying sample size, power and target effect size. Recall that there were 10,000 pairs of trials run in sequence, following a Normal distribution with varying "true" mean and a fixed standard deviation. From Table 8, the conditions for the simulations can be gathered. For the evaluation of a varying power and constant effect size, the power is increased in 5% increments whilst the effect size is held constant at 0.2. This was based on a mean of 10, and a standard deviation of 50. This resulted in the sample size increasing as the power increased.

For the evaluation of varying the effect size whilst power remained constant, the power was set to 80% and the effect size was increased from 0.2 to 0.8, whilst keeping the standard deviation at 50. This was achieved by increasing the mean from 10 to 40 in stages.

The results for these simulations are extensive and have not been included. Figure 1 shows the comparison of the adjusted estimates with the "true" treatment effect. It can be observed that there is some variation between the adjusted values, the Maximum Likelihood adjustment seems the most stable to the "true" mean difference. Figure 2 shows the comparison of the various adjustments applied to simulation data with varying effect size.

[Insert Table 8 here]

[Insert Figure 1 here.]

Figure 1 shows how the various adjustment methods behave in relation to the true mean difference (10) for varying powers. As the power increases, the unadjusted mean difference gets closer to the true mean difference, mirroring the results shown earlier in the paper. The Maximum Likelihood adjustment is the most stable across all powers, with only minor fluctuations due to the random nature of simulations. It is interesting to note that as the power increases the Wang adjustment tends towards the true mean whilst the Kirby adjustment, which is a constant value tends away from the true mean difference.

[Insert Figure 2 here.]

Figure 2 shows a comparison of the adjustments with the true effect size for 80% power. We can see that both Maximum Likelihood and Kirby adjustments follow the true mean difference very closely since the Kirby adjustment would multiply the observed effect size by 0.90, whereas the Maximum Likelihood adjustment would multiply the observed by 0.89. The unadjusted line shows the effect that regression to the mean has on the observed result.

Worked Example

A new study is being planned which will follow a first trial where a statistically significant result was the observed. The wish is to use the effect from the observed study to assist in the design of the new study. Now suppose the first study was designed with 80% power and the observed treatment effect for the primary outcome was 19.1. If we apply the adjustment from Table 5, we get the following result

$$19.1 \times 0.89 = 16.98$$

Therefore, when designing the second study, we should use 16.98 as the estimate of effect from the first study to assist in the quantification of the effect size.

Now suppose the first study was a preliminary Phase II trial and it was powered on an early time point. The later time point was also assessed and the effect of 16.98 was the observed treatment difference. If we assume the primary and secondary outcomes have the same variance, and that the correlation between them was ρ =0.80 we can then make use adjustments in Table 7. The adjustment in Table 7 which corresponds to a correlation of 0.80 and the adjustment from Table 5 (0.89) is 0.91.

Thus, if we apply the adjustment from Table 7, we get the following result to account for bias,

$$19.1 \times 0.91 = 17.38$$

Thus, assuming that there is a correlation of 0.8 between the primary endpoint and the secondary endpoints in the first study we should use 17.38 to guide us in quantifying the target estimated effect size for the main trial.

Discussion

The implication of the results in the paper is that if a point estimate from a trial is being used to assist in the design of a confirmatory trial then the observed effect should be interpreted with caution as the estimate. Even if the primary outcome for the confirmatory trial is a secondary endpoint in the first trial then the observed effects should be interpreted with care.

This paper has demonstrated how inherent bias is introduced when using a point estimate from one trial to design the sample size for the confirmatory trial. The reason the first trial is biased is that only results which are supportive (evidence of effect) would result in a secondary confirmatory trial.

The implication of this is if the point estimate from the first trial is then used as an effect size for the main trial, it will be overestimating the effect for the planned study. If the observed effect is being used to assist the quantification of an effect – complementing subjective clinical opinion for example – then the issue of bias needs to be considered but the consequences are less.

There have been simple comparable adjustments described previously for this regression to the mean effect when moving from one trial to the next. The other adjustments mentioned in this manuscript either apply a standard “rule-of-thumb” method, or are considering combining study results (study 1, study 2) and comparing the combined observed with an expected difference. The focus of the work documented in this paper is the design of the second study based on the results of the first study. This has been expanded to a range of design scenarios.

The simple adjustment approach described in this paper compares well with these previous adjustments and is more stable for various target effect sizes. To assist researchers, tables have also been provided which adjust for the bias in the first study, depending on the power of the first study and the power of the second study (and the correlation between outcomes if the primary outcome in the main is being used as a secondary outcome in the first.)

The results in the paper were highlighted using continuous data assumed to take a Normal form. However, they can be extended to other distributional forms if the test statistic has a Normal approximation.

The work in the paper assumes that the true treatment difference is assumed known, it does not cover the scenario where the true treatment difference is unknown. This could be considered a limitation. However, point estimate and a 95% confidence interval will contain the true treatment difference 95% of times if the estimates are unbiased. If the estimates are biased, this will not be the case. We have shown when the estimates from the first study done in sequence will be biased and will need an adjustment, this ensures the 95% confidence interval has an appropriate coverage.

A limitation of the approach we are proposing is that it is a simple one. Alternative approaches could be considered extending the work of group sequential trials[19] or sequential meta analyses[20]. Kirby et al also describes alternative methods[10,13]. The methods in this paper could be used to guide researchers on plausible effects to assist in the decision of a future study. Especially if this quantification is complemented by alternative approaches such as clinical judgement. Further recent work has investigated the regression to the mean when moving from small proof-of-concept studies to larger trials, as we have discussed in this paper, and considered adjustments for selection bias from a frequentist and Bayesian perspective[21].

The results could also be extended to when a systematic review has been used as a justification to undertake a definitive trial. If the consequent meta-analysis suggests further work should be undertaken, then the results of this meta-analysis should be interpreted with caution when designing the definitive trial. Extending the results from this work applied to meta-analyses being used to assist in the design of future trials needs investigating further.

Conclusion

It has been shown in this paper that there is a bias when the point estimate from a first trial is being used in the design of a second trial, even if the outcome of interest for the main trial is a secondary outcome in the initial trial. The recommendation from this work is to implement the adjustment method proposed in this paper. Tables are provided to assist researchers to apply the results in practice.

Acknowledgements

We wish to thank the University of Sheffield Clinical Trials Research Unit, and the Medical Statistics Group in ScHARR for their invaluable advice and knowledge during the development of this research.

Work from this project was presented at the annual PSI conference in 2017.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

1. Rothwell, J. C., Julious, S. A., Cooper, C. L. A study of target effect sizes in randomised controlled trails published in the Health Technology Assessment journal. Trials, 19 (544), 2018.
2. Pereira, T. V., Horwitz, R. I., & Ioannidis, J. P. A. (2012). Empirical Evaluation of Very Large Treatment Effects of Medical Interventions. JAMA, 308(16), 1676-1684. doi:10.1001/jama.2012.13444
3. Ioannidis, P. A. J. (2008). Why Most Discovered True Associations Are Inflated. Epidemiology, 19(5), 640-648. doi:10.1097/EDE.0b013e31818131e7
4. Krum, H., & Tonkin, A. (2003). Why do phase III trials of promising heart failure drugs often fail? The contribution of "regression to the truth". In J Card Fail (Vol. 9, pp. 364-367). United States.
5. Chuang-Stein, C., & Kirby, S. (2014). The shrinking or disappearing observed treatment effect. Pharm Stat, 13(5), 277-280. doi:10.1002/pst.1633
6. Julious, S. A. (2010). Sample sizes for clinical trials [electronic resource]. Boca Raton: Boca Raton : CRC Press/Taylor & Francis, c2010.
7. McCall, W. V., D'Agostino, R., Jr., Rosenquist, P. B., Kimball, J., Boggs, N., Lasater, B., & Blocker, J. (2011). Dissection of the factors driving the placebo effect in hypnotic treatment of depressed insomniacs. Sleep Med, 12(6), 557-564. doi:10.1016/j.sleep.2011.03.008
8. Morton, V., & Torgerson, D. (2005). Regression to the mean: treatment effect without the intervention. Journal Of Evaluation In Clinical Practice, 11(1), 59-65.
9. Whitehead J. Group sequential trials revisited: Simple implementation using SAS. Statistical Methods in Medical Research 2011 20(6): 635-56 https://doi.org/10.1177/0962280210379036
10. Kirby S, Li J, Chuang-Stein C. Selection bias for treatments with positive Phase 2 results. Pharm Stat. 2020 Sep;19(5):679-691. doi: 10.1002/pst.2024. Epub 2020 Apr 14. PMID: 32291941.
11. Zhang, J. J., He, K., Tang, S., Sridhara, R., Blumenthal, G. M., & Cortazar, P. (2012). Overestimation of the effect size in group sequential trials. Clinical Cancer Research, 18(18), 4872-4876. doi:10.1158/1078-0432.CCR-11-3118
12. Wang, S. J., Hung, H. M. J., & Neill, R. T. (2006). Adapting the sample size planning of a phase III trial based on phase II data. Pharmaceutical Statistics, 5(2), 85-97. doi:10.1002/pst.217
13. Kirby, S., Burke, J., Chuang-Stein, C., & Sin, C. (2012). Discounting phase 2 results when planning phase 3 clinical trials. Pharmaceutical Statistics, 11(5), 373-385. doi:10.1002/pst.1521
14. Johnson, A. and Thomopoulos, N. (2002). Use of the left-truncated normal distribution for improving achieved service levels. In Proceedings of the 2002 Annual Meeting of the Decision Sciences Institute, pages 2033{2041.
15. Senn, S. (1993). Cross-over trials in clinical research. Chichester : Wiley, c1993, Chichester.
16. Cook, J. A., Hislop, J., Adewuyi, T. E., Harrild, K., Altman, D. G., Ramsay, C. R., Fraser, C., Buckley, B., Fayers, P., Harvey, I., Briggs, A. H., Norrie, J. D., Fergusson, D., Ford, I., and Vale, L. D. (2014). Assessing methods to specify the target difference for a

randomised controlled trial: Delta (difference elicitation in trials) review. Health technology assessment (Winchester, England), 18(28):v{vi,1-175}.

17. Whitehead, A. L., Julious, S. A., Cooper, C. L., Campbell, M. J. (2016) Estimating the sample size for a pilot randomised trial to minimise the overall trial sample size for the external pilot and main trial for a continuous outcome variable. Statistical Methods in Medical Research. 2016 Jun;25(3):1057-73. doi: 10.1177/0962280215588241

18. Whitehead, J. (1986) Supplementary analysis at the conclusion of a sequential clinical trial. Biometrics pp. 461-471.

19. Jennison, C., & Turnbull, B. W. (1999). Group sequential methods with applications to clinical trials. CRC Press.

20. Whitehead, J. (1997). *The design and analysis of sequential clinical trials*. John Wiley & Sons.

21. Qu Y, Du Y, Zhang Y, Shen L. Understanding and adjusting for the selection bias from a proof-of-concept study to a more confirmatory study. Statistics in Medicine. 2020 Dec 30;39(30):4593-604.

Appendix 1

This Appendix contains further details on the derivation of the adjustment described in this paper.

Finding the Truncation Point

Based on the results discussed up to this point, we can deduce that it is possible to calculate the truncation point using the results from the distribution.

Consider many t-test results, forming a t-distribution. The results for a t-distribution of t-statistics is given by

$$t = \frac{\left(\bar{x}_1 - \bar{x}_2\right) - \left(\mu_1 - \mu_2\right)}{s/\sqrt{n}}$$

We can define $d = \bar{x}_1 - \bar{x}_2$. Under the null hypothesis, $\mu_1 - \mu_2 = 0$ so the equation becomes

$$t = \frac{d}{s/\sqrt{n}}$$

If the number of subjects in each group can be assumed to be equal, the degrees of freedom are $2n-2$. Therefore the truncation point can be given by the proportion of trials excluded due to having $P \geq 0.05$ corresponding to the value $t_{2n-2,1-\alpha/2}$. Therefore, the truncation point for small samples ($n < 30$) could be calculated by taking the inverse cumulative density function of a t-distribution with mean d, standard deviation s and $2n-2$ degrees of freedom.

As previously discussed, for large sample sizes ($n \geq 30$), the t-distribution tends to a standard Normal distribution. The adjustments proposed in this section are based on the Normal distribution but could be generalised.

Using the standard result for a non-central t-distribution, the power of a trial can be calculated as

$$1\text{-}\beta = 1\text{-}T^{\text{-}1}\left(t_{1\text{-}\frac{\alpha}{2}, n_A(r+1)\text{-}2}, n_A(r+1)\text{-}2, \sqrt{\frac{rn_A d_S^2}{(r_1)\sigma^2}}\right)$$

Where $T^{\text{-}1}(\dots)$ is the cumulative density function of a non-central t-distribution with non-centrality parameter[15]

$$\sqrt{\frac{rn_A}{r+1}}.$$

In this paper the focus is on two-arm trials with $r = 1$, therefore the non-centrality parameter becomes

$$\sqrt{\frac{\mathrm{n}\mathrm{d}_{\mathrm{S}}^{2}}{2\sigma^{2}}}.$$

It can be observed that

$$\sqrt{\frac{\mathrm{d}_{\mathrm{S}}^{2}}{\sigma^{2}}}$$

is the standardised effect size, denoted ES. Therefore, the non-centrality parameter becomes

$$\mathrm{ES} \times \sqrt{\frac{\mathrm{n}}{2}}$$

The distribution of the effect sizes multiplies by

$$\sqrt{\frac{\mathrm{n}}{2}}$$

to give a Normal distribution

$$\mathrm{N}\left(\mathrm{ES}\sqrt{\frac{\mathrm{n}}{2}}, 1\right).$$

Let $\mathrm{E(Y)}$ be denoted μ^{*}, which is the mean of the truncated Normal distribution. Since the truncation point, a, can be calculated using $\mathrm{t}_{2\mathrm{n}\text{-}2,1\text{-}\alpha/2}$, and the truncated mean μ^{*} is known, we can rearrange the equation (5) in terms of the true mean to be

$$\mu = \mu^{*}\text{-}\sigma\frac{\Phi(\mathrm{A})}{1\text{-}\Phi(\mathrm{A})}$$

where $\mathrm{A} = \frac{\mathrm{a}\text{-}\mu}{\sigma}$.

Table 1 The adjustment values for the detectable difference. Note: x is the value by which the target difference, d, should be multiplied.

Power	**X**
80	0.700
81	0.691
82	0.682
83	0.673
84	0.663
85	0.654
86	0.645
87	0.635
88	0.625
89	0.615
90	0.605
95	0.544
99	0.457

Table 2 Comparison of mathematically calculated truncation points and ratios of mean differences with simulated values for various powers, having multiplied the ES by $\sqrt{2/n}$

Trials in sequence							
		Truncation Point		Mean difference		Ratio	
Power	Sample Size (n)	a_{det}	a	μ	μ^*	$\frac{\mu}{\mu^*}$	Simulations $\frac{\mu}{\mu^*}$
80	393	1.962	1.963	0.2	0.225	0.889	0.885
85	450	1.962	1.963	0.2	0.218	0.916	0.926
90	526	1.962	1.962	0.2	0.212	0.943	0.945
95	651	1.963	1.962	0.2	0.206	0.971	0.970
99	950	1.960	1.961	0.2	0.201	0.994	0.994

Table 3 Comparison of mathematically derived truncation points and the associated ratio of mean differences with the simulated values for various effect sizes having multiplied the ES by $\sqrt{2/n}$

Trials in sequence (80% Power)							
		Truncation Point		Mean difference		Ratio	
Effect	Sample Size (n)	a_{det}	a	μ	μ^*	$\frac{\mu}{\mu^*}$	Simulations $\frac{\mu}{\mu^*}$
0.2	393	1.962	1.963	0.2	0.225	0.889	0.886
0.3	175	1.962	1.963	0.3	0.338	0.889	0.891
0.4	99	1.962	1.963	0.4	0.450	0.889	0.884
0.5	64	1.962	1.963	0.5	0.561	0.891	0.880
0.6	45	1.962	1.963	0.6	0.672	0.892	0.892
0.8	26	1.962	1.963	0.8	0.893	0.896	0.896

Table 4 Comparison of mathematically calculated truncation points and ratios of mean differences with simulated values for various effect sizes for pilot study to main trial.

Pilot to Main trial						
		Truncation Point	Mean difference		Ratio	
Effect	Sample Size (n)	a	μ	μ^*	$\frac{\mu}{\mu^*}$	Simulations $\frac{\mu}{\mu^*}$
0.2	20	0	0.2	0.340	0.588	0.591
0.3	14	0	0.3	0.440	0.682	0.682
0.4	11	0	0.4	0.533	0.751	0.745
0.5	10	0	0.5	0.610	0.820	0.823
0.6	10	0	0.6	0.680	0.883	0.881
0.8	10	0	0.8	0.837	0.955	0.959

Table 5 The adjustment for trials in sequence for various powers and effect sizes. Note: x is the value by which the observed difference d_{T1} should be multiplied.

Trials in Sequence	
Effect Size=0.2	
Power	**Adjustment (x)**
80	0.89
85	0.92
90	0.94
95	0.97
99	0.99
Power = 80%	
Effect	**Adjustment (x)**
0.2	0.89
0.3	0.89
0.4	0.89
0.5	0.89
0.6	0.89
0.8	0.90

Table 6 The adjustment for pilot study to main trial designs. Note: x is the value by which the observed difference d_{pilot} should be multiplied.

Pilot study to main trial	
Effect	**Adjustment (x)**
0.2	0.59
0.3	0.68
0.4	0.74
0.5	0.82
0.6	0.88
0.8	0.96

Table 7 Ratio of Point Estimates for the Secondary Outcomes from Studies Run in Sequence

	Bias in the primary outcome								
	Extending Table 6 Adjustments				Extending Table 5 Adjustments				
ρ	0.59	0.68	0.74	0.82	0.89	0.92	0.94	0.97	0.99
0.1	0.96	0.97	0.97	0.98	0.99	0.99	0.99	1.00	1.00
0.2	0.92	0.94	0.95	0.96	0.98	0.98	0.99	0.99	1.00
0.3	0.88	090	0.92	0.95	0.97	0.98	0.98	0.99	1.00
0.4	0.84	0.87	0.90	0.93	0.96	0.97	0.98	0.99	1.00
0.5	0.80	0.84	0.87	0.91	0.95	0.96	0.97	0.99	1.00
0.6	0.75	0.81	0.84	0.89	0.93	0.95	0.96	0.98	0.99
0.7	0.71	0.78	0.82	0.87	0.92	0.94	0.96	0.98	0.99
0.8	0.67	0.74	0.79	0.86	0.91	0.94	0.95	0.98	0.99
0.9	0.63	0.71	0.77	0.84	0.90	0.93	0.95	0.97	0.99

Table 8 All adjustment methods results for trials in sequence, by power. Note: x is the value by which the observed difference d_{T1} should be multiplied. The value y should be subtracted from the observed difference.

Trials in Sequence

Varying Power, Constant Effect Size (0.2)

Power	**X (Maximum Likelihood)**	**X (Kirby)**	**Y (Wang)**
80	0.89	0.90	0.139
85	0.92	0.90	0.126
90	0.94	0.90	0.113
95	0.97	0.90	0.097
99	0.99	0.90	0.074

Varying Effect Size, Constant Power (80%)

Effect Size	**X (Maximum Likelihood)**	**X (Kirby)**	**Y (Wang)**
0.2	0.89	0.90	0.139
0.3	0.89	0.90	0.309
0.4	0.89	0.90	0.551
0.5	0.89	0.90	0.853
0.6	0.89	0.90	1.207
0.8	0.90	0.90	2.093

Figure 1

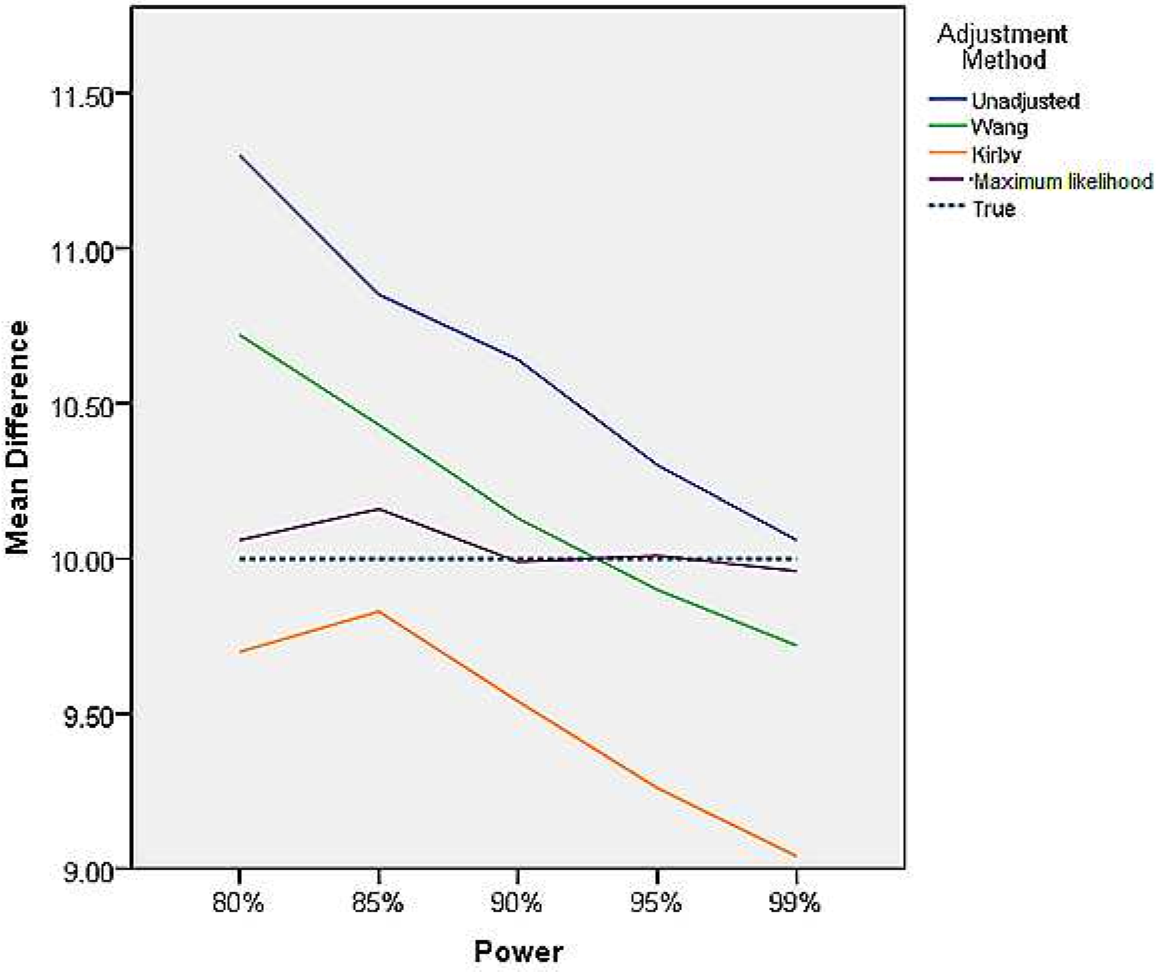
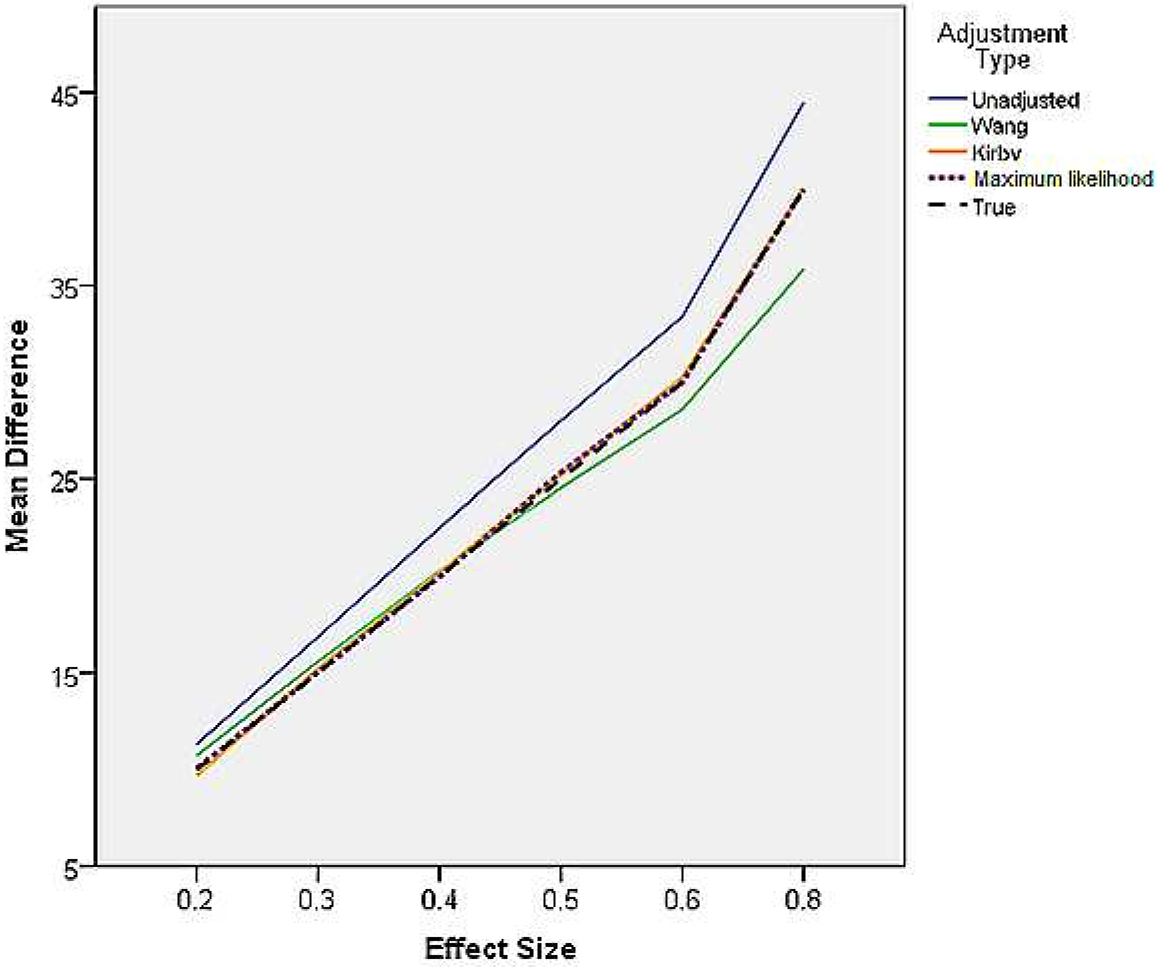

Adjustment Method
Unadjusted
Wang
Kirby
Maximum likelihood
True
Mean Difference
11.50
11.00
10.50
10.00
9.50
9.00
80%
85%
90%
95%
99%
Power

Figure 2


Adjustment Type
Unadjusted
Wang
Kirby
Maximum likelihood
True
Mean Difference
45
35
25
15
5
0.2
0.3
0.4
0.5
0.6
0.8
Effect Size