Noname manuscript No.
(will be inserted by the editor)

Unequal sample sizes and the use of larger control groups pertaining to power of a study

Dr Marie Oldfield

**DSTLTR92592 P2PP2R-
2016-02-23T13 (UK Ministry of
Defence)**

the date of receipt and acceptance should be inserted later

Abstract To date researchers planning experiments have always lived by the mantra that 'using equal sample sizes gives the best results' and although unequal groups are also used in experimentation, it is not the preferred method of many and indeed actively discouraged in literature. However, during live study planning there are other considerations that we must take into account such as availability of study participants, statistical power and, indeed, the cost of the study. These can all make allocating equal sample sizes difficult, and sometimes near impossible. This, some might say, means that the study would not adhere to rigorous statistical standard (Rosenbaum and Rubin, 1985). However, here we present evidence that, not only is this a false assumption, but that we may actually gain more power in the study by actually using unequal groups. Here, data from a Sepsis Biomarker study is used, in which the aim is to predict, by biomarker level and presence, whether the patient would go on to develop sepsis. It was found that larger control groups may give more power to studies looking for an effect in the mid range but not for large or small effects. This study shows merit in the hypothesis that more power can be achieved when a larger control group is used.

Keywords Unequal sample sizes · Power of a study · Sepsis · Biological · Medical · Biomarker

1 Introduction

It is often difficult in observational clinical studies to achieve the numbers of treatment samples needed (Sharma *et al.*, 2019; Guo and Luh, 2013; Dibao-Dina *et al.*, 2014). This then impacts the power of the analysis making it difficult to detect a scientifically meaningful difference of interest (Dibao-Dina *et al.*, 2014).

Address(es) of author(s) should be given

The rational behind this study was that the rarity of clinical subjects resulted in difficulties obtaining a sample large enough to conduct biomarker profiling. Therefore, the methods in this paper were examined in order to find a reasonable way forwards with clinical experimentation on rare subjects.

The power of a test is defined as $1 - \beta = P(\text{accepting H1 given H1 true})$; this is also called the sensitivity (Lan and Lian, 2010). To be able to detect a minimum difference (for example, between two means) with a given significance level, it is therefore necessary to have a sufficiently powered test. In reality, this means that the sample size must be sufficiently large, as noted above. Given a specified significance level (often 5%) and power (often 80%) it is possible to calculate the sample size required to allow detection of a stated minimum significant difference. Power is the probability of rejecting a false null hypothesis and is equal to $1 - \beta$. β is the probability of a type-II [1] error, which occurs when a false null hypothesis is not rejected. In this procedure, a type-II error occurs when you fail to reject the null hypothesis of nonequivalent means when in fact the means are equivalent.

The current approach of using equal groups can be problematic from a cost and study design perspective, i.e. being able to afford enough subjects or even being able to obtain enough subjects. Cohen (1988) noted that "One does not ordinarily plan to use samples of unequal size (since equal sample sizes are optimal) but unequal n's can occur in planning when... one sample's size is necessarily fixed by circumstances, so that the researchers freedom in setting sample size is restricted to only one of the two samples". As stated by Hsu (1993) "... the availability of a very large second sample may not compensate for a constraint in the size of the first sample". It is possible though, that a larger control group may compensate somewhat for the smaller size of the treatment group. This type of problem is particularly relevant to medical studies, especially when only a small percentage may go on to develop a condition in question. If a larger control group could be used, it could have implications on cost and design of studies and this could be exploited in clinical studies. This leads to discussion of ethics in relation to this type of study.

This paper examines the hypothesis that, in some cases, the use of a large control group may provide a study with a greater power than equal sample sizes could. The assumption here is that sample sizes and, indeed treatment and control ratios may be varied for better performance and that increasing the control group gives more power (Tichy and Chytry, 2006). This way forward is supported by Guo and Luh (2013) in the following statement: "Equal allocation design is popular because of convenience and efficiency, but it is not

[1] Type I Error: In a hypothesis test, a Type I error occurs when the null hypothesis is rejected when it is in fact true; that is, H_0 is wrongly rejected. A Type I error is often considered to be more serious, and therefore more important to avoid, than a Type II error. The hypothesis test procedure is therefore adjusted so that there is a guaranteed 'low' probability of rejecting the null hypothesis wrongly; this probability is never 0. A Type I error can also be referred to as an error of the first kind.
Type II Error: A Type II error occurs when the null hypothesis H_0, is not rejected when it is in fact false. This is frequently due to sample sizes not being large enough to identify the falseness of the null hypothesis (especially if the truth is very close to hypothesis).

practical". In the next two sub sections we discuss areas where unequal sample sizes have been more appropriate within studies and why this is the case.

The rest of the paper is organised as follows. Firstly, an overview of the area in which this paper sits is presented. Experimentation is then conducted to evaluate the claim of this paper. The findings are presented and a discussion is conducted around recent work in this area. This paper then concludes with a discussion about limitations and possible extensions of this work.

1.1 Unequal randomisation in studies

Dibao-Dina *et al.* (2014) conducted a study to discover what the most important reasons were for unequal randomisation, i.e. when there are unequal numbers of patients in each group. This study went further to discuss whether sample size calculations had been performed and whether they were clear enough to be understood. This was examined in the area of Medical Research.

Dibao-Dina *et al.* (2014) states that statistical power is "usually maximal with equal groups". However, this is not substantiated with analysis or reference. Due to the nature of medical trials, unequal randomisation could lead to more patients being allocated to the experiment rather than the control group. This could influence the response in the placebo group. This response may then appear exaggerated in comparison to an equal balanced design and produce bias in the effect estimate. This can be caused by psychological factors where patients were conscious that they were more likely to receive an active ingredient than a placebo. A major problem with unequal randomisation can be that that the experimental group is larger than the control. What happens in the opposite circumstance where it is not possible to have a large treatment group?

Dibao-Dina *et al.* (2014) states that the most common justification for unequal randomisation in the review by (Dumville *et al.*, 2006) was "gaining experience with treatment". As safety issues (i.e. side effects of the drug such as adverse events, withdrawals due to adverse events and severity data) are relatively rare this is another reason why unequal randomisation may be chosen. Industry was the main sponsor of many of the studies cited in this paper and 12% of the studies were related to infectious diseases. Of 106 reports from the medical area: 29.2% did not give a sample size calculation and in a further 4.7% the calculation was unclear. Among the 70 reports for which a sample size calculation was reported, unequal randomisation was not reported to have been taken into account in 18 (25.7%) and in a further 4, (5.7%) the calculation was unclear. In the 70 reports for which a sample size calculation was reported , the authors explicitly stated that they considered only equal size groups for the sample size calculation. 77.4% did not report any justification for using unequal randomisation (Dibao-Dina *et al.*, 2014) This is an extremely worrying set of statistics – especially in the field of medicine. This could point to a lack of understanding of study planning or, indeed, that older statistical work has been taken on face value and applied without the assumptions having been examined. Below are listed the justifications given where unequal randomisation was used.

In the 24 of 106 reports that justified their use of unequal randomisation, the major justifications were (Dibao-Dina *et al.*, 2014):

- Provide safety data (48.8%)
- Patient acceptability (26.8%)
- Only 4 trials had more patients in the control group and the justifications were cost (2 studies) and patient acceptability (1 study); one study had no justification for their design.

Dibao-Dina *et al.* (2014) Among the reports for which obtaining safety data was a justification for using unequal randomisation, four reports did not report on adverse events and seven did not describe one or more adverse events, severity data or withdrawals due to adverse events.

Unequal randomisation occurs here mostly when there appears to be a bias towards including more people in the treatment group, possibly in the hope that more treatment samples may give a "more robust result or make the treatment seem more effective" (Lan and Lian, 2010). It does not appear that the basic assumptions of power and sample sizes have been looked at in the context of an overall experiment, where any group number could be changed to give more power. It appears that that the treatment group can be over inflated in a bid to attain the required number of subjects to determine whether a drug is safe or to include more people into the treatment group to influence patient acceptability. As stated in Dibao-Dina *et al.* (2014) the patients were "aware of a greater probability of receiving an active treatment than a placebo". These are very different aims from the ones purported to have been tested in the studies themselves. Had the study set out a primary aim of testing a drug for patient safety, that would be a completely different study plan to one which tested the efficacy of a drug. Here it could be seen that two aims are being merged into one but this could lead to a waste of resources when two mini studies would be more cost effective and provide more useful results (Dibao-Dina *et al.*, 2014).

1.2 The impact of cost on group sizing

"When conducting research with controlled experiments, sample size is one of the most important decisions that researchers have to make" (Guo and Luh, 2013).

Allocation of sample size in the case of maxmin[2] theory (Guo and Luh, 2013) is not well explored in literature. The maxmin theory examines the financial cost of subjects for a study and allocates the groups in an optimal manner according to cost as a constraint. When financial costs and weightings of samples come into play, analysis is needed that takes this into account

[2] in this context, this would be maximizing the power whilst minimizing the cost or group sizes.

when allocating samples. For example, if a particular subject is rare Pisano *et al.* (1998), then this may be associated with a high financial cost. When cost is not considered and adequate preparation and analysis is not undertaken, it can be seen that “insufficient or excessive sample sizes” will result (Guo and Luh, 2013). As noted in previous sections, some studies can “result in less power or greater cost” (Levin, 1997). Within experimentation one would prefer the sample size to provide maximal precision with minimal resources. Also, groups with expensive treatments or rarity of subjects, and/or patients, would inherently contain fewer subjects than the control. Lan and Lian (2010) discuss using affordable groups but do not enter into analysis on cost of subjects/patients.

Guo and Luh (2013) again make the point that existing analysis concerns homogenous groups and raise the point that groups with heterogeneous variances need to be analysed also. When costs are included in an experiment then equal allocation may no longer be feasible. In the case of medical trials this can certainly be the case. Four scenarios were considered in Guo and Luh (2013):

- Fixed research budget
- Fixed statistical power
- Fixed total sample size
- Uneven incremental costs

For costs and/or weightings to be applied a maxmin concept (concerning the minimisation of cost and maximisation of group sizes, or subjects) was used in order to maximise the power whilst minimising the cost. The resulting allocation should:

- Minimise total sample sizes
- Minimise variable cost for designated power
- Maximise results for a fixed cost

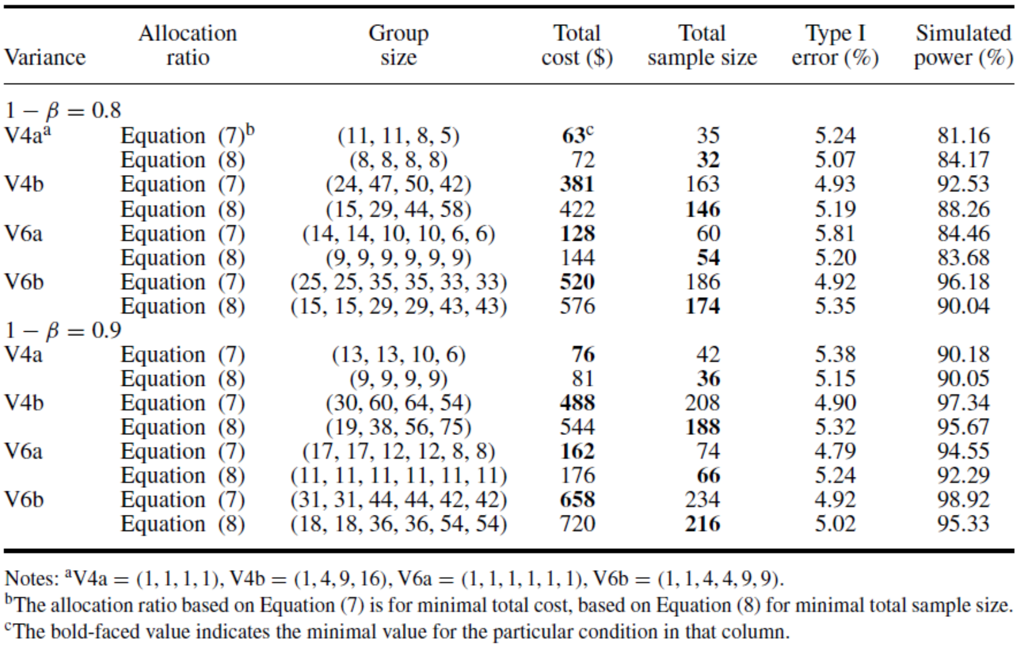
To compare the effects by using different allocation ratios, two sample size tables (see Figure 1 and 2 below) were produced by considering:

- The number of groups as two different sizes, as 4 and 6 (with subset a being equal variance and b being unequal variance)
- The variance pattern
- The cost for each observation
- The designated power as 0.8 and 0.9.

Guo and Luh (2013) used two equations (Equation (7) and Equation (8) in their paper) for calculating the efficient sample size allocation ratio. For ease, these equations are reproduced below:

$$\gamma_j = \frac{n_j}{n_1} = \frac{s_j}{s_1}\sqrt{\frac{c_1}{c_j}} \tag{7}$$

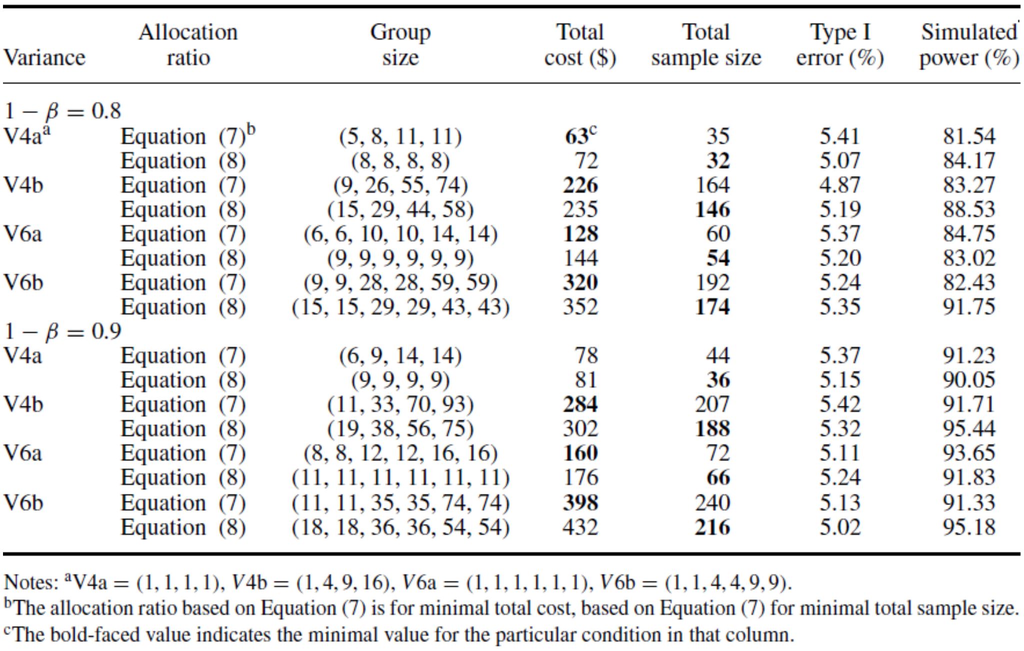
where j is the sample/group index, γ_j represents the allocation ratio, n_1 and n_j are the sample/group sizes, s_1 and s_j are the standard deviations of the groups



Variance	Allocation ratio	Group size	Total cost ($)	Total sample size	Type I error (%)	Simulated power (%)
$1-\beta = 0.8$						
V4a[a]	Equation (7)[b]	(11, 11, 8, 5)	**63**[c]	35	5.24	81.16
	Equation (8)	(8, 8, 8, 8)	72	**32**	5.07	84.17
V4b	Equation (7)	(24, 47, 50, 42)	**381**	163	4.93	92.53
	Equation (8)	(15, 29, 44, 58)	422	**146**	5.19	88.26
V6a	Equation (7)	(14, 14, 10, 10, 6, 6)	**128**	60	5.81	84.46
	Equation (8)	(9, 9, 9, 9, 9, 9)	144	**54**	5.20	83.68
V6b	Equation (7)	(25, 25, 35, 35, 33, 33)	**520**	186	4.92	96.18
	Equation (8)	(15, 15, 29, 29, 43, 43)	576	**174**	5.35	90.04
$1-\beta = 0.9$						
V4a	Equation (7)	(13, 13, 10, 6)	**76**	42	5.38	90.18
	Equation (8)	(9, 9, 9, 9)	81	**36**	5.15	90.05
V4b	Equation (7)	(30, 60, 64, 54)	**488**	208	4.90	97.34
	Equation (8)	(19, 38, 56, 75)	544	**188**	5.32	95.67
V6a	Equation (7)	(17, 17, 12, 12, 8, 8)	**162**	74	4.79	94.55
	Equation (8)	(11, 11, 11, 11, 11, 11)	176	**66**	5.24	92.29
V6b	Equation (7)	(31, 31, 44, 44, 42, 42)	**658**	234	4.92	98.92
	Equation (8)	(18, 18, 36, 36, 54, 54)	720	**216**	5.02	95.33

Notes: [a]V4a = (1, 1, 1, 1), V4b = (1, 4, 9, 16), V6a = (1, 1, 1, 1, 1, 1), V6b = (1, 1, 4, 4, 9, 9).
[b]The allocation ratio based on Equation (7) is for minimal total cost, based on Equation (8) for minimal total sample size.
[c]The bold-faced value indicates the minimal value for the particular condition in that column.

Fig. 1 Group size, total cost, total sample size, and the corresponding Type I error and the simulated power for cost ($1, $1, $2, $5) and ($1, $1, $2, $2, $5, $5) (Guo and Luh, 2013).



Variance	Allocation ratio	Group size	Total cost ($)	Total sample size	Type I error (%)	Simulated power (%)
$1-\beta = 0.8$						
V4a[a]	Equation (7)[b]	(5, 8, 11, 11)	**63**[c]	35	5.41	81.54
	Equation (8)	(8, 8, 8, 8)	72	**32**	5.07	84.17
V4b	Equation (7)	(9, 26, 55, 74)	**226**	164	4.87	83.27
	Equation (8)	(15, 29, 44, 58)	235	**146**	5.19	88.53
V6a	Equation (7)	(6, 6, 10, 10, 14, 14)	**128**	60	5.37	84.75
	Equation (8)	(9, 9, 9, 9, 9, 9)	144	**54**	5.20	83.02
V6b	Equation (7)	(9, 9, 28, 28, 59, 59)	**320**	192	5.24	82.43
	Equation (8)	(15, 15, 29, 29, 43, 43)	352	**174**	5.35	91.75
$1-\beta = 0.9$						
V4a	Equation (7)	(6, 9, 14, 14)	78	44	5.37	91.23
	Equation (8)	(9, 9, 9, 9)	81	**36**	5.15	90.05
V4b	Equation (7)	(11, 33, 70, 93)	**284**	207	5.42	91.71
	Equation (8)	(19, 38, 56, 75)	302	**188**	5.32	95.44
V6a	Equation (7)	(8, 8, 12, 12, 16, 16)	**160**	72	5.11	93.65
	Equation (8)	(11, 11, 11, 11, 11, 11)	176	**66**	5.24	91.83
V6b	Equation (7)	(11, 11, 35, 35, 74, 74)	**398**	240	5.13	91.33
	Equation (8)	(18, 18, 36, 36, 54, 54)	432	**216**	5.02	95.18

Notes: [a]V4a = (1, 1, 1, 1), *V*4b = (1, 4, 9, 16), *V*6a = (1, 1, 1, 1, 1, 1), *V*6b = (1, 1, 4, 4, 9, 9).
[b]The allocation ratio based on Equation (7) is for minimal total cost, based on Equation (7) for minimal total sample size.
[c]The bold-faced value indicates the minimal value for the particular condition in that column.

Fig. 2 Group size, total cost, total sample size, and the corresponding Type I error and the simulated power for cost ($5, $2, $1, $1) and ($5, $5, $2, $2, $1, $1) (Guo and Luh, 2013).

and c_j is the cost of obtaining a single observation from group j (similarly c_1 is the cost of obtaining a single observation from group 1).

$$\gamma_j = \frac{s_j}{s_1} \tag{8}$$

where γ_j represents the allocation ratio and s_1 and s_j are the standard deviations of the groups. This is a special case of Equation (7) when $c_1 = c_j$ for all j (Guo and Luh, 2013).

Based on the research purpose, the allocation ratio was presented to minimize the total cost (based on Equation (7)) or to minimize the total sample size (based on Equation (8)), respectively Guo and Luh (2013).

In the case of minimal total cost, the resulting cost is generally less than the cost of minimal total sample size as expected. For example, in Figure 1, when power = 0.8 and the variance = (1, 1, 4, 4, 9, 9), the cost is \$520 by using Equation (7) , whereas the cost is \$576 by using Equation (8), which wastes 10.77%($(576 - 520)/520 = 10.77\%$) of the cost. For all the conditions studied in figure 1, the average waste of the cost is about 10.56%. Finally, in regard to the total sample size, the resulting size by using Equation (8) is smaller than the size by using Equation (7), as expected. The average waste of subjects if Equation (7) is used is about 10.85% across conditions. Therefore, the impact of selecting different allocation ratios is clearly shown in these two tables, and especially, substantial cost savings are noted (Guo and Luh, 2013).

In Figure 1, it is not always the case that equal group sizes produce a higher power in a study. The examples in this study show that, based on the above constraints and parameters, unequal group sizes can both cost less and generate a higher study power. The difference between equal variances (subscript a) and unequal variances (subscript b) further compounds the analysis. Tests using unequal variances invariably needed a larger sample size to attain the desired power; equal variances did not.

In Figure 2, the costs have changed but quite similar patterns can be seen, overall the costs are lower, the Type I error is generally reduced but the power overall is lower in this figure. This seems to be as a result of changing configurations of the groups. However, there are changes in the desired power level too.

It is not clear how the experiment has been conducted in its entirety as the following statements show.

– The more expensive the cost for each observation, the smaller the group size allocated.
– If the variances are unequal, the larger the variance, the larger the group size allocated.
– If the large variance is concurrent with large cost the allocated group size is moderated to being relatively not too large
– When the large variance is concurrent with small cost; the allocated size is moderated to relatively large.

Guo and Luh (2013)

In statement two there appears to be an assumption that a larger variance requires a larger group. This does not seem to have been explored in the analysis. The statements surrounding moderation of the group have no further explanation within the paper. Therefore it is difficult to conclude what has actually been done by the researcher. The paper is very in depth and it would benefit from further explanation surrounding any assumptions made earlier on in the paper. The work looks promising but in order to critique it fairly a more detailed experiment would need to be run.

1.3 Study Assumptions

Due to the complexity and number of factors that could be accounted for within this study, a number of assumptions have been used to constrain this investigation.

- This study will be generic in that complex sampling designs will not be considered and will therefore not take into account methods such as clustering/stratifying etc.
- This study will be based on a simple data set where treatments are added or not added to individual patients. This may produce results that will not work with groups.
- Treatment effects discovered on subgroups may not be indicative of the entire population or behaviour of the population.
- The results will only be valid for the range of data points considered.
- As this was a simulation study, the results are specific to the conditions investigated. While a range of likely values and variables were included in conducting the simulations, not all ranges or variables could be modelled.

2 Experimental Scenario

The scenario here was that Porton Down[3] wanted to test sepsis patients and determine if there was a way of forecasting the likelihood of a patient becoming septic based on biomarker profiles. The data is very rare (von Knebel Doeberitz and Lacroix, 1999), due to patients either becoming septic quickly or the signs not being spotted in time in order to take a sample from the patient. Additionally there are relatively few patients across the UK who do become septic. This led to discussions of the 'cost' or 'rarity' of the samples that we were able to obtain in order to do the study. The debate focused around whether a larger control group or unequal size groups would alleviate this issue but it was determined that the power might be affected in such a way as to produce a poor power thereby negatively affecting the study and potentially causing poor results. This assumption was a very useful one for the experiment at

[3] The UK's centre for Chemical, Biological, Radiological and Nuclear defense (CBRN), part of the Defence Science and Technology Laborator (DSTL) and the Ministry of Defence (MOD)

hand. However, it challenged fundamental statistical theory. Therefore, it was thought prudent to produce an investigation into this area. As patients for this study are rare (Yeh *et al.*, 2009), this analysis was conducted through a simulation study. Data and original study are not available due to experimental sensitivity at Porton Down.

2.1 Methodology

In examining the utility of a larger control group to lend more power to a study, the t-test [4] and Fisher's test [5] were chosen as common tests that are used in this area of experimentation by dstl. The experiments were set up to show the power gained by a study as the control group increases. In this experiment the control group was increased by iterations to simulate the movement form equal to unequal group sizes. The panels of each graph show the different effect sizes sought. The following experiments were conducted using R.

2.2 t-test experiment

As stated in Zimmerman and Zumbo (1993) the t-test is robust with heterogeneous variances provided that the sample sizes are equal. According to their paper, the Type I error becomes hugely inflated by unequal variances. In cases where population distribution is normal but variances are unequal there have been many suggestions of modifications to the t-test by a number of authors such as Satterthwaite (1946), Smith (1938), Welch (1938). The modifications these authors proposed however, all have problems controlling the Type I error in non normal distributions.

A T-test is a statistical examination of two population means (Horne, 1998). A two-sample t-test examines whether two samples are different and is commonly used when the variances of two normal distributions are unknown and when an experiment uses a small sample size.

[4] A T-test is a statistical examination of two population means. A two-sample t-test examines whether two samples are different and is commonly used when the variances of two normal distributions are unknown and when an experiment uses a small sample size.

[5] Fishers test is designed to test if two population variances are equal. It does this by comparing the ratio of two variances. So, if the variances are equal, the ratio of the variances will be 1. All hypothesis testing is done under the assumption the null hypothesis is true. (Hoffman, 2015)

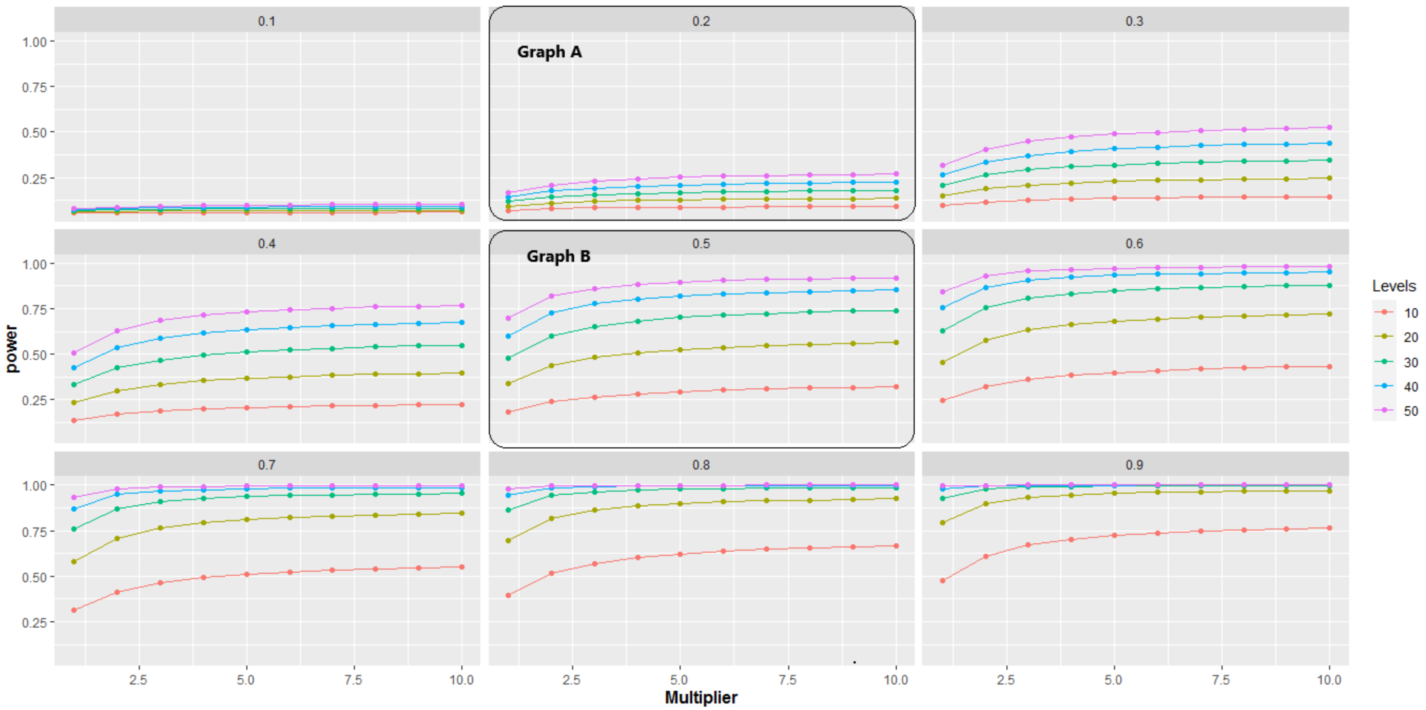


Fig. 3 T-test experiment results for effect sizes from 0.1 to 0.9. Graph A represents no further improvement in power, Grpah B demonstrates where an improvement in power can be acheieved.

In Figure 3 the following is represented:

- Orange indicates the lowest number of control subjects (10) and Purple the highest (50);
- Power is represented on the vertical;
- The horizontal titles on each panel indicate the change in effect size. In each panel, as the size of the control group is increased, the effect on the power is shown.

The t-test experiment results displayed in Fig. 3 show the following:

- It can be seen that the smaller the effect size the lower the power and the larger the effect size the higher the power
- Label A denotes the area where it can be seen that it may be useful to increase the control group slightly. After this level there is no real increase in power per increase in control group
- Label B denotes where it can be seen that it could be extremely useful to increase the control group to obtain a better power level and that the control could be increased substantially

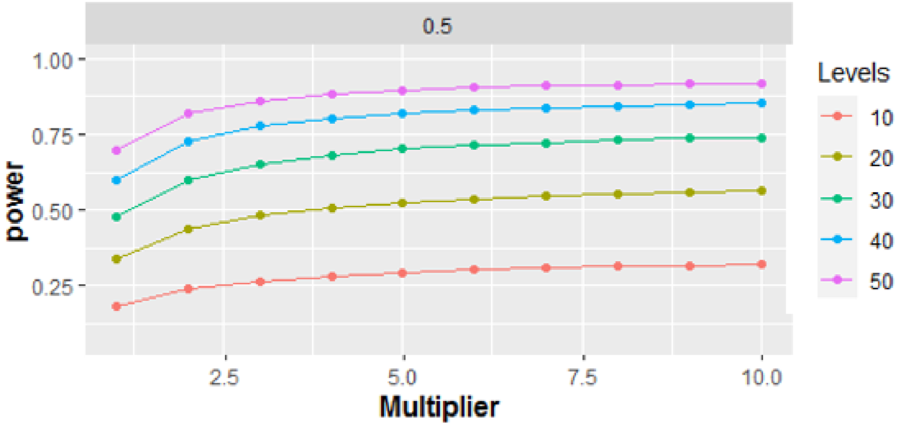


Fig. 4 T-test graph for effect size of 0.5. The multiplier represents the increase in the control group and the levels indicate the colour bar representing each group

This methodology could produce a higher power for the study. The t-test experiment (see Fig. 3) shows that small effect sizes may not particularly benefit from increasing the control group, neither do large effect sizes. But effect sizes that are in the panels representing 0.4, 0.5 and 0.6 may benefit from some increase in the control group. The larger and smaller effect sizes show very little or no increase in power for the increase of the control group.

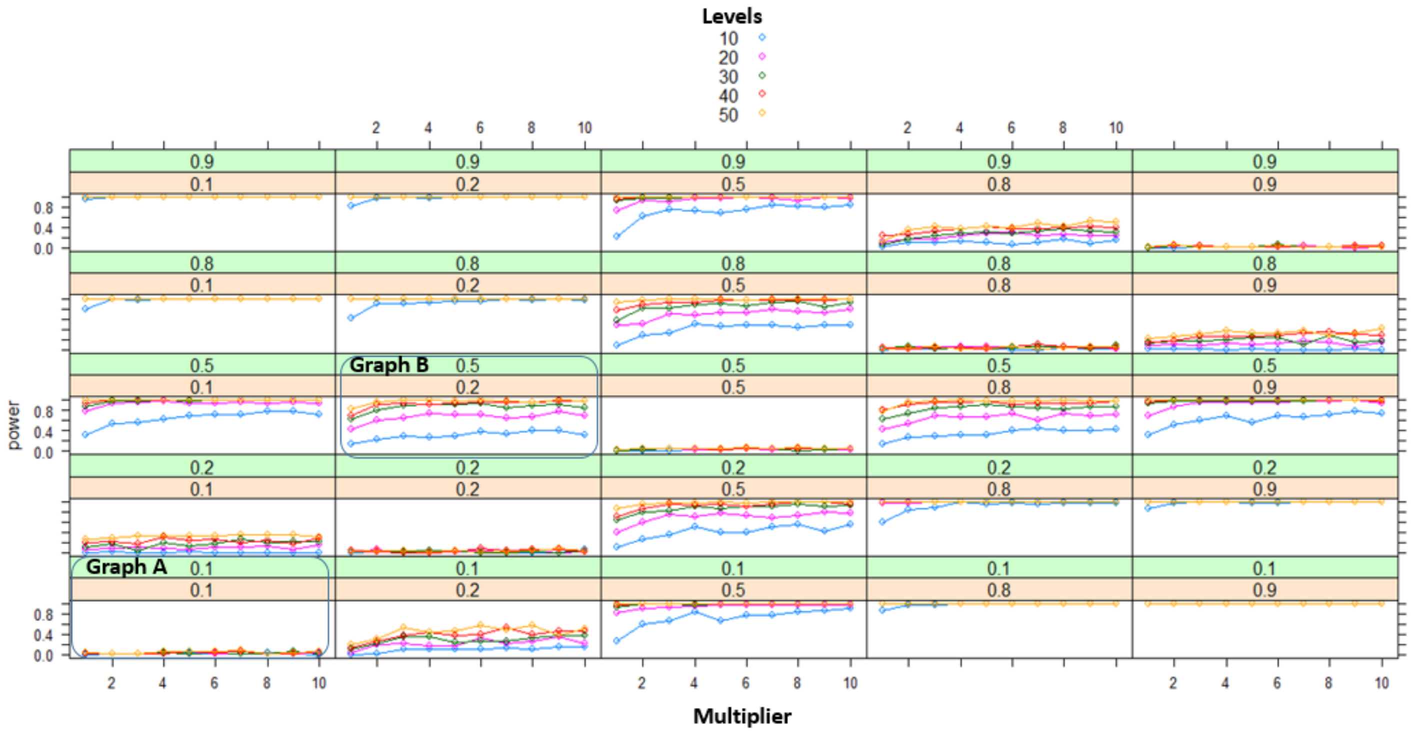


Fig. 5 Fisher test graph for showing changes in power. The multiplier represents the increase in the control group and the levels indicate which colour represents which size group

2.3 Fishers Test experiment

In order to visualise what might happen to power during Fisher's test [6] when group sizes changed an experiment was used to produce a generic output. The experiment had examined the change in power as the control group is increased, in Fig 5.

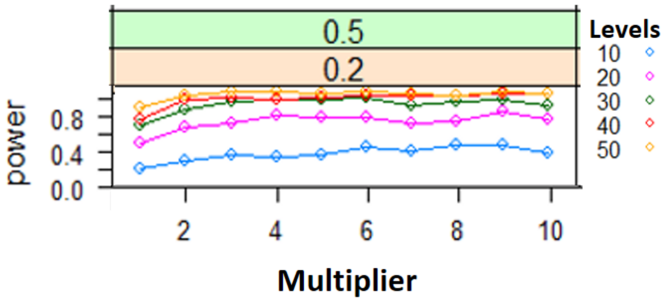


Fig. 6 Fisher test graph for showing increases in power for proportions of 0.5 and 0.2. The multiplier represents the increase in the control group and the levels indicate the colour bar representing each group

Power is represented on the vertical. The horizontal axes on each panel indicate the effect size and corresponding proportions. It can be seen, in each panel, as the size of the control group is increased, what happens to the power. It can be seen that the smaller the effect size the lower the power and the larger the effect size the higher the power. Blue indicates the lowest number of control subjects (10) and yellow the highest (50).

In Figure 6 the following is represented:

- Blue indicates the lowest number of control subjects (10) and yellow the highest (50);
- Power is represented on the vertical;
- The horizontal titles on each panel indicate the proportions and the change in effect size as the multiplier is increased. If we examine Graph A we do not see an improvement on power related to group size but on Graph B we see an improvement on Power in relation to a larger control group size.

The Fisher's test experiments displayed in Fig. 5 indicate the following:

- Label A denotes the area where it can be seen that it may be useful to increase the control group slightly. After this level there is no real increase in power per increase in control group

[6] The F-test is designed to test if two population variances are equal. It does this by comparing the ratio of two variances. So, if the variances are equal, the ratio of the variances will be 1. All hypothesis testing is done under the assumption the null hypothesis is true.

– Label B denotes where it can be seen that it could be extremely useful to increase the control group and that it could be increased substantially. This could produce a higher power for the study

The results above relate only to the experiment above and may not apply to any study in particular. However, it is a worthwhile illustration that shows that where there are medium effect sizes it could be useful to increase the control slightly but after a certain point there is no further benefit to increasing the size of the control group.

This shows that in certain circumstances there may be merit in increasing the size of the control group but as with other studies, this study has a list of assumptions, which mean that the results would have to be applied very carefully and with justification.

3 Discussion

Ethics is clearly a large part of any trial or live study and the main questions arising from this paper are, why do so few papers carry well discussed reasons for choosing their methods, why is power not frequently used to determine the size of the effects that can be expected within the study and what is the stance of ethics on high risk studies on rapidly deteriorating patients? (Suresh and Chandrashekara, 2012). This is concerning, as is the claim that a lot of studies are performed under the guise of “safety”. There is investigation required here to determine the circumstances surrounding statistical studies and what can be done to improve them.

The Welch and t-test have been investigated as popular tests and De Winter (2013), Ahad and Yahaya (2014), Rusticus and Lovato (2014) all describe similar results. Equal size groups under positive pairing with homogeneous variances can be analysed by the t test and Welch test with a robust error rate. However, negative pairing, heterogeneous variance and heterogeneous group size lead to an inflated Type I error. This leads to the question, what theory is out there or being developed to address this issue, as it is uncommon to have a ‘nice’ data set. Guo and Luh (2013) are looking at this issue and also looking at the financial implications of studies.

The studies of the t-test and Fisher’s test indicated the benefits of increasing the size of the control group where the effect size is in the middle range. This hypothesis should be investigated in an actual trial to determine whether the conclusions from the simulated results are valid. The experiment study showed the sample size at which the benefits of an increase in the control group were no longer as clear cut due to diminishing returns. This hypothesis should also be validated through running trials. It is certain that, with further research, this area of statistics will flourish and provide the community with tools that have not been seen before and that are able to solve more of our problems that either cannot currently be solved or are solved in an unsatisfactory manner.

Many studies have published results of large effect sizes but lack the precision to detect differences of interest. Such shortcomings have led some to argue for reform of current sample size conventions in order to avoid misinterpretation of completed studies and harm to scientific research as seen in Jia and Lynn (2015). This statement is extremely important for the future of the work contained within this paper. There must be an investigation into current practise and indeed education on current practise. The risk of this not happening is that truly misleading information could be disseminated to the public or into important decisions for business and government.

Bacchetti *et al.* (2011) states "conventional power calculations provide precise sample sizes – but only using precise assumptions". Another important issue is that many, if not all, statistical tests are based on assumptions, as is much of mathematics. Therefore the practitioner must be acutely aware of the data they have and how tests can be used to provide the answers they seek. Blindly applying theory or partially understood theory can lead to errors that put the study at risk of being misleading or completely incorrect. The chosen theory/test should be investigated fully i.e. the assumptions and pre requisites for using a particular theory are extremely important when applying theory to practise. It may be that the theory does not fit the application and new theory may need to be sought out. As the t-test and Welch test are very commonly used tests the question is are they being applied correctly to the correct situations? It has been seen in previous work that mistakes have been made in applications of theory or indeed the theory has not been applied at all (Dibao-Dina *et al.*, 2014) (Blanchin *et al.*, 2013). If this is the case with two of the most commonly used tests then action must be taken by professional bodies to generate guidance to ensure that this does not continue (Bradley and Schaefer, 1998).

4 Positioning within recent work

The impact of sample size is shown to impact power and Type 1 error rates in a study by Ahad and Yahaya (2014). The experimental work in this paper is beginning to formalise issues that have been found in other work. The fact that Fishers test and t-tests are impacted by the choice of sample sizes and the resulting power does not appear to have been considered overmuch due to the theory of using equal sample sizes as being the best practice. Looking at the Welch test is another example of how group sizes can impact a standard statistical test in potentially large, and undesirable ways. If, as is found in the study by Ahad and Yahaya (2014), power and Type 1 error rates are affected then the power of the study may not be suitable for the objective and, even worse, there may be errors that significantly affect the results. This has the potential to render a very expensive study either worthless or misleading to a point that it may not be used or relied upon.

In this section two studies will be analysed that show:

- The impact of sample size and variability on power and Type 1 error rates

- The impact of group sizes on a Welch Test

4.1 Impact of sample size and variability on power and Type I error rates

Bacchetti *et al.* (2005) states that "The average projected burden per participant remains constant as the sample size increases, but the projected study value does not increase as rapidly as the sample size if it is assumed to be proportional to power or inversely proportional to confidence interval[7] width." This implies that the value per participant declines as the sample size increases and that smaller studies therefore have more favourable ratios of projected value to participant burden. The reality is that researchers are far from the "mega trial of 10,000 subjects", Ioannidis (2013). Due to restrictions on "time, budget or ethical considerations", De Winter (2013) large samples may not be accessible. The main issue with using a small sample is the higher risk of Type I or Type II errors. A small sample size can imply low statistical power or high Type II Error. This puts the researcher at risk of a false positive result, Cohen *et al.* (1965).

Siegel (1956) states that traditional parametric tests should not be used with small sample sizes due to the underlying required assumptions:

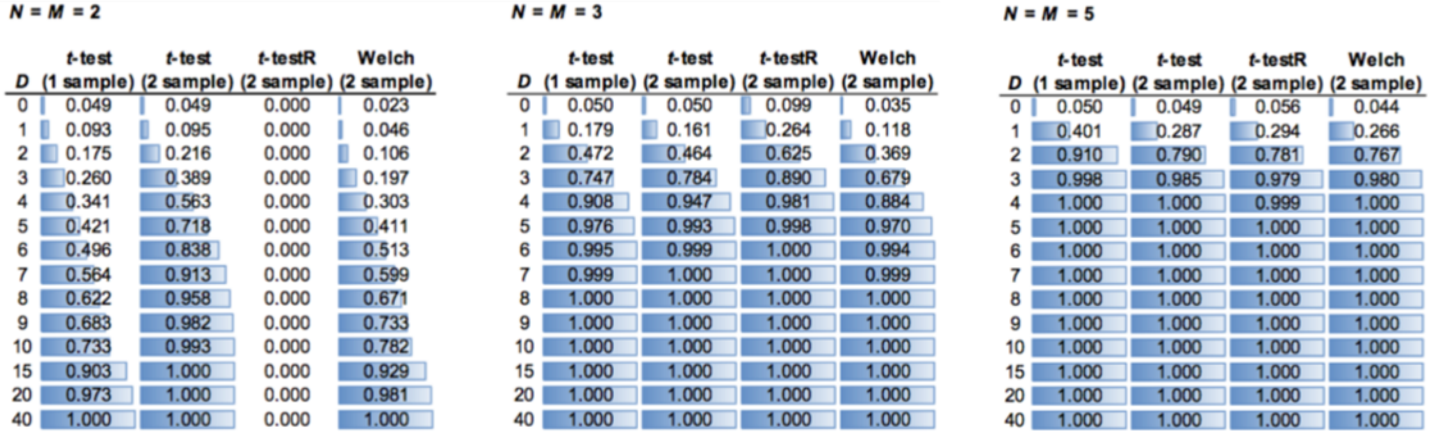
- t-test requires observations to be drawn from a normally distributed population
- Two sample t-test requires that the two populations have the same variance

Siegel (1956) stated that these assumptions cannot be tested when there is a small sample size. Therefore, the t-test should be avoided in favour of a non parametric test when dealing with small samples. Intuitively we would believe that a small sample size would only be able to show large effects within the data, not smaller size effects due to a lack of data points. With small samples low statistical power must therefore be accepted in this situation. Studies have been conducted that counter this assertion. Siegel counters this argument in Siegel (1956).

In the following output from a study conducted by De Winter (2013) experiments have been conducted to determine the statistical power and Type I error rate of the one and two sample t-tests, De Winter (2013). The experiments were carried out for effect sizes (D) between 0 (i.e. Null Holds) and 40 (i.e. alternative hypothesis holds with large effect) and for N[8]=2, N=3, N=5. Each case was simulated 100,000 times.

[7] The confidence level describes the uncertainty associated with a sampling method. Suppose we used the same sampling method to select different samples and to compute a different interval estimate for each sample. Some interval estimates would include the true population parameter and some would not. A 90% confidence level means that we would expect 90% of the interval estimates to include the population parameter; A 95% confidence level means that 95% of the intervals would include the parameter; and so on.

[8] sample size



***N* = *M* = 2**

D	*t*-test (1 sample)	*t*-test (2 sample)	*t*-testR (2 sample)	Welch (2 sample)
0	0.049	0.049	0.000	0.023
1	0.093	0.095	0.000	0.046
2	0.175	0.216	0.000	0.106
3	0.260	0.389	0.000	0.197
4	0.341	0.563	0.000	0.303
5	0.421	0.718	0.000	0.411
6	0.496	0.838	0.000	0.513
7	0.564	0.913	0.000	0.599
8	0.622	0.958	0.000	0.671
9	0.683	0.982	0.000	0.733
10	0.733	0.993	0.000	0.782
15	0.903	1.000	0.000	0.929
20	0.973	1.000	0.000	0.981
40	1.000	1.000	0.000	1.000

***N* = *M* = 3**

D	*t*-test (1 sample)	*t*-test (2 sample)	*t*-testR (2 sample)	Welch (2 sample)
0	0.050	0.050	0.099	0.035
1	0.179	0.161	0.264	0.118
2	0.472	0.464	0.625	0.369
3	0.747	0.784	0.890	0.679
4	0.908	0.947	0.981	0.884
5	0.976	0.993	0.998	0.970
6	0.995	0.999	1.000	0.994
7	0.999	1.000	1.000	0.999
8	1.000	1.000	1.000	1.000
9	1.000	1.000	1.000	1.000
10	1.000	1.000	1.000	1.000
15	1.000	1.000	1.000	1.000
20	1.000	1.000	1.000	1.000
40	1.000	1.000	1.000	1.000

***N* = *M* = 5**

D	*t*-test (1 sample)	*t*-test (2 sample)	*t*-testR (2 sample)	Welch (2 sample)
0	0.050	0.049	0.056	0.044
1	0.401	0.287	0.294	0.266
2	0.910	0.790	0.781	0.767
3	0.998	0.985	0.979	0.980
4	1.000	1.000	0.999	1.000
5	1.000	1.000	1.000	1.000
6	1.000	1.000	1.000	1.000
7	1.000	1.000	1.000	1.000
8	1.000	1.000	1.000	1.000
9	1.000	1.000	1.000	1.000
10	1.000	1.000	1.000	1.000
15	1.000	1.000	1.000	1.000
20	1.000	1.000	1.000	1.000
40	1.000	1.000	1.000	1.000

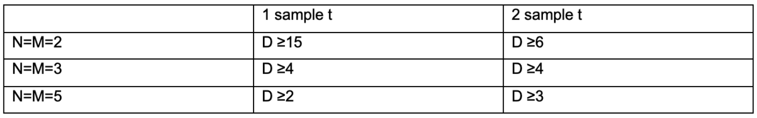
Fig. 7 Proportion of 100,000 repetitions yielding p¡0.05 for various mean distances D and Equal sample sizes, (De Winter, 2013)

The study also investigates the following scenarios:

- Unequal variances
- Unequal sample sizes
- Unequal sample sizes and unequal variances

Only the t-test and Welch Test columns (t-test 1 sample, t-test 2 sample & Welch) are relevant here.

In Figure 7 the group sizes are very small and it is debatable whether, using group sizes of 2 would actually yield these results in live trials. For the one sample t-test, acceptable statistical power (1-Type II error rate ¿80%) is reached for $D \geq 15$. The t-test provides acceptable power for small sample sizes, provided the effect is large. The Welch 2 sample test indicates an increased ability to detect effects as the group sizes increase. This shows that there comes a point at where more effort yields nominal results. The results are summarised below.



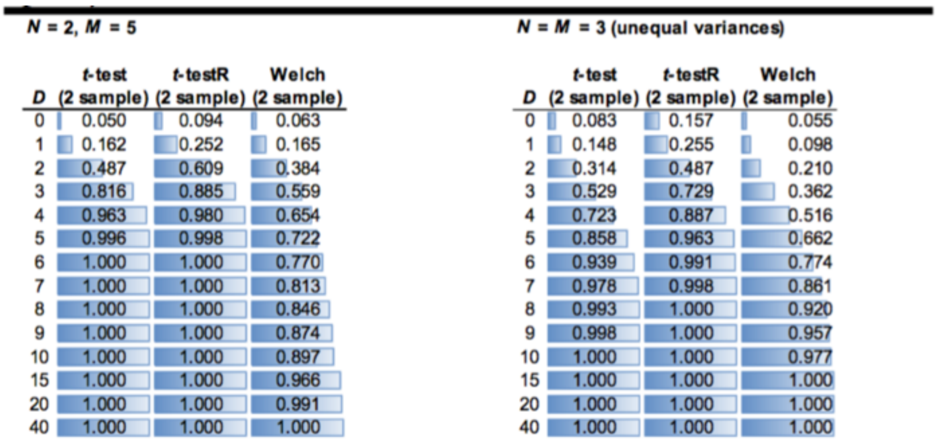
	1 sample t	2 sample t
N=M=2	D ≥15	D ≥6
N=M=3	D ≥4	D ≥4
N=M=5	D ≥2	D ≥3

Fig. 8 Proportion of 100,000 repetitions yielding $p < 0.05$ for various mean distances D. Equal sample sizes, (De Winter, 2013)

It can be seen that the larger the group, the easier it is to detect small effects. When the group is smaller then it would be expected that we would only be able to see the largest effects. However, by using a two sample experiment rather than a one sample there is a greater chance of detecting medium to small effects with lower samples sizes.

For unequal groups (N=2, M=5) acceptable statistical power is reached at $D \geq 4$, for unequal variances the corresponding figure is $D \geq 6$. It can be seen that unequal variances have more of an effect on whether the power can be correctly calculated than unequal groups. The Welch test begins to perform poorly under unequal group sizes and variances. Indeed the t-test performs better here.

For the case N=2 with small variance and M=5 with large variance the t-test reaches acceptable statistical power at $D \geq 6$. For the case N=5 (small variance) and M=2 with large variance the t-test reaches acceptable power at $D \geq 4$. The Welch Test out performs the t-test in the first case but fails considerably in the second. Therefore the best case scenario for detecting smaller effects, i.e. the case where the highest power is observed, is a large group with small variance compared with a smaller group with larger variance. It seems that variance has a much larger effect on all of the tests above than group size. Ahad and Yahaya (2014) suggest that the t-test can be unsatisfactory with regard to Type I error rates when data are from populations with unequal variances. experiments conducted with N=3, D=0 for estimating the Type I



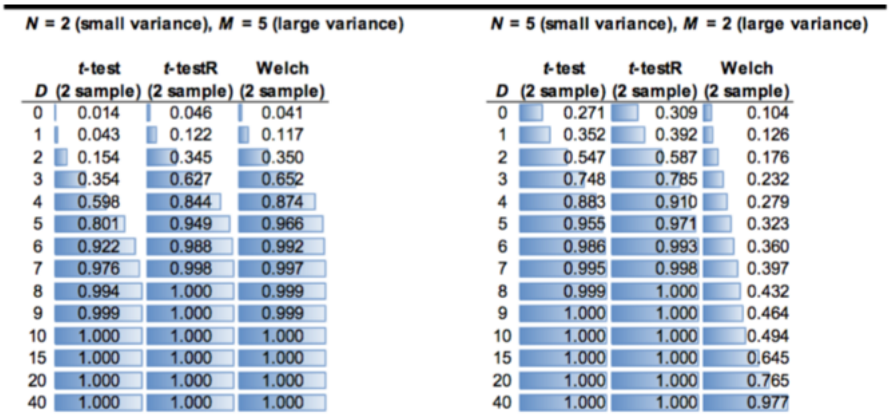
***N* = 2, *M* = 5**

D	*t*-test (2 sample)	*t*-testR (2 sample)	Welch (2 sample)
0	0.050	0.094	0.063
1	0.162	0.252	0.165
2	0.487	0.609	0.384
3	0.816	0.885	0.559
4	0.963	0.980	0.654
5	0.996	0.998	0.722
6	1.000	1.000	0.770
7	1.000	1.000	0.813
8	1.000	1.000	0.846
9	1.000	1.000	0.874
10	1.000	1.000	0.897
15	1.000	1.000	0.966
20	1.000	1.000	0.991
40	1.000	1.000	1.000

***N* = *M* = 3 (unequal variances)**

D	*t*-test (2 sample)	*t*-testR (2 sample)	Welch (2 sample)
0	0.083	0.157	0.055
1	0.148	0.255	0.098
2	0.314	0.487	0.210
3	0.529	0.729	0.362
4	0.723	0.887	0.516
5	0.858	0.963	0.662
6	0.939	0.991	0.774
7	0.978	0.998	0.861
8	0.993	1.000	0.920
9	0.998	1.000	0.957
10	1.000	1.000	0.977
15	1.000	1.000	1.000
20	1.000	1.000	1.000
40	1.000	1.000	1.000

Fig. 9 Proportion of 100,000 repetitions yielding p¡0.05 for various mean distances D. (Left) - Unequal Sample Sizes, (Right) Unequal Variances, De Winter (2013)



***N* = 2 (small variance), *M* = 5 (large variance)**

D	*t*-test (2 sample)	*t*-testR (2 sample)	Welch (2 sample)
0	0.014	0.046	0.041
1	0.043	0.122	0.117
2	0.154	0.345	0.350
3	0.354	0.627	0.652
4	0.598	0.844	0.874
5	0.801	0.949	0.966
6	0.922	0.988	0.992
7	0.976	0.998	0.997
8	0.994	1.000	0.999
9	0.999	1.000	0.999
10	1.000	1.000	1.000
15	1.000	1.000	1.000
20	1.000	1.000	1.000
40	1.000	1.000	1.000

***N* = 5 (small variance), *M* = 2 (large variance)**

D	*t*-test (2 sample)	*t*-testR (2 sample)	Welch (2 sample)
0	0.271	0.309	0.104
1	0.352	0.392	0.126
2	0.547	0.587	0.176
3	0.748	0.785	0.232
4	0.883	0.910	0.279
5	0.955	0.971	0.323
6	0.986	0.993	0.360
7	0.995	0.998	0.397
8	0.999	1.000	0.432
9	1.000	1.000	0.464
10	1.000	1.000	0.494
15	1.000	1.000	0.645
20	1.000	1.000	0.765
40	1.000	1.000	0.977

Fig. 10 Proportion of 100,000 repetitions yielding p¡0.05 for various mean distances D (Left) - Unequal Sample Sizes (Right) Unequal Variances, De Winter (2013)

error rate and D=2 was used for estimating statistical power, (De Winter, 2013).

Figure 11 shows that the Type I error rate is quite low in this study but the Type II error rate can get quite high for the t-test. A high Type I error was observed for unequal variances combined with unequal sample sizes. The experiments further clarified that when the sample size is extremely small, Type II errors can only be avoided if the effect size being detected is large. The high Type I error rate for the t-test is caused by the pooled standard deviation being mostly determined by the larger sample size having the lower variability, while the difference in sample size is determined mostly by the smaller sample size having the higher variability. Therefore “the t statistic is

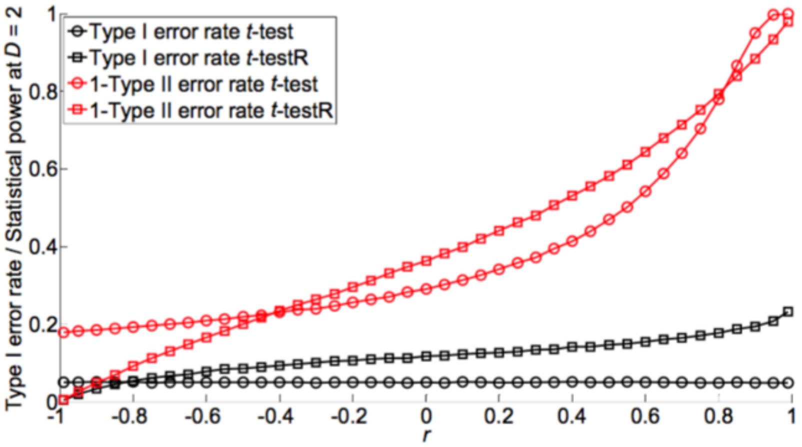


Fig. 11 Type I error rate and Statistical Power (1- Type II error rate) for paired t-test. experiments conducted with N=3, D=0 for estimating the Type I error rate and D=2 was used for estimating statistical power, De Winter (2013)

inflated", De Winter (2013). Consequently, conducting a t-test with a small sample size is acceptable providing the effect size is large. The study also found that there could be a high false positive rate on the one sample t-test on non-normal data. This means that the high false positive rate could also affect conclusions drawn from the data and render them misleading. The primary take away from this analysis is that it is important to know the characteristics of the data and the assumptions attached to the tests. If this is not known then the wrong test can be chosen or a test chosen that does not suit the parameters of the data, thereby invalidating the experiment. In the case of experiments with costly or rare subjects, this can end the experiment with poor, or misleading conclusions.

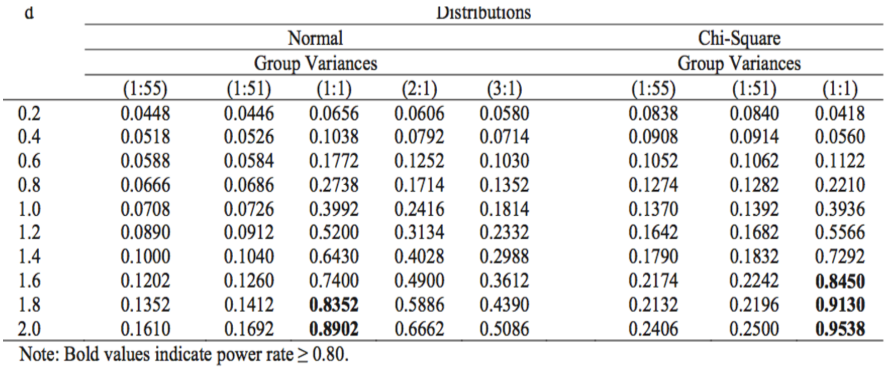
4.2 Welch Test (Heterogeneous variances and group sizes)

The following analysis by Ahad and Yahaya (2014), looks at the robustness of the Welch test when variances are unequal and also under the alternative, the Chi Squared. The Welch test is used to compare means between two independent groups without assuming equal population variances. The Welch test however is not robust in the following circumstances:

- Distribution is non-normal
- Variance is heterogeneous and unequal size groups occur together

In these cases the Type I error will inflate. This is supported by the earlier analysis in sub section 4.1, Impact of sample size and variability on power and Type I error rates.

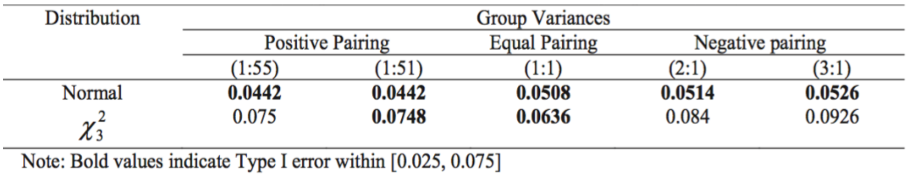
In the investigation by Ahad and Yahaya (2014), various conditions such as sample sizes, types of distributions and unequal group variances were manipulated. Normal distribution was used and for non normal chi squared was chosen. Alpha[9] was set at 0.05 and 0.8 was used as a desired power level. Different ratios of positive and negative pairings were examined. Positive meaning the variance and sample sizes are directly associated, negatively paired being inversely associated.



d	Distributions							
	Normal					Chi-Square		
	Group Variances					Group Variances		
	(1:55)	(1:51)	(1:1)	(2:1)	(3:1)	(1:55)	(1:51)	(1:1)
0.2	0.0448	0.0446	0.0656	0.0606	0.0580	0.0838	0.0840	0.0418
0.4	0.0518	0.0526	0.1038	0.0792	0.0714	0.0908	0.0914	0.0560
0.6	0.0588	0.0584	0.1772	0.1252	0.1030	0.1052	0.1062	0.1122
0.8	0.0666	0.0686	0.2738	0.1714	0.1352	0.1274	0.1282	0.2210
1.0	0.0708	0.0726	0.3992	0.2416	0.1814	0.1370	0.1392	0.3936
1.2	0.0890	0.0912	0.5200	0.3134	0.2332	0.1642	0.1682	0.5566
1.4	0.1000	0.1040	0.6430	0.4028	0.2988	0.1790	0.1832	0.7292
1.6	0.1202	0.1260	0.7400	0.4900	0.3612	0.2174	0.2242	**0.8450**
1.8	0.1352	0.1412	**0.8352**	0.5886	0.4390	0.2132	0.2196	**0.9130**
2.0	0.1610	0.1692	**0.8902**	0.6662	0.5086	0.2406	0.2500	**0.9538**

Note: Bold values indicate power rate ≥ 0.80.

Fig. 12 Power of Welch test when group size (5,15) (Ahad and Yahaya, 2014)



Distribution	Group Variances				
	Positive Pairing		Equal Pairing	Negative pairing	
	(1:55)	(1:51)	(1:1)	(2:1)	(3:1)
Normal	**0.0442**	**0.0442**	**0.0508**	**0.0514**	**0.0526**
χ_3^2	0.075	**0.0748**	**0.0636**	0.084	0.0926

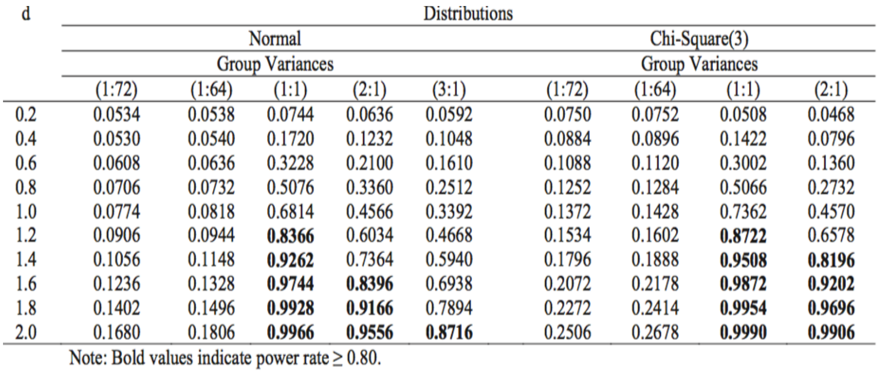
Note: Bold values indicate Type I error within [0.025, 0.075]

Fig. 13 Type I error rates of Welch test when group size (5,15) (Ahad and Yahaya, 2014)

In Figure 12, it is illustrated that only large effects can be detected at a statistically significant power level when the variances are the same in the case of varying group sizes. In all other cases where there is heterogeneity of variance and group size, not even large effects can be detected with any certainty. When the normal criterion is violated it can be seen again that only in the case of equal variances that any effect can be detected and the only effect

[9] This option specifies one or more values for the probability of a Type-I error. A Type-I error occurs when a true null hypothesis is rejected. In this procedure, a Type-I error occurs when you reject the null hypothesis of nonequivalent means when in fact the means are nonequivalent. Values must be between zero and one. Historically, the value of 0.05 has been used for alpha. This means that about one test in twenty will falsely reject the null hypothesis. A value is chosen for alpha that represents the risk of a type-I error you are willing to take in your experimental situation.

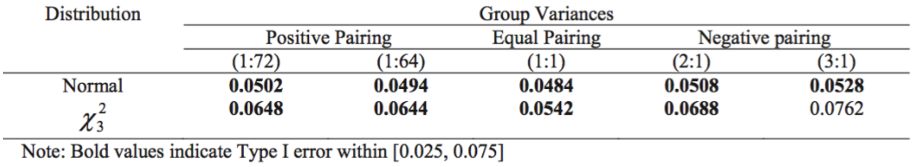
that can be detected is a large one. The test showed lower rates of performance under variance heterogeneity, failing faster with negatively paired under the chi squared.



d	Distributions								
	Normal					Chi-Square(3)			
	Group Variances					Group Variances			
	(1:72)	(1:64)	(1:1)	(2:1)	(3:1)	(1:72)	(1:64)	(1:1)	(2:1)
0.2	0.0534	0.0538	0.0744	0.0636	0.0592	0.0750	0.0752	0.0508	0.0468
0.4	0.0530	0.0540	0.1720	0.1232	0.1048	0.0884	0.0896	0.1422	0.0796
0.6	0.0608	0.0636	0.3228	0.2100	0.1610	0.1088	0.1120	0.3002	0.1360
0.8	0.0706	0.0732	0.5076	0.3360	0.2512	0.1252	0.1284	0.5066	0.2732
1.0	0.0774	0.0818	0.6814	0.4566	0.3392	0.1372	0.1428	0.7362	0.4570
1.2	0.0906	0.0944	**0.8366**	0.6034	0.4668	0.1534	0.1602	**0.8722**	0.6578
1.4	0.1056	0.1148	**0.9262**	0.7364	0.5940	0.1796	0.1888	**0.9508**	**0.8196**
1.6	0.1236	0.1328	**0.9744**	**0.8396**	0.6938	0.2072	0.2178	**0.9872**	**0.9202**
1.8	0.1402	0.1496	**0.9928**	**0.9166**	0.7894	0.2272	0.2414	**0.9954**	**0.9696**
2.0	0.1680	0.1806	**0.9966**	**0.9556**	**0.8716**	0.2506	0.2678	**0.9990**	**0.9906**

Note: Bold values indicate power rate ≥ 0.80.

Fig. 14 Power of Welch test when group size (10,20) (Ahad and Yahaya, 2014)



Distribution	Group Variances				
	Positive Pairing		Equal Pairing	Negative pairing	
	(1:72)	(1:64)	(1:1)	(2:1)	(3:1)
Normal	**0.0502**	**0.0494**	**0.0484**	**0.0508**	**0.0528**
χ_3^2	**0.0648**	**0.0644**	**0.0542**	**0.0688**	0.0762

Note: Bold values indicate Type I error within [0.025, 0.075]

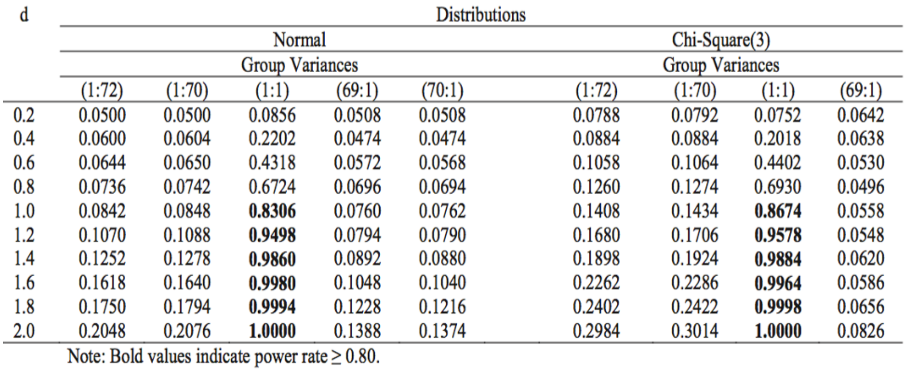
Fig. 15 Type I error rates of Welch test when group sizes (10, 20) (Ahad and Yahaya, 2014)

Figure 15, shows that as the groups come closer to being the same size:

- In the case of homogeneous variance more medium sized effects are detected. In the case of heterogeneous variance larger effects are detected.
- Where large differences in variance are seen, still no effect can be detected with any certainty

In the chi square test the same results are seen. In the Type I error table the test is robust for all but the Chi Squared distribution under negatively paired conditions, with heterogeneous variance. This demonstrates that there could be a clear cut off point for the performance of the test.

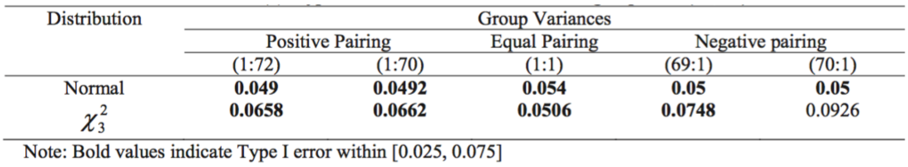
Figure 15 shows results that seem to suggest that an optimal sample size comparison has been reached. The Welch test produced robust Type I error rates for all combinations of group sizes and variances, under chi squared the test appears robust under homogeneous variances and positive pairings. However, when the variance heterogeneity increased with unequal groups under negatively paired conditions, the test produced higher Type I error rates. In



d	Distributions								
	Normal					Chi-Square(3)			
	Group Variances					Group Variances			
	(1:72)	(1:70)	(1:1)	(69:1)	(70:1)	(1:72)	(1:70)	(1:1)	(69:1)
0.2	0.0500	0.0500	0.0856	0.0508	0.0508	0.0788	0.0792	0.0752	0.0642
0.4	0.0600	0.0604	0.2202	0.0474	0.0474	0.0884	0.0884	0.2018	0.0638
0.6	0.0644	0.0650	0.4318	0.0572	0.0568	0.1058	0.1064	0.4402	0.0530
0.8	0.0736	0.0742	0.6724	0.0696	0.0694	0.1260	0.1274	0.6930	0.0496
1.0	0.0842	0.0848	**0.8306**	0.0760	0.0762	0.1408	0.1434	**0.8674**	0.0558
1.2	0.1070	0.1088	**0.9498**	0.0794	0.0790	0.1680	0.1706	**0.9578**	0.0548
1.4	0.1252	0.1278	**0.9860**	0.0892	0.0880	0.1898	0.1924	**0.9884**	0.0620
1.6	0.1618	0.1640	**0.9980**	0.1048	0.1040	0.2262	0.2286	**0.9964**	0.0586
1.8	0.1750	0.1794	**0.9994**	0.1228	0.1216	0.2402	0.2422	**0.9998**	0.0656
2.0	0.2048	0.2076	**1.0000**	0.1388	0.1374	0.2984	0.3014	**1.0000**	0.0826

Note: Bold values indicate power rate ≥ 0.80.

Fig. 16 Power of Welch test when group sizes (15,25) (Ahad and Yahaya, 2014)



Distribution	Group Variances				
	Positive Pairing		Equal Pairing	Negative pairing	
	(1:72)	(1:70)	(1:1)	(69:1)	(70:1)
Normal	**0.049**	**0.0492**	**0.054**	**0.05**	**0.05**
χ_3^2	**0.0658**	**0.0662**	**0.0506**	**0.0748**	0.0926

Note: Bold values indicate Type I error within [0.025, 0.075]

)

Fig. 17 Type I error rates of Welch test when group sizes (15,25)) (Ahad and Yahaya, 2014)

this experiment Welch's test produced a consistent Type I error rate for all combinations of group sizes and group variances, Ahad and Yahaya (2014).

In De Winter (2013) it can be seen that only large effects were able to be detected when using the t-test, where the groups were small but as the groups reached the size N=M=5 smaller effects could be detected. Smaller sample sizes affected the performance of the test negatively. In the following cases the test performed poorly: small samples sizes, unequal variances, unequal group sizes and unequal variance with unequal group sizes. De Winter (2013) investigated the Welch Test under the same conditions as the t-test and found that it performed well under assumptions of homogeneous variances and group sizes but as the group sizes and variances became more heterogeneous the test performed poorly. The Welch test simulation in Ahad and Yahaya (2014) performed well under the normal and chi squared distributions when variance was homogenous and groups were similar in size under positively paired conditions. The test performed poorly under the conditions of heterogeneous variance, heterogeneous group size and under the condition of negative pairing. The two studies support each other's findings with De Winter (2013) perhaps being the most informative in this investigation as a comparison visual was produced for the t-test and Welch test. This, along with the study within this paper shows that equal sample sizes may not be possible, desirable or even optimal. Indeed, under varying conditions it is prudent to determine your exact experimental conditions, such as cost or difficulty in obtaining a sample, the same or unequal variances for groups, same or unequal sample sizes for groups, etc (Schulz and Grimes, 2002), and then choose your method of analysis and examine the potential power calculation based on the characteristics of the experiment. As we have seen within the experimental analysis part of this paper, the power calculation can be optimised purely on group size and as we have seen in (Guo and Luh, 2013) the calculation can also be optimised for cost or rarity of subject. Therefore, it is worth considering the aims and the study conditions intently before starting a study.

5 Ethical Concerns

For a study to be ethical in its design, its projected value must outweigh the projected risks to participants (Bacchetti *et al.*, 2011). Does this mean that if these conditions are met then the study is ethical?

As we have entered the territory of patient acceptability and medical trials, it seems prudent that a discussion on ethics should take place. Ethics is an extremely important consideration when running any trial and it is what keeps a trial within the bounds of acceptability in terms of pain, safety and acceptability, to name but a few parameters. Patient acceptability is a term that could be used to suggest that patients have no possibility of access to treatment unless included in study and randomised to treatment. Another reason for using the justification of a larger treatment group was to expose fewer patients to inferior treatment,(Dibao-Dina *et al.*, 2014). Neither of these reasons can be seen as ethical as they "presuppose a high degree of certainty regarding risks and benefits of the intervention, (Dibao-Dina *et al.*, 2014). Indeed Dibao-Dina *et al.* (2014) states that "adverse events were not fully reported in about half of the reports for which safety issues were the justification". One would expect to see a full report of any safety or safety related issues should this be the aim of the study and especially where this was given as justification for altering the parameters of the study. However Dibao-Dina *et al.* (2014) stuck to the familiar adage that "statistical power is usually maximal with equal sized groups" and raised this as a concern with the studies examined.

Another ethical consideration is when unequal randomisation leads to allocating more patients to the experimental group than the control group. This could influence the response in the control group. Dibao-Dina *et al.* (2014) stated that "the placebo response could then be exaggerated... inducing a bias in the treatment effect estimate". The result may be explained by patient expectations: "they tend to expect a good treatment response with the un- equal randomization because they were aware of a greater probability of receiving an active treatment than a placebo" (Dibao-Dina *et al.*, 2014).

Dibao-Dina *et al.* (2014) reports that the unequal randomisation method in 96.2% of cases recruited more patients to the intervention than the control and this was taken into account in the sample size calculation in 46.2% of reports. This is concerning as every statistical plan should be fully explained, not just where it is envisaged that a departure from the usual approach has taken place. Bacchetti *et al.* (2005) counters this by stating "we see no valid ethical argument against small, high risk/high payoff studies as have been recently advocated for rapidly fatal disease...a more legitimate ethical issue regarding sample size is whether it is too large". This relates to the balance between the value of a study and the burdens accepted by the patients, as the burden would not necessarily improve as the sample size increased. This is an extremely valid point in the world of ethics; however, this may not counteract the need for well designed statistical studies in cases other than rapidly fatal diseases.

Bacchetti *et al.* (2011) states "conventional power calculations provide precise sample sizes – but only using precise assumptions". In this paper it is argued that if the precise assumptions cannot be met then the study should not be rejected out of hand as there may still be value to it. The question here is whether this view is truly ethical. Bacchetti *et al.* (2005) makes the argument that in first time intervention studies or studies requiring nonhuman primate participation, to increase the size of the study in the quest for power would then make the study inherently unethical. Here the point can be raised again that each statistical study is unique and will have its own set of characteristics and assumptions. While most will be amenable to power calculations, some may not. What matters is that the study is taken on its own merits, conducted in an ethical manner and robust analysis is conducted with whichever method is the most appropriate. There is a very important lesson here that despite the current furore surrounding power and analysis – it may not always be relevant to the study.

6 Conclusion

The question of whether unequal groups is a useful route to take is a complex one. This paper has investigated the work surrounding this area and produced experiments to investigate the results.

This study was used to move forward modelling on sepsis patients at Porton Down. The rarity of these patients, and indeed, ones who were in the first stages of sepsis, was such that it was extremely hard to get a sample from them and this led to small sample sizes overall. In order to conduct a meaningful study, the results of this paper have shown that we can obtain more power in some circumstances by using a larger control group of unequal size groups. This enabled the research to be designed in a cost effective way and also in a way where small sample sizes could be used but not to the detriment of the study.

However, this is not to say that this could be applied to any live study. It would be extremely useful to apply this theory to live trials to investigate whether the experiments and conclusions drawn here stand up in reality. There are some limitations to this study as the results are specific to the conditions investigated. The results from the individual papers discussed also adhere to a set of assumptions and particular conditions. The initial question this paper set out to answer was "Can more power be gained from a study by increasing the control group?". It has been determined that, dependent on the following points, more power could potentially be gained from a study by using a larger control group:

- Has the theory been correctly applied?
- Does the data have the correct format for the tests in the study?
- Has an initial power calculation been done and aims of the study determined?
- Has an exploratory data analysis been conducted on the data?

Further work needs to be done in this area as it is stated frequently that equal sample sizes perform the best in many papers. This may be true but this luxury is not always available and cost is always a factor in any study. Much research is being done currently on developing theory such as minimax in statistics, tests that can deal with heterogeneity of variances and unequal sample size. This is promising but worrying that it is only now that this issue is being recognised. The theories put forward here will have lasting impact on the way study design is done and indeed on medical and educational papers where statistics is concerned. More needs to be done to make sure that – especially in a trials field- the relevant theory is applied correctly. Where it is necessary, power analysis should be applied and used correctly.

It is fair to conclude that small sample sizes and/or unequal variances will render studies unable to detect all but the largest effects and in the case of the two occurring together it is possible that no effect at all may be detected. This must be addressed when initial calculations are being made to determine the nature and construct of the study. Cost and sample size is a new area that is evolving and allowing us to maximise gains whilst minimising cost and/or group size. This has applications in most areas as studies can become extremely expensive and provides a precision not seen before by merging Operational Research and Statistics in a beautiful way. Further work could be undertaken to investigate the areas explored in this paper.

Acknowledgements Alison Berry, Helen Keeling CStat and Ralph Mansson at DSTL. For technical help, mentorship and being a pleasure to work with.

References

Ahad, N. A. and Yahaya, S. S. S. (2014). Sensitivity analysis of welch'st-test. In *AIP Conference Proceedings*, volume 1605, pages 888–893. American Institute of Physics.

Bacchetti, P., Wolf, L. E., Segal, M. R., and McCulloch, C. E. (2005). Ethics and sample size. *American journal of epidemiology*, **161**(2), 105–110.

Bacchetti, P., Deeks, S. G., and McCune, J. M. (2011). Breaking free of sample size dogma to perform innovative translational research. *Science translational medicine*, **3**(87), 87ps24–87ps24.

Blanchin, M., Hardouin, J.-B., Guillemin, F., Falissard, B., and Sébille, V. (2013). Power and sample size determination for the group comparison of patient-reported outcomes with rasch family models. *PloS One*, **8**(2), e57279.

Bradley, W. J. and Schaefer, K. C. (1998). *The uses and misuses of data and models*. Sage.

Cohen, J. (1988). Statistical power analysis.

Cohen, J. *et al.* (1965). Some statistical issues in psychological research. *Handbook of clinical psychology*, pages 95–121.

De Winter, J. C. (2013). Using the student's t-test with extremely small sample sizes. *Practical Assessment, Research, and Evaluation*, **18**(1), 10.

Dibao-Dina, C., Caille, A., Sautenet, B., Chazelle, E., and Giraudeau, B. (2014). Rationale for unequal randomization in clinical trials is rarely reported: a systematic review. *Journal of clinical epidemiology*, **67**(10), 1070–1075.

Dumville, J., Hahn, S., Miles, J., and Torgerson, D. (2006). The use of unequal randomisation ratios in clinical trials: a review. *Contemporary clinical trials*, **27**(1), 1–12.

Guo, J.-H. and Luh, W.-M. (2013). Efficient sample size allocation with cost constraints for heterogeneous-variance group comparison. *Journal of Applied Statistics*, **40**(12), 2549–2563.

Hoffman, J. I. (2015). *Biostatistics for medical and biomedical practitioners*. Academic press.

Horne, A. D. (1998). Statistics, use in immunology.

Hsu, L. M. (1993). Using cohen's tables to determine the maximum power attainable in two-sample tests when one sample is limited in size. *Journal of Applied Psychology*, **78**(2), 303.

Ioannidis, J. P. (2013). Mega-trials for blockbusters. *JAMA*, **309**(3), 239–240.

Jia, B. and Lynn, H. S. (2015). A sample size planning approach that considers both statistical significance and clinical significance. *Trials*, **16**(1), 213.

Lan, L. and Lian, Z. (2010). Application of statistical power analysis–how to determine the right sample size in human health, comfort and productivity research. *Building and Environment*, **45**(5), 1202–1213.

Levin, J. R. (1997). Overcoming feelings of powerlessness in" aging" researchers: A primer on statistical power in analysis of variance designs. *Psychology and aging*, **12**(1), 84.

Pisano, E. D., Fajardo, L. L., Tsimikas, J., Sneige, N., Frable, W. J., Gatsonis, C. A., Evans, W. P., Tocino, I., and McNeil, B. J. (1998). Rate of insufficient samples for fine-needle aspiration for nonpalpable breast lesions in a multicenter clinical trial: The radiologic diagnostic oncology group 5 study. *Cancer: Interdisciplinary International Journal of the American Cancer Society*, **82**(4), 679–688.

Rosenbaum, P. R. and Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, **39**(1), 33–38.

Rusticus, S. A. and Lovato, C. Y. (2014). Impact of sample size and variability on the power and type i error rates of equivalence tests: A simulation study. *Practical Assessment, Research, and Evaluation*, **19**(1), 11.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics bulletin*, **2**(6), 110–114.

Schulz, K. F. and Grimes, D. A. (2002). Unequal group sizes in randomised trials: guarding against guessing. *The Lancet*, **359**(9310), 966–970.

Sharma, M., Nazareth, I., and Petersen, I. (2019). Observational studies of treatment effectiveness: worthwhile or worthless? *Clinical Epidemiology*, **11**, 35.

Siegel, S. (1956). Nonparametric statistics for the behavioral sciences, new york, 1956. *Sisdpalvelusohjesaant& (The manual of Interior Duty), Helsinki*.

Smith, H. F. (1938). An empirical law describing heterogeneity in the yields of agricultural crops. *The Journal of Agricultural Science*, **28**(1), 1–23.

Suresh, K. and Chandrashekara, S. (2012). Sample size estimation and power analysis for clinical research studies. *Journal of human reproductive sciences*, **5**(1), 7–13.

Tichy, L. and Chytry, M. (2006). Statistical determination of diagnostic species for site groups of unequal size. *Journal of Vegetation Science*, **17**(6), 809–818.

von Knebel Doeberitz, M. and Lacroix, J. (1999). Nucleic acid based techniques for the detection of rare cancer cells in clinical samples. *Cancer and Metastasis Reviews*, **18**(1), 43–64.

Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika*, **29**(3/4), 350–362.

Yeh, I.-T., Martin, M. A., Robetorye, R. S., Bolla, A. R., McCaskill, C., Shah, R. K., Gorre, M. E., Mohammed, M. S., and Gunn, S. R. (2009). Clinical validation of an array cgh test for her2 status in breast cancer reveals that polysomy 17 is a rare event. *Modern pathology*, **22**(9), 1169–1175.

Zimmerman, D. W. and Zumbo, B. D. (1993). Rank transformations and the power of the student t test and welch t'test for non-normal populations with unequal variances. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, **47**(3), 523.