International Journal of Statistics and Probability; Vol. 12, No. 3; May 2023
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

The Just-About-Right Pilot Sample Size to Control the Error Margin

Scholastica C. Obodo[1], Deirdre Toher[1], Paul White[1]

[1] Department of Data Science and Mathematics, University of the West of England, Bristol, United Kingdom

Correspondence: Paul White, Department of Data Science and Mathematics, University of the West of England, Bristol, United Kingdom

Received: February 6, 2023 Accepted: April 20, 2023 Online Published: May 23, 2023

doi:10.5539/ijsp.v12n3p1 URL: https://doi.org/10.5539/ijsp.v12n3p1

Abstract

In practice, the required sample size for a two-arm randomised controlled trial cannot always be determined pre-study with great accuracy. This lack of accuracy has economic, ethical and scientific implications. The sample size for a pilot study is an important consideration in helping the decision making for the sample size of a follow-on trial. Consideration of under- and over-estimation of the sample size results in the idea of a Just-About-Right (JAR) sample size. For studies involving a minimally clinical important difference (MCID) we present the pilot sample sizes to meet investigator desired JAR considerations.

Keywords: just-about-right, overestimation, pilot study, power, sample size, underestimation

1. Introduction

1.1 Incorrect Estimation of Sample Sizes

A pilot study is a small-scale investigation designed to test the feasibility of methods and procedures for later use on a larger scale (Thabane et al, 2010). In clinical studies, a pilot randomised controlled trial (RCT) could be used to help in the planning of a proposed substantive RCT (power = 0.8) or definitive RCT (power >= 0.9). The pilot RCT provides a means to collect preliminary data on safety, is used to assess the recruitment rate and the degree of participant retention, provides data on willingness to be randomised, and crucially, to provide estimates of variation in outcomes measures to assist the decision-making process for the sample size of the follow-on trial (Lancaster et al, 2004, Ln 2005, Arnold et al, 2009). This latter consideration begs the question, on how to determine the optimal RCT pilot sample size for any given context, with the aim of being able to estimate accurately the sample size requirements of the proposed follow-on RCT.

One of the most common errors in any type of empirical scientific research is an insufficient sample size (Makin, 2019). Small sample sizes can lead to Type Two errors (false negatives) and in practice this is especially true when combined with moderately low or low effect sizes. Small sample sizes can leave a research community in some doubt as to whether effects are real. There is also the position that it is unethical to ask participants to commit to taking part in a study which is insufficiently powered to meet objectives (Altman 1980, Halpern et. al. 2002). In addition, any such study would be an uneconomic use of resources. On the contrary, having too large a sample size could also be problematic. A sample size might be considered too large if the same quality of conclusions could have been obtained with a much smaller sample size. If the sample size is too large, then this too may be considered an uneconomic use of resources and it may be deemed unethical to be randomly allocating any excess sample size to control or intervention irrespective of whether intervention confers a benefit or not. In summary, for any substantive or definitive trial, the sample size should be sufficient to achieve worthwhile results, but not so large as to involve unnecessary recruitment of participants. Guidance is needed to allow research teams, ethics committees, funding panels, data monitoring committees, and protocol reviewers to evaluate whether a study intends to recruit too many participants (overpowered) or too few participants (underpowered) and it is important to get a just-about-right (JAR) sample size which is not too small, not too large but just-about-right.

A well conducted pilot study could be instrumental in helping to determine a JAR sample size for the follow-on study. Extant literature provides some rules-of-thumb for pilot sample size estimation. Julious (2005) noted that the marginal additional sample information content decreases with each unit increase in sample size and recommended a sample size of at least n = 12 per group. Similarly, Birkett and Day (1994) suggest 20 per arm, Kieser and Wassmer (1996) suggest between 20 to 40 per arm be used when the main trial requires between 80 to 250, Teare et al, (2014) suggest ≥70, and Browne (1995) indicates that n = 30 per arm is commonplace practice. However, the prevailing sentiment is that a simple one-size-fits-all solution or one rule-of-thumb would be inadequate when context specific considerations apply.

In terms of context specific considerations, Browne (1995), considered determining sample size for a two-arm parallel RCT study when (a) a pilot study is used to collect preliminary data on outcome variation and (b) the minimum clinically important difference (MCID) is pre-specified and (c) the follow-on study is to be adequately powered to detect an effect and (d) an assumption of normally distributed outcome data can be made. For these situations, the required per arm sample size, n, is given by

$$n = \frac{2\left(Z_{1-\alpha/2} + Z_{1-\beta}\right)^2 \sigma^2}{(\mu_1 - \mu_2)^2} \tag{1}$$

where $\mu_1 - \mu_2$ is the true mean difference or MCID, $Z_{1-\alpha/2}$, and $Z_{1-\beta}$ are standardised normal deviates for two-sided significance testing with nominal significance level α and required power $1-\beta$, and σ^2 is the population variance for the outcome measure assumed to be equal between arms. Although the MCID might be specified by hypothesis, the true population variance σ^2 would be unknown. The pilot study would provide a sample estimate for σ^2, but this sample estimate s^2, would most likely underestimate the population variance σ^2, since $(m_1 + m_2 - 2)\, s^2/\sigma^2 \backsim \chi^2_{m_1+m_2-2}$ where m_1 and m_2 denotes the sample sizes in the two arms of the pilot study. It is well known that chi-square distributions are positively skewed, hence using s^2 in place of σ^2 in the above formula would typically produce an estimated sample size lower than truly required. For this reason, Browne (1995) cautiously suggested estimating and replacing σ^2 in the sample size formula with the estimated $100(1-\gamma)$ per cent one-sided upper confidence limit (UCL) for σ^2. Specifically, the sample size per arm, for 1:1 randomisation under Browne's suggested approach is given by

$$\hat{n} = \frac{2\left(Z_{1-\alpha/2} + Z_{1-\beta}\right)^2 k s^2}{(\mu_1 - \mu_2)^2} \tag{2}$$

where s^2 is the sample pooled variance and ks^2 is the $100(1-\gamma)$ percent one-sided upper confidence limit (UCL) for σ^2. The quantity $100(1-\gamma)$ is the "coverage" i.e., the percentage of times that the predicted sample size per arm, $\hat{n}$, would exceed the true required sample size per arm n. From a practical perspective, Browne advocated a coverage of 80% (0.8) or a coverage of 90% (0.9).

1.2 Browne's Method

Simulation work conducted by Browne (1995) and Obodo et al, (2021) confirms that the approach considered by Browne has merit, achieving the required coverage of 0.8 or 0.9 as appropriate, for $\alpha = 0.01, \alpha = 0.05$, $\beta = 0.2, \beta = 0.1$, and for a range of effect sizes (small, medium, large) and for a range of pilot sample sizes between 5 and 100. However, Obodo et al, (2021) show that the procedure can produce underpowered studies, or frequently produce an intolerably large degree of excess, and that the extent of the problem depends on pilot sample size per arm (m), level of coverage $(1-\gamma)$ but not on significance level $\alpha = 0.01, 0.05$, nor on power $1-\beta = 0.8, 0.9$, nor on MCID. Both coverage and pilot sample size are at the control of an investigator at the trial planning stage. We therefore sought to quantify the relationship between pilot sample size and JAR requirements for coverage of 0.8 and 0.9 separately.

We operationalise an investigator chosen JAR interval to be $[n - \lambda_1 n, n + \lambda_2 n]$ where $\lambda_1, \lambda_2 \in [0,1]$, are investigator chosen parameters to prevent the degree of underpowering (λ_1) and degree of overpowering (λ_2). We aim for trialists to be able to justify pilot sample size and to make a statement to the effect of "The proposed two group pilot study will have a sample size of m per arm. This sample size is chosen so that the resultant power calculations for a larger study will have $100(1-\gamma)\%$ chance of exceeding the minimum required sample size and which in a two-sided test with significance level α will have $100(1-\beta)\%$ power for detecting a difference between arms assuming a MCID of $(\mu_1 - \mu_2)$. This proposed pilot sample size of m per arm will ensure that the estimated sample size will lie in the interval $(1-\lambda_1)n$ to $(1+\lambda_2)n$, with probability π providing a safeguard for under- and over- powering." In this statement we consider $\alpha = 0.01, 0.05$, power $(1-\beta) = 0.8, 0.9$, coverage $(1-\gamma) = 0.8, 0.9$, any value for MCID, lower bounds $\lambda_1 = 0.1, 0.2$ and upper bounds $\lambda_2 = 0.1, 0.2, 0.3, 0.4$ for any chosen level of π.

2. Monte Carlo Simulation Design

The Monte Carlo simulations are informed by Browne (1995) and mimic the design given by Obodo et al, (2021). In brief, we consider the two-arm parallel RCT with 1:1 randomisation which is to be analysed using the independent samples t-test (equal variances assumed, two-sided, alpha = 0.05, 0.01). The true sample size for the RCT is calculated for desired power (0.8 or 0.9), for a specified MCID corresponding to a small, medium or large effect (0.1, 0.4, 0.75) assuming equal variances $(\sigma^2 = 1)$ under an assumption of normality.

For pilot samples sizes (m = 5, 10, 30, 50, 100) the 80% and 90% upper one-sided confidence limit for the pooled sample variance is used in Browne's formula. The percentage of times that the estimated sample size, $\hat{n}$ is in the interval $[n - \lambda_1 n, n + \lambda_2 n]$ is recorded for λ_1 = 0.1, 0.2, and λ_2 = 0.1, 0.2, 0.3, 0.4, 0.5. Table 1 summarises the factor levels for the 2 by 2 by 2 by 3 by 5 fully crossed design.

Table 1. Parameter combinations

FACTOR	Number of Levels	LEVELS
Power	2	0.8, 0.9
Significance level	2	0.01, 0.05
Coverage level	2	0.8, 0.9
Effect size	3	0.10, 0.40, 0.75
Pilot sample size	5	5, 10, 30, 50, 100

Simulation was done using the R programming language with 100,000 replicates (as against Browne 1995 who used 2,000 replicates) for each cell of the design to obtain more precise simulation values.

3. Results

Table 2 summarises the percentage of times the estimated sample size, $\hat{n}$, for the follow-on study would be in the interval $[n - \lambda_1 n,\ n + \lambda_2 n]$ for λ_1 = 0.1, 0.2, λ_2 = 0.1, 0.2, 0.3 for m = 5(5)100, and for coverage $(1 - \gamma)$ = 0.8, 0.9. Simulation percentages are aggregated over significance level α = 0.01, 0.05, over prior reasoned statistical power $(1 - \beta) = 0.8, 0.9$ and assumed effect size $\mu_1 - \mu_2$ = 0.1, 0.4, 0.75 as it is known that these factors do not affect the estimated sample size (Obodo et al, 2021).

Inspection of Table 2 and Figure 1, clearly shows the percentage within any given interval monotonically increases with increasing pilot sample size for each of coverage = 0.8 and for coverage = 0.9. It is also clear that the percentage in any given interval is greater for coverage = 0.8 compared with coverage = 0.9 and this is only to be expected since, for any estimated sample size, the sample size for when coverage is 0.9 must be greater than the sample size when a tolerance for coverage is set to be equal to 0.8. Table 2 and Figure 1 show that the percentage of instances within an interval is particularly sensitive to the upper bound λ_2 which naturally follows from the positively skewed chi-square distribution used in the estimation process.

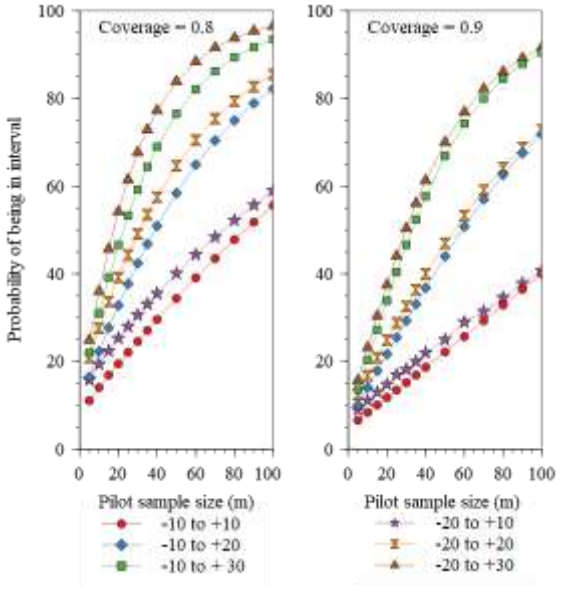

Figure 1. Percentage of simulation instances $100\hat{\pi}$ in the interval $[n - \lambda_1 n, n + \lambda_2 n]$ for λ_1 = 0.1, 0.2, λ_2 = 0.1, 0.2, 0.3, for m = 5(5)100, and for coverage $(1 - \gamma)$ = 0.8, 0.9

Table 2. Percentage of simulation instances $100\hat{\pi}$ in the interval $[n - \lambda_1 n,\ n + \lambda_2 n]$ for λ_1 = 0.1, 0.2, λ_2 = 0.1, 0.2, 0.3, for m = 5(5)100, and for coverage $(1 - \gamma)$ = 0.8, 0.9

Coverage = 0.8						
	λ_1 0.1	λ_1 0.1	λ_1 0.1	λ_1 0.2	λ_1 0.2	λ_1 0.2
m	λ_2 0.1	λ_2 0.2	λ_2 0.3	λ_2 0.1	λ_2 0.2	λ_2 0.3
5	11.1	16.5	22.0	15.8	20.9	24.9
10	14.1	22.3	31.0	19.4	27.6	36.0
15	16.9	27.7	39.2	22.4	33.7	45.7
20	19.5	32.8	46.6	25.3	39.2	54.2
25	22.1	37.7	53.3	28.0	44.3	61.5
30	24.6	42.4	59.2	30.6	49.1	67.7
35	27.1	46.8	64.4	33.1	53.5	72.9
40	29.6	50.9	69.0	35.5	57.5	77.3
50	34.4	58.4	76.5	40.1	64.6	83.9
60	39.1	64.9	82.1	44.4	70.5	88.4
70	43.5	70.4	86.2	48.4	75.4	91.6
80	47.8	75.0	89.4	52.2	79.4	93.8
90	51.8	78.9	91.7	55.7	82.7	95.4
100	55.6	82.2	93.5	59.0	85.5	96.5
Coverage = 0.9						
m	λ_1 0.1	λ_1 0.1	λ_1 0.1	λ_1 0.2	λ_1 0.2	λ_1 0.2
	λ_2 0.1	λ_2 0.2	λ_2 0.3	λ_2 0.1	λ_2 0.2	λ_2 0.3
5	6.6	10.1	13.6	9.0	12.5	15.7
10	8.4	14.1	20.4	11.1	16.8	23.2
15	10.1	17.9	27.2	13.0	20.9	30.4
20	11.8	21.7	33.9	14.8	24.9	37.5
26	13.5	25.5	40.4	17.0	28.8	44.1
30	15.2	29.3	46.6	18.2	32.6	50.4
35	16.9	33.1	52.4	20.0	36.4	56.1
40	18.7	36.8	57.7	22.0	40.0	61.3
50	22.2	44.0	66.9	25.0	47.0	70.0
60	25.7	50.8	74.3	29.0	53.3	76.9
70	29.3	57.0	80.0	31.4	59.1	82.2
80	32.9	62.6	84.5	34.6	64.2	86.1
90	36.5	67.5	87.9	37.8	68.8	89.2
100	40.0	71.9	90.5	40.8	72.8	91.6

The monotonic trends between $\hat{\pi}$ and pilot per arm sample size m, for each interval $[n - \lambda_1 n,\ n + \lambda_2 n]$ and each level of coverage has been modelled using linear regression with the functional form $\ln(\hat{\pi}) = b_0 + b_1\sqrt{m}$. Thus, for instance, when coverage = 0.8 and the interval $n \pm 0.1n$ is considered then it is readily verified that $\ln(\hat{\pi}) = -2.745 + 0.297\sqrt{m}$ and that the overall goodness-of-fit, $100R^2$, is 96.3%. Table 3 provides the estimated intercepts, gradients and goodness of fit for λ_1= 0.1, 0.2; $\lambda_2 = 0.1, 0.2, 0.3, 0.4\ 0.5$ for coverage 0.8 and coverage 0.9.

Table 3. Regression equations of the form $\ln(\pi) = b_0 + b_1\sqrt{m}$ giving estimated intercept (b_0), gradient (b_1), coefficient of determination (R-squared) for λ_1 = 0.1, 0.2, and λ_2 = 0.1, 0.2, 0.3, 0.4, 0.5

Coverage = 0.8				
Lower Percentage $(100\ \lambda_1)$	Upper Percentage $(100\ \lambda_2)$	Intercept	Gradient	R- Squared
10	10	-2.745	.297	.963
10	20	-2.531	.406	.988
10	30	-2.399	.506	.993
10	40	-2.094	.543	.981
10	50	-1.697	.527	.952
20	10	-2.256	.262	.954
20	20	-2.228	.400	.989
20	30	-2.375	.569	.997
20	40	-2.613	.759	.997
20	50	-2.557	.853	.998
Coverage = 0.9				
Lower Percentage $(100\ \lambda_1)$	Upper Percentage $(100\ \lambda_2)$	Intercept	Gradient	R- Squared
10	10	-3.306	.290	.957
10	20	-3.082	.402	.984
10	30	-3.029	.528	.995
10	40	-3.028	.656	.998
10	50	-2.827	.712	.991
20	10	-2.872	.250	.955
20	20	-2.795	.378	.986
20	30	-2.856	.524	.995
20	40	-3.108	.716	.996
20	50	-3.450	.919	.993

For any level of coverage and any interval, any regression equation in Table 3 may be re-written in terms of pilot sample size i.e., $m = ([\ln(\hat{\pi}) - b_0]/b_1)\text{^}2$. Solution of this will give an estimated pilot sample size per arm, m, for any required percentage for the given interval.

Table 4 shows the pilot sample size per arm (m) needed to have a required probability (π) of being in a given interval $[n - \lambda_1 n,\ n + \lambda_2 n\ ]$ for coverage of 0.8 or coverage 0.9. Thus, for instance, if an investigator requires an 80% chance of not being underpowered for a definitive trial (coverage = 0.8) and requires a 70% chance $(\pi = 0.7)$ of being within $\pm$ 10% of the true required sample size $(\lambda_1 = 0.1,\ \lambda_2 = 0.1)$ then a sample size per arm (m) of 65 is needed for any given MCID.

Table 4. Pilot sample size (m) required for a required proportion (π) to be in the interval $[n - \lambda_1 n,\ n + \lambda_2 n\,]$ for a given coverage.

π	λ_1 0.1 λ_2 0.1	λ_1 0.1 λ_2 0.2	λ_1 0.1 λ_2 0.3	λ_1 0.2 λ_2 0.1	λ_1 0.2 λ_2 0.2	λ_1 0.2 λ_2 0.3
Coverage = 0.8						
0.50	48	20	11	36	15	9
0.55	52	23	13	40	17	10
0.60	56	25	14	44	18	11
0.65	61	27	15	49	20	12
0.70	65	29	16	53	21	13
0.75	68	31	17	56	23	13
0.80	72	32	18	60	25	14
0.90	79	36	21	67	28	16
Coverage = 0.9						
0.50	81	35	20	76	31	17
0.55	87	38	21	83	34	18
0.60	93	41	23	89	37	20
0.65	98	43	24	95	39	22
0.70	103	46	26	101	42	22
0.75	108	48	27	107	43	24
0.80	113	50	28	112	46	25
0.90	121	54	31	122	51	27

4. Discussion and Conclusion

Pilot studies are conducted for a variety of reasons. One such reason is to help determine variation in outcome measures to help plan the required sample size for a large-scale substantive or definitive follow-on study. The preceding sections consider the situation where the MCID can be pre-specified for a scale outcome variable and an assumption of normality is reasonable.

Sample size may be calculated if parameters are either known or can be reasonably estimated. For instance, in a two-arm study, if for example MCID = 0.2, variance = 1, alpha = 0.05, beta = 0.10, then the required sample size may be verified to n = 526 per arm (complete data set after any missing data). In practice the variation of the outcome measure may not be known but may be estimated by collecting pilot data. In these regards, Browne's method, may be used to estimate a required sample size with either 80% or 90% coverage i.e. the estimated sample size has an 80% or 90% chance of exceeding the required minimum sample size. A problem with this approach is the chance of underestimating the required sample size, or in having an estimated sample size which far exceeds the required sample size (see Obodo et al, 2021). We considered a strategy to curb these excesses so that estimated sample sizes would be "not too small" and "not too large" in comparison to the true required but unknown sample size, by considering a just-about-right (JAR) sample size. The chosen coverage (say 80% or 90%) is not dependent on pilot sample size. However, with a given level of coverage a researcher may wish to ensure that the probability of the margin of error attached to any estimate is pre-specified to be within an interval around the true required sample size e.g. 70% chance of being within 10% of the required sample size. By inspection of Table 4, if 80% coverage is required with a 70% chance of being within +/- 10% of true sample size, then the pilot study would require at least $m = 65$ per arm. The protocol may then contain a summary "The proposed two group pilot study will have a sample size of 65 per arm. This sample size is chosen so that the resultant power calculations for sample size in a larger study will have an 80% chance of exceeding the minimum required sample size and which in a two-sided test with significance level α will have $100(1 - \beta)\%$ power for detecting an effect assuming an MCID of $(\mu_1 - \mu_2)$. This proposed pilot sample size of 65 per arm will ensure that the estimated sample size will have a 70% chance of being in an interval of +/- 10% of the true required sample size providing a safeguard over under- and over-

powering."

The pilot sample sizes given in this article (Table 4) is predicated on an MCID. If the true effect size exceeds the MCID then the follow-on study is likely to be overpowered to detect a difference (the lesser of the two possible errors). If the true effect is smaller than the MCID then any effect smaller than the MCID is not of clinical interest and may go undetected.

The pilot sample sizes given in this article are based on assumptions of normality and equal variance. In these regards, the practical utility of the pilot sample size recommendations needs further investigation for variance heterogeneity and non-normal distributions including binary outcomes. In a similar way, other simulations may consider the two arm pre- post- RCT design with repeated measures ANCOVA as the analysis strategy. As such the given pilot sample sizes are restricted to the stated assumptions with a direct parametric comparison between the two groups.

References

Arnold, D. M., Burns, K. E., Adhikari, N. K., Kho, M. E., Meade, M. O., & Cook, D. J. (2009). The design and interpretation of pilot trials in clinical research in critical care. *Critical care medicine*, *37*(1), S69-S74. https://doi.org/10.1097/CCM.0b013e3181920e33

Altman, D. G. (1980). Statistics and ethics in medical research: III How large a sample? *British medical journal*, *281*(6251), 1336. https://doi.org/10.1136/bmj.281.6251.1336

Birkett, M. A., & Day, S. J. (1994). Internal pilot studies for estimating sample size. *Statistics in medicine*, *13*(23-24), 2455-2463. https://doi.org/10.1002/sim.4780132309

Browne, R. H. (1995). On the use of a pilot sample for sample size determination. *Statistics in medicine*, *14*(17), 1933-1940.

Halpern, S. D., Karlawish, J. H., & Berlin, J. A. (2002). The continuing unethical conduct of underpowered clinical trials. *Jama*, *288*(3), 358-362. https://doi.org/10.1001/jama.288.3.358

In, J. (2017). Introduction of a pilot study. *Korean journal of anesthesiology*, *70*(6), 601-605. https://doi.org/10.4097/kjae.2017.70.6.601

Julious, S. A. (2005). Sample size of 12 per group rule of thumb for a pilot study. *Pharmaceutical Statistics: The Journal of Applied Statistics in the Pharmaceutical Industry*, *4*(4), 287-291. https://doi.org/10.1002/pst.185

Kieser, M., & Wassmer, G. (1996). On the use of the upper confidence limit for the variance from a pilot sample for sample size determination. *Biometrical journal*, *38*(8), 941-949. https://doi.org/10.1002/bimj.4710380806

Lancaster, G. A., Dodd, S., & Williamson, P. R. (2004). Design and analysis of pilot studies: recommendations for good practice. *Journal of evaluation in clinical practice*, *10*(2), 307-312. https://doi.org/10.1111/j..2002.384.doc.x

Makin, T.R. and Orban de Xivry, J.J., 2019. Ten common statistical mistakes to watch out for when writing or reviewing a manuscript. *Elife*, *8*, p.e48175. https://doi.org/10.7554/eLife.48175

Obodo, S., Toher, D., & White, P. (2021). Estimation of the two-group pilot sample size with a cautionary note on Browne's formula. *Journal of Applied Quantitative Methods*, *16*(3).

Teare, M. D., Dimairo, M., Shephard, N., Hayman, A., Whitehead, A., & Walters, S. J. (2014). Sample size requirements to estimate key design parameters from external pilot randomised controlled trials: a simulation study. *Trials*, *15*, 1-13. https://doi.org/10.1186/1745-6215-15-264

Thabane, L., Ma, J., Chu, R., Cheng, J., Ismaila, A., Rios, L. P., ... & Goldsmith, C. H. (2010). A tutorial on pilot studies: the what, why and how. *BMC medical research methodology*, *10*, 1-10. https://doi.org/10.1186/1471-2288-10-1

Copyrights