Towards improving quality of the evidence base for medical decision-making

Jansen, M.S.

Part I.

CHALLENGES RELATED TO GENERATION AND DISSEMINATION OF EVIDENCE

Chapter 2

The power of sample size calculations

Marieke S. Jansen, Rolf H.H. Groenwold, Olaf M. Dekkers

European Journal of Endocrinology. 2024 Oct 29;191(5):E5-E9

Abstract

Researchers frequently come across sample size calculations in the scientific literature they read, in projects undertaken by their peers, and likely within their own work. However, despite its ubiquity, calculating a sample size is often perceived as a hurdle and not fully understood. This paper provides a brief overview of sample size estimation to guide readers, researchers, and reviewers through its fundamentals.

Significance

Sample size estimation not only plays a key role in the design of confirmatory studies, but also in the interpretation of their results. This brief guide aims to help clinical researchers understand the basics and avoid common pitfalls.

Introduction

Sample size estimation is an essential component of the design of any medical study, it is required for ethics approval and should be reported in the publication.[1-3] However, its relevance and the implications of inadequately calculated sample sizes are not always understood. What is more, the multitude of options to choose methods, parameters, programs or web applications for sample size calculations is huge and sometimes confusing. In this paper we provide readers with a basic introduction into sample size calculations. With a focus on experimental research (clinical trials), we discuss the relevance of the sample size calculations with brief illustrations. Thereafter, we discuss several pitfalls, as well as the role of sample size calculations in observational research.

Why calculate a sample size?

A sample size calculation informs how many participants should be included to detect a particular treatment effect (or harmful effect, in case of a risk factor), should it exist. Including too many participants in experimental research has clear drawbacks: it would unnecessarily increase costs and patient efforts, and could unnecessarily expose them to unwanted adverse effects, as well as to a less effective treatment option should the experimental treatment turn out inferior to the standard of care.[4] However, including too few participants could lead to failure to detect an effect, even if it exists.[4] If, nevertheless, an effect is detected with (too) few participants, the effect is likely inflated.[5] After all, for an effect to exceed the threshold of statistical significance in an underpowered study, the effect must be very large or (at least partly) be the result of chance. A study too small may thus yield inconclusive or even misleading results. An optimal sample size is crucial to contribute meaningfully to scientific literature and prevent research waste. This particularly holds for confirmatory studies, which aim to confirm or refute a hypothesised effect. If – upfront – researchers lack clarity on the potential effect's direction and magnitude, it will be difficult to design a study with an appropriate sample size, and this will limit the subsequent interpretation of the results. In contrast, exploratory studies are typically aimed at generating new hypotheses rather than confirming or refuting them, and therefore often require different considerations regarding sample size and interpretation of study results (see penultimate section).[6]

How to calculate a sample size?

The first step when calculating a sample size concerns determining the research objective (e.g., superiority or non-inferiority) and (primary) endpoint of interest.[7] This should ideally be translated into an 'estimand', which is a precise, formal description of the treatment effect to be estimated in the study.[8,9] A conventional sample size calculation (i.e., superiority two-arm parallel trial with 1:1 treatment allocation) requires four main ingredients:

1. Target difference

The target difference is the smallest difference between the effects of two treatments, that should be reliably detectable, if it indeed exists in the population of interest (or an effect more extreme). The target difference is the main driver of the required sample size; if it is reduced by half, the sample size will quadruple.[7] Obviously, estimating this difference is the primary objective of the study and therefore by default unknown; hence its magnitude is based on assumptions. It is important that it represents a difference that is considered clinically relevant to one or more stakeholder groups (such as patients and clinicians), and that it is realistic (based on existing evidence, or expert opinion).[10] There are multiple strategies to inform the target difference. For some endpoints and disease areas, previous research and/or guidelines have helped define minimal

clinically important differences, such as change in HbA1c or change in body weight for type 2 diabetes.[11] Other strategies may for example include seeking expert opinion (e.g., patient representatives and clinicians), review of literature, and results of pilot studies.[10,12]

2. Variability in the population

The variability, or variance, indicates how much variation is observed in the endpoint. For continuous endpoints, such as HbA1c, this variation can be captured by the standard deviation. Existing literature or pilot studies can be useful to inform this parameter for sample size calculations. If these are unavailable or of low quality, researchers could opt for a planned interim analysis to re-estimate the required sample size based on the variation observed in the first data of the trial (note, in such interim analysis no treatment effect is estimated, nor is a formal statistical test of the treatment effect performed).

3. Type I error (alpha)

The type I error probability, or alpha, is the probability of incorrectly rejecting the null hypothesis when it is true (i.e., false-positive result). It corresponds with the significance level of a hypothesis test and should thus reflect the intended significance level of the statistical analysis plan. Traditionally, (two-sided) alpha is set at 5%, but other values, such as 2.5%, could also be considered.

4. Type II error (beta)

The type II error probability, or beta, is the probability of failing to reject the null hypothesis when the alternative hypothesis is true (i.e., false-negative result). A related concept, 'power', represents the probability of detecting a true effect, should it exist, and is calculated as 1 – beta. Type II errors are often considered less critical than type I errors; beta is typically set at 10 or 20%.

In **Box 1** and **Box 2**, examples are given of how to calculate the required sample size for a continuous and a binary endpoint by hand. Most statistical software programs, such as SPSS, STATA, and R include options to calculate sample sizes. PASS is reliable software entirely dedicated to sample size and power calculations. For simple sample size calculations, certain web applications can also offer reliable, easy-to-use alternatives.[13,14] It is advisable to have sample size calculations reviewed by others, as errors in the initial calculation are not easily corrected once the study is underway (or worse, impossible once finished). Additionally, consulting a statistician should be considered, particularly in case of more complex designs. For further reading on sample size estimation, including calculations for other endpoints (e.g., time-to-event), objectives (e.g., equivalence, non-inferiority) and allocation ratios, a plethora of resources are available.[7,15,16]

What if the required sample size is too large?

After having calculated the required sample size, the resulting number of participants might be higher than considered feasible. There are different ways to proceed. First, a reasonable option could be to simply conclude that the current research objective is infeasible, and that research resources are better spent elsewhere. Second, researchers may explore whether options exist to increase recruitment (for example, by including more study sites or relaxing eligibility criteria). Alternatively, some may be tempted to 'tweak' the parameters used for the sample size calculation so that the required number of participants is reduced. We generally advise against doing this. Increasing the target difference (or reducing the expected variation) results in the issues outlined before; clinically important effects are

Box 1. Example of a sample size calculation for a continuous endpoint

Suppose we are interested in comparing the effect of two antidiabetic medications on Hb1Ac in patients with type 2 diabetes. The primary outcome of interest is percentage change of Hb1Ac at 40 weeks of treatment (a continuous endpoint). On the basis of existing literature, we consider a difference of 0.5 percentage point between drug A and B in Hb1Ac reduction to be realistic and clinically relevant. Prior studies have further shown that the percentage change of Hb1Ac after treatment with drug A and B are normally distributed with a standard deviation of approximately 1.0 percentage point. We accept a conventional type I error probability of 5% (for a two-sided hypothesis) and a type II error probability of 20% (which corresponds with 80% power). We can then calculate the required sample size using **Formula 1**:

$$\frac{2(0.84 + 1.96)^2 1.0^2}{0.5^2} \approx 63,$$

which is the number of participants per treatment arm. Accounting for potential dropouts in the study, can be done by dividing the sample size by 1 – attrition probability (assuming non-informative censoring). For an expected attrition of 10%, this implies 63/0.9 = 70 participants per arm, or 140 participants in total.

Formula 1

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

n	=	sample size per treatment arm
d	=	target effect size (difference)
σ	=	(pooled) standard deviation
α	=	type I error probability (significance level)
β	=	type II error probability (power = $1 - \beta$)
$Z_{1-\beta}$	=	corresponding Z-value for $1 - \beta$ of the normal cumulative distribution function
$Z_{1-\alpha/2}$	=	corresponding Z-value for $1 - \alpha/2$ of the normal cumulative distribution function

more likely to go undetected, while significant effects (or even close to significance) are likely to be severely exaggerated.[17] Tweaking the sample size increases the probability of imprecise, difficult-to-interpret study results (e.g., is there truly no relevant effect or did we miss it?). If there are compelling arguments to conduct a confirmatory study with a (too) small sample size, we urge researchers to be transparent about this and opt for a compromise in power (i.e., increasing beta), instead of choosing an unrealistically large (overly optimistic) target difference to mask sample size problems and mislead readers.

Box 2. Example of a sample size calculation for a binary endpoint

Suppose we are interested in comparing the effects of two antidiabetic medications on the percentage of type 2 diabetic patients that achieve Hb1Ac of <6.5% at 40 weeks (a binary endpoint). Based on previous studies we expect that the proportion of patients that will achieve the outcome under standard of care is 50%. A difference of 10 percentage points was considered to be clinically relevant and realistic by experts. Therefore, we hypothesize that in the experimental arm, the proportion of patients with the outcome will be 60%. We accept a conventional type I error probability of 5% (for a two-sided hypothesis) and a type II error probability of 20%. Using **Formula 2**, this yields the required sample size:

$$\frac{(0.84 + 1.96)^2(0.5(1 - 0.5) + 0.6(1 - 0.6))}{(0.5 - 0.6)^2} \approx 385$$

In this scenario, we would require 385 participants per arm, or 770 participants in total.

Formula 2

$$n = \frac{(Z_{1-\beta} + Z_{1-\alpha/2})^2(\pi_A(1 - \pi_A) + \pi_B(1 - \pi_B))}{(\pi_A - \pi_B)^2}$$

n	=	sample size per treatment arm
π_A	=	proportion of participants with the outcome in experimental arm
π_B	=	proportion of participants with the outcome in control arm
α	=	type I error probability (significance level)
β	=	type II error probability (power = $1 - \beta$)
$Z_{1-\beta}$	=	corresponding Z-value for $1 - \beta$ of the normal cumulative distribution function
$Z_{1-\alpha/2}$	=	corresponding Z-value for $1 - \alpha/2$ of the normal cumulative distribution function

Early termination: now what?

Early termination of studies occurs frequently; about one in four trials discontinues prematurely.[18] While sometimes this is part of the design (e.g., a pre-specified decision based upon a planned interim-analysis), most studies stop early due to recruitment failure or lack of funding. In such scenario, stopping before reaching the target sample size is problematic. It leads to imprecise estimates and an increased risk of both missing important effects as well as inflated effect sizes. Ideally, researchers should prevent stopping early, for example with a pilot study testing feasibility. By no means should researchers look at the data before the study is completed and stop early because the results look convincing (except for a formal interim analysis). Regardless, sometimes early termination may be inevitable. In that case, we caution against standard interpretation of the results and drawing conclusions about the magnitude of the effect of interest. Second, we encourage researchers to nevertheless disseminate their results, since these can contribute to meta-analyses, and experiences regarding recruitment and feasibility may be valuable to future researchers.

Post-hoc power calculations

Sometimes, after finishing the study, researchers report how much power their study had to detect a specified effect given the achieved sample size. These 'post-hoc power calculations' may also sometimes be requested by peer reviewers and editors. However, post-hoc power calculations are meaningless and lead to flawed interpretations.[19] It is irrelevant to calculate the probability of an event after the event has already been observed (i.e., what is the point of calculating the chance of getting heads or tails, after the coin has landed on tails?). A post-hoc power calculation does not provide an answer to whether the observed effect is indeed 'real' or close to the truth. It also doesn't add anything on top of the magnitude of the effect size and the width of the confidence interval which captures the statistical uncertainty of the effect (in fact, 'post-hoc power' is a direct function of the observed p-value).[19,20] Post-hoc power calculations have no practical value; they do not change the study results or their interpretation and are not an alternative or justification for not doing a proper sample size calculation a priori.

Role of sample size calculations in observational studies

Formal sample size calculations are less frequently performed in observational research, and their role is context dependent, where a distinction should be made between exploratory and confirmatory objectives. For exploratory objectives, sample sizes are typically determined by pragmatic considerations and convenience. The aim is to generate hypotheses about potential effects, with further research needed for confirmation (or rejection).[6] For confirmatory objectives, such as estimating causal effects to inform clinical practice, it is relevant to consider whether data need to be collected or are already available. If the data need to be collected, sample size calculations are useful in providing the number of subjects that should be included to detect a meaningful effect (after which we can decide on the study's feasibility). Note that sample size calculations for observational research require additional assumptions regarding confounders (and potentially other issues such as missing data and measurement error), which results in more uncertainty in the required sample size.[21] If the data have already been collected, a power calculation could still be considered to determine whether a specified target difference can be reliably detected with the available sample size. Additionally, it is possible to calculate the detectable differences given the sample size and a set power level. This information is valuable for assessing whether the dataset is suitable to address the research question(s) of interest and for determining whether pursuing additional datasets may be necessary, if possible. Although a formal power calculation might seem redundant when data are already available, determining beforehand which differences are clinically meaningful, will help to prevent twisting the interpretation of the results (i.e., what is relevant) afterward, particularly in the context of enormous datasets ('big data') for which small non-relevant differences likely become significant.

Final remarks

Calculating a sample size is essential for the design of confirmatory studies. Chosen parameters and their rationale should be clearly reported in the protocol and publication. In particular, the target difference is the main driver of the sample size, which should be considered clinically relevant and realistic (plausible).[10] Common pitfalls, such as tweaking of sample size parameters to reduce participant numbers as well as post-hoc power calculations, should be avoided. While we did not discuss sample size calculations for particular designs (e.g., equivalence or non-inferiority objectives, cluster RCTs, case-control studies, prediction models), the basic principles and pitfalls discussed apply to these situations as well.[15,22-24]

References

1. Chan AW, Tetzlaff JM, Altman DG, et al. SPIRIT 2013 Statement: defining standard protocol items for clinical trials. *Rev Panam Salud Publica*. Dec 2015;38(6):506-14.
2. Schulz KF, Altman DG, Moher D, Group C. CONSORT 2010 Statement: Updated guidelines for reporting parallel group randomised trials. *J Clin Epidemiol*. Aug 2010;63(8):834-40. doi:10.1016/j.jclinepi.2010.02.005
3. von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Epidemiology*. Nov 2007;18(6):800-4. doi:10.1097/EDE.0b013e3181577654
4. Altman DG. Statistics and ethics in medical research: III How large a sample? *Br Med J*. Nov 15 1980;281(6251):1336-8. doi:10.1136/bmj.281.6251.1336
5. Ioannidis JP. Why most discovered true associations are inflated. *Epidemiology*. Sep 2008;19(5):640-8. doi:10.1097/EDE.0b013e31818131e7
6. Luijken K, Dekkers OM, Rosendaal FR, Groenwold RHH. Exploratory analyses in aetiologic research and considerations for assessment of credibility: mini-review of literature. *BMJ*. May 3 2022;377:e070113. doi:10.1136/bmj-2021-070113
7. Julious SA. *Sample Sizes for Clinical Trials*. 2nd ed. Chapman and Hall/CRC; 2023.
8. ICH E9 (R1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials. 2019. Accessed Aug 28, 2024. https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e9-r1-addendum-estimands-and-sensitivity-analysis-clinical-trials-guideline-statistical-principles-clinical-trials-step-5_en.pdf
9. Lawrance R, Degtyarev E, Griffiths P, et al. What is an estimand & how does it relate to quantifying the effect of treatment on patient-reported quality of life outcomes in clinical trials? *J Patient Rep Outcomes*. Aug 24 2020;4(1):68. doi:10.1186/s41687-020-00218-5
10. Cook JA, Julious SA, Sones W, et al. DELTA(2) guidance on choosing the target difference and undertaking and reporting the sample size calculation for a randomised controlled trial. *Trials*. Nov 5 2018;19(1):606. doi:10.1186/s13063-018-2884-0
11. Dankers M, Nelissen-Vrancken M, Hart BH, Lambooij AC, van Dijk L, Mantel-Teeuwisse AK. Alignment between outcomes and minimal clinically important differences in the Dutch type 2 diabetes mellitus guideline and healthcare professionals' preferences. *Pharmacol Res Perspect*. May 2021;9(3):e00750. doi:10.1002/prp2.750
12. Cook JA, Hislop J, Adewuyi TE, et al. Assessing methods to specify the target difference for a randomised controlled trial: DELTA (Difference ELicitation in TriAls) review. *Health Technol Assess*. May 2014;18(28):v-vi, 1-175. doi:10.3310/hta18280
13. Sealed Envelope Ltd. Power calculator for continuous outcome superiority trial. 2012. Accessed Aug 28, 2024. https://www.sealedenvelope.com/power/continuous-superiority/
14. Dupont WD, Plummer WD. PS: Power and Sample Size Calculation. 2018. Accessed Aug 28, 2024. https://cqsclinical.app.vumc.org/ps/
15. Flight L, Julious SA. Practical guide to sample size calculations: non-inferiority and equivalence trials. *Pharm Stat*. Jan-Feb 2016;15(1):80-9. doi:10.1002/pst.1716
16. Flight L, Julious SA. Practical guide to sample size calculations: superiority trials. *Pharm Stat*. Jan-Feb 2016;15(1):75-9. doi:10.1002/pst.1718
17. van Zwet E, Gelman A, Greenland S, Imbens G, Schwab S, Goodman SN. A New Look at P Values for Randomized Clinical Trials. *NEJM Evid*. Jan 2024;3(1):EVIDoa2300003. doi:10.1056/EVIDoa2300003

18. Speich B, Gryaznov D, Busse JW, et al. Nonregistration, discontinuation, and nonpublication of randomized trials: A repeated metaresearch analysis. *PLoS Med*. Apr 2022;19(4):e1003980. doi:10.1371/journal.pmed.1003980

19. Hoenig JM, Heisey DM. The abuse of power: The pervasive fallacy of power calculations for data analysis. *Am Stat*. Feb 2001;55(1):19-24. doi:Doi 10.1198/000313001300339897

20. Althouse AD. Post Hoc Power: Not Empowering, Just Misleading. *J Surg Res*. Mar 2021;259:A3-A6. doi:10.1016/j.jss.2019.10.049

21. Haneuse S, Schildcrout J, Gillen D. A two-stage strategy to accommodate general patterns of confounding in the design of observational studies. *Biostatistics*. Apr 2012;13(2):274-88. doi:10.1093/biostatistics/kxr044

22. Riley RD, Ensor J, Snell KIE, et al. Calculating the sample size required for developing a clinical prediction model. *BMJ*. Mar 18 2020;368:m441. doi:10.1136/bmj.m441

23. Hemming K, Eldridge S, Forbes G, Weijer C, Taljaard M. How to design efficient cluster randomised trials. *BMJ*. Jul 14 2017;358:j3064. doi:10.1136/bmj.j3064

24. Groenwold RHH, van Smeden M. Efficient Sampling in Unmatched Case-Control Studies When the Total Number of Cases and Controls Is Fixed. *Epidemiology*. Nov 2017;28(6):834-837. doi:10.1097/EDE.0000000000000710