Research Techniques Made Simple: Sample Size Estimation and Power Calculation

Sigrun A.J. Schmidt[1], Serigne Lo[2,3] and Loes M. Hollestein[4,5]

Sample size and power calculations help determine if a study is feasible based on a priori assumptions about the study results and available resources. Trade-offs must be made between the probability of observing the true effect and the probability of type I errors (α, false positive) and type II errors (β, false negative). Calculations require specification of the null hypothesis, the alternative hypothesis, type of outcome measure and statistical test, α level, β, effect size, and variability (if applicable). Because the choice of these parameters may be quite arbitrary in some cases, one approach is to calculate the sample size or power over a range of plausible parameters before selecting the final sample size or power. Considerations that should be taken into account could include correction for nonadherence of the participants, adjustment for multiple comparisons, or innovative study designs.

Journal of Investigative Dermatology (2018) **138,** 1678–1682; doi:10.1016/j.jid.2018.06.165

CME Activity Dates: 19 July 2018
Expiration Date: 18 July 2019
Estimated Time to Complete: 1 hour

Planning Committee/Speaker Disclosure: All authors, planning committee members, CME committee members and staff involved with this activity as content validation reviewers have no financial relationships with commercial interests to disclose relative to the content of this CME activity.

Commercial Support Acknowledgment: This CME activity is supported by an educational grant from Lilly USA, LLC.

Description: This article, designed for dermatologists, residents, fellows, and related healthcare providers, seeks to reduce the growing divide between dermatology clinical practice and the basic science/current research methodologies on which many diagnostic and therapeutic advances are built.

Objectives: At the conclusion of this activity, learners should be better able to:
- Recognize the newest techniques in biomedical research.
- Describe how these techniques can be utilized and their limitations.
- Describe the potential impact of these techniques.

CME Accreditation and Credit Designation: This activity has been planned and implemented in accordance with the accreditation requirements and policies of the Accreditation Council for Continuing Medical Education through the joint providership of Beaumont Health and the Society for Investigative Dermatology. Beaumont Health is accredited by the ACCME to provide continuing medical education for physicians. Beaumont Health designates this enduring material for a maximum of 1.0 *AMA PRA Category 1 Credit(s)*™. Physicians should claim only the credit commensurate with the extent of their participation in the activity.

Method of Physician Participation in Learning Process: The content can be read from the Journal of Investigative Dermatology website: http://www.jidonline.org/current. Tests for CME credits may only be submitted online at https://beaumont.cloud-cme.com/RTMS-Aug18 – click 'CME on Demand' and locate the article to complete the test. Fax or other copies will not be accepted. To receive credits, learners must review the CME accreditation information; view the entire article, complete the post-test with a minimum performance level of 60%; and complete the online evaluation form in order to claim CME credit. The CME credit code for this activity is: 21310. For questions about CME credit email cme@beaumont.edu.

INTRODUCTION

Sample size and power calculations may involve estimating (i) the number of participants (sample size) required to test the prespecified hypothesis, (ii) the power to detect a given association with a fixed sample size, or (iii) the association possible to detect given a prespecified power and sample size (Case and Ambrosius, 2007).

Although many (clinical) researchers outsource the sample size calculation of study to a statistician, their expertise is required to specify outcomes to be measured and the time points and difference(s) that would be meaningful. Understanding the methodology is of utmost importance to ensure that plausible assumptions are used in the sample size calculation.

[1]*Department of Clinical Epidemiology, Aarhus University Hospital, Aarhus, Denmark;* [2]*Melanoma Institute Australia, The University of Sydney, North Sydney, New South Wales, Australia;* [3]*Institute for Research and Medical Consultations, University of Dammam, Dammam, Kingdom of Saudi Arabia;* [4]*Department of Dermatology, Erasmus MC University Medical Center, Rotterdam, The Netherlands; and* [5]*Department of Research, Netherlands Comprehensive Cancer Center, Utrecht, The Netherlands*

Correspondence: Loes Hollestein, Department of Dermatology, Erasmus MC University Medical Center, PO Box 2040, 3000 CA Rotterdam, The Netherlands. E-mail: l.hollestein@erasmusmc.nl

Abbreviations: 5-FU, 5-fluorouracil; AK, actinic keratosis

SUMMARY POINTS

- Sample size and power calculations help determine if a study is feasible based on a priori assumptions about the study results and available resources.
- Calculations require specification of the null hypothesis, the alternative hypothesis, type of outcome measure and statistical test, one or two-sided α-level, β, effect size, and variability (if applicable).
- Limitation: assumptions about the expected effect size and variability may have to be made without prior knowledge.

HYPOTHESIS TESTING

Calculation of sample size and/or study power requires precise specification of the statistical hypothesis to be tested. In the hypothesis testing procedure, two mutually exclusive assertions (the null and the alternative hypotheses) are evaluated to determine which assertion is best supported by the sample data. The logical purpose of a clinical trial is to disprove this null hypothesis (denoted H_0) in favor of an alternative hypothesis denoted H_1. The alternative hypothesis is either a two-sided hypothesis when it covers both sides of the null hypothesis or one sided when it covers only one side of the latter.

When performing hypothesis testing, researchers face two potential types of errors as shown in Figure 1. Committing a type I error is to reject the null hypothesis when it is actually true (a false positive association). The probability of this happening is equal to the statistical significance level (α), which also corresponds to the *P*-value. A type II error occurs when we fail to reject a false null hypothesis (a false negative association). This probability is termed β. Statistical power (1 − β) refers to the probability of detecting a difference if there is one.

SAMPLE SIZE CALCULATIONS

Table 1 provides an overall algorithm that can be extended to sample size calculations for most studies.

Research question

A well-formulated research question contains essential information for the sample size calculation. For example, in the Veterans Affairs Keratinocyte Carcinoma Chemoprevention (i.e., VAKCC) Trial, the investigators ran a randomized controlled trial to respond to the question *Does the use of 5-fluorouracil (5-FU) decrease the incidence rate of new actinic keratoses (AKs) among patients with AK compared with placebo during the first 2 years?* (Walker et al., 2017). This question contains relevant information about the patient population to be investigated, intervention, control group, and outcome measure (i.e., PICO), which are needed for the sample size calculation.

Study hypotheses

The next step is to state the null and alternative hypotheses. The null hypothesis for testing equality is most frequently used. In the VAKCC trial, the null hypothesis (H_0) was *The incidence rate of AK is equal between the 5-FU group and the placebo group*. The alternative hypothesis (H_1) was *The incidence rate of AK is not equal between the 5-FU group and the placebo group* (a two-sided hypothesis).

Choose outcome and corresponding statistical test

The outcome measure determines the design of the study and the type of statistical test. Therefore, an essential question when designing a study is *What is/are the most relevant outcome measure(s), and how are you going to measure it/them?* The nature of data (e.g., dichotomous, continuous, or time-to-event), number of groups, (un)paired groups, and time points of measurement will then determine the type of statistical test (Kim et al., 2017).

Effect size and variability

Infinite samples can detect any small difference, but these may not be clinically or biologically relevant. It is therefore recommended that the sample size calculation be based on the minimal (clinical) important difference. If there is no literature on the minimum relevant effect size, it should be based on expertise. Sample size calculations for continuous outcome measures require an estimate of the variability (or standard deviation). Large variability requires larger sample sizes. Methods for identifying the standard deviation for a continuous outcome include a literature search, consulting colleagues, or performing a pilot study (Hulley and Cummings, 2013).

		Reality	
		No effect	Effect
Study	No effect	✓ Fail to reject H_0 when it is true	Type II error (β) Fail to reject H_0 when it is false
	Effect	Type I error (α) Reject H_0 when it is true	✓ (Power = 1 – β) Reject H_0 when it is false

Figure 1. Hypothesis testing. Researchers face two potential types of error, α and β.

Table 1. Algorithm for sample size estimation in analytical studies

1. Formulate the research question
2. State the null hypothesis and a one- or two-sided alternative hypothesis
3. Choose the primary outcome measure and corresponding type of statistical test
4. Consider a range of plausible effect sizes and, if applicable, the variability
5. Select α and β, based on the objective, clinical considerations, and/or phase of the study
6. Use steps 1–5 to compute the sample size with a statistical package or an online calculator

Significance level (α) and power (1 − β)

Values of α and β should suit the objective, but they typically depend on the phase of the study. For example, a large false positive rate (type I error) may be more acceptable for a phase II study (Case and Ambrosius, 2007). It is important to realize that both the significance level and power are quite arbitrary figures, and thus one approach is to select a range of values and compute different sets of sample size estimates to identify the most appropriate trade-off (Case and Ambrosius, 2007).

Calculate the sample size

Based on the assumptions specified in steps 1–5, the next step is to calculate the sample size over a range of plausible parameters before selecting the final sample size. Specific formulas exist for each statistical model, and most are supported by statistical packages and various free online repositories.

POWER CALCULATIONS

Some studies have a predetermined fixed sample size. This typically includes studies based on routinely collected data. In these situations, either (i) the detectable effect size based on a given power can be estimated or (ii) the power to detect a given effect can be estimated (Hulley and Cummings, 2013). Researchers may consider plotting a power curve, with the power plotted against the effect size for their fixed sample size. If the population size is too small, the minimal detectable effect estimate will be very high, and the study may not be worthwhile. Power and sample size calculations must be performed a priori (i.e., during the study design phase). In some special circumstances, researchers may want to run post hoc analyses, but post hoc power calculations are debated and should be dealt with cautiously.

TYPES OF STUDIES

In vitro and animal studies

The concepts presented in the clinical example and Table 1 also apply to in vitro and animal studies. The expected effect size is generally larger in these studies, and thus the required sample size is smaller. As in human studies, it is important to define the end points in advance, decide how they will be measured, and identify the additional sources of variability within the experiment to ensure that the appropriate design and statistical approach have been chosen (Neuberg, 2017). In studies with cell lines, it is important to distinguish biological replicates (e.g., cells from multiple people or animals) and technical replicates (e.g., the same cell line of the same conditions measured multiple times). Technical replicates reduce the variability due to measurement error but should still be counted as a single measurement.

Genetic studies

In a genome-wide association study, hundreds of thousands of single nucleotide polymorphism markers are evaluated for an association with the outcome of interest. The association of every single nucleotide polymorphism with the outcome is considered testing of an independent hypothesis, and therefore a correction for testing multiple hypotheses should be applied. For 1 million single nucleotide polymorphism markers, a *P*-value less than 5×10^{-8} is typically considered statistically significant, which has been calculated by the Bonferroni correction (0.05/number of independent single nucleotide polymorphism markers). Because the low α level, very large sample sizes are needed to achieve adequate statistical power. The sample size for genome-wide association studies is also known to be highly affected by disease prevalence, disease allele frequency, linkage disequilibrium, and inheritance models (e.g., additive, dominant, and multiplicative models) (Hong and Park, 2012). Online sample size and power calculators can be used to take this into account.

Equivalence and noninferiority trials

Sometimes, the objective of a clinical study is to show that a new intervention is equally effective as (i.e., equivalence) or

Table 2. Examples of null hypotheses, alternative hypotheses and α-levels for different study types[1]

Type of Study	Null Hypothesis (H_0)	Alternative Hypothesis (H_1)	α Level	Reference
Equality (often referred to as superiority)	The incidence rate of new AKs is **equal** between the 5-FU and placebo groups	The incidence rate of new AKs is **NOT equal** between the 5-FU and placebo groups	Two sided	Walker et al. (2017)
Equivalence	Humira (AbbVie, Chicago, IL) is **NOT equivalent** to biosimilar BI 695501 in patients with active RA	Humira (AbbVie) is **equivalent** to biosimilar BI 695501 in patients with active RA	Two sided	Cohen et al. (2018)
Noninferiority	5-FU is **inferior** to MAL-PDT by **MORE than** 10% for superficial BCC	5-FU is **inferior** to MAL-PDT by **LESS than** 10% for superficial BCC	One sided	Jansen et al. (2018)

Abbreviations: 5-FU, 5-fluorouracil; AK, actinic keratosis; BCC, basal cell carcinoma; MAL-PDT, methyl aminolevulinate photodynamic therapy; RA, rheumatoid arthritis.

[1]Words in boldface type highlight the differences between the null and alternative hypotheses.

not worse than (i.e., noninferior) the standard (or control) treatment with similar or fewer adverse effects. In noninferiority studies, only one side of the alternative hypothesis (H_1) is of interest (Jansen et al., 2018) (Table 2). Because the sample size can be based on a one-sided α level, a smaller sample size is typically required than in an equivalence trial. Regardless, large sample sizes are typically required, because a high power and small effect size are needed for the credibility of the study.

Descriptive and diagnostic studies

To calculate the sample size in descriptive studies, the researcher should specify (i) the expected proportion or mean and standard deviation, (ii) the width of the confidence interval (the distance from the lower confidence limit to the upper confidence limit), and (iii) the confidence level (calculated as $1 - \alpha$, typically a 95% confidence interval). Based on this, the required sample size can be computed.

For diagnostic studies, the sample size is calculated to achieve either an adequate sensitivity or an adequate specificity. The calculation also includes the width of the confidence interval and the prevalence of the disease (Jones et al., 2003).

SPECIAL CONSIDERATIONS

Efficient study designs

Various techniques are available to increase efficiency and thus provide optimal sample size (Hulley and Cummings, 2013). Possibilities include reducing measurement error (smaller standard deviation), paired measurements (reduced interindividual variability), using a continuous measurement (more efficient than a dichotomous variable), increasing the number of controls, or increasing the frequency of the outcome measure (e.g., restricting to high-risk study populations). However, some of these possibilities may affect the generalizability and inferences of the study. When possible, innovative study designs should be considered to adequately address the trial objectives.

Nonadherence

Trial participants may not adhere to their therapeutic group. Patients who are randomized to the control treatment can start taking the experimental treatment (drop-in), or patients can drop out of the experimental group. Nonadherence makes the two groups more similar and could make a study underpowered (Wittes, 2002). The total sample size should be adjusted by an inflation factor, 1/(1 − drop-in rate − drop-out rate), to prevent underpowered studies (Table 3).

Table 3. Sample size inflation factors for various drop-in and drop-out rates in a 2-arm randomized controlled trial[1]

		Drop-In Rate (Control Group → Experimental Group)			
		0%	5%	10%	15%
Drop-out rate (from experimental group)	0%	1	1.11	1.23	1.38
	5%	1.11	1.23	1.38	1.56
	10%	1.23	1.38	1.56	1.78
	15%	1.38	1.56	1.78	2.04

[1]To read the table, specify the percentages of people you expect to drop in and drop out. Suppose one expects 15% each to drop out and drop in. The sample size necessary to achieve the prespecified α level and power would be more than double (2.04 times) the size needed if all participants adhered to their assigned treatment.

MULTIPLE CHOICE QUESTIONS

1. What is statistical power?
   - A. Probability of detecting an effect when it truly exists
   - B. Failure to detect an effect when it truly exists
   - C. Probability of detecting an effect when there is no true effect
   - D. Not observing any effect when there is no true effect.
2. Which information for the sample size calculation should be derived from a good research question?
   - A. Type of statistical test and power
   - B. Type of statistical test
   - C. Type of outcome measurement
   - D. Type of outcome measurement, α, and β
3. The null and alternative hypotheses of a noninferiority trial are as follows: H_0, treatment B is worse than treatment A by more than a prespecified difference and H_1, treatment B is worse than treatment A by less than a prespecified difference. H_1 implies which of the following?
   - A. A one-sided α level
   - B. A two-sided α level
   - C. A one-sided β level
   - D. A two-sided β level
4. Which parameters are needed to calculate the sample size for a trial with two independent groups and a binary outcome measure?
   - A. α, β, expected difference, and standard deviation
   - B. Type 1 error level, type II error level, one- or two-sided α level, expected difference, and the control group success rate
   - C. α, β, and expected difference
   - D. α, power, and expected difference
5. In which situation is a power calculation appropriate?
   - A. After a trial for secondary outcome measures
   - B. Before analyzing available data to calculate the detectable effect size
   - C. Before analyzing available data to calculate the power of detecting a specified effect
   - D. Situations B and C

Multiple comparisons

An α level of 0.05 implies that 1 in every 20 tests will be statistically significant by chance when there is nothing to find (false positive). Examples of situations in which the α level may need to be adjusted include studies with more than two treatment arms, studies with multiple outcomes, interim analyses in trials, and genome-wide association studies. Comprehensive multiple testing correction procedures are provided by the US Food and Drug Administration and the European Medicines Agency (Dmitrienko and D'Agostino, 2017). The guidelines include, among other procedures, the Bonferroni correction (dividing the α level by the number of independent hypotheses test), the Benjamini-Hochberg method (controlling the false discovery rate), or classifying the hypotheses as primary and secondary.

LIMITATIONS OF SAMPLE SIZE AND POWER CALCULATIONS

Limitations include that the specification of the parameters (e.g., effect size) involves some guesswork (Hulley and Cummings, 2013). Second, assumptions (e.g., completely random errors, correctly specified models) are almost implausible in reality, and thus the sample size may be underestimated (Rothman et al., 2008). In addition, researchers may reduce inference to dichotomy at an arbitrary level of statistical (rather than clinical) significance ($P < 0.05$), although according to good epidemiological practice, precision is best quantified by the width of the confidence interval.

SUGGESTED READING AND TOOLS

We provide a brief description of the most important aspects of sample size and power calculations. We recommend the references for a detailed discussion of the aforementioned topics. In the PowerPoint slides, we provide suggestions for power and sample size calculations.

CONFLICT OF INTEREST

The authors state no conflict of interest.

AUTHOR CONTRIBUTIONS

SS, SL, and LH all contributed to drafting, editing, and finalizing the manuscript and teaching material.

SUPPLEMENTARY MATERIAL

Supplementary material is linked to this paper. Teaching slides are available as supplementary material.

REFERENCES

Case LD, Ambrosius WT. Power and sample size. Methods Mol Biol 2007;404:377–408.

Cohen SB, Alonso-Ruiz A, Klimiuk PA, Lee EC, Peter N, Sonderegger I, et al. Similar efficacy, safety and immunogenicity of adalimumab biosimilar BI 695501 and Humira reference product in patients with moderately to severely active rheumatoid arthritis: results from the phase III randomised VOLTAIRE-RA equivalence study. Ann Rheum Dis 2018;77:914–21.

Dmitrienko A, D'Agostino RB Sr. Editorial: Multiplicity issues in clinical trials. Stat Med 2017;36:4423–6.

Hong EP, Park JW. Sample size and statistical power calculation in genetic association studies. Genomics Inform 2012;10:117–22.

Hulley SB, Cummings SR. Designing clinical research. Philadelphia, PA: Lippincott Williams and Wilkins; 2013.

Jansen MHE, Mosterd K, Arits A, Roozeboom MH, Sommer A, Essers BAB, et al. Five-year results of a randomized controlled trial comparing effectiveness of photodynamic therapy, topical imiquimod, and topical 5-fluorouracil in patients with superficial basal cell carcinoma. J Invest Dermatol 2018;138:527–33.

Jones SR, Carley S, Harrison M. An introduction to power and sample size estimation. Emerg Med J 2003;20:453–8.

Kim N, Fischer AH, Dyring-Andersen B, Rosner B, Okoye GA. Research techniques made simple: choosing appropriate statistical methods for clinical research. J Invest Dermatol 2017;137:e173–8.

Neuberg D. How many mice? Design considerations for murine studies [podcast]. Blood Advances 2017;1:1466.

Rothman KJ, Greenland S, Lash TL. Precision and statistics in epidemiologic studies. In: Modern epidemiology. Philadelphia, PA: Lippincott Williams & Wilkins; 2008. p. 148–67.

Walker JL, Siegel JA, Sachar M, Pomerantz H, Chen SC, Swetter SM, et al. 5-Fluorouracil for actinic keratosis treatment and chemoprevention: a randomized controlled trial. J Invest Dermatol 2017;137:1367–70.

Wittes J. Sample size calculations for randomized controlled trials. Epidemiol Rev 2002;24:39–53.