


ORIGINAL ARTICLE

WILEY

Beyond statistical power in designing clinical trials: Shiny/R apps to the rescue

Pedro Sandoval^{1,2} | Ester VilaprinYO^{1,3} | Rui Alves^{1,3} | Albert Sorribas^{1,3} 

¹Biomodels Research Group, University of Lleida & Institute of Biomedical Research of Lleida (IRBLleida), Lleida, Spain

²Department of Mathematics and Statistics, Universidad Técnica Nacional (UTN), Balsa de Atenas, Costa Rica

³Mathematical Systems & Synthetic Biology in Biomedicine and Biotechnology (MathSy2Bio), University of Lleida & Institute of Biomedical Research of Lleida (IRBLleida), Lleida, Spain

Correspondence

Albert Sorribas, Biomodels Research Group, University of Lleida & Institute of Biomedical Research of Lleida (IRBLleida), Lleida, Spain.
Email: albert.sorribas@udl.cat

Funding information

Instituto de Salud Carlos III, Grant/Award Number: PI20/00377; Direcció General de Recerca, Generalitat de Catalunya

Abstract

Medical students must understand statistical reasoning and sample size selection to design and interpret clinical trials. Beyond achieving sufficient statistical power, ensuring meaningful precision in treatment effect estimates is equally important. We developed free, interactive Shiny/R tools that let learners explore how varying sample sizes influence both power and precision in common study designs. As an example, we discuss a Shiny/R app that explores a two-group design with a binary outcome, allowing sample size computation, effect estimation, and simulation of experiments in different scenarios. By engaging with these simulations, students gain a practical grasp of clinical trial design, ultimately enhancing their research skills and improving patient care.

KEYWORDS

clinical trial, confidence interval precision, sample size, Shiny/R app, statistical literacy, statistical power

1 | BACKGROUND

Clinical trials are crucial for medical progress, providing the scientific foundation for the development of new treatments and advancements in biomedicine.^{1, 2} The validity of this knowledge largely relies on the appropriateness of the trial design, where the use of an appropriate sample size plays a critical role.^{3–5} Therefore, a thorough understanding of sample size selection is essential, as an appropriate sample size is crucial to ensure that a study has sufficient statistical power to detect clinically significant differences. This prevents the financial and ethical costs associated with a study failing to achieve its intended objectives.⁶

In clinical trials, minimizing bias is essential for generating reliable and valid results.

Randomization is a cornerstone of trial design, ensuring that participants are allocated to treatment or control groups in a way that balances both known and unknown confounding factors, thereby reducing selection bias.^{1, 7} Alongside this, obtaining ethical approval is critical; it guarantees that the study meets rigorous ethical standards, protects the rights and well-being of participants, and maintains public trust in the research process.⁸ Together, these practices underpin the integrity of clinical trials and are fundamental to advancing evidence-based medicine.

When teaching medical students, introducing certain concepts through mathematics can be challenging. Here,

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2025 The Author(s). *Teaching Statistics* published by John Wiley & Sons Ltd on behalf of Teaching Statistics Trust.

we propose using simulation tools to facilitate discussion of these challenges and enhance the understanding of key concepts essential for informed decision-making. Although future healthcare professionals may not be responsible for performing the calculations themselves, it is crucial that they comprehend the principles behind sample size selection when collaborating with the statisticians in the research team and when evaluating new evidence.

To contribute to medical training, we developed a set of simulation tools that can be used both in an introductory course in statistical methods and in assisting in the preparation of clinical trials. These tools are designed not for analyzing clinical trials but for exploring the implications of various trial designs before their initiation. The primary goal is to provide guidance, optimize resource allocation, and prevent undertaking trials that are unlikely to yield meaningful results.

In developing those tools, we consider two major concerns when preparing projects in biomedical research. First, we address the problem of assuring appropriate statistical power. A clinical trial that lacks sufficient statistical power is more likely to fail in detecting beneficial effects, which can result in incorrect conclusions about the treatment's effectiveness. Sample size is critical for ensuring adequate power, as using an inappropriate sample size can result in resource losses and ethical problems, including exposing more participants than necessary to experimental treatments or placebos.^{9, 10} Second, in addition to achieving sufficient statistical power, the study design and sample size should also guarantee that treatment effects are estimated with appropriate precision.^{11–16} In addition, in many cases a final decision on sample size will also involve considering the available budget.¹⁷ Determining the optimal ratio between treatment and control groups requires careful evaluation of ethical concerns, subject availability, cost–benefit considerations, and other factors. Notably, if the tested treatment is more expensive than the control (e.g., placebo), using a larger number of controls can help reduce the overall cost of the trial.¹⁸

The precision in estimating the effect size is determined by achieving a specified width of the confidence intervals. Poor precision indicates the inability to draw reliable conclusions about the treatment effect, even if statistical significance is demonstrated.^{14, 15, 19} In general, small sample sizes lead to wider confidence intervals, which reduce the reliability of estimated effects and may hinder the clinical application of trial results.²⁰ Therefore, achieving a careful balance is necessary, as the precision of estimates is not only a statistical concern but

also crucial for guiding treatment decisions and patient care.^{21–24}

We present a Shiny/R app that can help bridge theory and practice by allowing users to experiment with key design parameters and that serves as a practical guide for medical students and early-stage researchers, providing a robust framework to inform decision-making in clinical research.

2 | METHODS

2.1 | Statistical power as a design goal

Statistical power is defined as the probability of correctly rejecting a false null hypothesis. In the context of a clinical trial, it is the likelihood that a study will detect an effect when there is an actual effect to be detected. High statistical power means that the trial has a high probability of finding a true effect. A clinical trial with inadequate statistical power may fail to detect a potentially beneficial treatment effect, leading to the erroneous conclusion that the treatment is ineffective.¹³ This not only misguides future research but can also have significant implications for patient care and policymaking.

Several key factors influence statistical power:

- i. **Effect Size:** The minimum magnitude of the effect that you want to be able to detect. Larger effects are easier to detect, thus requiring smaller sample sizes to achieve the same power.
- ii. **Sample Size:** Larger sample sizes increase the power of a study, reducing the impact of random variation.
- iii. **Significance Level (α):** The threshold for rejecting the null hypothesis. A lower α (such as 0.01 instead of 0.05) requires stronger evidence to reject the null hypothesis, which reduces power.
- iv. **Variability in the Population:** Greater variability within data, for instance, due to low accuracy in measurements or to heterogeneity in the population decreases power. Trials with less variability (more precise measurements, less experimental errors, etc.) have higher power.

Before initiating a trial, a desired statistical power, typically set at 80% or 90%, must be established during the design phase.²⁵ Once determined, this statistical power is used to compute the necessary sample size. Calculating the sample size also involves defining the minimum clinically relevant effect size and the desired significance level.^{12, 13, 16, 26, 27}

2.2 | Precision in estimating treatment effects in clinical trials as a design goal

Precision is often defined as the width of confidence intervals for an effect.²⁸ A confidence interval represents the range of parameter values that are consistent with the observed data. Statistically, we expect the true parameter value to fall within the calculated interval in a specified proportion of samples, corresponding to the confidence level. Hence, for any given sample, the confidence interval provides an estimate that likely contains the true parameter value. Narrower confidence intervals indicate greater precision and reduce uncertainty about the parameter estimate.¹⁴

Precision plays a pivotal role in clinical trials, as it directly impacts the quality of evidence available for clinical decision-making. Clear and precise confidence intervals enable healthcare professionals to better interpret treatment outcomes, guiding patient care and informing therapeutic strategies with greater reliability.

As in the case of statistical power, several factors influence the precision of effect estimates in clinical trials:

- i. **Sample Size:** Larger sample sizes provide more precise estimates as they reduce the impact of random variation in the data.
- ii. **Variability in the Population:** High variability in participant responses or careless measurements can decrease precision. It is therefore important to train trial staff and participants in proper measurement techniques. Additionally, trials involving populations with less variability in treatment response tend to yield more precise estimates.
- iii. **Measurement Quality:** The accuracy and consistency of the methods used to measure outcomes play a significant role in precision. Higher quality measurement tools and techniques typically yield more precise estimates.
- iv. **Study Design:** Certain study designs, such as cross-over designs or matched case-control studies, can increase precision by reducing variability or by controlling confounding variables.

2.3 | Development of applications

In preparing the different applications that may facilitate exploring specific designs, we consider the following common structure:

- Set up a scenario by defining appropriate parameters that represent an educated guess of the actual research situation.

- Using the defined scenario, compute the appropriate sample size for attaining a given statistical power.
- Evaluate the required sample size if we focus on precision when estimating treatment effects.
- Perform simulations using the computed sample size and compare the results so that the user can understand the concept of statistical power and precision.
- Allow discussing the implications of selecting sample size and the interplay among statistical power, precision, and clinical significance.
- Incorporate the apps into classroom activities—for example, use them in student projects, to introduce key statistical concepts, or as a basis for discussions with research groups on planning parallel two-arm trials.

So far, the available tools include (see the Availability of data and materials Section):

- Two-arms parallel clinical trial with binary outcome.
- Multi-arms parallel clinical trial with Normal distributed outcome
- Multicentric (block) design of a multi-arms parallel clinical trial with Normal distributed outcome
- Repeated Latin Squares designs
- Comparison of treatment responses in a longitudinal study with replicated measurements for each subject and subject variability (mixed-model)
- Nested designs

2.4 | Parallel two-arms clinical trial with binary outcome

As an example of the organization and usefulness of our simulation tools, in this paper we concentrate on a parallel two-arms clinical trial with a binary outcome (e.g., comparing the proportion of subjects who improve with treatment to those in a control or reference group). This approach allows us to emphasize core concepts while laying a solid foundation for understanding more complex designs.

A parallel two-arms clinical trial with binary outcome is a basic design in clinical research. In this design, participants are randomly assigned to one of two groups: typically, a treatment group (i.e., the treatment we want to evaluate), and a reference group (either a reference treatment or a placebo). The outcome of interest in these trials is binary, meaning it has two possible outcomes, often denoted as “success” or “failure,” “effective” or “ineffective,” “improved” or “not improved,” etc. In these trials, the primary focus is to compute the proportions of success (or failure) between the two groups and evaluate if the data can support the existence of differences on the population probabilities.

Being a simple design, it has all the elements that will allow discussing the most relevant concepts that lead to selecting a sample size. Also, through simulations, we will see the implications of this decision and the possible drawbacks when drawing conclusions from a trial.

When using this design, we must carefully identify which is the primary goal in comparing the two groups. These are the alternative cases (see Figure 2 and corresponding text)^{29, 30}:

- i. **Test Equivalence:** Equivalence trials are designed to demonstrate that two treatments, typically a new treatment and an existing standard, have no clinically significant difference in their effects. These trials are crucial when the new treatment offers other benefits, such as reduced cost, fewer side effects, or improved patient compliance. In a parallel two-arms trial with binary outcomes, the goal is to show that the success rates in both groups fall within a prespecified equivalence margin.
- ii. **Test Superiority:** Superiority trials are perhaps the most common type of clinical trial. They aim to demonstrate that one treatment is better than another (often a placebo or standard treatment). In the context of parallel two-arms trials with binary outcomes, superiority is demonstrated when the success rate in the treatment group is significantly higher than in the control group, beyond what could be attributed to chance alone. These trials are fundamental in establishing new treatments as effective.
- iii. **Test Clinical Superiority:** Beyond showing superiority, in many cases the goal is to establish that the new treatment improves the effect of the old treatment above a certain threshold, making it clinically relevant. In that case, the minimum increment in effectivity must be fixed in advance.
- iv. **Test Non-Inferiority:** Non-inferiority trials are designed to show that a new treatment is not worse than an existing treatment by more than a specified margin. These trials are important when the new treatment may offer other advantages, such as safety, cost-effectiveness, or ease of use. In parallel two-arms trials with binary outcomes, demonstrating non-inferiority involves showing that the success rate of the new treatment is not significantly lower than that of the existing treatment, within a predefined margin.

The decision to conduct equivalence, superiority, or non-inferiority trials should be based on research questions, ethical considerations, and the clinical relevance of the treatments being compared.²⁹ Setting appropriate margins for equivalence (equivalence range) or non-inferiority is a critical aspect of designing this kind of trial.^{16, 27} These margins must be clinically justified and

should not be too small to be meaningless or too large to be unachievable. Results from these trials must be interpreted carefully. For instance, a finding of non-inferiority does not imply equivalence, and failing to demonstrate superiority does not necessarily mean the treatments are equivalent.

Understanding the distinctions and appropriate applications of equivalence, superiority, and non-inferiority trials is crucial in the context of a clinical trial. Each type addresses different research questions and offers unique insights into the comparative effectiveness of treatments. Proper design, execution, and interpretation of these trials are key to advancing medical knowledge and improving patient care.

3 | IMPLEMENTATION

3.1 | Exploring scenarios in a parallel two-arms clinical trial with binary outcome

As indicated, as a proof-of-concept example we will concentrate on a Shiny/R app developed for understanding, planning, and simulating the effect of the various design factors that must be evaluated when computing the required sample size in a parallel two-arms clinical trial with a binary outcome³¹ (Figure 1).

The menu options of this application are planned to fulfill three objectives. First, in the “What’s your goal?” option, the user can see the effect of changing equivalence margins, sample sizes, and proportion of the effect in each branch of the trial between alternative treatments. The interface also explains what conclusions can be drawn about the difference between treatment arms of the trial and their clinical significance.

Second, in the “Plan a trial” option the user can plan an experiment by defining a scenario that suits the available knowledge on the specific problem. The app will compute the sample size required for different goals under the requirement of either statistical power or precision in estimation of the treatment’s effect.

Third, in the “Simulate trials” option, the user visualizes the simulation of different experiments, which will facilitate understanding the expected results when performing an actual trial.

3.2 | What’s your goal? Equivalence, non-inferiority, or superiority

Before discussing the actual computation of sample size, it is important to explore the concept of Equivalence Range (Δ) and the interpretation of Superiority,

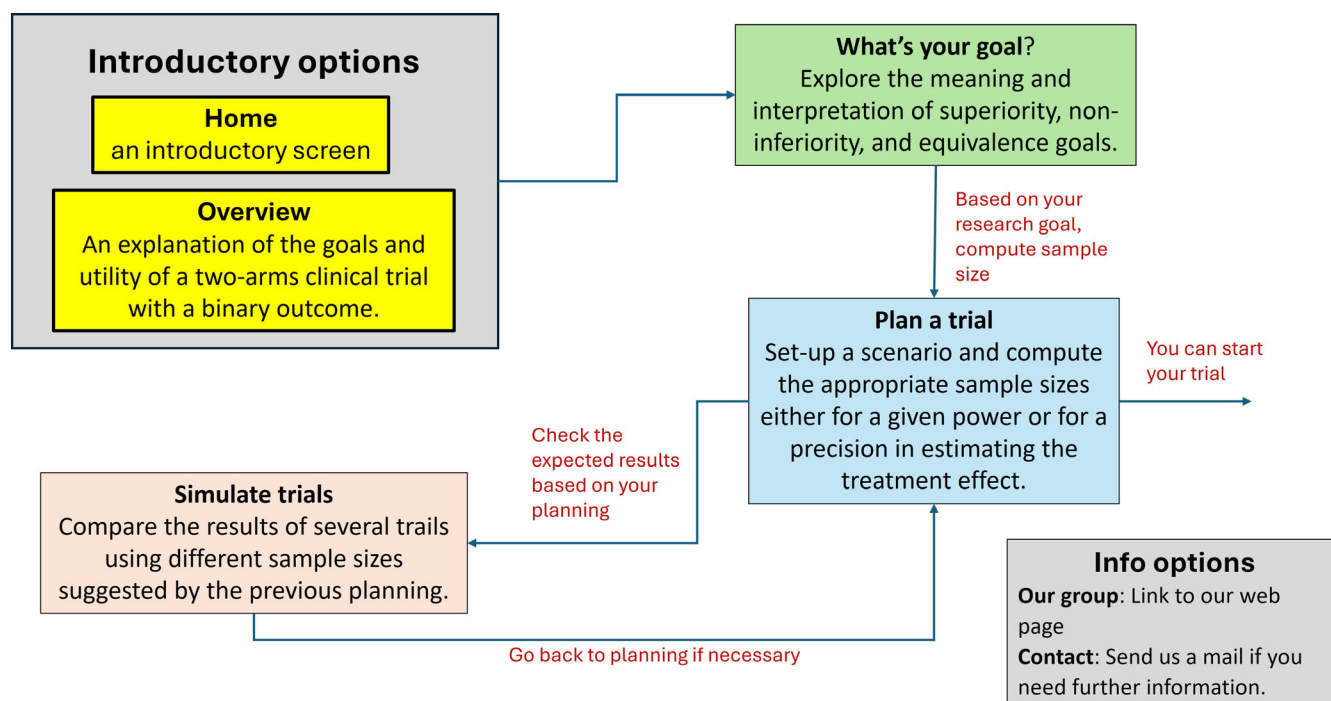


FIGURE 1 Workflow of the Shiny/R application for sample size decision on planning a parallel two-arms clinical trial with binary outcome. [Colour figure can be viewed at wileyonlinelibrary.com]

Non-Inferiority, and Equivalence. We define the Equivalence Range as the change on the difference on outcomes' probabilities (treatment effect) that we consider nonrelevant. For instance, an equivalence range of 0.1 means that a difference of probability of ± 0.1 is considered too low for concluding a practical difference between the treatments compared. This option is illustrated in Figure 2 and can be explored in the menu option: What's your goal?

In practice, the equivalence range must be defined in accordance with the goal of the study. For instance, if a trial is designed for evaluating superiority of a new treatment, then Δ represents the minimum difference in outcome's probabilities to conclude in favor of that treatment. In Figure 2 we present graphically the possible interpretations of a trial result. As an example, suppose that we obtain a confidence interval (a, b) for the difference of probabilities in the two arms (either treatment vs. control, or between a new treatment and a reference treatment). Basically, we have the following situations.

- If $-\Delta < (a, b) < \Delta$ then we conclude that the result indicates practical equivalence
- If $0 > a > -\Delta$ and $b > \Delta$ then we conclude non-Inferiority
- If $a > 0$ and $a < \Delta$ then we conclude Superiority
- If $a > \Delta$ then we conclude Clinical Superiority

Consider the situation of a trial that results in a 78% improvement in the treatment group ($n = 50$) and 56% in

the reference (control) group ($n = 50$) (Figure 2). These sample proportions can be used for computing the confidence interval for the treatment effect (difference in population probabilities). Now, let us consider that we are interested in concluding that our treatment is clinically superior to the reference (i.e., the improvement is greater than a minimum value, say $\Delta = 0.1$). In this example, although the difference of improvement proportions is 0.22, the CI suggests we can conclude superiority, that is probability of improvement in the treatment is greater than of the reference, but the result does not meet the threshold of clinical relevance. The user can play with different scenarios by changing the sample sizes and the proportions of improvement observed in both samples. For instance, it is important to understand that the interpretation of a specific result will depend on the equivalence range considered. Also, you should explore the effect of increasing sample size in each scenario. After playing with this panel, the user is ready to plan an actual trial.

3.3 | Sample size considerations in a parallel two-arms clinical trial with a binary outcome

Before planning a trial, the determination of an appropriate sample size requires considering the following issues²⁸:

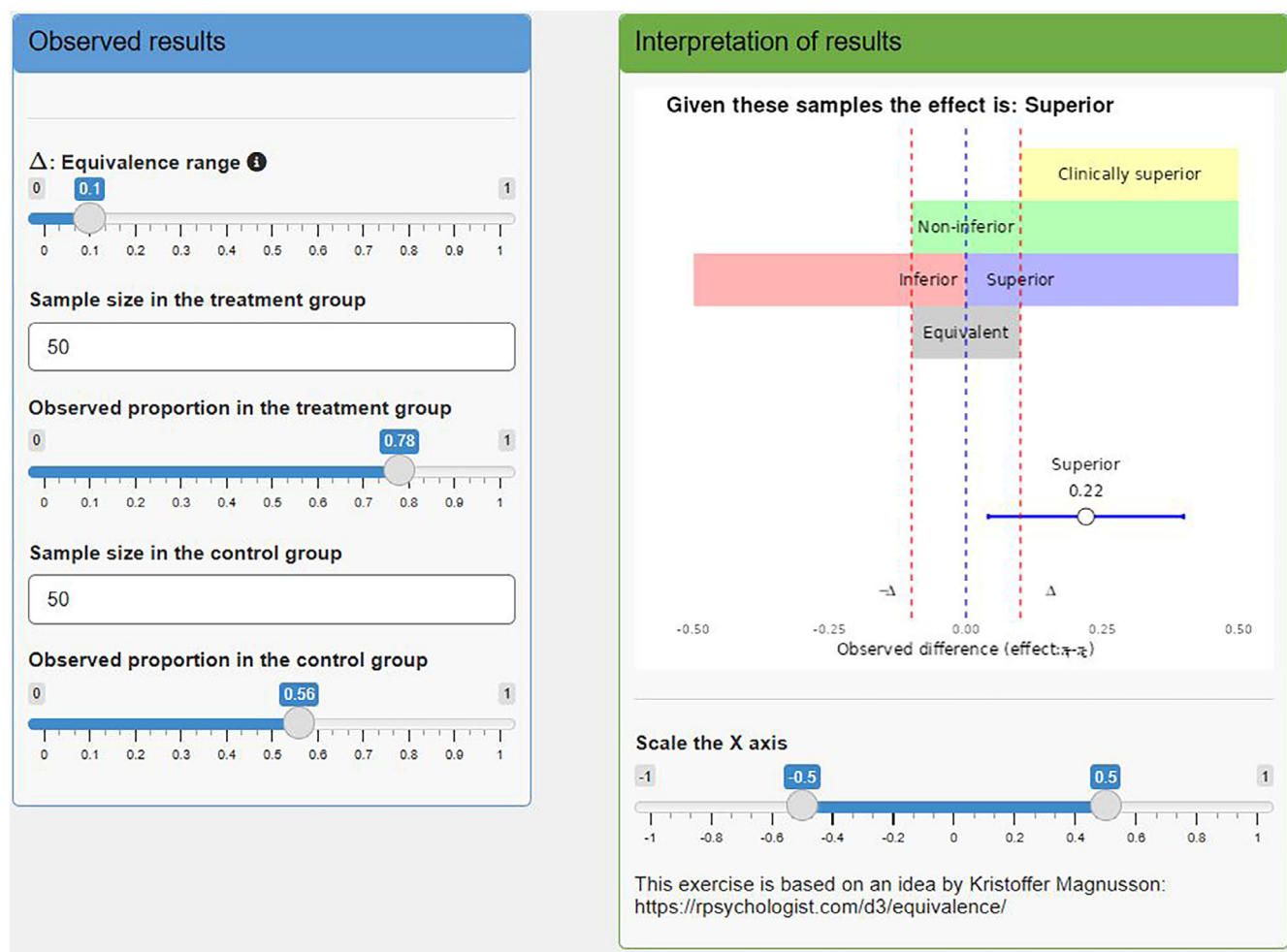


FIGURE 2 Exploring different scenarios when evaluating the results of a parallel two-arms clinical trial with a binary outcome. [Colour figure can be viewed at [wileyonlinelibrary.com](https://onlinelibrary.wiley.com/doi/10.1111/tesl.12403)]

- **Determining the effect of minimum desirable differences in probability for the trial:** The required sample size depends on the actual difference in probabilities one sets as minimal to be considered of practical significance. In a scenario where the difference in the probabilities of the event between groups is low, it will be difficult or even impossible to prove that difference in a trial (see below).
- **Understand the practical meaning of statistical power:** Statistical power is the probability of detecting the minimum effect size that is considered meaningful in practical applications. Although a value of 0.8 is currently used, it is important to stress that this means that 20% of the trials will not be able to detect this effect. Can we increase power at will? What are the implications?
- **Importance of the minimum desirable effect size for clinical relevance (Δ : difference of probabilities):** The computed sample size will assure that you will attain the desired statistical power and be able to

identify the treatment effect if that effect is equal or greater than Δ .

- **Effect of changing the ratio between the number of treatments and controls:** The optimal ratio depends on various factors—including ethical considerations, subject availability, and cost-benefit analysis. For simplicity and demonstrative purposes, we will determine the optimal decision based solely on the costs of treatments and controls.
- **Effect of changing significance level:** To reduce the probability of rejecting the null hypothesis when it is true, a lower significance level must be used. Increasing significance levels also increases the sample size required for attaining the desired power.
- **Effect of choosing between a Superiority, Non-inferiority, or a Clinically relevant result?** The decision on sample size will be different for each case.

In Figure 3, we show a typical calculation of sample size using our app (Menu option: “Plan a trial”). First, we

Scenario to explore

Probability in treatment group ⓘ

Probability in control group

Minimum effect size for a clinically relevant effect

Ratio of the sample sizes treatment/control

Power

Significance level

Cost of both interventions

Cost per treatment

Cost per control

Sample size required

The required sample sizes will depend upon the hypothesis you want to prove. If the minimum effect for clinical relevance is greater than the difference between the probabilities in treatment and reference groups, then you cannot design the trial to prove clinically relevance (in that case, the sample size computed is meaningful)

Goal	Treatment	Control	Cost
Superior	62	62	682
Non-inferiority	28	28	308
Clinically superior	248	248	2728

FIGURE 3 Computation of sample size in a two-branch clinical trial with binary outcome. [Colour figure can be viewed at [wileyonlinelibrary.com](https://onlinelibrary.wiley.com/doi/10.1111/est.12403)]

must indicate a reasonable value of the probability in both groups. Let us assume a 60% improvement for the control group, and 80% for the treatment group. Also, we fix a minimum increment of 0.1 to consider a clinically relevant result. What are the sample size requirements in that case? Can we reach a correct conclusion in a trial? We will compute the appropriate sample size for attaining a statistical power of 0.8 (which is a typical value) and a significance value of 0.05, and an allocation rate of 1. Furthermore, we consider that the cost for a treatment is 10€, while the cost of a control is 1€. In the considered conditions, the app indicates that we need 62 subjects per group to be able to conclude superiority if the difference between the improvement probability among treatment and control is at least $0.8 - 0.6 = 0.2$ and the minimum effect size to detect is 0.1. For non-inferiority, a sample size of 28 per group is enough, while we need 248 per group to show clinical superiority (a difference greater than the minimum of 0.1).

At this point, we have different costs for the trial (682 € for a Superiority trial, 308€ for a Non-inferiority, and

2728€ for showing Clinical superiority). Given that the cost of treatment is 10 times the cost of a control, we may ask if we could consider fewer treatments while compensating by increasing the number of controls. In Figure 4 we show the optimization for the ratio of the number of treatments versus the number of controls. In the considered conditions, the optimum ratio is around 0.27, reducing the cost for the different possible trials. For instance, concluding clinical superiority with a statistical power of 0.8 would require 140 treatments and 515 controls.

The app computes also the necessary sample size if we aim to have a precise estimate of the treatment effectiveness. You can see that in this scenario precision of ± 0.05 with a confidence of 95% requires a much higher sample size, with 1612 controls and 436 treatments.

Once the appropriate sample size based on statistical power has been determined, one may inquire about the expected outcomes of the trial. Simulation enables us to assess the potential results of a trial conducted under optimal conditions (Menu option: Simulate trials). Using this option, we can perform several simulations

comparing typical results of many trials and compare them. In Figure 5 we show simulations with the specific conditions to conclude superiority. We have computed a

sample size of 35 treatments and 129 controls (power of 0.8, minimum effect of 0.1 for the scenario considered). The simulated results show several interesting issues:

- After evaluating the 2000 trials, 77.75% conclude with superiority (i.e., a statistical power of about 80%)
- Each trial produces a relatively wide confidence interval. Thus, for a given trial, we cannot obtain a precise estimation of the treatment effect. In that situation, we cannot conclude clinical superiority (only 37.4% of the trials show clinical superiority, hence a poor statistical power for this goal). We should increase the sample size to 140 treatments and 515 controls to achieve an 80% power for clinical superiority.
- Approximately 96.7% of the trials show non-inferiority (we are using a much higher sample size than the minimum for an 80% statistical power in that case).
- It is important to stress that those are results of many trials. In practice, we would obtain one of those results. Thus, this makes especially relevant the understanding of the appropriate sample size after considering all the issues involved and minimizing the risk of obtaining misleading results.

FIGURE 4 Optimal ratio treatment/control for a minimum cost (see text). [Colour figure can be viewed at wileyonlinelibrary.com]

The user can test the different conditions and decide which are the appropriate settings for a specific situation.

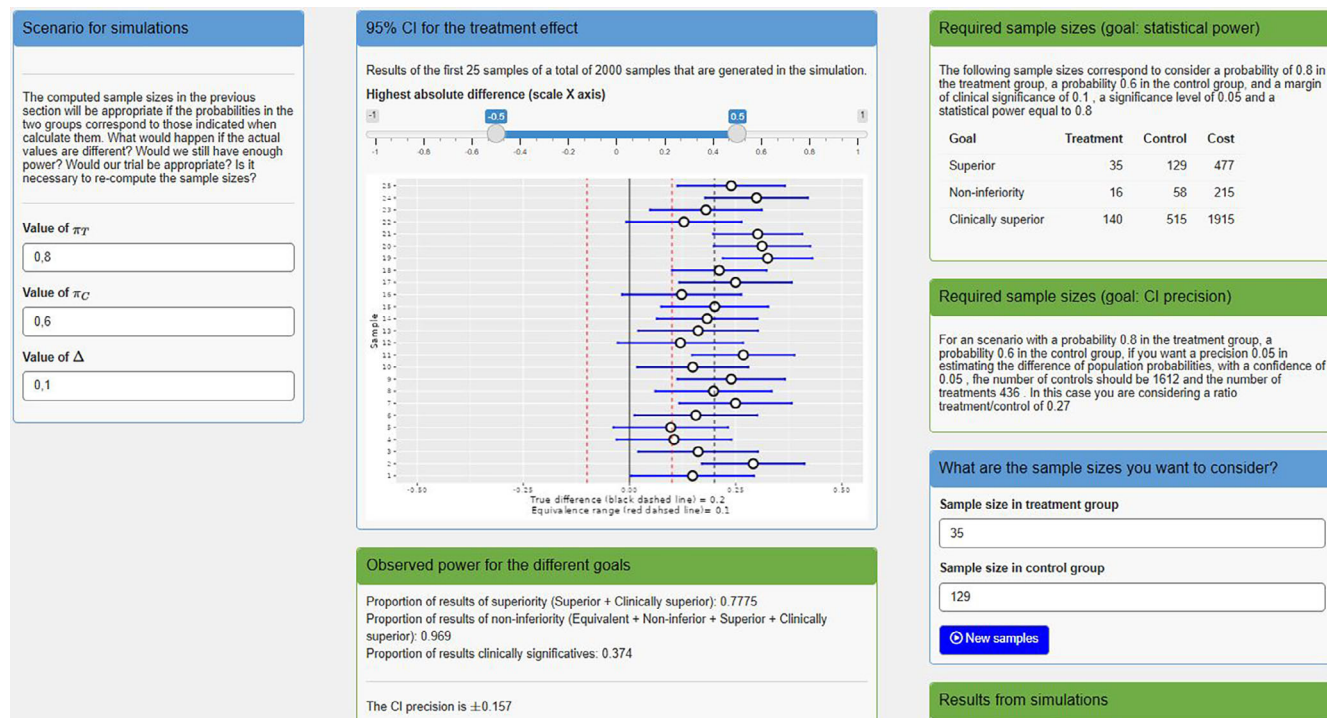


FIGURE 5 Simulation of different trials. Using the optimal sample size for a statistical power of 0.8 produces results that estimate the treatment effect with poor precision. (Vertical red line): Equivalence range; (Vertical black line): Actual treatment effect in the considered scenario. [Colour figure can be viewed at wileyonlinelibrary.com]

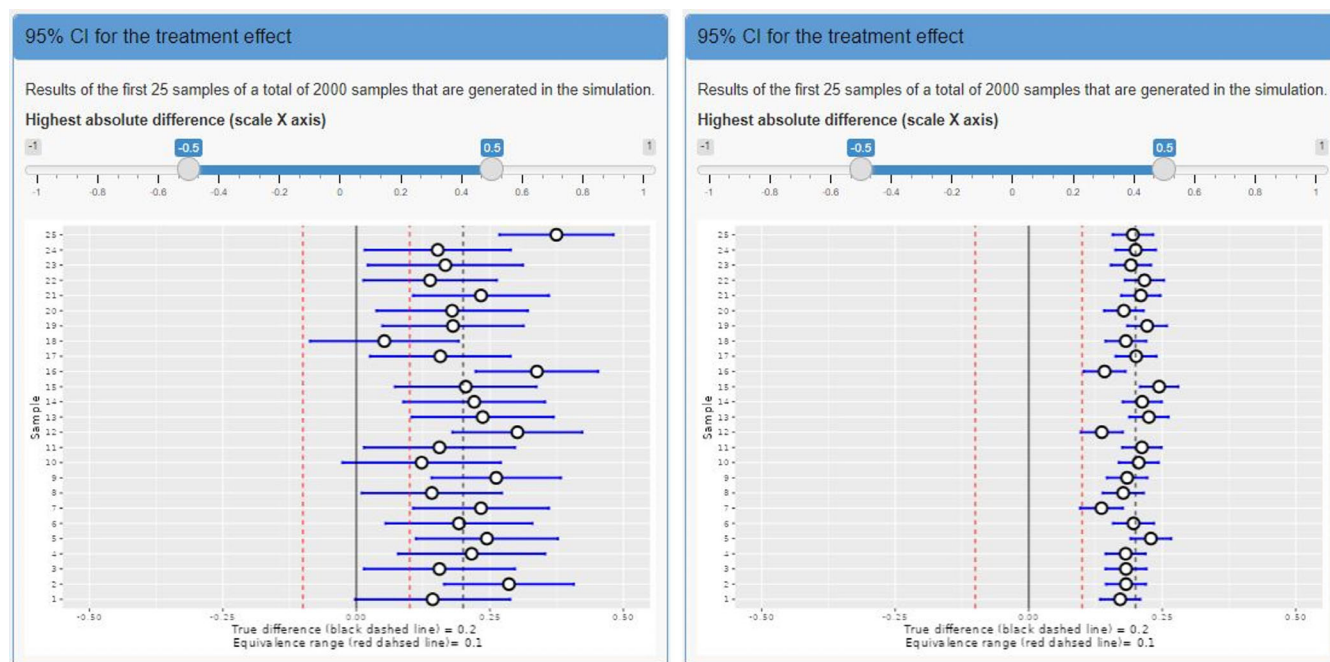


FIGURE 6 Comparison of simulated trials. Left: using the sample size for a 0.8 statistical power (superiority goal). Right: using the sample size for a precision of ± 0.05 . [Colour figure can be viewed at wileyonlinelibrary.com]

In the previous results, it is important to emphasize that statistical power does not assure an appropriate precision in effect estimation. In that example, the precision is ± 0.157 and may be insufficient for practical conclusions. As shown before, if we want to have a precision of ± 0.05 we need to consider 1612 controls and 436 treatments. The difference is large, and the results are shown in Figure 6.

3.4 | Using the app in lecturing medical students

We have extended experience of teaching introductory biostatistics to medical, nursing, bioscience, and biotechnology students in groups that range from 30 to about 70 people. In general, our students have a limited background in statistical reasoning. As those are basic courses, we cannot dig into formal issues, and we make extensive use of simulations for introducing the fundamental concepts and related methods. As we are training future professionals on health-related activities, we focus on the interpretation and utility of statistical methods.

As commented in the introduction, we started a program on interactive simulations based on tools that go beyond the simple computations available on many web pages and applications. Our goal is to provide an integrated framework that facilitates developing a sound understanding of concepts. Here, as an example of the

different tools, we presented one of these applications designed to facilitate the discussion of sample size selection in a clinical trial.

In the example commented on in this paper, the students must be able to achieve the following goals:

- The students should be able to understand the interplay between significance level, statistical power, sample size, and confidence interval.
- The students will master the difference between the statistical significance of a treatment effect and the clinical significance of that effect.
- The students will master the difference in planning a two-group clinical trial for evaluating superiority, non-inferiority, equivalence, and clinical relevance of the treatment being tested.
- The students should be able to plan a two-group clinical trial, decide on the different elements that must be considered for deciding an appropriate sample size, and defend the proposed election.
- The students should be able to perform simulations for exploring the possible outcomes of different trials and discuss their practical consequences.

A fruitful plan for taking advantage of our tool requires several steps:

1. An introductory session on the elements of a two-group clinical trial (2 h). Discuss examples and

consider how to randomize participants in the groups. It is interesting to develop the idea of a control (reference) group and the use of placebo if required. Also, focus on the outcome—in that case, the number of subjects that show good progression, or in the absence of adverse effects. Make the students think of other examples.

2. Introduce the concept of confidence interval as a tool for estimating treatment effect (2 h). Discuss the possible results by using the interactive tool. Make the students guess the practical consequences of each case. Play with the sample size so that they can realize the importance of considering appropriate choices. Here the question should be: How much is enough?
3. Review the concept of statistical power (2 h). This is an important and difficult question both for teaching and learning. Introduce the concept of the minimum effect to detect as this is important for understanding statistical power.
4. Propose some of the suggested activities and let students run simulations and compute results (Let the students define a simple project and ask for a report discussing the appropriate decisions in designing a parallel two-arms clinical trial with a binary outcome). Make them write down their own interpretations. When finished, let them share their results and draw useful conclusions.

3.4.1 | Utility of the Shiny/R apps

Our tools have been used in courses and research support since 2022 in the following cases:

1. **Basic Statistical Courses:** These tools have been integrated into courses in Biotechnology (around 50 students per year), Nutrition (around 20 students per year), and Medicine (around 150 students per year) at the University of Lleida (Spain). Their use significantly enhances the understanding of fundamental statistical concepts and fosters open discussions in class.
2. **Statistical Training for Medical Professionals:** The apps have been incorporated into statistical courses for medical doctors and clinical research groups at the University Hospital Arnau de Vilanova in Lleida (Spain), providing an interactive and practical approach to learning key biostatistical principles relevant to clinical research. In these sessions, focusing on the interpretation of different scenarios greatly facilitates participation and discussion based on actual cases.
3. **Statistical consulting and research support:** The apps are particularly valuable in planning clinical

trials, assisting research teams in refining study designs, determining sample sizes, and evaluating statistical power. They have been actively used by the biostatistical service of the Biomedical Research Institute of Lleida, where direct interaction with researchers has led to continuous improvements in the tools.

Students' suggestions regarding organization, clarity, statistical functionality, and pedagogical utility were evaluated, leading to several improvements and refinements in the apps. Furthermore, discussions with research groups highlighted the need for additional applications, prompting the expansion of our tools to cover stepwise, adaptive, nested, and other advanced clinical trial designs, thereby broadening their applicability in biomedical research. Our experience in these activities demonstrates the practical utility of interactive statistical tools in supporting both teaching and research. Their integration into educational settings and clinical trial planning has proven to be an effective approach for enhancing statistical understanding and improving study design methodologies.

3.5 | Proposed activities

We suggest the following activities for exploring the problem of deciding on sample size in a two-group clinical trial with a binary outcome. The proposed activities do not exhaust the possibilities, but they are a good starting point for further proposals that can be adapted to the specificity of each class. The duration of these activities will depend on the background of the students. In a basic course, parts of the activity can be developed as a practical session after introducing the theory. In more advanced courses, each activity can be suggested as a project and allow time for discussing the design, deciding parameters, and conditions, etc. These activities can be done in a computer room with other students or be part of a personal project. As the Shiny/R app is freely available, it provides flexibility for students to use it at their convenience.

3.5.1 | Does a new treatment outperform the standard treatment?

A pharmaceutical company has developed a new drug, Drug A, intended to improve recovery rates in patients with a certain viral infection. The goal of the clinical trial is to compare the recovery rates of patients treated with Drug A to those treated with the standard treatment, Drug B.

Study Design:

Population: Patients with confirmed viral infection (age 18–65).

Groups

- Group 1 (New Treatment—Drug A): 500 patients receive the new treatment.
- Group 2 (Standard Treatment—Drug B): 500 patients receive the standard treatment.

Outcome:

Recovery (yes/no) 30 days after treatment initiation.

Goal:

The minimum effect the researchers want to detect is a 10% increase in recovery rates with Drug A compared to Drug B.

Trial Results

- In the Drug A group, 375 patients (75%) recovered.
- In the Drug B group, 310 patients (62%) recovered.

Questions for Discussion

1. **Sample Size:** Was a sample size of 500 patients per group sufficient to detect a 10% improvement in recovery rates? Would a smaller or larger sample size be appropriate to reach a similar conclusion?
2. **Statistical Power:** What is the statistical power of this trial to detect a 10% difference in recovery rates? Would a higher power (e.g., 90% instead of 80%) require a larger sample size?
3. **Effect Size:** Is the 10% improvement considered clinically meaningful, or should the trial be powered to detect smaller improvements (e.g., 5%)?
4. **Treatment Cost:** If the new treatment has a cost five times greater than the standard one, would you recommend it?

3.5.2 | We developed a new treatment; can we confirm that it is not worse than the in-use treatment?

A new medication, Drug C, has been developed to treat chronic conditions. It is cheaper and potentially has fewer side effects than the standard treatment, Drug D. However, before Drug C can replace Drug D, researchers need to demonstrate that Drug C is not worse than Drug D in terms of effectiveness.

The goal of this trial is to show that Drug C is non-inferior to Drug D, meaning that it is at least as effective as Drug D in terms of patient outcomes. The outcome measured is the proportion of patients who achieve

disease control after 6 months of treatment (binary outcome: success/failure).

Study Design:

- Population: Patients with chronic condition (age 18–65).

Groups:

- Group 1 (New Treatment—Drug C): 400 patients receive the new treatment.
- Group 2 (Standard Treatment—Drug D): 400 patients receive the standard treatment.

Non-Inferiority Margin: A maximum difference of 5% in disease control rates is considered clinically acceptable for non-inferiority (i.e., Drug C must not be more than 5% worse than Drug D).

Outcome: Binary outcome (disease control achieved: yes/no).

Trial Results

- In the Drug C group, 308 patients (77%) achieved disease control.
- In the Drug D group, 320 patients (80%) achieved disease control.

Questions for Discussion

1. **Non-Inferiority Margin:** Was the 5% non-inferiority margin appropriate for this trial? Would using a narrower or wider margin affect the interpretation of the results?
2. **Sample Size and Power:** Was the sample size of 400 patients per group sufficient to detect non-inferiority within the 5% margin? How would increasing the sample size affect the confidence interval and the ability to prove non-inferiority?
3. **Confidence Interval Interpretation:** The 95% confidence interval for the difference in effectiveness was (−1%, 7%). How does this CI impact the conclusion regarding non-inferiority? Would a 90% CI have led to a different conclusion about non-inferiority?
4. **Type I and Type II Errors:** What is the risk of a Type II error (failing to demonstrate non-inferiority when Drug C is, in fact, non-inferior)? How does the p-value for non-inferiority relate to the probability of a Type I error (incorrectly concluding non-inferiority)?
5. **Clinical Relevance:** Is a 3% lower success rate with Drug C clinically relevant? Could the potential benefits (e.g., lower cost, fewer side effects) justify this difference in effectiveness?

4 | DISCUSSION AND CONCLUSION

The most common question a clinical research group asks a biostatistical advisor is, “How many subjects do we need to include in the trial?” As we have discussed in this paper, this is far from a simple question and requires consideration of various factors. While it may seem straightforward for a statistician to address, our experience in collaborating with clinical groups reveals that understanding the process can be challenging for them.

Unfortunately, some of the basic questions that are required for deciding on sample size (variability of data, effect size to detect, difference between statistical power and treatment effect, etc.) are not well grasped by researchers, even after their courses in statistical methods. There are many computer tools to help compute sample size in clinical trials, but few of them allow for easily simulating the practical implications of a specific selection. For instance, the question on the importance of defining an equivalence range is not trivial. The importance of defining the value of this range on the trial design is facilitated by simulating different scenarios. To help in understanding the process of sample size selection, we presented a set of interactive applications that complemented sample size computation with the simulation of different scenarios to help decide this important issue.

Although we have focused on a basic design, our results are common to many other designs and highlight the importance of this approach and the need to consider precision in estimating treatment effects in the decision process. The possibility of anticipating results that are not robust enough when using inappropriate sample sizes is an added value to our proposal. While more sophisticated computation must be used for complicated clinical trials, training in understanding the concepts discussed in this work sets a sound basis for further training.

AUTHOR CONTRIBUTIONS

AS defined the scope of this work. AS and PS developed the Shiny/R app. EV and RA tested the apps in different scenarios and suggested technical improvements. All the authors participated in writing the manuscript and approved the final version.

ACKNOWLEDGMENTS

The authors are grateful to Dr. Erik Cobo and Dr. Jose Antonio Gonzalez from the Polytechnic University of Catalonia for their commentaries and suggestions in early versions of the Shiny/R app.

FUNDING INFORMATION

This work has been supported by grant PI20/00377 from Instituto de Salud Carlos III (Spain). Grups de Recerca SGR-Cat 2021 Reconegets per la Generalitat de Catalunya.

CONFLICT OF INTEREST STATEMENT

The authors declare no competing interests.

DATA AVAILABILITY STATEMENT

Project name: Study Design Project home page: <https://www.biomodels.udl.cat/en/research-lines/study-design/>
Operating system: Platform independent; Programming language: R (Shiny/R app); Other requirements: None; License: GNU GPL.

ORCID

Albert Sorribas  <https://orcid.org/0000-0002-7407-4075>

REFERENCES

1. V. W. Berger, L. J. Bour, K. Carter, J. J. Chipman, C. C. Everett, N. Heussen, C. Hewitt, R. D. Hilgers, Y. A. Luo, J. Renteria, Y. Ryznik, O. Sverdlov, D. Uschner, and R. A. Beckman, *A roadmap to using randomization in clinical trials*, BMC Med. Res. Methodol. **21** (2021), no. 1, 168. <https://doi.org/10.1186/s12874-021-01303-z>.
2. R. Kay, *Statistical principles for clinical trials*, J. Int. Med. Res. **26** (1998), 57–65.
3. P. Charles, B. Giraudeau, A. Dechartres, G. Baron, and P. Ravaud, *Reporting of sample size calculation in randomised controlled trials: Review*, BMJ **338** (2009), b1732. <https://doi.org/10.1136/bmj.b1732>.
4. D. Lakens, *Sample size justification*, Collabra Psychol. **8** (2022), no. 1. <https://doi.org/10.1525/collabra.33267>.
5. S. Nemes, J. M. Jonasson, A. Genell, and G. Steineck, *Bias in odds ratios by logistic regression modelling and sample size*, BMC Med. Res. Methodol. **9** (2009), 56. <https://doi.org/10.1186/1471-2288-9-56>.
6. T. Clark, U. Berger, and U. Mansmann, *Sample size determinations in original research protocols for randomised clinical trials*, BMJ **346** (2013), f1135. <https://doi.org/10.1136/bmj.f1135>.
7. P. M. Spieth, A. S. Kubasch, A. I. Penzlin, B. M. W. Illigens, K. Barlinn, and T. Siepmann, *Randomized controlled trials – A matter of design*, Neuropsychiatr. Dis. Treat. **12** (2016), 1341–1349. <https://doi.org/10.2147/NDT.S101938>.
8. D. Moher, S. Hopewell, K. F. Schulz, V. Montori, P. C. Gøtzsche, P. J. Devereaux, D. Elbourne, M. Egger, and D. G. Altman, *CONSORT 2010 explanation and elaboration: Updated guidelines for reporting parallel group randomised trials*, BMJ. **340** (2010), c869. <https://doi.org/10.1136/bmj.c869>.
9. M. Tracy, *Methods of sample size calculation for clinical trials*, Ph.D. Thesis (2012). <https://theses.gla.ac.uk/id/eprint/671>.
10. R. M. Turner, S. D. Walter, P. Macaskill, K. J. McCaffery, and L. Irwig, *Sample size and power when designing a randomized trial for the estimation of treatment, selection, and preference effects*, Med. Decis. Mak. **34** (2014), 711–719.

11. J. M. Bland, *The tyranny of power: Is there a better way to calculate sample size?* *BMJ* **339** (2009), 1133–1135.
12. J. A. Cook, J. Hislop, D. G. Altman, P. Fayers, A. H. Briggs, C. R. Ramsay, J. D. Norrie, I. M. Harvey, B. Buckley, D. Fergusson, I. Ford, and L. D. Vale, *Specifying the target difference in the primary outcome for a randomised controlled trial: Guidance for researchers*, *Trials* **16** (2015), 12. <https://doi.org/10.1186/s13063-014-0526-8>.
13. J. A. Cook, S. A. Julious, W. Sones, L. V. Hampson, C. Hewitt, J. A. Berlin, D. Ashby, R. Emsley, D. A. Fergusson, S. J. Walters, E. C. F. Wilson, G. MacLennan, N. Stallard, J. C. Rothwell, M. Bland, L. Brown, C. R. Ramsay, A. Cook, D. Armstrong, D. Altman, and L. D. Vale, *Practical help for specifying the target difference in sample size calculations for RCTs: The DELTA2 five-stage study, including a workshop*, *Health Technol Assess (Rockv)*. **23** (2019), 60. <https://doi.org/10.3310/hta23600>.
14. M. J. Gardner and D. G. Altman, *Confidence intervals rather than P values: Estimation rather than hypothesis testing*, *Br. Med. J. (Clin. Res. Ed.)* **292** (1986), 746–750.
15. M. J. Gardner and D. G. Altman, *Estimating with confidence*, *Br. Med. J. (Clin. Res. Ed.)* **296** (1988), no. 6631, 1210–1211. <https://doi.org/10.1136/bmj.292.6522.746>.
16. W. Sones, S. A. Julious, J. C. Rothwell, C. R. Ramsay, L. V. Hampson, R. Emsley, S. J. Walters, C. Hewitt, M. Bland, D. A. Fergusson, J. A. Berlin, D. Altman, L. D. Vale, and J. A. Cook, *Choosing the target difference and undertaking and reporting the sample size calculation for a randomised controlled trial - the development of the DELTA2 guidance Suzie Cro*, *Trials* **19** (2018), 542. <https://doi.org/10.1186/s13063-018-2887-x>.
17. L. Everest, B. E. Chen, A. E. Hay, M. C. Cheung, and K. K. W. Chan, *Power and sample size calculation for incremental net benefit in cost effectiveness analyses with applications to trials conducted by the Canadian cancer trials group*, *BMC Med. Res. Methodol.* **23** (2023), no. 1, 179. <https://doi.org/10.1186/s12874-023-01956-y>.
18. J. Nay, A. Haslam, and V. Prasad, *Justification for unequal allocation ratios in clinical trials: A scoping review*, *Contemp Clin Trials* **139** (2024), 107484.
19. K. B. Freedman, S. Back, and J. Bernstein, *Sample size and statistical power of randomised, controlled trials in orthopaedics*, *J Bone Joint Surg Br.* **83 B** (2001), 397–402.
20. C. C. Serdar, M. Cihan, D. Yücel, and M. A. Serdar, *Sample size, power and effect size revisited: Simplified and practical approach in pre-clinical, clinical and laboratory studies*, *Biochem. Med. (Zagreb)*. **31** (2021), 1–27.
21. B. M. Cesana, *Sample size for testing and estimating the difference between two paired and unpaired proportions: A “two-step” procedure combining power and the probability of obtaining a precise estimate*, *Stat. Med.* **23** (2004), 2359–2373.
22. B. M. Cesana, G. Reina, and E. Marubini, *Sample size for testing a proportion in clinical trials: A “two-step” procedure combining power and confidence interval expected width*, *Amer. Statist.* **55** (2001), 288–292.
23. T. Daimon, *Bayesian sample size calculations for a non-inferiority test of two proportions in clinical trials*, *Contemp Clin Trials* **29** (2008), 507–516.
24. B. Jia and H. S. Lynn, *A sample size planning approach that considers both statistical significance and clinical significance*, *Trials* **16** (2015), 213. <https://doi.org/10.1186/s13063-015-0727-9>.
25. H. Kang, *Sample size determination and power analysis using the G*power software*, *J Educ Eval Health Prof* **18** (2021), 17. <https://doi.org/10.3352/jeehp.2021.18.17>. Epub 2021 Jul 30.
26. E. C. Lee, A. L. Whitehead, R. M. Jacques, and S. A. Julious, *The statistical interpretation of pilot trials: Should significance thresholds be reconsidered?* *BMC Med. Res. Methodol.* **14** (2014). <https://doi.org/10.1186/1471-2288-14-41>.
27. R. A. Parker and J. A. Cook, *The importance of clinical importance when determining the target difference in sample size calculations*, *Trials* **24** (2023), 485. <https://doi.org/10.1186/s13063-023-07532-5>.
28. S. A. Julious, *Tutorial in biostatistics: Sample sizes for clinical trials with Normal data*, *Statist Med* **23** (2004), 1921–1986. <https://doi.org/10.1002/sim.1783>.
29. D. T. Dunn, A. J. Copas, and P. Brocklehurst, *Superiority and non-inferiority: Two sides of the same coin?* *Trials* **19** (2018), 499.
30. P. Vavken, *Rationale for and methods of superiority, noninferiority, or equivalence designs in orthopaedic, controlled trials*, *Clin. Orthop. Relat. Res.* **469** (2011), 2645–2653.
31. A. Sorribas, E. Vilapinyó, and P. Sandoval. *Two-arms clinical trials with a binary outcome: A shiny app*. 2024 <https://irblleida-biostatistics.shinyapps.io/Clinical-Trial-Two-Arms-Binomial/>.

How to cite this article: P. Sandoval, E. Vilapinyó, R. Alves, and A. Sorribas, *Beyond statistical power in designing clinical trials: Shiny/R apps to the rescue*, *Teach. Stat.* (2025), 1–13, DOI [10.1111/test.12403](https://doi.org/10.1111/test.12403)