ScientificScholar®
Knowledge is power
Publisher of Scientific Journals

www.clinicalimagingscience.org

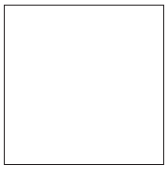
Journal of Clinical Imaging Science

Review Article *Education*

Sample size matters: A step-by-step guide for radiologists

Ramon Gheno[1], Rogério Boff Borges[2], Rodrigo Citton Padilha dos Reis[3]

[1]Department of Radiology, Moinhos de Vento Hospital, [2]Department of Biostatistics, Biostatistics and Data Analysis Unit, Hospital de Clínicas de Porto Alegre, [3]Institute of Mathematics and Statistics, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil.



***Corresponding author:**
Ramon Gheno,
Department of Radiology,
Hospital Moinhos de Vento,
Porto Alegre, Rio Grande do Sul, Brazil.

gheno@pm.me

Received: 04 March 2025
Accepted: 07 June 2025
Published: 18 September 2025

DOI
10.25259/JCIS_36_2025

Quick Response Code:

ABSTRACT

Sample size is an essential step in any research study because it directly affects precision and statistical power. This article describes the main factors that determine the number of observations needed (power of a hypothesis test, significance criterion, minimum expected difference, variability, and asymmetry of the hypothesis test) and techniques for minimizing these factors. Our paper clearly presents examples of sample size calculations in radiology related to descriptive (mean and proportion) and comparative (two means, two proportions, intraclass correlations, and analysis of variance) studies. These examples are solved both manually and computationally using free R-based software.

Keywords: Analysis of variance, Intraclass correlation, Radiology, Sample size, Statistics

INTRODUCTION

Sample size is an important element of research studies, as it affects the statistical power and precision of the results. The goal of this procedure is to select an appropriate sample size that will correctly detect a specified difference at a given level of statistical significance or estimate an unknown parameter with a desired level of precision.[1-3]

In the field of radiology, inadequate sample size is one of the most common statistical errors found in articles submitted for publication. Unfortunately, these errors are rarely reported and few readers are aware of their importance.[1,4] On the one hand, large samples can lead to a waste of time, effort, and resources, especially if availability is limited, on the other hand, small samples can generate low statistical power or inaccurate estimates.[1,5] Furthermore, sample size is also a fundamental issue in experiments involving human beings or animals for ethical reasons.[1]

Researchers often decide on the sample size based on previous studies, either arbitrarily or according to a conventional rule (e.g., selecting 30 or more observations).[2,6-8] This last approach is usually sufficient for maintaining the central limit theorem and using approximations to the normal theory for measures such as the standard error of the mean.[7,9] However, this number may not be suitable in many situations because the appropriate number of observations depends on various characteristics, such as the statistical test used, the desired level of precision, and the study design.[2,8,10]

In general, the sample calculation can be based on an analysis of precision or power, which is usually carried out by controlling for type I (significance level) and type II (power) errors,

which occur when we test hypotheses.[3] We usually test two hypotheses, a null hypothesis (H0), which states that there are no differences between the groups in terms of mean or proportion, and an alternative hypothesis (H1), which contradicts H0, which states that there are such differences. If the null hypothesis is rejected when it is true, a type I error occurs. If the null hypothesis is not rejected when it is false, there is a type II error. The probabilities of making type I and type II errors are denoted by α and β, respectively [Figure 1] An upper limit for α is the significance level of the test, while the power of the test is defined as the probability of correctly rejecting the null hypothesis when the null hypothesis is false, i.e., power = 1 - β. Typically, the aim of research is to avoid a type I error and, at the same time, reduce a type II error. In general, when the sample size is fixed, it decreases α as it increases β and increases α as it decreases β. Thus, the only approach to decrease α and β simultaneously is to increase the sample size.[3,10]

The aim of this paper is to describe in a pragmatic way the factors that determine the number of observations needed in radiological studies and ways to obtain optimal sample sizes in descriptive (mean and proportion) and comparative (two means, two proportions, intraclass correlation [ICC], and analysis of variance [ANOVA]) studies. Practical examples will also be presented, both solved manually and using R-based free statistical software.

FACTORS DETERMINING SAMPLE SIZE

There are several approaches to calculating the sample size, the main factors that influence its size are: The power of a hypothesis test, the significance criterion, the minimum expected difference, the variability, and the asymmetry of the hypothesis test.[1,10,11]

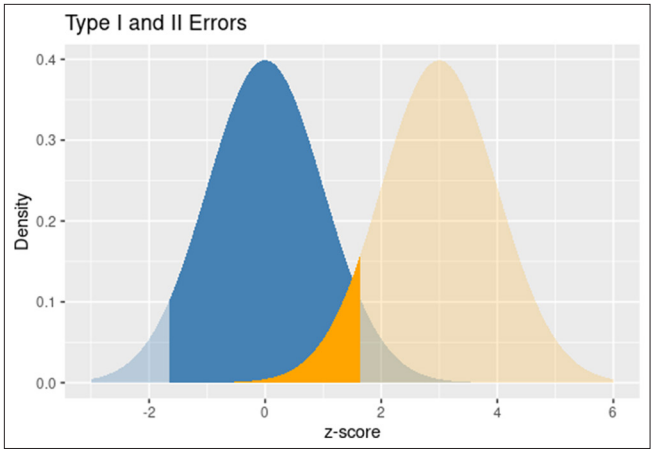


Figure 1: Type I error (light blue, bicaudal) and type II error (dark orange), with probabilities α and β, respectively; 1 - Type II error (dark blue), 1 - Type II error (light orange).

The statistical power of a hypothesis test is the probability that a statistical test will indicate a significant difference when it actually exists and is usually at least 0.8.[12] However, many clinical trial experts advocate for 90% power as it further reduces the Type II error rate by half compared to 80% and offers greater flexibility if assumptions about variability or recruitment prove incorrect, although it necessitates a larger sample size.[13] In a study with insufficient power, there is a risk of mistakenly accepting the null hypothesis when the alternative is true, which we call a type II or β error. Type I or α error occurs when we reject the null hypothesis when it is true. As the sample size increases, the power increases, while keeping the other conditions fixed [Figure 1]. Although high power is always desirable, there is an obvious trade-off between the number of individuals that can be feasibly studied, usually considering a fixed period of time, and the resources available to carry out a study.[11,12,14]

The significance criterion is the maximum *P*-value for which a difference should be considered statistically significant, usually defined as 0.05, where the *P*-value is the probability of observing data as extreme or more extreme than that observed in the study, purely by chance, assuming the hypothesis is true. As the significance criterion is reduced, made more strict, the sample size needed to reject the null hypothesis increases.[11,12] Although this threshold is described as arbitrary and not based on scientific validity, the 5% significance level is the standard that currently exists in the literature.[10] A value <0.05 is primarily used when it is important to avoid a Type I error, e.g., drug studies, one sided-tests.[15,16] The 95% confidence interval (CI) is also related to significance at the 0.05 level; if a 95% CI does not include the null value, it indicates a statistically significant difference at this level.[16,17]

The minimum expected difference is the smallest measured difference between the comparison groups that the researcher would like the study to detect or between the null and alternative hypotheses, depending on the type of study.[11,12]

The measure of variability is represented by the expected standard deviation (SD) of the measurements taken in each comparison group. As statistical variability increases, so does the necessary sample size to detect the minimum difference. Variability can be obtained from the literature or from pilot studies.[1,11,14]

The asymmetry of hypothesis testing refers to whether the approach is one-tailed or two-tailed. Studies involving one-tailed tests generally require a smaller sample size than those involving two-tailed tests. However, one-tailed tests should only be used when the direction of the test is evident. For example, this would be the case if the value of the alternative hypothesis was greater than or less

than that of the null hypothesis rather than simply being different.[11]

Choosing realistic values for effect size, SD, and desired CI width is a crucial yet often challenging step in sample size determination.[1,18] This value is often subjective and should be based on clinical judgment, experience, and expertise in the specific research area, rather than solely on arbitrary conventions or overly optimistic assumptions.[12,15,18] Since these parameter values involve assumptions and potential guesswork, a practical approach is to perform sample size calculations using a range of plausible values to understand the impact on the required sample size and aid in selecting the most appropriate trade-off given study resources.[19]

SOFTWARE

In recent years, various software programs and websites have been developed that can calculate sample sizes for different types of studies.[16,18] The support offered by these programs varies, as do their interfaces, mathematical formulas, and assumptions.[18]

Among these alternatives, R is one of the most popular software, because, in addition to being a programming language and an environment for carrying out statistical analyses, it is freely available as part of the GNU project, meaning users are free to run, copy, distribute, study, change, and improve the software.[20] R has a basic set of packages that provide a substantial collection of useful functions for calculating sample size. Its active community of users continually extends these functionalities by implementing commonly used statistical and computational tasks.[20,21]

Although using R requires some basic computer skills, many internet applications have been created with R packages that perform various functions related to sample calculation, without the need to install the program or have programming skills.[22,23] One of these applications that have been gaining popularity is power and sample size for health researchers (PSS Health), written in R with added packages (presize, stats, EnvStats, ICC. Sample Size, etc.), which can be used directly in R, with the PSS Health package, or through the Internet, through the website: https://hcpa-unidade-bioestatistica.shinyapps.io/PSS_Health.[21,22,24-26]

TYPES OF STUDY

This article presents methods for determining sample size in two broad categories: Descriptive studies, which aim to describe one or more characteristics of a group using means or proportions, and comparative studies, where the objective is to analyze and evaluate variables in different subjects to detect correlations and relationships [Table 1].

DESCRIPTIVE STUDIES

In descriptive studies, the aim is simply to describe one or more characteristics of a group, using means or proportions/percentages.[11,12] In these studies, the sample size is important because it affects the degree of precision of the estimates of means and proportions. The minimum clinically expected difference in a descriptive study reflects the difference between the upper and lower values of an interval and can be expressed as a percentage.[12] There are three important uses of this type of study: Hypothesis generation, planning, and trend analysis.

Table 1: Summary of the main aspects related to the types of studies presented in the article.

Study type	Estimated parameter	Inputs
Descriptive studies	Mean	- Critical value - Standard deviation - Width of the confidence interval
	Proportion	- Critical value - Estimate of the proportion - Amplitude of the interval
Comparative studies	Two means (dependent groups)	- Standard deviation - Critical value for significance - Critical value for power - Range of difference of means
	Two proportions (Chi-squared or Z test based on normal approximation)	- Estimations of the proportions to be compared - Amplitude - Average proportion - Critical value for significance - Critical value for power
	ICC (measures consistency/ reliability)	- Critical value for significance - Critical value for power - Number of evaluators - ICC value under the null hypothesis - ICC value of the alternative hypothesis
	Three or more independent means: Analysis of variance	- Means for each group - Standard deviation - Significance level - Power

ICC: Intraclass correlation

Estimating a mean parameter

Example 1. A radiologist wants to describe the mean range of the joint space between the scaphoid and the semilunar in a group of patients with connective tissue disorders who had wrist radiographs. It is desired that the limits of the 95% CI should be no more than 0.1 mm above or 0.1 mm below the mean interval. According to the literature, the SD of this interval is 0.4 mm.

In the case where the mean parameter is to be estimated, the following formula can be used to obtain the appropriate sample size:

$$n = \frac{(4 \cdot Z_{crit}^2 \cdot \sigma^2)}{d^2} \quad (1)$$

where n = sample size, Z_{crit}= critical value ($= Z\frac{\alpha}{2}$), σ = SD (being σ^2 the population SD), and d = width of the CI; the critical value is defined by the researcher [Table 2].[12,27]

Thus, d = 0.2 mm, σ = 0.4 mm, and Z_{crit} = 1.96. According to equation 1, a sample of 61 radiographs from different individuals would be sufficient.

In R, we can load the PSS Health package and use the following syntax:

presize: prec_mean (mean = 0, SD = 0.4, conf.width = 0.1*2, conf.level = 95/100).

To use PSS Health to calculate the sample size for this example, the user simply accesses the application's main page and selects the "Averages" tab and the "One sample" item. Next, fill in the values shown in this example in the fields indicated. The software indicates that a sample of 64 radiographs would be sufficient.

The difference found between the manual and computational methods is due to the fact that the program considers the unknown SD and uses the t distribution instead of the Z distribution. Similar differences due to this reason are also observed in some of the following examples.

Estimating a proportion parameter

Table 2: Standard normal deviation (Z_{crit}) corresponding to selected significance criteria.

Significance criterion	Z_{crit}-value
0.01	2.58
0.02	2.33
0.05	1.96
0.1	1.65

Example 2. A researcher wants to estimate the accuracy of chest radiographs in diagnosing an adult community-acquired pneumonia, with a 95% CI for ±10%. Based on a previous study, the accuracy is 76%.

To estimate the proportion of a descriptive study, we can use the following equation:

$$n = \frac{4 \cdot Z_{crit}^2 \cdot p(1-p)}{d^2} \quad (2)$$

Equation 2. n = sample size, Z_{crit} = critical value, p = estimate (pre-study) of the proportion to be measured, d = amplitude of the interval.[12,27]

If the researcher does not know the expected value, they should opt for the worst-case scenario and choose one of 0.5, given the characteristic of the proportion, which varies between 0 and 1. Therefore, d = 0.2, p = 0.8, and Z_{crit} = 1.96. Applying equation 2, we obtain an n of 70.

In R, we can load the PSS Health package and use the following syntax:

presize: prec_prop (P = 76/100, conf.width = 20/100, conf. level = 95/100, method = "wald").

To perform this calculation using the PSS Health application, go to the main page and select the "Proportions" tab and the "One sample" item. Then, fill in the values used in the formula above in the fields indicated. The sample size obtained using the software was 71.

COMPARATIVE STUDIES

The objective of a comparative study is to analyze and evaluate one or more variables in different subjects using quantitative and qualitative methods to detect associations, correlations, and relationships.

Two means: Two dependent groups

Example 3. A radiologist wants to compare the cross-sectional area of brain aneurysm using CT and MRI scans. Based on the literature, an estimated 5% difference is expected between the two methods, with MRI having a mean cross-sectional area of 61 mm^2 and CT 58 mm^2, with an SD of 8 mm^2. The significance criterion was set at 0.05, and the power was set at 0.8.

For studies comparing two means, we can use the following equation:

$$n = \frac{4 \cdot \sigma^2 \cdot (Z_{crit} + Z_{pow})}{d^2} \quad (3)$$

Equation 3. n = sample size, σ = SD, Z_{crit} = critical value,

Z_{pow} = power, d = range of difference of means.[28,29] Z_{pow} can be obtained from R using the qnorm () function, in this case just type: qnorm(0.8).

Therefore, $d = 3$, $\sigma = 8$, $Z_{crit} = 1.96$, and $Z_{pow} = 0.842$. Applying equation 3, we get an n of 77.

In R, we can load the PSS Health package and use the following syntax:

stats: power.t.test (n = NULL, power = 80/100, sig. level = 5/100, delta = abs(3), SD = 8, type = "paired," alternative = "two.sided").

If we choose to perform this calculation using the PSS Health application, we simply need to go to the main page and select the "Averages" tab, followed by the "Two dependent groups" option. Then, we enter the values used in the formula above into the indicated fields. The sample size for the software was 58.

Two proportions: Two groups

Example 4. A researcher wants to evaluate the effectiveness of using artificial intelligence software to detect fractures on radiographs in an emergency department. According to the literature, the diagnostic sensitivity for detecting fractures in conventional radiographs is approximately 65%. For this study, two sets of radiographs will be selected, one for conventional evaluation and one for evaluation with the aid of artificial intelligence. The researchers believe that sensitivity will increase to approximately 75%, which would justify the cost/benefit of this technique. A significance criterion of 0.05 and a power of 0.8 were chosen.

For studies where two proportions are compared with an X^2 or Z test, which is based on a normal approximation of the binomial distribution, the following formula can be used:

$$n = \frac{2 \cdot \left[Z_{crit} \cdot \sqrt{2 \cdot p' \cdot (1 - p')} + Z_{pow} \cdot \sqrt{p1(1 - p1) + p2(1 - p2)} \right]^2}{d^2} \quad (4)$$

Equation 4. $p1$ and $p2$ = (pre-study) estimations of the proportions to be compared, d = amplitude ($p1$ - $p2$ or minimum expected difference), $p' = (p1+p2)/2$, Z_{crit} = critical value, Z_{pow} = power.[12,28]

Therefore, $p1 = 0.65$, $p2 = 0.75$, $d = 10$, $p' = 70$, $Z_{crit} = 1.96$, and $Z_{pow} = 0.842$.

Applying equation 4, we get an n of 658, which will be divided into two groups of 329 samples.

In R, we can load the PSS Health package and use the following syntax:

EnvStats: propTestN (p.or.p1 = 75/100, p0.or.p2 = 65/100, alpha = 5/100, power = 80/100, sample type = "two sample," alternative = "two sided," ratio = 1, correct = FALSE, warn = FALSE).

To perform this calculation using the PSS Health application, we simply go to the main page, select the "proportions" tab, followed by the "two independent groups" option. Then, we fill in the formula's values in the indicated fields. Using the software, the sample size was 658, with two groups of 329.

Two or more quantitative measures: ICC

Example 5. Two radiologists want to investigate the agreement of their measurements of urinary bladder neoplasms in hospitalized patients. They considered a significance criterion of 0.05, a power of 0.8, and ICC values of 0 under the null hypothesis and 0.5 under the alternative hypothesis

The ICC measures the consistency between two or more measures, is considered an important indicator of reliability, and is commonly used in measures of agreement, both intraobserver and interobserver.[29,30] The ICC is a value between 0 and 1, in which reliability can be defined as low (<0.5), moderate (≥0.5 and <0.75), good (≥0.75 and <0.9), and excellent (≥0.9).[31]

The sample size can be calculated using the following derived formula:

$$n = 1 + \frac{2 \cdot (Z_{crit} \cdot Z_{pow})^2 \cdot k}{(InCo)^2 \cdot (k - 1)},$$

$$Co = \frac{1 + k\theta 0}{1 + k\theta 1},$$

$$\theta 0 = \frac{R0}{1 - R0},$$

$$\theta 1 = \frac{R1}{1 - R1} \quad (5)$$

Equation 5. Z_{crit} = 1.96 and Z_{pow} = 0.842, k = number of evaluators, $R0$ = ICC value under the null hypothesis, $R1$ = ICC value of the alternative hypothesis.[32]

Using equation 5, a sample of 27 radiographs of different lower limbs should be obtained.

In R, we can load the PSS Health package and use the following syntax:

ICC Sample Size: calculate Icc Sample Size (P = 0.5, p0 = 0, k = 2, alpha = 5/100, power = 80/100, tails = 2).

In the PSS Health application, simply access the main page and select the "Concordance" tab, followed by the "ICC" item. Then, fill in the chosen values in the indicated fields. The calculated sample size was 28.

Three or more independent means: ANOVA

Example 6. A researcher wants to compare the degree of deep tumor invasion in patients with squamous cell carcinoma between three different methods, ultrasound, computed tomography, and magnetic resonance imaging. Based on the literature, the estimated means for the ultrasound, CT, and MRI groups are 3.6, 6.1, and 6.1 mm, respectively, with an SD of 2 mm. The desired power is 0.8 and the desired significance level is 0.05.

For studies comparing three or more means, we can use the formula:

$$n = \frac{\lambda}{\Delta},$$

$$\Delta = \frac{1}{\sigma^2}\sum_{i=1}^{k}(\mu_i - \overline{\mu})^2,$$

$$\overline{\mu} = \frac{1}{k}\sum_{j=1}^{k}(\mu_j) \qquad (6)$$

Equation 6. λ comes from Table 3, where we selected powers of 0.8 or 0.9 and significance levels of 0.01 or 0.05.[3]

Using equation 6 and Table 3, an n of 10 is recommended for each group.

In R, we can load the PSS Health package and use the following syntax:

EnvStats: aovN (mu.vec = c (3.6, 6.1, 6.1), sigma = 2, alpha = 5/100, power = 80/100, n.max = 1E5).

In the PSS Health application, simply access the main page and select the "Averages" tab, followed by the "One-way ANOVA" option. Then, fill in the chosen values in the indicated fields. The calculated sample size was 11 for each group.

Table 3: Values λ that satisfy the equation $X^2_{k-1}(X^2_{\alpha,k-1}|\lambda) = \beta$ k=number of groups you want to compare.

	1-β=0.80		1-β=0.90	
k	**α=0.01**	**α=0.05**	α=0.01	α=0.05
2	**11.68**	**7.85**	14.88	10.51
3	**13.89**	**9.64**	17.43	12.66
4	**15.46**	**10.91**	19.25	14.18
5	**16.75**	**11.94**	20.74	15.41
6	**17.87**	**12.83**	22.03	16.47
7	**18.88**	**13.63**	23.19	17.42
8	**19.79**	**14.36**	24.24	18.29
9	**20.64**	**15.03**	25.22	19.09
10	**21.43**	**15.65**	26.13	19.83
11	**22.18**	**16.25**	26.99	20.54
12	**22.89**	**16.81**	27.80	21.20
13	**23.57**	**17.34**	28.58	21.84
14	**24.22**	**17.85**	29.32	22.44
15	**24.84**	**18.34**	30.04	23.03
16	**25.44**	**18.82**	30.73	23.59
17	**26.02**	**19.27**	31.39	24.13
18	**26.58**	**19.71**	32.04	24.65
19	**27.12**	**20.14**	32.66	25.16
20	**27.65**	**20.56**	33.27	25.66

The values in bold were used only to differentiate the two approaches, with power between 0.8 and 0.9

DISCUSSION

This paper presented some relatively simple ways to calculate sample sizes in different contexts, using manual and computational methods.

The manual methods for calculating sample size presented in this article are based on samples that supposedly come from a normal distribution.[12,27] This normality can be visually assessed using histograms or tested statistically, for example, with the Shapiro–Wilk test.[33] In addition, when comparing groups, sample size calculations often assume that the variance is equal between the groups being compared (homoscedasticity).[12,33,34] This assumption is embedded in formulas for sample size calculation, such as those used for comparing two means, where a single SD is assumed to be equal for both comparison groups. The estimated measurement variability, represented by this assumed SD, is a critical parameter; as it increases, the required sample size to detect a specified difference also increases.[11,12]

When we compare the results of the sample calculation obtained by manual and computer methods, we notice variations in the values in some cases. Small variations can be explained by rounding issues, while larger variations may be due to how the software implemented the calculations. For example, the software may assume a Student's t distribution instead of a normal distribution when the population deviation is unknown.[35-38] It should be emphasized that a normal approximation to the t-distribution can be poor for small sample sizes, potentially leading to an overestimation of power or an underestimation of the required sample size. It has been suggested that the normal approximation is

acceptable if the sample size in each arm is at least 30.[15,34] The t-distribution is suitable when the sample is small and the population SD is unknown, requiring the use of the sample SD; the t-distribution, which is wider than the normal distribution and depends on the sample size, takes this additional uncertainty into account.[17] In addition, when performing manual calculations, it is important to retain as many significant digits as possible until the last step in a sequence of calculations and, when obtaining the result of the final step, round up to the appropriate number of digits.[39]

In addition to the variables affecting sample size, there are other strategies that can minimize it, such as using continuous variables, taking paired measurements and expanding of the minimum expected difference. Although radiological tests often present binary results with categorical answers, it is important to inform researchers that continuous variables add more statistical power because they incorporate mathematical properties more effectively than those related to proportions.[12] With regard to paired studies, they are more robust than unpaired studies because each measurement is paired with its own control, resulting in lower SD.[6,10,12] As for expanding the minimum expected difference, this can be increased in some cases, especially in preliminary studies used as the basis for larger studies.[12]

Other aspects of the study design that could be considered pitfalls can affect the sample size. These include correlated data and prospective studies. In the former, more than one observation is taken per patient. These observations of the same subjects are not statistically independent and are considered correlated. This requires a different approach to correctly calculate the sample size.[6,40] In the second case, researchers must consider the dropout rate. A large discrepancy between the calculated and obtained samples can distort the analysis and generalization of the results. This rate can be obtained from previous studies in the literature or by adjusting the sample calculation.[14,18,41]

In situations where it is not possible to achieve the minimum sample, either for economical or ethical reasons, other alternatives can be recommended, such as reducing the scope of the study (for example, keeping more factors fixed) or proposing it as part of a sequence of studies.[1,42] In the event of doubts or, above all, when evaluating more complex cases, close and honest collaboration between the researcher and the statistician becomes imperative.

R software is increasingly preferred over other tools such as GPower, Stata, Statistical analysis system (SAS), and the Statistical Package for the Social Sciences (SPSS) for several reasons. While GPower is highlighted as an easy-to-use and free tool specifically designed for sample size and power calculations across various statistical tests, it is not a complete statistical software like the others.[14] Other software such as SAS, SPSS, or STATA, because they are paid for and have proprietary codes, have a smaller community of users and developers, affecting the sharing of knowledge, innovation, testing, and feedback. R is presented as a more flexible and powerful programming language and environment. The open-source nature of R makes it freely available and supported by a large community of professionals and academicians who contribute to its extensive collection of well-documented packages and functions. This package system allows R to be tailored to meet individual statistical needs and offers extensive integration with complex statistical approaches. R facilitates the modeling of complex situations and integration with primary data analysis and is particularly well-suited for computationally intensive statistical and mathematical methods such as simulation analysis, Bayesian inference, and advanced parameter estimation techniques. Furthermore, R offers advantages in documentation, transparency, automation, troubleshooting, and reproducibility (as code can be easily shared and rerun) and provides superior graphing capabilities for creating publication-quality figures. These combined attributes, particularly its flexibility for integrating various stages of research from data analysis and modeling to simulation and reporting within a single environment, contribute to R's increasing popularity, being described as the fastest-growing software in some areas of health research.[16,20,22]

Sample size calculations are rarely reported by clinical investigators for diagnostic studies, including diagnostic accuracy studies and agreement studies in radiology. Instead of following a formal calculation process, sample size is often determined arbitrarily or based on convenience and available resources such as limitations in patient volume, research time, or money.[43] The clinical implications of inadequately sized studies are considerable and raise significant ethical concerns. Underpowered studies, that have too few participants, have a high risk of a Type II error (false negative). This can lead to the erroneous conclusion that no difference exists when one is merely hidden by the small sample size. Such studies can expose participants to potential risks or inconveniences without a sufficient probability of generating meaningful findings, representing a waste of valuable resources.[1,12,18] Conversely, overpowered studies, by enrolling more participants than necessary, can find statistically significant differences that are not clinically important. This risks misdirecting clinical practice based on trivial findings. They can unnecessarily expose more individuals to study interventions (which may carry risks or involve withholding a potentially beneficial treatment), and they consume resources that could be better used elsewhere.[1,15,18]

The study's limitations include the absence of Bayesian method for sample size calculation. This method explicitly incorporates prior information about the parameter of

interest through a prior probability distribution.[36] By integrating this prior information with the potential data that could be observed (through a predictive distribution), Bayesian approaches can lead to sample size criteria based on concepts like the average coverage probability or average length of credible intervals over all possible data sets, weighted by the predictive distribution. This more comprehensive and efficient use of prior information is often cited as the reason why Bayesian sample size estimates frequently suggest smaller sample sizes compared to corresponding frequentist estimates.[36,44] Implementing these Bayesian sample size criteria typically requires numerical methods because closed-form analytical solutions are often unavailable. Monte Carlo simulations are a commonly used technique for this purpose, which involves generating multiple sets of hypothetical data according to the predictive distribution (which is derived from the prior information and proposed sample size).[36,44,45] There were also no approaches to calculating sample size in studies involving the receiver operating characteristic curve due to its complexity.[6] Furthermore, of the numerous statistical software programs available, only R and applications that depend on it were analyzed, based on its characteristics, which have already been mentioned.

CONCLUSION

In summary, we have described various methods for calculating sample size, including manual and computational approaches that are straightforward and quick to implement. We have also outlined the primary factors influencing the results and strategies for optimizing them. Future studies could address more advanced sample size techniques in a way that is understandable to professionals from different backgrounds who need this knowledge to properly design their projects.

Ethical approval: The Institutional Review Board approval is not required.
Declaration of patient consent: Patient's consent is not required as there are no patients in this study
Financial support and sponsorship: Nil.
Conflicts of interest: There are no conflicts of interest.
Use of artificial intelligence (AI)-assisted technology for manuscript preparation: The authors confirm that they have used artificial intelligence (AI)-assisted technology to assist in the writing or editing of the manuscript or image creations.

REFERENCES

1. Lenth RV. Some practical guidelines for effective sample size determination. Am Stat 2001;55:187-93.
2. Verma JP, Verma P. Determining sample size and power in research studies: A manual for researchers. 1st ed. Singapore: Springer; 2020. p. 3-6
3. Chow SC. Sample size calculations in clinical research. 3rd ed. United States: CRC Press; 2017. p. 8-9.
4. Levine D, Bankier AA, Halpern EF. Submissions to radiology : Our top 10 list of statistical errors. Radiology 2009;253:288-90.
5. Cochran WG. Sampling techniques. 3rd ed. United States: Wiley; 1977. p. 72-3.
6. Hajian-Tilaki K. Sample size estimation in diagnostic test studies of biomedical informatics. J Biomed Inform 2014;48:193-204.
7. Daniel WW, Cross CL. Biostatistics: A foundation for analysis in the health sciences. 11th ed. United States: Wiley; 2019. p. 189-92.
8. Lakens D. Sample size justification. Collabra Psychol 2022;8:33267.
9. Kwak SG, Kim JH. Central limit theorem: The cornerstone of modern statistics. Korean J Anesthesiol 2017;70:144-56.
10. Beam CA. Strategies for improving power in diagnostic radiology research. AJR Am J Roentgenol 1992;159:631-7.
11. Rodríguez Del Águila M, González-Ramírez A. Sample size calculation. Allergol Immunopathol (Madr) 2014;42:485-92.
12. Eng J. Sample size estimation: How many individuals should be studied? Radiology 2003;227:309-13.
13. Flight L, Julious SA. Practical guide to sample size calculations: An introduction. Pharm Stat 2016;15:68-74.
14. Kang H. Sample size determination and power analysis using the G*Power software. J Educ Eval Health Prof 2021;18:17.
15. Mascha EJ, Vetter TR. Significance, errors, power, and sample size: The blocking and tackling of statistics. Anesth Analg 2018;126:691-8.
16. Serdar CC, Cihan M, Yücel D, Serdar MA. Sample size, power and effect size revisited: Simplified and practical approaches in pre-clinical, clinical and laboratory studies. Biochem Medica (Zagrab) 2021;31:010502.
17. Medina LS, Zurakowski D. Measurement variability and confidence intervals in medicine: Why should radiologists care? Radiology 2003;226:297-301.
18. Althubaiti A. Sample size determination: A practical guide for health researchers. J Gen Fam Med 2023;24:72-8.
19. Schmidt SA, Lo S, Hollestein LM. Research techniques made simple: Sample size estimation and power calculation. J Invest Dermatol 2018;138:1678-82.
20. Jalal H, Pechlivanoglou P, Krijkamp E, Alarid-Escudero F, Enns E, Hunink MG. An overview of R in health decision sciences. Med Decis Making 2017;37:735-46.
21. Core Team R. R: A language and environment for statistical computing; 2025. Available from: https://www.r-project.org [Last accessed on 2025 Jun 01].
22. Borges RB, Mancuso AC, Camey SA, Leotti VB, Hirakata VN, Azambuja GS, *et al*. Power and sample size for health researchers: A tool for calculating sample size and test power for health researchers da área da saúde. Clin Biomed Res 2021;40:247-53.
23. Doi J, Potter G, Wong J, Alcaraz I, Chi P. Web application teaching tools for statistics using R and shiny. Technol Innov Stat Educ 2016;9:1-32.
24. Haynes AG, Lenz A, Stalder O, Limacher A. Presize: An R-package for precision-based sample size calculation in clinical research. J Open Source Softw 2021;6:3118.
25. Millard S, Kowarik A. EnvStats: An R package for environmental statistics; 2025. Available from: https://cran.r-

project.org/web/packages/envstats/index.html [Last accessed on 2025 Jun 01].
26. Rathbone A, Saurabh S, Kumbhare D. ICC sample size: Calculation of sample size and power for ICC; 2015. Available from: https://cran.r/project.org/web/packages/icc.sample.size/index.html [Last accessed on 2025 Jun 01].
27. Snedecor GW, Cochran WG. Statistical methods. 8th ed. United States: Iowa State University Press; 1989. p. 102-5.
28. Feinstein AR. Principles of medical statistics. Boca Raton: Chapman and Hall/CRC; 2002.
29. Ionan AC, Polley MY, McShane LM, Dobbin KK. Comparison of confidence interval methods for an intra-class correlation coefficient (ICC). BMC Med Res Methodol 2014;14:121.
30. Armstrong RA. Should Pearson's correlation coefficient be avoided? Ophthalmic Physiol Opt 2019;39:316-27.
31. Bobak CA, Barr PJ, O'Malley AJ. Estimation of an inter-rater intra-class correlation coefficient that overcomes common assumption violations in the assessment of health measurement scales. BMC Med Res Methodol 2018;18:93.
32. Walter SD, Eliasziw M, Donner A. Sample size and optimal designs for reliability studies. Stat Med 1998;17:101-10.
33. Anvari A, Halpern EF, Samir AE. Statistics 101 for radiologists. Radiographics 2015;35:1789-801.
34. Harrison DA, Brady AR. Sample size and power calculations using the noncentral t-distribution. Stata J Promot Commun Stat Stata 2004;4:142-53.
35. Ryan TP. Sample size determination and power. United States: John Wiley and Sons; 2013. p. 114.
36. Adcock CJ. Sample size determination: A review. Statistician 1997;46:261-83.
37. Altman DG. Practical statistics for medical research. Boca Raton: Chapman and Hall/CRC; 1999. p. 455-60.
38. Kirkwood BR, Sterne JA. Essential medical statistics. 2nd ed. United States: Blackwell Science; 2009. p. 413-7.
39. Zar JH. Biostatistical analysis. 5thed. United States: Prentice Hall; 2010. p. 6.
40. Liu G, Liang KY. Sample size calculations for studies with correlated observations. Biometrics 1997;53:937-47.
41. Moore CM, MaWhinney S, Forster JE, Carlson NE, Allshouse A, Wang X, *et al*. Accounting for dropout reason in longitudinal studies with nonignorable dropout. Stat Methods Med Res 2017;26:1854-66.
42. Wittes J. Sample size calculations for randomized controlled trials. Epidemiol Rev 2002;24:39-53.
43. Bachmann LM, Puhan MA, Riet GT, Bossuyt PM. Sample sizes of studies on diagnostic accuracy: Literature survey. BMJ 2006;332:1127-9.
44. Joseph L, Du Berger R, Bélisle P. Bayesian and mixed bayesian/likelihood criteria for sample size determination. Stat Med 1997;16:769-81.
45. Cheng D, Branscum AJ, Stamey JD. A Bayesian approach to sample size determination for studies designed to evaluate continuous medical tests. Comput Stat Data Anal 2010;54:298-307.

How to cite this article: Gheno R, Borges RB, Dos Reis RC. Sample size matters: A step-by-step guide for radiologists. J Clin Imaging Sci. 2025;15:34. doi: 10.25259/JCIS_36_2025