Symposium

Sample size matters: A guide for urologists

Peter Hinh, Steven E. Canfield

Division of Urology, University of Texas Medical School at Houston, Houston, Texas, USA

ABSTRACT

Understanding sample size calculation is vitally important for planning and conducting clinical research, and critically appraising literature. The purpose of this paper is to present basic statistical concepts and tenets of study design pertaining to calculation of requisite sample size. This paper also discusses the significance of sample size calculation in the context of ethical considerations. Scenarios applicable to urology are utilized in presenting concepts.

Key words: Evidence-based practice, research, statistics

INTRODUCTION

Urology is evolving to emphasize evidence-based practice. This movement is driven by the desire to deliver scientifically substantiated, optimal healthcare approaches applied at patient centred and population wide levels. Decisions based on evidence are driving medical guidelines, healthcare policies, and insurance reimbursements. An integral part of developing and utilizing evidence-based practice is understanding what makes an individual study valid, and an important component of validity is the role of sample size in study design and critical appraisal.

For researchers, sample size impacts the scale of the trial needed to answer research questions. Once a sample size is calculated, investigators can determine if they will have the resources or be able to recruit enough patients for the study. Investigators also balance ethical components for unknown clinical interventions, which risk both passive harm from inferior therapy as well as active harm from unknown adverse effects, and weigh exposure to these unknowns against the needed sample size to answer the proposed question. This can be particularly relevant in a surgical field such as urology, where the clinical intervention may result in irreversible modifications to a person's body or functioning.

For practitioners, detailed reporting of the study designs and methods, including descriptions of sample size calculations, are necessary in published papers. This practice allows the reader to determine whether the arrived at conclusion was based on proper principles or assumptions. Sample size is especially relevant when study findings are negative, and a consideration should be if the study possessed adequate sample size to declare equivalence. Recognizing the importance of sample size in study design, the Consolidated standards of reporting trials (CONSORT) guidelines recommend detailing the methodologies used in sample size calculation.[1]

The purpose of this article is to present principles underlying sample size calculation, and to illustrate how these factors interact and drive the dynamics of statistical power. We will show how a practical understanding of these factors helps formulate questions, devise research strategies, and critically appraise conclusions. This paper will focus on aspects of study designs that are directly relevant to sample size calculation, and to the many components, which influence this calculation [Figure 1].

CLINICAL SCENARIO

A 52-year-old woman presents with stress urinary incontinence (SUI). After standard work-up, you determine that she would benefit from a mid-urethral sling procedure. You offer her a transobturator tape (TOT) procedure. She wonders if you could do a "mini-sling" for her instead, which her friend has had done recently. After a literature search using the "PICO" (problem, intervention, comparison,

For correspondence: Dr. Steven E. Canfield,
6431 Fannin, MSB 6.018, Houston, TX 77030, USA.
E-mail: steven.canfield@uth.tmc.edu

Access this article online	
Quick Response Code:	Website: www.indianjurol.com
	DOI: 10.4103/0970-1591.91442

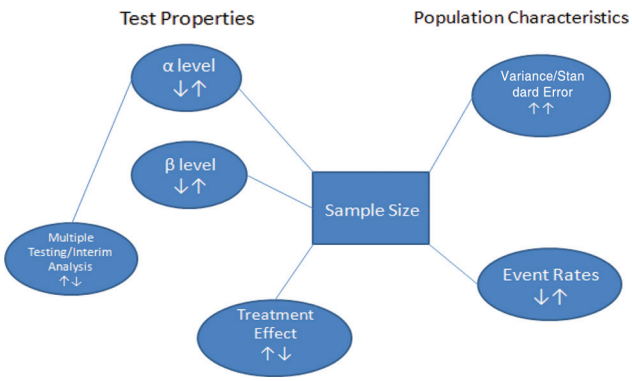

Figure 1: Variable and their relationship to sample size

outcome) question, "In women with SUI, how does the "mini-sling" compare to the TOT procedure for resolution of stress incontinence?", you find a clinical trial reporting the efficacy of the mini sling (MiniArc) compared to traditional transobturator tape (Monarch).[2] The investigators performed a nonrandomized, retrospective cohort study in which 75 patients had a MiniArc and 55 patients had a traditional Monarch midurethral sling placement for treatment of stress urinary incontinence. They report results at 6 weeks and 1 year for the endpoints of cough stress test, quality of life, pad usage, and symptoms score. At one year, 85% of the mini sling (MiniArc) group experienced resolution of incontinence based on negative cough stress test, versus 89% of the TOT (Monarch) group, with a reported *P* value of 0.60. The authors concluded that no difference existed between the two groups for SUI treatment efficacy. Before proceeding with the mini sling, you wonder if this negative study possessed an adequate sample size to draw this conclusion.

STATISTICAL BACKGROUND

The basics of variables and variability

The primary goal of sampling is to take a smaller, but still truthful representation of a specific population, for which the effects of the studied intervention can then be generalized to that population.[3] This review focuses on factors for arriving at the requisite sample size, rather than on the intricacies of selecting a truly representative sample.

Data is evaluated on properties of both accuracy and precision, where accuracy is defined as the reproducibility of obtaining the results and precision is a measure of how close to the true value the results are. Measurements of both will inherently contain errors, though the goal of a well designed study is to minimize random and systemic errors.

The intrinsic variability that comes with measuring (sampling) an outcome's known or unknown attributes is termed the standard error (SE).[4] The SE can be derived from recognized patient characteristics, baseline occurrence rates, previous studies, or a pilot study. Outcome variables can be defined as binary (for example, incontinent or continent), categorical (classification groups such as poor, moderate, or severe), or continuous (objectively measurable, such as flow rate). Binary and categorical variables need proportional calculations, while continuous data requires means testing. The standard error in binary variables is directly calculated from the baseline rate, whereas for categorical or continuous variables, the standard error is calculated from the observed variations from the mean.[4] This variability fundamentally shapes sample size calculation. When variability is low, sampling will be more consistent and more likely to be representative, and the sample size can be consequently smaller due to the smaller likelihood of sampling error. Conversely, large variability creates a higher sampling error, which needs to be offset with a larger sample size.

Most variables studied in human populations occur around a central tendency, measurable as a mean, median, or mode. In statistical terms, this is labelled as a normal distribution. A smaller standard deviation denotes values are well clustered around a mean, whereas a larger standard deviation implies that values are spread out. For any value with normal distribution, the variance can be mitigated by incorporating a larger sample size. However, the sample size should not be arbitrarily large, and the proper sample size should be tempered by others factors, including resource scarcity and a desire to minimize potential harm or exposure to inferior therapy. Additionally, if an intervention is truly unknown, researchers may elect to perform a 2-stage study design, where the initial stage is examined with a lower threshold of therapy response and a second stage continued if therapy response is confirmed. This approach minimizes exposure to harm from inferior therapy.[5]

In contrast to variability, the confidence interval reflects the range of data, including derived results such as relative risks or odds ratios. We may look to confidence interval as a surrogate resides within the presented range. A wide confidence interval indicates lack of precision on measurement of the effect, and a confidence interval that covers zero effect denotes a lack of statistical significance. The same factors that influence confidence interval also directly relate to sample size, including variability of the outcome, the significance level selected, and the magnitude of the effect.

Effect size and clinical significance

Studies are designed around the principle question being investigated. This seemingly obvious statement carries many important implications and corollaries. Endpoints need to be carefully considered and congruent to the question.

Effect size characterizes the primary endpoint studied, and expected treatment effects or event rates must be deliberated beforehand. For example, expected mortality can vary dramatically across different urological diseases. Localized transitional cell carcinoma of the bladder has a

five-year survival of approximately 74.3%, whereas clinically localized prostate cancer has an overall five-year survival of 99.7% (not differentiating stage or grade).[6] As such, a study of a proposed therapeutic intervention with mortality as the primary end point would necessarily require a much larger scale in a prostate cancer trial as opposed to a bladder cancer trial due to the lower mortality rates (lower expected event rates) associated with localized prostate cancer.

The magnitude of the expected difference in outcome between two interventions will shape the trial size. When the treatment effect is anticipated to be large, it will be more readily apparent, and a smaller sample size is needed. Conversely, if the treatment effect is anticipated to be smaller, a compensatory increase in sample size is needed. Taken to extremes, a treatment effect may be so large that a comparative study would be moot, and sample size could theoretically be increased so high that statistical significance could be shown irrespective of its clinical utility. In a urological context, one might think of a novel intervention to treat stress urinary incontinence compared to current mid-urethral sling procedures. In theory, at a large enough sample size, a difference in continence rates of 2% could achieve statistical significance. However, it is doubtful that any urologist would consider this difference to be clinically relevant. It is important to consider that significance is also contextual. A 2% difference may be clinically significant if the endpoint was simultaneously rare and grave, such as with fatal pulmonary embolism. Thus, studies should declare *a priori* what the researchers deem to be a clinically significant difference, and they must be able to justify how they arrived at this figure. Investigators should define a treatment effect or outcome that represents the minimal change in magnitude, which would provide a clinically significant improvement.

Random error assignments

Studies are designed to test a hypothesis. By convention, the hypothesis is often stated as a null. That is, for any intervention, the null hypothesis states that the intervention does not affect the outcome. The data is then examined and the conclusion is made whether to accept or reject that null hypothesis.[7] At this point, incorrect conclusions may be reached in opposing directions. A type I error exists when the null hypothesis is incorrectly rejected, i.e., stating that a treatment effect does exist when in truth it does not. The type I error is related to the *P value* (represented by α), which defines the threshold needed to achieve significance. A type II error occurs when the null hypothesis is erroneously accepted, i.e., concluding that no treatment effect is observed when in actuality one does exist. The type II error is represented by β, and is directly related to the power of the study to detect a treatment effect. This relationship is mathematically expressed as: Power = 1- β. The paramount goal of proper sample size estimation is to minimize the possibility for these two errors to occur.

An acceptable type I error threshold is usually set at 5%. This implies that the likelihood of the outcome occurring by chance alone is less than one in 20. Type II error is usually set at 20%, meaning that the likelihood of not seeing a real effect would occur in one in five. Its direct corollary power is 80%, using the previously noted power calculation. These established conventions signify that researchers are more disinclined to make an erroneous association of effect than they are to fail to recognize a true effect. These assumptions may or may not hold true for a particular research question, so in some instances further emphasis will be placed on finding the association. For example, a study may be directed to find equivalence, and the power of the study is correspondingly increased. Depending on the research question, expected occurrence rates, and other mitigating factors, the values selected for type I and type II errors may be adjusted. Choosing a higher type I error rate may compromise the legitimacy on any positive findings, and choosing a lower type I error may have the effect of lowering the statistical power of the study.

Setting the P value, testing for benefit versus harm, and using multiple end points

While the most conventionally selected *P* value is 0.05, this number is arbitrary. Certain clinical questions may warrant either a more or less stringent significance level (for example, *P*=0.01 or 0.1). This decision should be driven by the severity of the outcome and weighed against the risks associated with intervention.

Another consideration is if the study will only look for a positive difference or if it will also examine the possibility of a detrimental effect. This is correspondingly termed either a unidirectional one-sided significance test or a non-directional, two-sided test. Performing the significance test in a non-directional manner effectively halves the α level, such that an α of 0.05 actually sets the significance level at a 0.025 probability for correctly (or erroneously) concluding that the intervention is significantly better than the comparative treatment. The two-sided test therefore requires a larger sample size and is appropriate when an unknown intervention may in fact be harmful compared to the control. It is generally the more favored significance test to use and, indeed, the CONSORT guidelines recommend two-sided test methodologies be used for comparative studies.[1]

Statistical significance is also influenced by examining multiple endpoints. Multiple hypotheses testing can compromise the alpha level and increases the likelihood of committing a type I error. A urology study on voiding, for example, may choose to evaluate multiple endpoints, including flow rate, post void residuals, symptoms scores, detrussor pressure, retention rates, surgical intervention rates, and more. It should be recognized that as more endpoints are considered, it becomes more likely that the

intervention will be found to achieve statistical significance for one of these endpoint, due to chance alone. This can be imperfectly adjusted for by computational corrections. The most common is the Bonferroni correction, where the threshold of significance level is lowered by dividing the alpha by *n*, where *n* is the number of hypothesise tested.[8] Depending on the number of endpoints, this may greatly reduce the significance of each statistical test. Indeed, some argue that the correction is too conservative and drives alpha to a prohibitive level.[8] Power is thus reduced as a larger sample may be needed to achieve this significance.

Compliance

Patient recruitment, compliance, and loss to follow-up are key concerns. The robustness of the sample size is also compromised by events of unrelated mortality, patient drop outs, contamination, and unplanned crossover. Thus, the planned sample size may need upward adjustment by 10% to 15% to account for these events.[9] Other assumptions used in sample size calculation such as treatment effects or event rates may prove to be incorrect, and it is also often prudent for researchers provide a mechanism for interim analysis, so sample size or target enrolment can be revised if needed, or treatment stopped if unexpected harm is detected.[10]

Power and sample size with example calculations

Power is directly related to the factors of sample size, effect size, and significance level. The effect size and the significance level are intrinsic to the clinical question being researched, where only a certain range is justifiable. This leaves sample size as the most accessible variable to the investigator. For any circumstance, increasing the sample size effectively increases the power of the study, although the power should be tempered to detect only differences that are clinically significant.

For dichotomous variables, required sample size can be estimated once the values of baseline events, appropriate level of significance, desired power of the study, and the clinical effect size are known. For continuous variables, an estimation of the variance in the sample data is also needed.

Consider these examples to illustrate how these factors interrelate. A randomized trial evaluates a novel technique for stress urinary incontinence with comparison to traditional mid-urethral sling. The endpoint is social continence. Suppose that traditional, mid-urethral slings have an efficacy of 80% in our defined study population, based on prior evidence, and we define a clinically relevant difference to be 10%. We select a standard *P* value of 0.05 and a standard power of 80%. Sample size for this example can be calculated from available formulas.[11,12] Z values are found by referring to tables of probability distribution; in this example we use a normal distribution. Inserting values, we arrive at a calculated sample size of 83 subjects in each arm. However, if we also wish to consider that this novel therapy may in fact be worse than the mid-urethral sling, we need to use non-directional significance testing. This slightly different equation[11,12] shows we need 108 subjects in each arm. We should plan for recruiting in excess of this goal by 10% to 15% to account for possible loss to follow-up or data loss, which is approximately 118 subjects in each arm. If we change our definition of a clinically relevant difference to 20%, we now find that we would need approximately half the sample size, all other parameters being equal.

Looking at a continuous variable, we investigate a new medication in the treatment of benign prostatic hypertrophy, and our selected endpoint is flow rate. In our study population, the average flow rate is 12 ml/s with a standard deviation of 4 ml/s. We determine that a clinically significant improvement would be at least 3 ml/s, and we again select a *P* value of 0.05 and a power of 80%. Sample size calculation for this example[11,12] (size estimation for comparing means from a single population with a directional test) yields a study population of 11 patients per arm. If the standard deviation is 6 ml/s, the measurement of the outcome is much less precise, and the sample size needed for each group would become 25.[11,12] Inherent properties of measuring means can have tremendous influence in determining sample size.

Sample size in urology literature

Others have examined urological literature to determine how frequently negative studies were underpowered to find statistical significance. Breau *et al.* in 2006 looked at clinical trials in Urology where a negative conclusion was reached.[13] They identified 127 trials providing enough information for post hoc power calculation. Assuming a threshold for a treatment difference of 50%, they found that 33% to 65% of studies were adequately powered (>80%), depending if the outcome were respectively continuous or dichotomous. When the threshold was changed to look at a treatment effect of 25%, (a more reasonable expectation) only 23% to 33% of those studies were adequately powered to detect a difference. They also noted that the prevalence of adequately powered studies did not improve over time. Another group found that in randomized controlled trials there was a higher prevalence of sample size calculation in 2004 as compared to 1996 (47% versus 19%),[14] suggesting an increasing understanding within the urology community of its importance. This still revealed that the majority of randomized controlled trials in the urologic literature did not justify sample size. A significant portion of published comparative studies with negative findings may be underpowered to validly reach this conclusion, and such conclusions should be approached with caution.

Resolution of the clinical scenario

For clinicians, understanding sample size calculations is essential in analyzing the veracity of a study finding when there is a positive finding, but also when there is a finding of no difference between therapeutic interventions. At times, the failure to detect a difference stems from a lack of power

in the study and not from true equalities of the treatments. Therefore, it is vitally important for study authors to state a priori the sample size calculation, and delineate the underlying assumptions and expectations that went into that calculation.

In a negative finding study, such as the one comparing mini-sling to TOT, one of the concerns is whether the study was adequately powered to detect a difference if one existed. As the comparison was performed retrospectively, no sample size calculation was performed, yet we can still use this study to illustrate post hoc power calculation. We know from the literature that the TOT sling has an efficacy of approximately 85%. We can define a clinically significant difference to be 10%, a reasonable estimate based on the clinical scenario, prior studies, and potential benefits or harms of the procedure. We select the customary alpha values of 0.05 (running a two-tailed test) and decide on a beta value of 0.20, giving us a power of 80%. From this post hoc power calculation,[11,12] the study would have needed at least 78 patients in each arm to draw a valid conclusion, and appears to be slightly underpowered to detect a 10% difference in efficacy between these two modalities. Based on these clinical assumptions, the conclusion of equivalence may be in fact invalid. You counsel your patient appropriately and she is able to make an informed decision.

SUMMARY

An understanding of sample size estimation and how various factors interrelate into its calculation is invaluable knowledge to any person seeking to perform a clinical study or to critically appraise evidence. Sample size estimation helps ensure responsible usage of resources and the ethical treatment of human subjects. It also obliges researchers to define key assumptions including projected occurrence rates or effect size, levels of statistical significance, and acceptable power of the study so that the readers may make determination on the merits of these factors. This is imperative for studies with negative findings or for studies aiming to prove equivalence.

REFERENCES

1. Saarni SI, Gylling HA. Evidence based medicine guidelines: A solution to rationing or politics disguised as science? J Med Ethics 2004;30:7171-5.
2. De Ridder D, Berkers J, Deprest J, Verguts J, Ost D, Hamid D, *et al*. Single incision mini-sling versus a transobutaror sling: A comparative study on MiniArc and Monarc sling. Int Urogynecol J Pelvic Floor Dysfunct 2010;21:773-8.
3. Glantz SA. Primer of Biostatistics. 5th ed. New York: McGraw-Hill; 2002.
4. Forthofer RN, Lee ES, Hernandez M. Biostatics: A Guide to Design, Analysis, and Discovery. 2nd ed. Amsterdam: Elsevier Academic Press; 2007.
5. Simon R. Optimal two-stage designs for phase II clinical trials. Control Clin Trials 1989;10:1-10.
6. SEER Cancer Registry, 1975-2007. Available from: http://www.seer.cancer.gov. [Last accessed on 2010 Dec 27].
7. Ewens WJ, Grant GR. Statistical Methods in Bioinformatics: An Introduction. Berlin: Springer-Verlag; 2001.
8. Perneger TV. What's wrong with Bonferroni adjustments? BMJ 1998;316:1236-8.
9. Scales DC, Rubenfeld GD. Estimating sample size in critical care literature. J Crit Care 2005;20:6-11.
10. Pocock SJ. When to Stop a Clinical Trial. BMJ 1992;305:235-40.
11. Riffenburgh R. Statististics in Medicine. 2nd ed. Amsterdam: Elsevier Press; 2006.
12. Rollin B. Online simple power/sample size calculations. Available from: http://www.stat.ubc.ca/~rollin/stats/ssize/ [Last accessed on 2010 Dec 27].
13. Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, *et al*. Improving the quality of reporting of randomized controlled trials. The CONSORT statement. JAMA 1996;276:637-9.
14. Scales CD Jr, Norris RD, Preminger GM, Vieweg J, Peterson BL, Dahm P. Evaluating the evidence: Statistical methods in randomized controlled trials in the urological literature. J Urol 2008;180:1463-7.

How to cite this article: Hinh P, Canfield SE. Sample size matters: A guide for urologists. Indian J Urol 2011;27:503-7.

Source of Support: Nil, **Conflict of Interest:** None declared.