

Chapter 14

Basic Statistical Considerations

Eileen M. Stock and Kousick Biswas

Introduction

In recent years, significant strides in medical research have led to improvements in disease treatment and patients' quality of life. In any research, it is critically important to formulate a research question that adequately addresses the aims of a study. An example of a research question might be "Does laparoscopic cholecystectomy differ from open cholecystectomy in hospital length of stay" (ACTIVE trial) [1]. This research question will ultimately dictate the study design and methodology that will be employed. The reliability and validity of the results will depend on the proper selection of a research approach and design.

Hypothesis Testing

Following the establishment of a research question, a *null* and *alternative* hypothesis, denoted H_0 and H_a , should be formulated. The hypotheses will be determined by the research question, what groups and how many groups are to be compared, and at what time points an outcome will be measured such as cross-sectional occurring at one point in time or longitudinal to measure differences over time as in a prospective clinical trial. The alternative (research) hypothesis

E.M. Stock (✉) · K. Biswas
Cooperative Studies Program Coordinating Center, Office of Research
and Development, U.S. Department of Veterans Affairs, VA Medical Center,
5th Boiler Street, Perry Point, MD 21902, USA
e-mail: Eileen.Stock@va.gov

K. Biswas
e-mail: Kousick.Biswas@va.gov

corresponds to the primary purpose of the trial and what the researcher is trying to prove. The null hypothesis is the hypothesis being tested, the complement of H_a , always contains equality, and assumed to be true until it is decided to either *reject* or *fail to reject* H_0 .

The usual scenario in hypothesis testing is demonstration of a difference (between two procedures). For example, if testing for a difference in the mean hospital length of stay for open cholecystectomy (μ_1) versus laparoscopic cholecystectomy (μ_2), the hypothesis is $H_0 : \mu_1 = \mu_2$ vs. $H_a : \mu_1 \neq \mu_2$. For a one-sided upper or lower tail test, $H_a : \mu_1 > \mu_2$ or $H_a : \mu_1 < \mu_2$. Here μ_1 and μ_2 represent the unknown “true” mean hospital length of stay for treatment group 1 and treatment group 2. Depending on the scientific hypothesis, the design of a trial can be superiority, non-inferiority, or equivalence. The objective of a superiority trial is to find a procedure to be better than the established alternative and was the framework proposed for the original randomized clinical trials. The objective of a non-inferiority trial is to find a new procedure that is not inferior to another procedure. Lastly, an equivalence trial aims to determine whether a new procedure is neither worse nor better than the established intervention.

Study Design

While RCTs are the gold standard for establishing safety and efficacy of a therapeutic intervention, there are many challenges in the design of trials assessing a procedure. Barriers forfeiting the reliability of trials, hence clinical evidence influencing surgical practice, can entail areas related to planning and design, eligibility criteria, choice of treatment comparator, benefit-to-harm ratio, as well as experience of the study team [2, 3]. Unfortunately, only about 15% of trials are in surgery of which nearly half are discontinued (43% vs. 27% in medicine), mostly for slow recruitment (18%) [4, 5], wasting already scarce resources, and raising ethical concerns if results are never reported to inform practice. An unexpected lower recruitment can result from the approach for obtaining consent, the randomization scheme, or the treatment comparator.

Recruitment Approach and Consent

There are numerous reasons why recruitment in a RCT may be low. First, patients may be completely unaware of an ongoing trial applicable to their ailment. Also, providers may be unaware and not mention a trial because of time constraints in explaining the trial, treatments, risks, benefits, and alternatives [6]. Additionally, affecting recruitment is the added burden on medical staff along with regular clinical duty, especially if facilities are already understaffed.

The subject area may influence recruiting. When informed of a new malady or needed surgery, patients may feel overwhelmed, alone, and be hesitant to participate. The complexity of the enrollment process involving extensive screening of eligibility criteria, difficult terminology, travel expenses, insurance coverage, and an unknown or experimental treatment can be daunting.

The method of recruitment can also impact participation. For seven varying recruitment strategies of 1562 cancer patients and their caregivers, two were the most effective—online recruitment by researchers of patients waiting for radiotherapy and mailing study information with routine care letters to patients scheduled to receive radiotherapy that were later contacted by telephone if opted out [7]. Less effective approaches included those relying on hospital providers, recruitment at a rehabilitation center, newspaper advertising, flyers, internet, and social media.

Equally important is ensuring patients' understanding of the material provided during consent, which tends to be the most common element absent from the process [6]. Among 141 consent discussions for an orthopedic surgical intervention, only 12% evaluated patients' understanding [8]. While pamphlets, diagrams, videos, and audio may improve comprehension, it should not be a substitute for open dialogue between the patient and provider [9].

Treatment Comparator and Randomization

One aspect of the design that can largely influence patient accrual to the point where the desired sample size is not attainable or findings are so biased they are deemed unreliable is the choice of a treatment comparator. In a traditional two-arm RCT, patients are assigned to one of two arms. One arm may comprise a new experimental treatment and the other standard care or a sham (placebo). If a patient finds they are not randomized to the new treatment, they may react negatively and refuse participation. In a trial with equal allocation to a surgical procedure versus sham, patients may be reluctant to participate because of the high likelihood (50%) of not receiving treatment (Table 14.1). Ideally, treatment allocation should not be known in advance in order to preserve randomness and prevent potential manipulation and bias [2, 3]. This poses an additional challenge when comparing an operative and non-operative procedure.

When treatments differ greatly, patient preferences are likely to influence the balance of patient accrual in each arm [3] as was observed in the MIMOSA trial comparing two distinct initial treatment approaches (surgical vs. pharmacological and behavioral therapy), for women with mixed urinary incontinence despite both being standard therapy [10]. An unbalanced benefit-to-harm ratio may lead patients to favor one treatment more. Operative procedures may require multiple preoperative and postoperative visits, adding burden on the patient. In such trials, a feasibility phase should be considered.

Table 14.1 Challenges to the planning and conduct of randomized trials comparing a surgical procedure with different types of comparators

	Surgical versus sham (placebo)	Surgical versus similar procedure	Surgical versus different procedure	Surgical versus non-surgical
Patients reluctant to participate	Yes	Unlikely	Likely	Yes
Randomize in operating theater	Yes	Yes	Likely	No (providers vary)
Imbalanced surgical experience	No	Unlikely	Likely	No (providers vary)
Poor compliance with allocation	Yes	Unlikely	Yes	Yes
Contamination, lack fidelity	Unlikely	Yes	Unlikely	No

[Based on data from Ref. 2]

Randomization should occur as close to the time of the intervention as possible to avoid the effects of patients' preferences and knowledge of allocation leading to withdrawal [2, 3] such as in the operating theater for two surgical procedures. For substantially different treatments, participants may have to be informed of their randomization [2]. For multicenter trials, stratified randomization is important to offset variability (site-specific or surgeon-related) [3]. A number of alternative randomization approaches have been utilized to overcome difficulty in recruiting and meeting sample size, including the use of unequal treatment allocation ratios to account for dropouts and adaptive randomization in which the allocation ratio changes during the course of the trial [11, 12]. However, these methods still remain less commonly used in practice.

Blinding

Blinding of participants, investigators, physicians, or other caregivers plays a significant role in the removal of potential bias that might otherwise skew results and deem a RCT inferior and of poor quality. There are three types of blinding—*single* (participants), *double* (participants and physician), and *triple* (participants, physician, and others determining eligibility, compliance, or evaluating endpoints) [13]. The absence of blinding can result in several forms of bias. The first is *performance bias*, which refers to differences in the delivery of care between groups attributable

to behavioral responses by caregivers or participants from knowledge of treatment allocation. The comparison is confounded by the characteristics and preferences of caregivers and patients if one treatment is preferred over another; though, masking of surgeons, patients, and other caregivers is difficult and often impossible in surgical trials [2]. Another form of bias is *attrition bias*, resulting from differential withdrawal rates across groups. A surgical arm involving a waiting list or additional postoperative follow-up assessments is an example [2]. Lastly, *detection bias* refers to differences between groups in how outcomes are determined due to subjective evaluation of assessors such as self-reported outcomes when participants are unmasked [2].

Surgeon Characteristics

Most RCTs involve randomization administered by the same clinician, which is not possible when treatments are from different specialties (e.g., operative vs. non-operative) [3]. The delivery of a surgical procedure is influenced by attributes of the surgeon (e.g., skill, experience, preferences, decision-making ability), other team members (e.g., anesthesiologist, technicians, nurses), and those involved in preoperative and postoperative care (e.g., ED, ICU, imaging, recovery, rehabilitation) [2]. The learning curve for a procedure can confound results [3]. Outcomes such as symptomatology and functioning may be measured by surgeons and physicians differently (e.g., subjective assessments, unstandardized definitions). This variability in practice is unavoidable and, if great, can influence outcomes. To avoid criticism, the surgical procedure and care-practice measures should also be evaluated [2]. Comparing endoscopic versus open carpal tunnel release is an example of a non-operative and operative procedure with multiple interacting components and requiring a specific level of experience and training among surgeons [14]. Trials involving a varied benefit-to-harm ratio are sometimes difficult to recruit surgeons [15].

Analysis

When initially developing the idea for a RCT, a statistician should be consulted to help identify specific aims, hypotheses, the study design, analysis plan, and sample size. Hypotheses should focus on what the research is intended to demonstrate, clearly stating the outcomes of interest, groups to be compared, and the time period involved. Justification supportive of the hypotheses, such as pilot data, should also be included. Finally, it should be discussed how contingencies, such as missing data, which could bias findings will be handled.

Sample Size

For a clinical trial to be successful, sufficient planning is needed that should include sample size determination. This entails estimating how many participants should be enrolled in the study. The feasibility of the trial should also be assessed, identifying whether the proposed time and resources seem reasonable. Finally, a sample size should be estimated that achieves sufficient power to detect a specific treatment effect, factoring in the size or magnitude of the effect and its variability.

Outcome Measure

While a continuous outcome tends to result in improved statistical power (or alternately smaller sample size for the same level of power), a binary outcome is more easily interpretable. Most basic science and translational science studies use continuous outcomes, whereas RCTs usually involve binary or time-to-event outcomes. Observational studies may have either type of outcome. In a two-group parallel trial using a continuous outcome, the difference between groups is tested by comparing the means at some point in time using the t-test for two independent samples. For a binary outcome, the difference between groups is tested by comparing the proportion having the outcome at some point in time using the chi-square test for homogeneity of proportions or Fisher's exact test.

Baseline Assessment

While non-randomized studies try to account for pretreatment disparities between groups, the prospective design of a RCT helps provide some protection against biases resulting from baseline differences [2]. Although randomization on average produces homogeneity between groups, it does not guarantee balance. Accordingly, patient data should be collected at baseline before randomization during screening and after randomization prior to treatment. This information can then be used to check for balance [16]. Keep in mind that an insignificant association does not necessarily imply imbalance is not present. It merely suggests it was not detected (e.g., small sample size, low power). Furthermore, unless sample sizes are very large, rejecting the null hypothesis implies an imbalance that should be addressed in the analysis. These data may additionally be used to stratify (e.g., stratified randomization, subgroup analyses).

Intention-to-Treat Analysis

Under the intent-to-treat (ITT) principle (“once randomized, always analyzed”), patients remain in their assigned treatment group for the primary analysis, regardless of compliance or dropout. Even when a surgeon decides that a surgical procedure is inappropriate or unsafe after randomization, in which the patient may receive a different treatment from that originally assigned, these patients remain in their assigned treatment group regardless of receiving the alternative treatment. ITT analysis reflects the practical clinical scenario and allows generalizability by maintaining prognostic balance generated from the original random treatment allocation, producing an unbiased estimate of the treatment effect. Sample size is preserved, ensuring statistical power, and type I error is minimized [17]. Otherwise, the exclusion of non-compliant participants and dropouts from the final analysis may introduce bias. If allocation is disrupted, the study may no longer be considered a RCT. On the other hand, ITT analysis has been criticized for being too conservative (susceptible to type II error) and fails to answer the study question of whether the treatment works if used as intended.

Kaplan–Meier Estimator and Survival Curves

In RCTs, the outcome measure is typically binary or a time-to-event measure. Survival analysis is the branch of statistics for analyzing “time-to-event” data. Examples include time until a particular event (death), recurrence (revascularization), or time to a response (10% decrease in weight). Components necessary for the analysis include whether the event occurred (dichotomous) and the length of time from the start of follow-up to a precise endpoint, either when the event occurred or last known follow-up (censored). Censoring is when a participant does not experience the event prior to study closure, withdrawal, or loss to follow-up. Right-censoring is most common, i.e., the event has not been observed yet but might occur in the future. Left-censoring and interval-censoring are less common. The latter type of survival analysis applies when the time of the event is less-precisely known. When an event is noted to have occurred, it is assumed to have occurred in some interval since the last time event status was determined [18, 19].

Survival curves are estimated with Kaplan–Meier estimators to determine the probability of a patient surviving (or event-free) past a specified time. Curves are monotonically decreasing and stepwise (step-for-each event). When stratified by treatment arm, curves are estimated for each group separately and compared using a test for equality (parametric likelihood ratio test or nonparametric log-rank or Wilcoxon test). Rejection of equality indicates that the event rates differ between groups. However, these tests are less reliable when curves cross.

References

1. Catena F, Ansaloni L, Di Saverio S, et al. The ACTIVE (Acute Cholecystitis Trial Invasive Versus Endoscopic) study: multicenter randomized, double-blind, controlled trial of laparoscopic (LC) versus open (LTC) surgery for acute cholecystitis (AC) in adults. *Trials*. 2008;9:1.
2. Ergina PL, Cook JA, Blazeby JM, et al. Challenges in evaluating surgical innovation. *Lancet*. 2009;374:1097–104.
3. Cook JA. The challenges faced in the design, conduct and analysis of surgical randomised controlled trials. *Trials*. 2009;10:9.
4. Rosenthal R, Kasenda B, Dell-Kuster S, et al. Completion and publication rates of randomized controlled trials in surgery: an empirical study. *Ann Surg*. 2015;262:68–73.
5. Kasenda B, von Elm E, You J, et al. Prevalence, characteristics, and publication of discontinued randomized trials. *JAMA*. 2014;311:1045–51.
6. Cordasco KM. Chapter 39 obtaining informed consent from patients: brief update review. Evidence Reports/Technology Assessments, No. 211. Rockville (MD): Agency for Healthcare Research and Quality (US), 2013.
7. Sygna K, Johansen S, Ruland CM. Recruitment challenges in clinical research including cancer patients and their caregivers. A randomized controlled trial study and lessons learned. *Trials*. 2015;16:428.
8. Braddock C 3rd, Hudak PL, Feldman JJ, Bereknyei S, Frankel RM, Levinson W. Surgery is certainly one good option: quality and time-efficiency of informed decision-making in surgery. *J Bone Joint Surg Am*. 2008;90:1830–8.
9. Zuckerman MJ, Shen B, et al. Informed consent for GI endoscopy. *Gastrointest Endosc*. 2007;66:213–218.
10. Brubaker L, Moalli P, Richter HE, et al. Challenges in designing a pragmatic clinical trial: the mixed incontinence—medical or surgical approach (MIMOSA) trial experience. *Clin Trials*. 2009;6:355–64.
11. Dumville JC, Hahn S, Miles JN, Torgerson DJ. The use of unequal randomisation ratios in clinical trials: a review. *Contemp Clin Trials*. 2006;27:1–12.
12. Berry DA. Adaptive clinical trials: the promise and the caution. *J Clin Oncol*. 2011;29:606–9.
13. Bridgman S, Engebretsen L, Dainty K, Kirkley A, Maffulli N, Committee IS. Practical aspects of randomization and blinding in randomized clinical trials. *Arthroscopy*. 2003;19:1000–6.
14. Macdermid JC, Richards RS, Roth JH, Ross DC, King GJ. Endoscopic versus open carpal tunnel release: a randomized trial. *J Hand Surg Am*. 2003;28:475–80.
15. McCulloch P, Kaul A, Wagstaff GF, Wheatcroft J. Tolerance of uncertainty, extroversion, neuroticism and attitudes to randomized controlled trials among surgeons and physicians. *Br J Surg*. 2005;92:1293–7.
16. Senn S. Testing for baseline balance in clinical trials. *Stat Med*. 1994;13:1715–26.
17. Gupta SK. Intention-to-treat concept: a review. *Perspect Clin Res*. 2011;2:109–12.
18. Singh R, Mukhopadhyay K. Survival analysis in clinical trials: basics and must know areas. *Perspect Clin Res*. 2011;2:145–8.
19. Prinja S, Gupta N, Verma R. Censoring in clinical trials: review of survival analysis techniques. *Indian J Community Med*. 2010;35:217–21.