Published in final edited form as:

Surgery. 2009 June ; 145(6): 581–587. doi:10.1016/j.surg.2009.04.003.

RANDOMIZED TRIALS IN SURGERY

Annetine C. Gelijns, PhD^a, Deborah D. Ascheim, MD^a, Michael K. Parides, PhD^a, K. Craig Kent, MD^b, and Alan J. Moskowitz, MD^a

^aDepartment of Health Policy, Mount Sinai School of Medicine, New York, NY

^bDepartment of Surgery, University of Wisconsin School of Medicine and Public Health, Madison, WI

IN 1996, THE LANCET PUBLISHED A COMMENTARY “Surgical Research or Comic Opera: Questions, but Few Answers.”¹ The author reviewed the first issue of 9 general surgery journals in 1996 and found that only 7% of papers reported data derived from a randomized trial, whereas 46% were based on case series. A decade later (2006), we conducted a similar survey and found that the relative percentage of case series, operationally defined as a single institution’s experience with a single procedure, had decreased to 34%. At the same time, we found that the relative number of cohort studies (51%) and the relative percentage of articles reporting data derived from a randomized trial (10%) remained relatively similar.

Such a review raises several interesting questions about contemporary surgical research. The main issue that stands out is this one: Why the number of randomized trials, which had been identified as being low compared with other fields of medicine in the mid-1990s, has not increased over time in an environment where the demands for rigorous evidence have grown? In addressing this question, it is clear that different research designs---ranging from case series to randomized trials to meta-analyses---are needed for the evaluation of surgical procedures. The appropriate choice of research design depends on the effect size of the intervention, the specific clinical circumstances, the relevant clinical question, and the stage of the life cycle in which a surgical procedure is evaluated.

Case series, for example, are valuable in rare diseases, whereas prospective single-arm studies can provide important insights if the natural history of a disease is known and the outcome is unequivocal. Moreover, the first stage of clinical evaluation for a novel procedure is usually a single-arm prospective study in a diseased population to assess the feasibility of its use in humans. These studies, which are generally small in size, provide preliminary evidence about short-term safety and efficacy. The major value of these studies is that they offer easily obtainable prima facie evidence for designing large-scale, pivotal trials that are randomized or otherwise controlled. However, our review suggests that such large-scale confirmatory trials are not always carried out, especially randomized studies.

The advantage of randomized trials is well known; they offer a powerful means to evaluate the efficacy of procedures in that comparison groups are equally constituted with respect to known and unknown predictors of outcome. As such, these trials offer an unbiased assessment of the impact of a procedure. So why are such trials not more prevalent in surgery? We believe that such trials can be more difficult to design and conduct for surgical procedures than for new pharmaceutical agents. Moreover, because surgical innovation is characterized by a high

degree of incremental change and heavy dependence on technical skill and experience, statistical inference from trial data is often more challenging.

In this analysis, we first explore the methodological, ethical, and logistic challenges posed by conducting randomized trials of novel surgical procedures. We review several options for addressing these challenges in terms of trial design and analytic techniques. We then discuss issues involved in the interpretation of clinical trials and the value of so-called pragmatic randomized trials. We conclude by evaluating some of the larger infrastructure issues that need to be addressed before a robust, surgical clinical trial enterprise can be created.

CHALLENGES IN CONDUCTING TRIALS OF NOVEL PROCEDURES

Equipoise

A critical ethical issue in the design of any randomized trial is whether there is equipoise to conduct the trial, that is, if there is genuine uncertainty or controversy in the clinical community about the comparative therapeutic merits of each of the interventions being evaluated.² In pharmaceutical trials, which often entail double-blind comparisons of 2 medications, such equipoise may be easier to obtain and maintain than for procedural-based trials.

Surgical trials often involve the comparison of surgical procedures with medical therapies. Such different treatment approaches may trigger strong physician and patient preferences, which may unbalance equipoise for randomization, especially in cases of life-threatening illness. Patients, their families, and physicians may view an experimental procedure as their best hope (other therapies having failed) and would be devastated to learn, up front, that they would not receive this preferred intervention. This type of dilemma will deter some patients and physicians from entering a procedure-based trial; others might enroll but seek treatment outside the protocol if they do not receive the therapy they wanted. Such a loss to follow-up or out-of-protocol crossover could negatively impact the findings of a small-scale trial.

A salient example in recent years can be found in an experimental study of autologous bone marrow transplantation combined with high-dose chemotherapy as a potential therapy for advanced breast cancer. In the early 1990s, the procedure disseminated widely, driven by patient and physician demand, court litigation to gain access to the experimental procedure, and the subsequent decision by payers to cover the therapy outside the clinical trial setting. This latter decision substantially slowed trial enrollment; as a result, trial evidence did not emerge until the end of the decade. At that time, 4 trials showed no benefit for the experimental procedure; these results halted widespread use of this unsuccessful and highly toxic therapy.³

There are several design options that could reduce or minimize deterrents to enrollment, particularly if strong preferences exist for an investigational surgical procedure. One option is to combine randomized with non-randomized data in the evaluation of an intervention. The nonrandomized data could serve to supplement the control arm and so reduce the number of patients assigned to it, which would facilitate a higher randomization ratio (eg, 3:1) to the novel procedure. The effect, in other words, would be to increase the chances that any enrollee would be assigned to the investigational therapy. Pocock⁴ provides a general discussion of this approach, and proposes a set of conditions that need to be met to justify using historical controls. In particular, the control population should have been enrolled in a clinical study with similar eligibility criteria and undergone similar treatment; in addition, the outcome evaluations should be the same as the control patients in the randomized trials.

Strong assumptions are required to, without bias, combine historical data with randomized data that are unlikely to hold in practice. Statistical models may help to alleviate bias; for example, propensity scores could be used for risk adjustment of known and measured factors and

Bayesian models that account for bias through a prior distribution could be implemented. Begg and Pilote⁵ propose a meta-analytic approach to combining different sources of data via a random effects model that also allows for a preliminary test of bias and can incorporate heterogeneity of treatment effect. Temporal bias from using historical data may be addressed by the use of concurrent controls.

Other design options that would increase the likelihood that patients will receive the preferred therapy and, thereby, may enhance enrollment include response-adaptive randomization (randomized play-the-winner models). In these randomization schemes, the allocation probabilities are continuously updated on the basis of previous patient outcomes to increase the likelihood that patients are assigned to the more successful treatment. Although this type of randomization scheme can reduce the number of patients assigned to an inferior or less successful treatment, they increase the complexity of a trial and may introduce bias if there are changes in the characteristics of patients enrolled over time.

Endpoints and small target populations

The choice of endpoint is a critical one in trial design, because it is an important factor in determining the size, duration, and clinical relevance of the trial. Ideally, one would select the most meaningful endpoint for clinical decision-making, capturing the impact of the proposed treatment on length and quality of life. For many diseases in which surgical therapies are used, mortality remains a critical endpoint. With the increasing prevalence of surgical interventions for chronic diseases, functional status and quality of life have become more important trial outcomes; however, their measurement may be more challenging.

Blinding, which controls observational bias when evaluating the safety and efficacy of a new clinical intervention, may be problematic in surgical trials. Obviously, the treating surgeons cannot be blinded, and patient blinding is rarely possible when one of the trial arms is a procedure or implantable device and the other is medical therapy. One option to minimize measurement bias is to have the assessment carried out by local clinicians who are uninvolved in the trial or by a core lab that is blinded, whenever possible, to treatment arm, clinical events, and outcomes. This method may especially work in the case of more objective outcomes, such as an echocardiogram or mortality. However, endpoints that require patient input, such as functional status (eg, measured by a 6-minute walk test) and quality of life, remain subject to bias related to patient preferences.

Another characteristic of surgical (in comparison to pharmaceutical) trials is that the target population is often smaller, which has implications for the ability of researchers to recruit a sufficient number of participants in a reasonable period of time. Smaller target populations often dictate the need for trials with more modest sample sizes. One way to achieve this is to adopt a composite endpoint. Such an endpoint highlights the differences between treatment arms by combining different endpoints (eg, survival with an important adverse event such as stroke, creating stroke-free survival). As such, a composite endpoint may increase the difference in event rates and, consequently, increase statistical power, which allows for a smaller trial.

Composite endpoints, however, can become controversial in terms of their interpretation, particularly if the individual component endpoints are not consistently superior. A recent case in point can be found in the SYNTAX trial.⁶ This noninferiority trial compared percutaneous coronary interventions (PCI) with a drug-eluting stent to coronary artery bypass grafting (CABG) in patients with 3-vessel and/or left main coronary artery disease. The primary endpoint was major cardiac and cerebrovascular events during 1 year. The trial demonstrated that the composite primary endpoint of all-cause mortality, stroke, myocardial infarction, or repeat revascularizations was significantly higher for PCI than for surgery. The authors

concluded that CABG remains the standard of care for patients with 3-vessel or left main coronary artery disease.⁶ However, if we look at the individual components (for which the trial was not adequately powered), there was not a statistical difference between treatment arms in the rates of death and myocardial infarction. There were fewer strokes in the PCI group; however, the PCI group had a much higher rate of revascularization. As is obvious, these results make interpretation of the outcome of this trial challenging.⁷

There are, of course, circumstances in which randomized or other prospectively controlled studies are not feasible in terms of enrolling sufficient numbers of participants within a reasonable period of time. Most pediatric diseases that involve implantable devices, for example, are low prevalence condition.⁸ Similarly, left ventricular assist devices (LVADs) implanted to support patients awaiting cardiac transplantation (the so-called bridge to transplant population) meet the U.S. Food and Drug Administration (FDA) definition of an orphan indication---around 500 LVADs are implanted annually. In this case, the FDA allows single-arm studies that use a performance goal derived from previous studies as a benchmark. Alternatively, one could design a comparative randomized study that would require a smaller sample size with a greater chance of random variation error---in other words, accept a higher P value to establish statistical significance.⁹ Such a trial would have measurably less precision, but would avert bias by maintaining a randomized design.

Learning curves

Surgical trials typically have to contend with a much more prominent learning curve among users than is the case with pharmaceutical trials.

Such learning curves have been demonstrated for minimally invasive procedures, such as laparoscopic cholecystectomy and endovascular aneurysm repair.^{10,11} Differences in outcomes over time reflect the complex interactions of knowledge gained in technical skill, patient selection, and patient management. One strategy to account for learning curves is to design a pilot trial or a run-in period that will not be counted in the final analysis. Another strategy is to prespecify in the analytic plan that investigators will explore whether significant changes occur in outcomes over the course of the trial. The Randomized Evaluation of Mechanical Assistance for the Treatment of Congestive Heart Failure (REMATCH) trial funded by the National Heart, Lung and Blood Institute, for example, evaluated the efficacy and safety of implantation with a LVAD versus optimal medical management in patients with chronic end-stage (Stage D) heart failure. Compared to optimal medical management (n = 61), LVAD implantation (n = 68) was found to double the 1-year survival (from 25% to 51%) in this terminally ill population.¹² Despite the small size of the trial, there was a learning curve phenomenon. If we compare patients enrolled in the first half of the trial to those enrolled in the second half, we see a significant improvement in survival in the LVAD arm (Figure). Such improvement was not seen in the medical management arm. Moreover, we see a similar significant improvement over time in terms of the adverse event profile of LVAD patients, with less device driveline infections, postoperative bleeding, and sepsis.¹³

Incremental change

In comparison to pharmaceutical trials, another challenge to designing trials of novel procedures is dealing with the high level of incremental change that characterizes surgical innovation. A pharmaceutical compound generally does not undergo substantial change as it moves through the phases of clinical trials, although development may be discontinued if the profile is undesirable, and a modified compound may enter preclinical and Phase I trials. By comparison, during the clinical evaluation process of surgical procedures, both the procedure itself and the clinical management of patients may be extensively modified and refined. For example, in the abovementioned REMATCH trial, several changes were implemented, such

as modification of the device driveline, introduction of a locking screw ring to prevent detachment of the blood-transport conduits to and from the pump, and a clinical protocol to prevent and manage driveline infections with antimicrobial agents and laminar flow operating rooms. Such modifications in the device or the clinical management of patients can be accommodated in the design of clinical trials. With these variations in protocol, the predetermined sample size in the REMATCH trial did not change.¹² However, if changes to the device design or to the clinical management of patients substantially alter the measures of outcome, additional patients may need to be recruited to satisfy specific subgroup analyses.

Differences in surgeon and team skills

Another issue, beyond a learning curve, is differences in the skills of providers in a trial. In contrast to pharmaceuticals, the efficacy of a surgically implanted device can vary with the skill of the surgeon and the clinical team. Substantial variation among trial investigators makes the results of the trial difficult to interpret. A positive average outcome for the experimental therapy may be positive only by virtue of a few exceptionally well-skilled clinical sites, and a negative average outcome may be negative only because of a few less-skilled clinical sites.

Clinical trialists must be on guard for such outcomes, and so trials typically include a separate randomization scheme in which each clinical site is balanced with respect to the number of experimental and control patients that it treats. Moreover, examining the effect of the study site on the primary outcome is a routine analytic step. Participation may also be limited to providers with particular skill levels. A trial of a surgical procedure in highly specialized centers with unique surgical expertise may yield success; the results, however, may not be generalizable or provide helpful information on the value of using the procedure in less-specialized centers.

INTERPRETATION OF EVIDENCE

Our focus, thus far, has been on the design, conduct, and analysis of randomized trials for novel procedures. These trials, if well designed, offer rigorous evidence about the value of a new procedure and, as such, are critical to sound clinical and policy decision-making. Conducting a randomized trial of a novel procedure, however, does not generally constitute the be-all and end-all of evaluative efforts. There is a need for further evaluation for various reasons. The interpretation of trial evidence generally requires clinicians and policymakers to balance the tradeoffs among benefits, risks, and costs of new procedures in a context of uncertainty. In part, uncertainty lingers because trials of novel procedures are based on a sampling process, have limited timeframes and limited patient heterogeneity, and proceed in specialized centers, all of which complicates generalizability to more widespread clinical use.

For example, in a national trial of carotid endarterectomy in the early 1990s (North American Symptomatic Carotid Endarterectomy Trial), 0.6% of the patients died peri-operatively (30 day); when the technology spread, however, this rate increased substantially, reaching rates 2--3 times higher in low-volume hospitals in California and Ontario.¹⁴

LEARNING BY DOING

But a deeper source of uncertainty remains: Surgical procedures entail much more ongoing innovation and “learning by doing” than do pharmaceuticals. After a new surgical intervention is introduced into practice, the medical profession typically expands and shapes the patient population within a particular disease category. Only 4% of patients who were treated with CABG surgery a decade after its introduction would have met the eligibility criteria of the trials that determined its initial value.¹⁵ These first trials excluded the elderly, women, and patients with a range of comorbidities, all of whom receive CABG surgery today. Moreover, a surgical

procedure itself may evolve and improvements may occur in patient management techniques; for example, CABG first involved using a saphenous vein and, later, the internal mammary artery, which significantly improved its outcomes.

In addition, there is a parallel evolution in the treatment alternatives available for managing specific conditions. The first CABG trials, for instance, compared the surgical procedure to medical management, such as beta-blockers. Over time, the relevant comparison treatment became alternative means of revascularization, such as percutaneous transluminal coronary angioplasty, and then PCI with stents. The stents themselves also continued to evolve (from bare metal stents to drug-eluting stents), which means that the results of a trial in a highly innovative field may have a limited useful life.

In sum, the enormous value of conducting a rigorous trial of a novel procedure is that one defines a solid foundation of knowing whether a procedure works or does not work in a specific patient population at a given time. Yet, it is clear that ongoing innovation requires ongoing evaluation. Such evaluative studies may be observational studies or, in some instances, another randomized trial.

PRAGMATIC CLINICAL TRIALS AND OBSERVATIONAL STUDIES

These observational studies and so-called pragmatic or practical randomized clinical trials can provide important insights about expanding patient populations, changes in management techniques, or changes in the procedure itself. This possibility is especially the case for high-cost, high-prevalence procedures and conditions. In New York State, for example, registries exist for all patients undergoing CABG surgery or PCI.¹⁶ Similarly, in the case of LVADs, a registry (INTERMACS) was established, which is supported by the National Institutes of Health (NIH), the FDA, the Centers for Medicare and Medicaid Services (CMS), and industry. Evaluating changes in outcomes and validating risk scores are important contributions from such registries, because such scores constitute critical tools for guiding patient selection decisions.

Clearly, if there is a substantial change in the patient population receiving a procedure or modifications in the procedure itself, then another randomized trial may be warranted. To return to the case of CABG, the extension of the use of bypass surgery for patients with angina pectoris to patients with heart failure stimulated the conduct of the recent NIH-supported Surgical Treatment for Ischemic Heart Failure (STICH) trial. Clinical investigators tend to make a distinction between initial trials that evaluate the efficacy of novel procedures and subsequent trials that compare the effectiveness of a procedure to alternative treatment options. These latter trials (also known as pragmatic trials) usually compare clinically relevant interventions, include diverse populations of patients, recruit participants from various practice settings, and collect data on a broad range of health outcomes.¹⁷ The continued evaluation of procedures, whether it involves observational studies or randomized trials, requires adequate resources because these studies often are large and expensive undertakings, which brings us to consideration of funding and institutional issues.

CONCLUDING OBSERVATIONS

In recent years, there has been an increasing demand for rigorous evidence about the efficacy, safety, and costs of new procedures to guide decisions on their adoption and use. Randomized trials offer a powerful means to evaluate efficacy, because they balance comparison groups with respect to potential predictors of outcome. However, designing trials to generate unbiased estimates with external validity is more difficult in the surgical than the pharmaceutical context, because not only trial populations are typically smaller, but incremental change and variations in how procedures are performed among centers are also important determinants of outcomes.

Creating a robust surgical trials enterprise requires solutions that lie within the analytic, institutional, and financial realms. Many challenges can be addressed by innovative trial designs, such as Bayesian techniques and adaptive trial designs. Such randomized trials can provide a foundation of know-how about the risks and benefits of a procedure for a particular patient population upon which clinical and payment decisions can be made. However, initial trials of surgical procedures will always be limited in their ability to provide insights about long-term safety and effectiveness, especially because surgical innovation, as mentioned previously, is characterized by a high degree of “learning by doing.” Ongoing innovation, therefore, requires ongoing evaluative efforts to capture changing outcomes in everyday practice. Advances in informatics may help capture clinical data both for practice and research purposes, which would facilitate the operation of less costly patient registries. Moreover, the movement toward large, simple trials, which seek to capture a broader representation of patients and practitioners, may make randomized trials more efficient and economical while holding bias at bay.

In addition to advances in the analytic realm, an adequate funding and institutional infrastructure is needed to create a robust surgical clinical trials enterprise. Trials of new surgical procedures (especially if they do not involve a new drug or device) are mostly supported by the public sector, including the NIH, Agency for Healthcare Research and Quality (AHRQ), and U.S. Department of Veterans Affairs (VA). The public sector, however, invests considerably less in this type of research than the private sector, which invests around 30% of its considerable research and development budgets in clinical trials. By comparison, the NIH invests around 30% of its budget in overall clinical research, but it is estimated that only about 10% is allocated to clinical trials and other evaluative studies.

Several positive developments are underway, however, that may ameliorate this situation. The NIH increasingly supports the creation of networks of hospitals and providers to design, conduct, and analyze clinical trials in a particular area. A case in point is the Cardiothoracic Surgical Trials Network, which has been established to conduct 5--7 proof-of-concept trials of novel cardiac surgery procedures over a 5-year period. These networks create an infrastructure of clinical centers that are geared toward conducting collaborative clinical trials research. Another promising development is the recent emphasis on comparative effectiveness research and the need to support such research. The funds made available in the 2009 stimulus package, for example, will support trials comparing surgical to medical treatments.

In conclusion, over the past few decades, surgery has been a highly innovative field, with advances emerging in such fields as minimally invasive procedures, robotics, and stem cell transplants, to name but a few. In the past, evaluating such novel surgical procedures, and comparing the effectiveness of existing surgical interventions to alternative modalities, has not relied heavily on randomized trials. With the emergence of new analytic approaches that can address the challenges of surgical trials, the promise of new funding opportunities, and the accumulation of experience with new institutional models for conducting clinical research in surgery, the timing is right for change. Randomized trials, we believe, have a prominent role to play in the rigorous evaluation of evolving surgical interventions for ever-changing patient populations.

Acknowledgments

Supported in part by grants HL088942 and HL77096 from the National Heart Lung and Blood Institute and by grant NF051566 from the National Institute of Neurological Disease and Stroke of the National Institutes of Health.

REFERENCES

1. Horton R. Surgical research or comic opera. *Lancet* 1996;347:984–985. [PubMed: 8606606]

2. Freedman B. Equipoise and the ethics of clinical research. *N Engl J Med* 1987;317:141–145. [PubMed: 3600702]
3. Rettig, RA.; Jacobsen, PD.; Farquhar, CM.; Aubry, WM. False hope: bone marrow transplantation for breast cancer. New York (NY): Oxford University Press; 2007.
4. Pocock SJ. The combination of randomized and historical controls in clinical trials. *J Chronic Dis* 1976;29:175–188. [PubMed: 770493]
5. Begg CB, Pilote A. A model for incorporating historical controls into a meta-analysis. *Biometrics* 1991;47:899–906. [PubMed: 1742445]
6. Serruys PW, Morice MC, Kappetein AD, et al. Percutaneous coronary intervention versus coronary artery bypass grafting for severe coronary artery disease. *N Engl J Med* 2009;360:961–972. [PubMed: 19228612]
7. Lange RA, Hillis RD. Coronary revascularization in context. *N Engl J Med* 2009;360:1024–1026. [PubMed: 19228611]
8. Gelijns, AC.; Killelea, B.; Vitale, M.; Mankad, V.; Moskowitz, AJ. Appendix: C. The dynamics of pediatric device innovation: putting evidence in context. In: Field, MJ.; Tilson, H., editors. Institute of Medicine Committee on Postmarket Surveillance of Pediatric Medical Devices. Board on Health Sciences Policy. Washington (DC): The National Academies Press; 2005. p. 302-326. Safe medical devices for children
9. Parides MK, Moskowitz AJ, Ascheim DD, Rose EA, Gelijns AC. Progress versus precision: challenges in clinical trial design for left ventricular assist devices. *Ann Thorac Surg* 2006;82:1140–1146. [PubMed: 16928569]
10. Moore MJ, Bennett CL. The learning curve for laparoscopic cholecystectomy. The Southern Surgeons Club. *Am J Surg* 1995;170:55–59. [PubMed: 7793496]
11. Egorova N, Giacomelli J, Gelijns A, Mureebe L, Greco G, Morrissey N, et al. Defining high risk patients for endovascular aneurysm repair: a national analysis. *Circulation*. Accepted for publication.
12. Rose EA, Gelijns AC, Moskowitz AJ, et al. Long-term mechanical left ventricular assistance for end-stage heart failure. *N Engl J Med* 2001;345:1435–1443. [PubMed: 11794191]
13. Park SJ, Gelijns AC, Moskowitz AJ, Frazier OH, Piccioni W, Raines E, et al. LVADs as destination therapy: a new look at survival. *J Thorac Cardiovasc Surg* 2005;129:9–17. [PubMed: 15632819]
14. Tu JV, Hannan EL, Anderson GM, et al. The fall and rise of carotid endarterectomy in the United States and Canada. *N Engl J Med* 1998;339:1441–1447. [PubMed: 9811920]
15. Hlatky MA, Lee KL, Harrell FE Jr, Califf RM, Pryor DB, Mark DB, et al. Tying clinical research to patient care by use of an observational database. *Stat Med* 1984;3:375–387. [PubMed: 6396793]
16. Hannan EL, Racz MJ, Walford G, et al. Long-term outcomes of coronary-artery bypass grafting versus stent implantation. *N Engl J Med* 2005;352:2174–2183. [PubMed: 15917382]
17. Tunis SR, Stryer DB, Clancy CM. Practical clinical trials: increasing the value of clinical research for decision making in clinical and health policy. *JAMA* 2003;290:1624–1632. [PubMed: 14506122]

Figure.
Kaplan-Meier survival curves for LVAD patients enrolled in 1998--1999 and those enrolled in 2000--2001 in the REMATCH Trial. $P = .00293$.