The Journal of Foot & Ankle Surgery 56 (2017) 417–420

Contents lists available at ScienceDirect

The Journal of Foot & Ankle Surgery

journal homepage: www.jfas.org

Investigators' Corner

Propensity Score Matching: Retrospective Randomization?

Daniel C. Jupiter, PhD

Assistant Professor, Department of Preventive Medicine and Community Health, The University of Texas Medical Branch, Galveston, TX

ARTICLE INFO

Keywords:
indication bias
propensity score matching
randomized controlled trial

ABSTRACT

Randomized controlled trials are viewed as the optimal study design. In this commentary, we explore the strength of this design and its complexity. We also discuss some situations in which these trials are not possible, or not ethical, or not economical. In such situations, specifically, in retrospective studies, we should make every effort to recapitulate the rigor and strength of the randomized trial. However, we could be faced with an inherent indication bias in such a setting. Thus, we consider the tools available to address that bias. Specifically, we examine matching and introduce and explore a new tool: propensity score matching. This tool allows us to group subjects according to their propensity to be in a particular treatment group and, in so doing, to account for the indication bias.

Randomized controlled trials (RCTs) are viewed as the best possible study design, the sine qua non of biomedical research… if we can do them. If they are appropriate. If they are ethical. If the other possible designs are not too tempting, for what are often very compelling reasons. Their strengths in terms of eliminating bias and allowing us to focus very closely on the one intergroup difference of interest are what make them appealing and what lead us to attempt to use this design in all possible situations. However, we often find ourselves presented with medical problems that are not amenable to exploration with this design. Alternatively, we are presented with already existing data, which, if already collected, cannot be post hoc redesigned or retooled to fit into the RCT model.

We explore the strengths of RCTs and why they are so attractive; some situations in which they are not possible or desirable and other designs should not be cast aside; and methods to cause other designs to look and behave more similar to RCTs. Specifically, for the latter, we expand on the matching we discussed in an earlier commentary (1) and explore a method referenced there, called propensity score matching.

Randomized Controlled Trials

An RCT is a prospectively designed and executed trial, usually aimed at exploring one particular intergroup difference. For example, we can imagine a trial that compares 2 methods of hallux valgus repair. We can imagine a trial comparing the outcomes of a novel laparoscopic surgery to the outcomes of traditional open surgery, a trial comparing medical versus surgical management of obesity (2), and a trial comparing various methods to heal diabetic foot ulcers (3).

Financial Disclosure: None reported.
Conflict of Interest: None reported.

Address correspondence to: Daniel C. Jupiter, PhD, Department of Preventive Medicine and Community Health, The University of Texas Medical Branch, 301 University Boulevard, 1.134G Ewing Hall, Galveston, TX 77555-1150.
E-mail address: dajupite@utmb.edu

1067-2516/$ - see front matter © 2017 Published by Elsevier Inc. on behalf of the American College of Foot and Ankle Surgeons.
http://dx.doi.org/10.1053/j.jfas.2017.01.013

Such trials are designed by randomizing patients into the various treatments groups, which often include a control comparator group, along with the experimental group or groups. Although this sounds simple enough, these 2 most important points of an RCT must be handled with care. We need very detailed consideration of what constitutes our various study groups, and we need a very strict idea and prescription of the protocols in the various groups. We must consider what we will do in the case of a protocol violation: will a patient who started in the (experimental) laparoscopic group but ended up needing (control) open surgery be analyzed as a part of the experimental or control group? (This and other topics will be discussed in a future commentary regarding evidence-based medicine, comparative effectiveness research, and pragmatic trials.) In terms of randomization, because the goal is to eliminate or reduce bias, we must ensure that the randomization does this as best as possible. Thus, we must know whether particular characteristics exist that we would like to ensure are balanced between the study groups, because we think they are important factors affecting our outcome of interest. If we are unwilling to trust simple randomization to ensure this balance, we might consider block or stratified randomization (4). In the former, we would break our patients into consecutive blocks, within each block we would perform randomization, ensuring a roughly even distribution between treatment groups at all times during the trial. In the latter, the blocks are determined by the combinations of factors we believe could influence the outcomes. For example, we could separately randomize within male diabetic patients, female diabetic patients, and so forth, such that we have a roughly even distribution between treatment groups within all strata at all times during the trial.

Because RCTs tend to be costly and difficult to design, we will often examine a primary outcome and also consider several secondary outcomes. Thus, we would not consider only correction of the first intermetatarsal angle 6 months after hallux valgus surgery but also the intermetatarsal angle 12 months postoperatively, examine the positions of the sesamoids, look for surgical complications, study hardware removal, attempt to collect functional data, measure the time to return to work, and so forth. This further increases the complexity and means that we must perform a careful design and accounting for all these secondary outcomes. One final wrinkle: if possible, RCTs should be double blinded, such that neither the patients nor the experimenters or evaluators know to which study group the patient belongs. This, we hope, eliminates an additional bias: that of the expectations of researchers and patients.

What Is So Good About RCTs?

Given that these requirements seem to be a lot of work and perhaps prohibitive in their complexity, why are we so enamored of the RCT? There are at least 2 answers. First, randomization, if done correctly, evens out all the differences between groups, except for the intervention. In fact, with inclusion and exclusion criteria and with blocking and stratification, we can even out many intergroup differences in terms of those variables that we believe could affect our outcomes of interest. Lurking in the background, however, are those factors that we did not know might affect the outcome and that we, thus, cannot account for. Randomization is the great equalizer: if we randomize, the differences in factors we have not explicitly accounted for, known or unknown, are likely to be evened out between the groups. Second, thanks to this evening out of factors, the RCT attempts to allow us to answer the following question. Here are 2 groups that are exactly the same—except for the single difference of interest, the intervention being studied. How do the outcomes differ? How would someone who received the control and not the novel treatment react, if they had instead received the novel treatment?

What About Paired *t* Tests?

The last question asked in the previous section should remind us of a much simpler study design. The comparison will be informative in our further discussion. We imagine a trial in which we are comparing analgesics for migraines. We give each patient analgesic A for their first headache. After a suitable washout period, in which we cruelly allow no pain relief, we have them use analgesic B for a subsequent headache. We can directly ask the question with which we ended the last section: how does each patient react if they received experimental treatment B, instead of control treatment A? We have precisely the data to answer this question: if our outcome is some form of pain score, the analysis proceeds as a basic paired *t* test. Perhaps we can tease out the effect of the other variables using multivariate regression. This is nothing compared with the complexity of the RCT. Thus, why do we not always use such a simple design? First, the design becomes more complicated if rather than 1 follow-up point for each patient, we are interested in longitudinal data. Second, although in the case of analgesic use, we can give one and then the other treatment, in the case of more invasive treatments, the use of the first treatment might preclude the use of the second. In hallux valgus repair, for example, once the closing base wedge osteotomy (CBWO) has been performed, we will not return to the same patient to perform an Akin osteotomy to determine whether that surgical approach results in better outcomes. We are led, inexorably, back to the RCT, in which we are able to accommodate mutually exclusive groups.

(A quick aside: life is not, of course, in any case, this simple. Such crossover trials have their own concerns. Should we give treatment A, then B? Or B, then A? Or alternate the order? Or randomize which patients receive which order? What happens if a patient only receives 1 of the treatments and withdraws from the study before taking the other? Because we are studying patients paired with themselves, we must account for that clustering, or random effects (5), which will also be further explored in a future commentary.)

But I Do Not Want To Perform an RCT!

RCTs are not always the answer. They might not be possible, they might not be ethical, and they might not be ideal. We can imagine several such possible situations. For example, we are exploring the 2 types of hallux valgus surgery (CBWO and Atkin), and our institution has a long history of performing these particular types of surgery. It also has a well-detailed electronic health record and good follow-up data available for the patients. Although we know that the indications for the 2 surgeries could be different, should we not attempt to extract as much useful information as possible from this wealth of data, before performing an RCT? Another scenario is that we are considering whether continued use of clopidogrel during foot or ankle surgery leads to more postoperative bleeding events compared with stopping such treatment (6). It would be neither possible nor ethical to randomize patients to receive continued treatment or not as they undergo surgery. How then can we retrospectively study the differences between these groups? Another situation is the case of the Framingham Heart Study, which was not randomized but has a surfeit of information available. How should we analyze these data?

When discussing randomization, we mentioned blocking and stratification as tools to help us achieve balance. Although these are prospective tools, we can think similarly in the above situations: we can block, as it were, retrospectively. This is the notion of matching (1). We hope we can eliminate some of the differences between groups. In the hallux valgus study, perhaps we can achieve some balance in terms of the factors that influence the outcomes independent of surgery and some balance in terms of the factors