Potential pitfalls of experimental design

Phillip Watkins MS

Abstract

Good experimental design begins with the end in mind. An early conversation with a statistician will both increase the chances of an experimental study contributing to the literature and minimize the risks to participating human subjects. Sir R.A. Fisher felt that "to consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination: he can perhaps say what the experiment died of." To this end, some questions from a statistician are presented along with the associated experimental study pitfalls to avoid during the study planning phase. Several concrete examples are provided to give some practical knowledge on how to improve an experimental study at the onset. Hypothesis formulation, sample size determination, randomization, and double-blinding are all explained from the viewpoint of a statistician's final analysis. Confounders, sampling, and missing data are also briefly covered through this hypothetical question and answer session.

Keywords: Experimental design, hypothesis formulation, sample size, randomization

Good experimental design should begin with the end in mind. Sir R.A. Fisher felt that "to consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination: he can perhaps say what the experiment died of." As such, let's explore the most common questions a statistician might ask about experimental design and the associated study pitfalls to avoid.

Q1: Why is this experimental study necessary in humans?

This question should show the body of evidence for this intervention is backed by observational studies and/or experimental animal studies, but has not been proven in human experiments. For a review of observational study designs and the associated pitfalls, see this article's precursor: Potential Pitfalls in Observational Study Design. Do not attempt to write a study protocol before conducting an exhaustive review of the literature[1], as it puts one at risk of "reinventing the wheel", designing an experiment that is theoretically flawed, or attempting to prove an effect that is unsupported by previously published data.

Q2: What is your Hypothesis?

That is, what intervention are you hoping to test? The starting point of a thoughtful study design is carefully defining what you hope to show. The word hypothesis was derived from the Greek hupothesis or "foundation" which may be literally translated as "under-placing." Just as one needs a firm foundation to build a good house, one also needs a rock solid hypothesis to build evidence for or against a scientific process or clinical practice.

Q3: What is the primary outcome of interest in this study?

In hypothesis formulation, one strives to make a statement regarding superiority* that is both testable

Corresponding author: **Philip Watkins**
Contact Information: Philip.watkins@ttuhsc.edu
DOI: 10.12746/swrccc2017.0517.226

*Generally speaking, the null hypothesis (of equivalence) cannot be testing using traditional statistical methods. There are techniques for testing an equivalence or non-inferiority hypothesis, but they require massive sample sizes and are beyond the scope of this article.

and specific. Therefore, it is critical to know the primary factor of interest. For example, "chicken soup is good for the soul," is not testable as one cannot quantify the state of a soul. While "chicken soup is good for a cold" is technically testable, "good" is still not specific enough to illustrate the outcome of interest. However, the statement "chicken soup reduces the duration of a cold" is both specific and testable, as we might compare the median of cold duration between the study and control group to test this belief.

Q4: Are there any secondary outcomes of interest in this study?

Considering secondary outcomes forces the investigator to focus on which outcome is of principal importance in formulating their hypothesis. Multiple primary outcomes mean multiple hypotheses that inflate the experiment-wide error rate. For example, two hypotheses tested at traditional significance (α=0.05) would have a nearly 10% chance of at least one type I error. As such, it is important to narrow the focus to a minimal number of testable statements or consult a statistician to appropriately handle multiple hypotheses. Also take care with double-barreled hypotheses, such as "Drug A is safer and more effective than Drug B", as this statement also contains two hypotheses to test.

Q5: Are there any known confounding variables that may affect your primary outcome of interest?

Confounding occurs when extraneous variables accounts for an observed relationship between the independent and dependent variable. As such, confounding factors are important considerations in thoughtful study design. At the very least, the initial table of any ensuing publication should compare baseline factors to confirm that the study and control groups are comparable with respect to known or suspected confounders.

Q6: How large an effect or difference do you expect to observe? Is it likely small, moderate, or large?

Estimating the effect size is required to power an experimental study. The purpose of powering is to determine the sample size needed to achieve a reasonable probability (~80%) of detecting a statistically significant difference (traditionally $p<0.05$). Failure to calculate power may put patients at risk with little-to-no chance of the study drawing any useful conclusions. Generally, effect size estimates should come from prior observational studies. If the literature on the topic isn't developed enough to estimate an effect size, then another observational study is needed before attempting an experiment design.

Q7: What is the population of interest and how are you planning to recruit subjects?

This explores inclusion and exclusion criteria to ensure that the sample collected yields representative study and control groups. Any rare factors known to affect the primary outcome of interest are ideal exclusion criteria. Do NOT use the results of another study as a historical control: the comparison group should not be systematically different from the intervention group with respect to time or physical location. A historically controlled study could not definitively show that the intervention is the true difference-maker between the study and comparison group.

Q8: If applicable, how are you planning to randomize?

While simple random samples are nice in theory, they are almost practically impossible to achieve in clinical research. Block or stratified randomization methods[2] may better control for known confounders and reduce the required sample size to achieve a reasonable power. A paired study design, if feasible, may be the best design as it results in subjects that are identical with respect to potential lurking or confounding variables. However, take great care to track the pairing information using the same study IDs (e.g. 001A and 001B) so the de-identified study subjects can still be matched appropriately when the study is finished.

Q9: Will you blind the study? If so, how will you conceal/protect the allocation algorithm?

Failure to blind may result in a placebo effect causing differences between the study and control groups rather than a true difference. Wherever possible, an inactive agent should be given that is visually indistinguishable from the drug/intervention. In some situations, one must be clever to achieve a

blind, such as using a double-dummy placebo when pills look different (like aspirin vs. ibuprofen[3]) or even employing some simulation to mimic the appearance of the intervention (e.g. acupuncture vs. sham acupuncture[4]). Double-blinding is the best practice of allowing neither the patient nor the administrator to discern the nature of the intervention administered, so every effort should be taken to achieve this end.

Q10: How much missing data or loss to follow-up do you anticipate and how will we deal with it?

In a perfect world, we would have data for all study subjects in all measured variables. However, instrument malfunctions, misplaced charts, patient drop-out, and a many other issues may result in missing data. While other articles[5] better address this subject, one may generally avoid replacing missing data with values obtained or computed from the non-missing data. A concrete *apriori* plan to minimize missing data is much better than an *ad-hoc* statistical fix. For example, setting a cutoff for acceptable missing data (5-10%) with contingencies for missing those benchmarks like retraining staff, recontacting patients, or instituting a protocol change can prevent issues with missingness. Furthermore, designing a patient-friendly study at the onset to minimize patient loss-to-follow-up may avoid such contingencies altogether. Again, the goal of planning for missing data is to maximize the chance that this study contributes to the literature.

Conclusions

The goal of good study design is to maximize the probability of success while minimizing the risk to study subjects. To this end, one needs a specific and testable hypothesis with an estimate of the effect size of the intervention in order to adequately power a study. Once powered, randomization with double-blinding is the best practice to reduce bias, but baseline characteristics should still be collected and compared to exclude the possibility of a confounding variable as the underlying causal factor. Having a concrete plan to acquire study subjects representative of the population of interest, along with careful consideration of the factors that might result in systematic missing data also require careful planning. Finally, building some stopping rules into your study may be useful to ensure that missing values and effect size estimates are in line with your expectations. In general, one should strive for a study design and write-up that is specific enough to be completely reproducible by an independent investigator. While this exploration was intended to demystify the statistical process of experimental study design, it is by no means a substitute for a conversation with a statistician. For statistical support or help in designing a study, please contact the corresponding author.

Article citation: Watkins P. Potential Pitfalls of Experimental Design. Southwest Respiratory and Critical Care Chronicles 2017;5(17):68-70.
Author Affiliation: Phillip Watkins is a statistician who works in the Clinical Research Institute at Texas Tech University Health Sciences Center in Lubbock, TX.
Submitted: 11/10/2016
Accepted: 12/9/2016
Conflicts of interest: none

References

1. Dawson B., Trapp R.G. Basic & Clinical Biostatistics, 4th ed. New York, NY: Lange Medical Books/McGraw-Hill; 2001: 332-343.
2. Kang M, Ragan BG, Park JH. Issues in outcomes research: an overview of randomization techniques for clinical trials. J Athl Train. 2008;43(2):215-21.
3. Nebe J, Heier M, Diener HC. Low-dose ibuprofen in self-medication of mild to moderate headache: a comparison with acetylsalicylic acid and placebo. Cephalalgia. 1995;15(6):531-5.
4. Moffet HH. Sham acupuncture may be as efficacious as true acupuncture: a systematic review of clinical trials. J Altern Complement Med. 2009;15(3):213-6.
5. Dziura JD, Post LA, Zhao Q, Fu Z, Peduzzi P. Strategies for dealing with missing data in clinical trials: from design to analysis. Yale J Biol Med. 2013;86(3):343-58.