Good Study Design and Analysis Plans as Features of Ethical Research with Humans

BY JANICE M. WEINBERG AND KEN P. KLEINMAN

This article highlights for investigators and institutional review board (IRB) members the importance of a methodologically well-designed study and statistically valid analysis plan for the ethical conduct of research with humans. We describe features of good study designs and statistical plans and explain their relevance to ethical research. Another author has considered this general area for a statistical audience.[1] We emphasize that our goal here is not to compare the ethical merits of competing statistical approaches; that is also material for a statistical audience.

We focus our attention on three categories of research methodology: study design; data analysis/statistical methods; and sample size considerations. For full details about study design and statistical methodology, we suggest texts on introductory biostatistics,[2] clinical trial design and methods,[3] and epidemiological methods.[4] Throughout, we use protocol examples based on our experiences working with our respective IRBs; we have changed some details to protect the confidentiality of the material.

Study Design

■ *Does the general study design enable the investigator to address the aims and hypotheses?* Investigators must choose a study design that allows them to address the hypotheses. For example, consider an educational intervention to reduce the distress caused by a false positive mammogram. The best study design would be a randomized trial that assigned patients to either receive the intervention or not. This would ensure that any changes seen over time are due to the educational intervention rather than some other factor. A deviation from this ideal design might be contemplated if practical considerations required, but the implications of this should be discussed and evaluated. If an inappropriate design is chosen, there is little chance that the study merits the risks to subjects because the study cannot answer the questions and hypotheses developed to meet its aims.

Consider a recently proposed pilot study designed to determine the safety of a new surgical technique. The investigators proposed randomizing subjects to receive the new surgical technique or the standard technique. Since this is a pilot safety study, randomization is probably not an appropriate design, in that randomization more commonly is used for efficacy comparisons. In this case, efficacy comparisons were specifically planned for a future study.

■ *Are outcomes clearly defined?* The investigator must put forward and defend a primary outcome measure by which the study's hypothesis is to be addressed. Returning to the example of an intervention to reduce the distress caused by false positive mammograms, the investigators must describe how they intend to measure distress and show that it can be measured using a valid and reliable method or instrument. If the outcome is not defined in advance, the study may be worthless: no outcome may ever be measured. Another danger is that the outcome may be defined post hoc; this raises the possibility that the investigators might report only the outcomes with the strongest results, which would invalidate statistical assessment.

For example, we recently reviewed a study that proposed a randomized trial to assess whether telemedicine could be used to reduce the length of hospitalization after surgery. Subjects in the telemedicine group would spend part of the usually hospitalized recovery period at home under regular remote contact with hospital staff. The study was intended to determine the cost-effectiveness of telemedicine. However, the protocol did not describe how cost would be measured. There are many aspects of cost, from equipment purchase to time lost from work for the patient. Without knowing the planned outcome, it is impossible to know whether cost could in fact be measured at all.

■ *What are the potential sources of bias and have they been addressed?* Bias is a systematic error that may result in invalid scientific conclusions. In clinical trials, bias may result in over- or under-estimating the true effect of treatment, perhaps leading to a beneficial new treatment being deemed ineffective or an ineffective treatment being

Janice M. Weinberg and Ken P. Kleinman, "Good Study Design and Analysis Plans as Features of Ethical Research with Humans," *IRB: Ethics & Human Research* 25 No. 5 (2003): 11-14.

deemed beneficial. In an observational study, bias may cause over- or under-estimation of the relationship between two variables, possibly resulting in inappropriate changes in patient care or recommendations for public health. The investigators must have carefully considered possible sources of bias and sought to eliminate them.

We provide a relatively simple example of bias. We recently reviewed a study of a new device aimed at improving the independence of subjects with limited mobility. The outcome was the time needed to perform a task with and without the device. It was not indicated whether the task would always be performed in the same order (e.g. without and then with the device). Suppose the device is always used in the second trial. Then an effective device may appear to perform poorly if there is fatigue on the second attempt, or an ineffective device may appear to perform well if subjects learn the environment in the first trial. (One possible way to avoid these biases is to randomize the order of tasks for each subject.)

We illustrate the following aspects of study design through intetervention studies and other planned experiments, though these have implications for observational studies as well.

■ *Is there an appropriate control group?* A control group provides an estimate of what would have happened to the experimental group had they not been experimented upon. For example, subjects' symptoms may improve over time without intervention. Without a control group, this improvement might falsely be attributed to the intervention under study. The choice of control group is related to several ethical considerations. We offer some brief guidelines useful in determining whether an appropriate control group has been proposed. First, is there a standard of care? If so, it should be considered as a treatment in the control group. If the standard of care is not chosen, this should be justified. Next, is the control group comparable to the intervention group in every way except the intervention? If it differs in other ways, the differences need to be carefully enumerated and their implications evaluated. Finally, the risks and benefits to the control arm must be weighed against the risks and benefits to the intervention arm.

■ *Were appropriate methods of blinding used and if so, were they ethically justified?* In a blinded intervention study, the identity of the assigned interventions can be hidden from the subjects, the treating clinicians, the investigators and/or the individuals responsible for monitoring the response variables. Blinding prevents the blinded parties from intentionally or unintentionally affecting the results through their knowledge of the intervention status. Blinding should be used in comparative trials whenever it is feasible and ethically appropriate. If it is not used, this should be justified, and the potential impact on results should be discussed.

For example, we recently reviewed a clinical trial comparing an infusion with a new medication to an infusion without it. The investigators claimed the study could not be blinded due to the unusual color of the new medication. In addition, the protocol stated that the investigators could administer concomitant medication if deemed necessary. Here, the lack of blinding means the decision to use concomitant medication could be influenced by the investigators' knowledge of the intervention group, leading to over- or under- estimation of the true treatment effect. (We question the investigators' assertion that blinding would be impossible in this study. While blinding may be more costly and problematic than in other studies, it would be possible and worthwhile.)

■ *Were appropriate methods of randomization used and if so, were they ethically justified?* Randomization prevents bias in treatment assignment and helps to ensure comparability of intervention groups in all ways except the intervention. Like blinding, it should be used in comparative trials whenever it is feasible and ethically appropriate. In addition, methods of randomization must be carefully considered. If randomization is not planned, this should be justified. Many proposals include randomization in the study design but do not describe how the randomization will be carried out.

We recently reviewed a randomized trial of an experimental treatment plus standard care versus standard care for chronic otitis media in two to seven-year-old children. Because children in this age range might react differently to the new treatment, it is important to ensure that adequate numbers of younger and older children are included in both treatment groups. However, the protocol proposed simple randomization without regard to age. In this case, it is possible that a more appropriate approach would be to randomize within age groups of younger and older children. Failing to do so risks imbalance in the ages between treatment groups, which could jeopardize the estimate of the overall treatment effect and may preclude age-specific estimates.

■ *Are there plans for interim monitoring and analysis of safety and/or efficacy data?* Interim monitoring should probably be done for any study of sufficient duration. For example, it may be unethical to continue to enroll subjects into a study if there is already sufficient information to determine efficacy or evi-

dence of significant toxicity. Continuing a trial that should be stopped, or stopping a trial prematurely, poses ethical problems. The issues associated with the statistical aspects of interim analyses can be complex and need to be carefully delineated in the study design.[5]

Data Analysis/Statistical Methods

■ *Is there a complete data analysis plan?* Investigators may be so focused on the background and/or study design that they give short shrift to the analysis plan. This may lead to analysis plans that effectively say: 'We will analyze the data.' A version of this problem occurs when only a subset of the aims is addressed in the analysis section. When the analysis plan does not provide enough detail, it is impossible to determine whether the investigator has paid sufficient attention to how statistics will be used to answer the study questions. The risk is that statistical techniques to analyze the data may not exist or may be impossible to carry out. It may also be impossible to determine if the study design and collected data are sufficient to address the study questions. If this is the case, the study may never provide useful results.

For example, one study we recently reviewed planned to administer an intervention to improve survival time. In such studies it is typical to use Kaplan-Meier curves and proportional hazards regression as a first approach to data analysis. However, the data analysis section named no statistical techniques. Without some assurance that the investigators were at least aware of these techniques, we could not know whether they would be able to interpret their data meaningfully. Without meaningful interpretation, no risk to subjects would be justifiable. There is also the additional risk that investigators may choose the analysis based on the results obtained, which would invalidate statistical assessment.

■ *Do the proposed analyses enable the investigator to address the study aims?* Even a sufficiently detailed analysis plan may fail to address the questions proposed in the study. The investigators may propose an analytic technique that will not provide the answer to the study question. For example, one study we reviewed was intended to compare the birth weight of infants born to women classified as "drinkers" or "non-drinkers" during pregnancy. The investigators proposed logistic regression as an analytic method. However, logistic regression is a technique designed for dichotomous endpoints; it cannot be applied to a continuous outcome such as birth weight. The proposal demonstrated that investigators had not obtained statistical advice during the study's design and that such advice may help them draw useful scientific conclusions.

■ *Are the proposed statistical methods correct or appropriate?* Finally, the proposed statistical methods may be incorrect or inappropriate. In this case the proposed analysis could appear to answer the question, but for technical reasons the answer obtained would not reflect reality as contained in the data. A classic example concerns the impact of side effects. In a placebo-controlled drug trial, side effects may cause sicker subjects to drop out of the drug arm, leaving only the healthiest subjects with complete data. In contrast, most subjects in the placebo arm are likely to complete the study, as they may not have side effects. An analysis of data from subjects who completed the trials will result in comparison of the healthiest subjects from the treatment arm with most of the subjects on the placebo arm. Although this analysis appears to address study drop out, it may not be an accurate answer. Statistical methods do exist to assess the impact of treatment effect and should be discussed in analysis plans when appropriate.

Sample Size Considerations

■ *Is there an adequate justification of the study sample size?* From a scientific perspective, there should be enough subjects to address the study question and to ensure that the correct conclusion has been reached. From an ethical perspective, if there are too few subjects, the investigator cannot adequately address the study question: the power may be so small that the investigator would be unlikely to detect an extant effect. Thus, the risk to study subjects has served no purpose. In addition, the investigator may reach an incorrect conclusion regarding the study questions. If more subjects are enrolled than are needed to address the study question, then too many subjects have been unnecessarily exposed to potential risk. Therefore, no more patients than are needed to answer the study question should be enrolled. However, the justification of sample size should be based on clinical significance as well as statistical significance. A good sample size assessment should discuss the power to detect a clinically or epidemiologically meaningful and plausible effect.

Some study proposals omit any discussion of how the proposed number of subjects was determined. Others provide incorrect justification. One error in the calculation of sample size is to base the calculation on the wrong outcome measure, as was the case in a protocol we recently reviewed for a new treatment for male erectile dysfunction. The primary outcome was the success rate of vaginal intercourse, defined as the percent of successful attempts at vaginal intercourse. This outcome

ranges between 0 and 100% for each subject. However, the sample size calculations for the study were based on whether or not there was a successful attempt at vaginal intercourse. This is a dichotomous or "yes/no" outcome. The sample size calculation did not reflect the primary outcome or the analysis that would be performed and was not applicable to the study.

Recommendations

We have described some basic features of good study design and statistical plans and have explained their relevance to ethical research. Based on this connection we make several recommendations.

First, we recommend that when initiating a study, investigators consult with a biostatistician, that is, someone with expertise in study design and statistical methodology, regardless of their formal training or professional position. In doing so, investigators can avoid designing studies with flaws in the design and analysis plan that make studies less ethically appropriate.

Second, we recommend that investigators learning methods of human research should specifically examine the ethical implications of a poorly designed study. Clinicians often take introductory courses in biostatistics or study design. In these courses, the scientific and ethical importance of statistical and study design concepts may be neglected or overshadowed by mathematical formulas. If investigators appreciate the ethical importance of statistical methodology and study design, they may pay greater attention to these topics when planning studies.

Some might contend our examples of real-life flaws in study design and statistical analysis plans are elementary mistakes. Yet, to the extent that IRBs fail to detect these flaws, research subjects may be exposed to unnecessary research risks. A statistical and design review of proposed research studies is important due to the need to balance the risk to subjects against the potential benefit to science and ultimately to society in general. This balancing may appear to go beyond the mandate of the IRB; however, the rights of the subject can be interpreted as the right to avoid risks that have little chance of benefiting science or society. This interpretation is formally supported by scholarly work,[6] by the Belmont Report,[7] and by the procedures of one of our own IRBs, which requires investigators to sign the following statement:

> We believe … the risks, if any, are outweighed by the possible benefit to the subject, and the importance of knowledge to be gained, and warrant the [IRBs] decision to approve these risks.

Thus, our final recommendation is that each IRB have a biostatistical reviewer as a member or advisor. A biostatistician can serve several functions. The most obvious is to review protocols that are submitted to the IRB and to make recommendations for changes or improvements. The biostatistician can also serve as an educator on these topics to the IRB.

We recognize that practical problems may arise when a protocol receives a scientific review as part of the IRB review process. A protocol may have received additional, possibly contradictory, scientific reviews from other regulatory or funding agencies or other IRBs. The issue of how an IRB review fits in with other scientific reviews is important and non-trivial, and deserves further analysis.

■ **Janice M. Weinberg, ScD,** is with the Department of Biostatistics, Boston University School of Public Health. **Ken P. Kleinman, ScD,** is with the Department of Ambulatory Care and Prevention, Harvard Medical School and Harvard Pilgrim Health Care.

Acknowledgment

We thank Dr. Susan Fish, Dr. Stephen Lagakos, Dr. Jim Sabin, Dr. Steve Pearson, and several anonymous reviewers for their helpful review and comments.

References

1. Freedman B, Shapiro SH. Ethics and statistics in clinical research: Towards a more comprehensive examination. *Journal of Statistical Planning and Inference* 1994;42:223-233.

2. Rosner B. *Fundamentals of Biostatistics*. Belmont, CA: Wadsworth, 1995.

3. Friedman LM, Furberg CD, DeMets DL. *Fundamentals of Clinical Trials*. New York: Springer, 1998; Meinert CL. *Clinical Trials Design, Conduct and Analysis*. New York: Oxford, 1986.

4. Hennekens CH, Buring JE. *Epidemiology in Medicine*. Boston/Toronto: Little, Brown and Company, 1987; Kelsey JL, Whittemore AS, Evans AS, Thompson WD. *Methods in Observational Epidemiology*. New York: Oxford, 1996; Rothman KJ, Greenland S. *Modern Epidemiology*. Philadephia: Lippicott-Raven, 1998.

5. See ref. 3, Friedman et al. 1998.

6. See ref. 1, Freedman et al. 1994.

7. The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. *The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research*; 18 April 1979. Available from http://206.102.88.10/ohsrsite/guidelines/belmont.html