Editorial

Common Design and Analysis Issues in Clinical Trials

Because I review manuscripts and protocols for clinical trials, I observe a number of recurring issues. These issues mostly relate to the statistical analysis but also include appropriate characterization of the trial itself.

Many reports of randomized controlled trials contain a table showing the baseline characteristics for each of the treatment groups, and some journals advocate use of significance tests to compare the groups at baseline. Indeed, the CONSORT (consolidated standards of reporting trials) statement supports inclusion of a baseline table; it also warns of the inappropriateness of using significance tests for comparison[1]. In statistics, hypothesis/significance tests concern population variables, and if indeed the allocation was randomized, a null hypothesis of no difference is true. Moreover, as Senn points out, an imbalance does not necessarily imply a problem with randomization; nor does a lack of imbalance prove randomization was successful[2]. Often, the results of baseline testing are used to decide which variables, if any, to include in an adjusted analysis of the outcome. Pocock, *et al* point out that baseline imbalance does not dictate the need for adjustment but rather it is the strength of the relationship between a baseline variable and the outcome[3]. Instead of p values, baseline comparability should be considered from the standpoint of clinical significance. Even then, imbalance should not be the criterion for inclusion in an adjusted model. In addition to the unadjusted analysis in the clinical trial, there may sometimes be a clinically compelling rationale to adjust for a small number of covariates irrespective of baseline imbalance.

Another setting where testing is inappropriately carried out involves normality of outcomes in preparation for a t-test or regression model. However, it is important to understand the precise role and importance of normality in these situations. Whether it is comparing 2 treatment groups or fitting a multiple regression model, it is the unexplained differences in the outcome (i.e., residuals) that should be normally distributed, not the raw data themselves. For example, if 2 groups have different means, even if the data are normally distributed in each group, the combined data will not be normally distributed. Figure 1 illustrates the situation using simulated data. After accounting for the different means, normality would be recovered in this example. Therefore, testing normality should not be a preliminary activity but part of model diagnostics. An undesirable consequence of testing for normality is the desire to transform a variable to achieve it. Interpretation of treatment effect of transformed data becomes very difficult clinically and it is not a simple matter of reversing the transformation on the treatment effect. This is unfortunate because inferences based on regression methods, including the t test, are quite robust to departures from normality. Normality testing should be conducted on residuals rather than on raw data, and transformations should be a last resort.

Another situation in which a less than ideal analysis is often conducted is when a continuous outcome is measured at baseline (before) randomization and again at some followup time. The implicit goal in such a design is to see whether one group changes (e.g., improves) more than the other. This leads to a simple comparison of the average within-patient change by means of a t test. As Vickers and Altman point out, there are problems with this approach and they recommend an ANCOVA approach[4]. Among other things, the ANCOVA correctly adjusts for any baseline differences. Further, if an ANCOVA is performed using change as the outcome and baseline as the covariate, the estimates and inference associated with the treatment variable are identical to an ANCOVA using followup as the outcome. Another benefit is that an ANCOVA can result in normal residuals when an outcome is non-normal (even after accounting for group differences in the mean). Figure 2 illustrates this using simulated data. For repeated postrandomization measurements of the outcome, the path is less clear. Mixed-effect models are needed in that case; however, the interpretation of the treatment effect is harder to understand.

See Yoga and arthritis RCT, *page 1194*

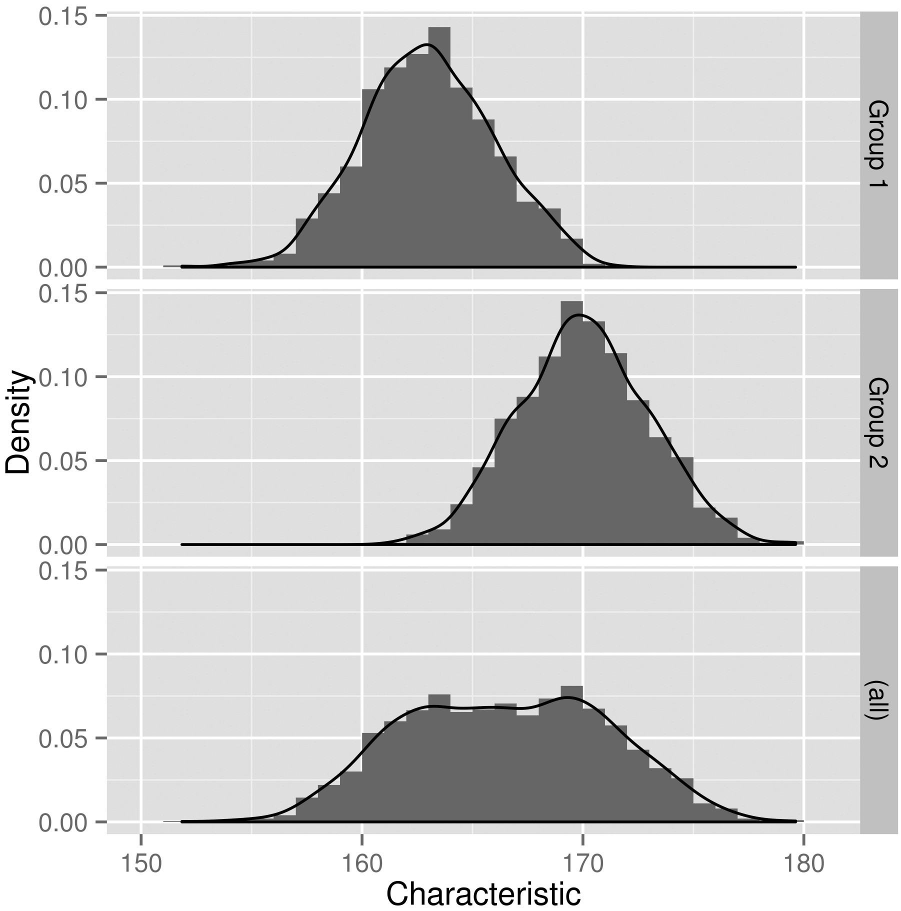

Figure 1. The top 2 panels show histograms of simulated, normally distributed data from 2 groups with different means. The bottom panel shows the histogram of the 2 groups combined and is very clearly not normally distributed.

It is far cleaner, in most cases, to choose a clinically relevant followup time for the primary analysis and perform the ANCOVA as the primary, and reserve the mixed-effect analysis as the secondary, especially to consider group-specific trajectories through the time-by-treatment interaction. However, it is often the case that an insufficient number of followup points are obtained to examine trajectories.

The preceding issues are all related to internal validity of trial results. More recently there has been a growing interest in the generalizability of trial findings. This has led to increased interest in categorizing trials as explanatory or pragmatic. Thorpe, *et al* describe in detail distinguishing characteristics of explanatory and pragmatic trials, but in the simplest terms possible, explanatory trials seek to answer the question, “Can this intervention work under the right circumstances in the right patients?” while pragmatic trials seek to answer the question, “Does this intervention work in practice?”[5] Pragmatic trials are often considered to have a more generalizable result than explanatory trials. Many trialists are familiar with the phase I, II, III, and IV designation of clinical trials in the drug development process. Although phase I-IV trials have their places in the explanatory-pragmatic spectrum, they do not cover all possible trial designs.

The trial reported by Moonaz, *et al* in this issue of *The Journal* is one such example[6]. One matter that adds to the complexity of categorizing trials is the existence of feasibility trials. The primary goal of a feasibility trial is to assess whether a full-scale trial can be conducted. Therefore, with the exception of the study outcomes, all other design elements of the feasibility trial protocol should mirror the design elements of the full-scale trial. If the full-scale trial is explanatory, the feasibility trial will also be explanatory in character, and if the full-scale trial is pragmatic, the feasibility trial will also be pragmatic in character. The term “feasibility” can be used as a qualifier for both explanatory and pragmatic trials.

One cannot hope to do justice to any of these topics, let alone all of them, in a single editorial. Nevertheless, I hope this will serve as useful guidance in analyzing and reporting clinical trials.

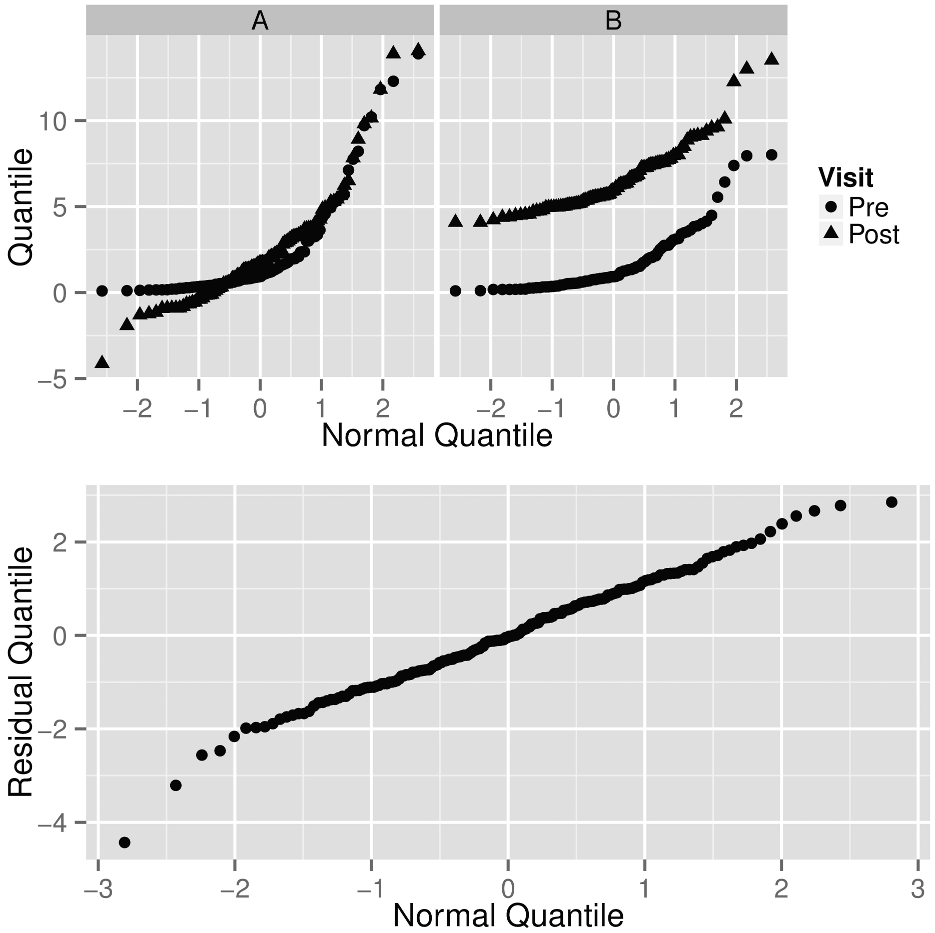

Figure 2. The graphs in the top row are normal QQ plots of simulated non-normal data that could arise in a clinical trial. Panel A is for one treatment group and panel B for the other. Each panel has the pre- and post-data plotted. The curved nature of the plots indicates quite severe lack of normality. The bottom panel shows the normal QQ plot of the residuals from a regression (ANCOVA) model applied to the simulated data. The plot indicates good agreement between the residuals and a normal distribution.

KEVIN E. THORPE, MMath,
Assistant Professor,
Dalla Lana School of Public Health,
University of Toronto;
Head of Biostatistics,
Applied Health Research Centre of the Li Ka Shing
Knowledge Institute of St. Michael's Hospital,
Toronto, Ontario, Canada.

Address correspondence to K.E. Thorpe, Dalla Lana School of Public
Health, 155 College St., 6th floor, Toronto, Ontario M5T 3M7, Canada.
E-mail: kevin.thorpe@utoronto.ca

REFERENCES

1. Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, Elbourne D, et al. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. Ann Intern Med 2001; 134:663-94.
2. Senn S. Testing for baseline balance in clinical trials. Stat Med 1994;13:1715-26.
3. Pocock SJ, Assmann SE, Enos LE, Kasten LE. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. Stat Med 2002; 21:2917-30.
4. Vickers AJ, Altman DG. Statistics notes: analysing controlled trials with baseline and follow up measurements. BMJ 2001;323:1123-4.
5. Thorpe KE, Zwarenstein M, Oxman AD, Treweek S, Furberg CD, Altman DG, et al. A pragmatic-explanatory continuum indicator summary (PRECIS): a tool to help trial designers. J Clin Epidemiol 2009;62:464-75.
6. Moonaz S, Bingham III CO, Wissow L, Bartlett S. Yoga in sedentary adults with arthritis: effects of a randomized controlled pragmatic trial. J Rheumatol 2015;42:1194-202.

J Rheumatol 2015;42:1075–7; doi:10.3899/jrheum.150461