

## Making Sense of Biostatistics: Group Sequential Designs and Interim Analyses

By J. Rick Turner

The trial designs we have discussed so far have been fixed-sample designs. The sample size is estimated during the trial's planning stage. The resulting number of subjects is stated in the study protocol. These subjects are then enrolled and randomized. Statistical analyses of the data are then conducted at the end of the trial. The focus of this column is another design, called the "group sequential design," in which interim analysis plays a crucial role. The fundamental purpose of interim analysis is to facilitate informed decision-making at various points about whether the trial should be terminated early.

The word "group" in the name "group sequential design" does not refer to a treatment group, or "arm." It refers to the group of all subjects who have completed a trial before an interim analysis is conducted. Imagine a study in which a maximum of 1,000 subjects will be recruited into each of two treatment arms. The study's design calls for three potential interim analyses and one final analysis if the trial has not been terminated. The first interim analysis will be conducted once 500 subjects (approximately 250 subjects in each arm) have completed their participation. If the trial continues, another 500 subjects will complete their participation and join the interim analysis group. The second interim analysis will use the data from all completed subjects, i.e., 1,000 subjects. If the trial continues, another 500 subjects will complete their participation and join the interim analysis group. If the trial continues, the final 500 subjects will complete their participation. At this point, all 2,000 subjects will have participated in the trial, and a final analysis will be performed.

Why would we want to terminate a trial before all of the potential subjects have completed their participation? There are several reasons, all of which relate to the fact that choosing the number of subjects for a clinical trial is a matter of estimation, not precise formulaic determination. A chapter from my book *New Drug Development* (Turner, 2007) that focuses on sample size estimation is available at [http://firstclinical.com/journal/2007/0711\\_Sample\\_Size.pdf](http://firstclinical.com/journal/2007/0711_Sample_Size.pdf), so I will be brief here.

The mathematics of statistical analysis mean that the likelihood of a successful result (from the viewpoint that "success" means obtaining a statistically significant result) can be increased by increasing the sample size. However, there are both ethical and business reasons why adding a large safety margin to the sample size is not an appropriate strategy. Each sponsor must therefore determine an "acceptable" sample size, where the term "acceptable" refers to balancing the likelihood of obtaining a statistically significant result with the costs of conducting the clinical trial. The strategy of interim analysis legitimately allows the analysis of data collected from less than the chosen maximum number of subjects, provided a well-delineated set of procedures, called "stopping rules," are followed for conducting such analyses. If a statistically significant result is obtained during one of these analyses, the logic is that a decision can legitimately be based on a smaller number of subjects.

One reason for terminating a trial based on the results of an interim analysis is that the test drug has been shown to be statistically better than placebo. Once this result is known, it is unethical to continue to give subjects the placebo treatment, since it is now known to be inferior to the test drug. A second reason is that there is compelling evidence that the drug has an unacceptable safety profile. A third reason is compelling evidence that, even if the

trial proceeds to its initially stated maximum number of subjects, it is probabilistically unlikely to generate a statistically significant result. In other words, a trial can be terminated early because of good news, bad news, or a forecast of no news.

We can now build upon the information discussed in last month's column about multiple comparisons to show how a sophisticated statistical approach is required when using interim analyses. Consider the scenario presented earlier, in which a maximum of 2,000 subjects will be recruited into a study, and the study design calls for three potential interim analyses and one final analysis if the trial has not been terminated. If we were using a fixed sample design, we would know in advance that we would only be conducting one statistical analysis and doing so once the entire sample of subjects has completed their participation. Therefore, the nominal significance level, typically the 5% p-level, can be chosen without having to consider the issue of multiplicity. However, with our group sequential design, we do not know in advance precisely how many times we will conduct the same analysis to test the same hypothesis. We know that the minimum number is one (the first interim analysis) and the maximum number is four (all three interim analyses plus the final analysis), but the actual number conducted could be one, two, three or four. The statistics are complicated further because we do not know in advance exactly how many comparisons we will conduct.

Two statistical considerations are pertinent. First, given that more than one analysis may be conducted, we need to lower the p-value or  $\alpha$ -level (the chance of a false positive) to make it more difficult to attain statistical significance, i.e., take a more conservative approach. Second, it is usual to place more faith in an analysis conducted on a larger sample than on a smaller sample. Accordingly, one approach that can be used in this context is the O'Brien-Fleming method. This method modifies the  $\alpha$ -level for each of the individual interim analyses by considering not only the number of analyses that may be conducted, but also the relative placement of each individual analysis in the string of possible analyses. This strategy effectively makes it considerably harder for the first interim analysis to obtain statistical significance, as would a very conservative  $\alpha$ -level, and relatively easier for the later ones to attain statistical significance, as would an  $\alpha$ -level that approaches the level that would be chosen if it were known that only one analysis would be conducted at the end of the trial.

O'Brien and Fleming's approach employs group sequential boundaries in defining the stopping rules. In the first interim analysis, the boundary values are extremely conservative because of the small sample size. That is, at this point, we need to be wary of efficacy and safety "results" that may not be borne out in larger sample sizes. The boundary values become less extreme later in the string of potential analyses, with the value at the final analysis, if reached, being close to the conventional p-value used in single analyses. Effectively, if all of the possible analyses are conducted, all p-values will be lower than 0.05 but will sum to 0.05 — the traditional level adopted if it is known in advance that only one analysis will be conducted. Note that, to make statistical room for the interim analyses, the p-value of the final analysis will be lower than 0.05.

## Reference

Turner, J.R., 2007, *New Drug Development: Design, Methodology, and Analysis*. Hoboken, NJ: John Wiley & Sons.

## Author

J. Rick Turner, PhD, MICR CSci, MTOPRA is the President and Chief Scientific Officer of Turner Medical Communications LLC. Contact him at RickTurnerUSA@nc.rr.com or 1.919.636-0701.