Making Sense of Biostatistics: Multiplicity in Statistical Testing

By Soha Elmorsy

When biostatisticians plan a study, they define the maximum probability of false positive results that is acceptable; they call that the alpha, and it is generally set at 5% or 0.05. When they analyze the results, they conclude that the hypothesis is proven if the probability of a false positive is less than the alpha (p<0.05). However, if a study makes multiple comparisons, e.g., of different endpoints or within different subgroups, there are multiple hypotheses and multiple chances for a positive outcome.

Imagine you have one chance to roll a die. If you roll a six, your chance of winning is one out of six. Now imagine you have three chances to roll the die; your chance of winning is much higher. The same principle applies to statistical “multiplicity” and false positives in clinical studies.

Multiplicity occurs when multiple comparisons are performed on the same data. The following are ways multiplicity can occur and how to deal with them:

- **Comparing three or more treatment arms.** For example, if a study has three arms — two active drugs and one placebo — biostatisticians can make three comparisons: Drug A vs. placebo, Drug B vs. placebo, and Drug A vs. Drug B. To counter this effect, biostatisticians can divide alpha by the number of comparisons (the Bonferroni correction). For example, if alpha = 5% and there are three comparisons, at least one comparison must have a p value less than 0.05/3 = 0.0167.

- **Comparing multiple treatment outcomes.** For example, a study on a drug for coronary artery disease might compare the incidence of angina, strokes and death between the drug and placebo. With multiple outcomes to compare, one of them might show significance just by chance. Therefore, a good approach is to designate a single comparison as the primary endpoint and leave the others as secondary endpoints of interest. Sometimes, researchers combine several endpoints into a single composite endpoint, but this approach is not suitable for all studies and has its own implications for interpretation of the results.

- **Comparing multiple subgroups.** For example, a study might assess the effect of a diabetes drug on the blood glucose levels of all participants, on male participants, and on female participants. If p>0.5 for the all-participant results, a subgroup might still show statistical significance. This might be a true difference between subgroups or it might be a chance finding because of multiple comparisons. We can believe the results if the subgroup analysis had been planned in the protocol, the number of subgroups is very limited, the results are biologically plausible, and the findings are consistent with external evidence. Otherwise, the results are viewed as preliminary, and further studies are recommended.

- **Assessing outcomes at multiple points in time (“repeated measures”).** For example, a study might measure liver enzymes monthly for two years. We could compare the measurements at each time point, but this would be difficult to interpret and will increase the probability of false positive results. We might decide to compare at only two or three time points, e.g., at baseline, at one year, and at two years. An alternative approach is to compare a summary measure, such as the mean or the median of all the readings. However, the best approach is to conduct an advanced type of analysis called “regression,” which enables us to compare trends in the readings between the two arms.
• **Analyzing interim data.** A data safety and monitoring committee (DSMC) might analyze the data at multiple interim points during a study. If the interim results demonstrate that there is a significant difference between the groups, the DSMC can stop the study early. Repeated interim analyses might increase false positive results. Biostatisticians have developed sophisticated methods to deal with this problem, such as the Pocock Rule, the O'Brien and Fleming Rule, and Lan and DeMet’s alpha spending function. Some methods divide alpha evenly over time or events. Some start with a very small alpha at the beginning and increase it as the study progresses. Alpha spending ties it to the proportion of information collected at the time of analysis. Each method has its own advantages, but all ensure that alpha is not exceeded. Ironically, since some of the alpha is used up by the interim reviews, the alpha available for the final data is smaller than it would have been without the interim reviews.

The issue of multiplicity must be addressed at the design stage of a study. However the endpoint(s) are defined and whatever statistical rules are employed, these decisions must be made before the study begins, or the number of possible comparisons becomes, in effect, infinite and the alpha for any comparison infinitesimal.

**References**


**Author**

Soha Elmorsy, M.D., Ph.D., M.Sc. in Clinical Trials, is a Research Consultant and Head of the Epidemiology and Statistics Department at King Abdullah Medical, Makkah, Saudi Arabia and Associate Professor of Medical Pharmacology, Faculty of Medicine, Cairo University. Contact her at sohaelmorsy@gmail.com.