The Clinical Significance of Statistical Significance

ROBERT C. KANE
Office of Oncology Drug Products, Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, Maryland, USA

Key Words. Statistical significance • Alpha • False positive inference

Disclosure: The content of this article has been reviewed by independent peer reviewers to ensure that it is balanced, objective, and free from commercial bias. No financial relationships relevant to the content of this article have been disclosed by the author, planners, independent peer reviewers, or staff managers.

LEARNING OBJECTIVES
After completing this course, the reader should be able to:

1. Explain the difference between exploratory (hypothesis-generating) analyses and confirmatory (hypothesis-testing) analyses.
2. Describe why a statistical test result (p-value) only expresses the likelihood that the result could have occurred by chance and does not prove the truth of the study.
3. Define what a hazard ratio expresses.

ABSTRACT
Modern clinical trials provide the evidence for most therapeutic advances, and that evidence, expressed in a statistical format, is used to draw inferences about a population from the study’s results. Clinician judgment translates these inferences for best individual patient care, but many clinicians struggle with the statistical interpretation of trial results. This review provides a clinical and non-Bayesian perspective on some key elements in the statistical design, analysis, and interpretation of randomized, comparative, phase III clinical trials intended to demonstrate a better outcome (superiority) than with a control treatment. The Oncologist 2008;13:1129–1133

INTRODUCTION
Considerable concern exists regarding the perceived high failure rate of phase III trials for new drugs. Statistical considerations, while crucial to successful study outcomes, are not the usual culprits. In a phase III trial, in contrast to the preliminary phase II setting, the endpoint of interest is often different and more clinically compelling. For example, in phase II, the response rate may be a useful endpoint to suggest drug activity, but in phase III, effects on overall survival, prolonged disease control, or abatement of disease-related symptoms are more appropriate clinical measures of the treatment effect on the entire treated group. Thus, phase III studies often impose a higher bar for success. Also, when phase III confirmatory or “pivotal” studies are planned, the...
true effect size of the treatment intervention on the clinical outcome is not known. The proposed study should be planned to demonstrate an effect judged as clinically meaningful as well as feasible to examine. It should be no surprise that some phase III trial designs may not demonstrate a positive result if actual treatment effects are smaller than projected. This does not mean that a study design is underpowered if it does not detect a smaller difference, and this does not imply that an increase in the sample size is automatically justified. Smaller effect sizes at some point are not clinically relevant, and studies should not be overpowered to achieve a statistically significant but clinically minuscule result.

**Effect Size, Sample Size, and Power**

In planning a study, the primary endpoint selected should be clinically relevant and objectively measurable. Also, the proposed effect size of the treatment on the primary endpoint should be clinically meaningful. Next, the sample size necessary to show this effect is estimated. The study sample size has an inverse relation to the effect size of the new therapy. (Note that the sample size is actually derived from the number of events needed for the analysis, which is inversely related to the square of the effect size.) The smaller the treatment effect, the larger the sample size needed to provide the number of events necessary for a comparison between study arms. The study should be designed to enable a high chance (power) of detecting a positive effect on that primary endpoint. The sample size estimation is vitally important in that it is not only futile but may be unethical to conduct a study that actually has little chance of demonstrating its purported result and which then leads to a false-negative conclusion, not because of lack of efficacy, but because of insufficient enrollment to discern the effect [1].

At times, following interim looks at ongoing study results, the sample size (study enrollment) may be increased to improve the power of the study. Such decisions require careful statistical planning to avoid falsely positive conclusions (see below).

Successful study planning requires an explicit definition of the clinically meaningful primary study endpoint(s), an estimate of the magnitude of the treatment effect on the endpoint, and an estimate of the sample size necessary to demonstrate the difference of interest. (Additional critical design factors such as selection of an appropriate comparator control arm, randomization, blinding, mandate for use of the intent-to-treat population for efficacy analysis, other efforts to minimize bias, censoring rules, and prespecified interim analysis plans are also important, but for this discussion are assumed to be in good order [1].) This advance planning allows the application of statistical tests after study completion to evaluate the results and to minimize the risk of drawing false inferences from the results. Such inferences may be falsely negative or positive. As noted above, studies should be designed with high power to detect a positive result if a treatment is truly effective. This power calculation is designed to provide control of the false-negative inference; that is, concluding that a treatment intervention is not effective because the study failed to demonstrate a benefit (when in fact the treatment is effective). Conversely, if a study endpoint is changed or expanded, the sample size originally estimated may require an adjustment. The Eastern Cooperative Oncology Group (ECOG) 1684 report of adjuvant interferon therapy of melanoma provides an example of some difficulties that may result from an attempt to analyze an endpoint not originally prespecified and for which a sample size recalculation was not described in the publication [2]. In this report, the stated original primary study endpoint, relapse-free survival, was statistically persuasive \((p = .0023)\). However, the study sample size did not enable a clear inference of the treatment effect on overall survival (OS) as shown by the variability over time in the nominal \(p\)-values for OS (see Table 2 in that report). Statistical procedures are available to re-estimate sample size in a blinded, planned, interim look as a study progresses.

The OS benefit of adjuvant chemotherapy for stage IB lung cancer patients remains uncertain based on the Cancer and Leukemia Group B (CALGB) 9633 study [3]. A survival benefit was reported initially but not sustained with additional observation time. While multiple factors confounded the results, a larger sample size (number of events) would have added more strength and precision to the survival effect [4]. If an important study is designed to demonstrate superiority but fails to do so, the most common reason is that the investigational therapy was ineffective in altering the outcome. However, the publication should include a statement estimating what difference the study, as performed, could actually have demonstrated. If this difference is clinically unimportant, the reader is immediately warned to be cautious in drawing any efficacy conclusions from the study.

**Alpha—The False-Positive Risk**

Conversely, it is also essential to estimate and to control for the risk of drawing a false-positive inference; that is, of erroneously concluding than an ineffective treatment is beneficial. A boundary is set to estimate and to limit the likelihood of making this error. This boundary is expressed numerically as a value, designated as alpha, \(\alpha\) or the type 1 error probability, because it represents the more serious inferential mistake that can occur from a study’s analysis. By
convention and historical precedent, the overall type 1 error can be no greater than 1 in 20, or 5%, or .05 in decimal form. As formulated, this calculation actually allows for the recognition of two directions (or sides) of difference, superiority or inferiority, so it is termed a two-sided $\alpha$ of .05. From the perspective of identifying a superior therapy, the side that confirms an inferior outcome is not of interest—only the superior half is of interest, and therefore the $\alpha$ barrier is effectively .025 or 2.5% for superiority (a one-sided $\alpha$ of .025). Thus, the acceptable risk can be no more than 1 in 40 for making a false-positive inference (that an ineffective treatment is effective).

The usual prespecified study hypothesis is that the study will show no difference in outcomes between the treatment groups being compared. By prespecifying this "null hypothesis," statistical testing can then evaluate an outcome difference. If a difference is found, the important questions then are: (a) How big is this difference? and (b) Is this a real difference?—How likely could a difference this large (or larger) arise merely by chance? After testing the difference statistically, one of two mutually exclusive conclusions may be reached: either (a) the difference reasonably could have arisen by chance alone (and therefore a difference of interest is not established by this study) or (b) the likelihood of chance producing this outcome difference is so small that we should conclude that there is a difference of interest.

A statistical test of the outcome difference, presented in the form of a $p$-value, expresses the probability of observing a difference this extreme if in fact the null hypothesis is true (i.e., despite this study finding a difference, in truth there is no difference). If the $p$-value (from the study results) is smaller than the prespecified $\alpha$ value (the prespecified acceptable risk for error), we then conclude that the observed difference demonstrated in the study is so unlikely to be from chance that an alternative explanation is required. Assuming the other study elements are in order, the inference is made that the difference is attributable to the intervention (treatment).

In this case, the observed difference found by the study is declared to be a statistically significant difference. Conversely, if the $p$-value for the result is larger than the prescribed boundary ($\alpha$), we can only conclude that the study, as performed, did not demonstrate a difference of interest. This failure to demonstrate a difference does not mean that there is no difference or that there is equivalence, only that this study, as conducted, did not demonstrate a difference and asserting a conclusion of equivalence is incorrect [5]. Similarly, demonstration of a statistically significant difference (superiority) does not prove that the difference is truly the consequence of the treatment, only that the difference is unlikely to be a result of chance (unlikely in this example meaning less than the 1 in 40 odds represented by the prespecified $\alpha$ of .025). While these odds may seem small and thus convincing, in some situations it is justified to insist on even more conservative results, such as a predetermined $\alpha$ (and thus a corresponding study result $p$-value) of .01 or .001, before concluding that a difference is "statistically significant" and is so unlikely to be the result of chance that clinical decisions and action should follow.

**STATISTICAL TESTING**

The statistical tests chosen for these analyses depend, among other factors, on the type of endpoint comparison proposed and what underlying assumptions have to be fulfilled for the test to be valid. For the comparison of proportions, such as response rates or frequencies of events, some form of $\chi^2$ procedure is usual and the test results are expressed as a specific $p$-value and associated 95% confidence interval. Comparisons of OS, disease-free survival, time to disease progression, or other "time-to-event" results usually are expressed in terms of both a significance level and a hazard ratio (HR). Here, statistical significance is usually tested by the log-rank procedure, from which a $p$-value is derived, again, for comparison with the prespecified $\alpha$ value. However, the $p$-value itself does not provide information about the size of the treatment effect. The HR, obtained from the Cox (proportional hazard) model, estimates the magnitude of the difference found—the treatment effect. This HR magnitude expresses the relative risk reduction achieved by the new therapy (versus the control), is usually expressed as a decimal value <1.0, and provides both the point estimate value for this specific study and the 95% confidence interval of plausible values around that point estimate HR (e.g., an HR of 0.70 is expressed as a 30% relative risk reduction in outcome). The importance of the confidence interval is that the true result of this new treatment for a population is more likely contained within the confidence interval than at the point estimate of this study sample, and the narrowness of the confidence interval indicates the precision of the estimate. Also, the relative risk reduction expressed by the HR has to be judged in the context of the treatment effect of the control arm and the absolute risk reduction achieved.

With time-to-event analyses, in which the null hypothesis is represented by a relative risk or HR of 1.0, HR confidence intervals that do not include 1.0, but are entirely above or below 1.0, also represent statistically significant differences. It is important to note that the risk reduction as expressed by the HR depends on the baseline risk. A treatment that produces a longer OS time or time to progression, from a control group median of 3 months to an intervention group median of 6 months, results in an HR of 0.50; the
same HR value occurs when a treatment doubles the outcome from a control median of 24 months to 48 months in the treatment group. (Note that the test is not comparing the median values but the entire distributions of both groups. Also, these comparisons assume that the proportional hazards model is applicable.) HRs cannot simply be compared numerically across studies. Assurance would be needed that the studies were performed in the same patient population (study eligibility), same disease condition of the illness (stage, prior therapy, etc.), and same control arm therapy, and that supportive care and other procedures including methods for assessing the study endpoint were similar. Time-to-event efficacy analyses should be performed only upon reaching a prespecified number of endpoint events (i.e., the analysis is event-driven, not time-driven). Primary analyses of treatment results generally should not be conducted at arbitrary time points (e.g., survival at 6 months or 1 year) even though these times may appear to have some clinical context. If the primary time-to-event analysis is favorable, these supplemental analyses of the proportion of patients with or without an event at certain time intervals may be informative clinically.

When time-to-event results are reported, median values usually are also reported, but comparison of the median values is not the intent. Median values provide a simple benchmark for descriptive clinical purposes but have no special provenance and are not helpful if the median time to event has not been reached for a study group. A description of the mean values of the outcomes could be meaningful but is almost never available because it usually requires final event data on all enrolled patients.

A critical component of study design, when specifying the study endpoint(s), includes defining the acceptable α risk (of making a false-positive inference) for the endpoint(s). When the study reports that the p-value for the observed difference is ≤.05, the reader should also be informed what prespecified α was determined to be appropriate for interpreting this result. If this false-positive risk has not been defined and stated in advance, for each test performed, then there is no defined process to declare the “statistical significance” of the study result. Statistical significance should not be implied; if the rules have not been followed, the report should state that the statistical significance of the result is not defined. In some circumstances, and prior to any knowledge of the study results, a statistical plan can be enacted or modified, but this can be hazardous to the credibility of the study and should be fully disclosed. When multiple endpoints are deemed clinically important, they may be prioritized as coprimary outcomes or categorized by their importance as primary and secondary outcomes; there are well-accepted rules for designing these options.

In disease conditions receiving considerable attention with numerous similar studies, it is likely that an occasional “statistically significant” study outcome (especially with a modest p-value, such as a p-value of .05) is in fact falsely positive and should be replicated before being accepted as a true result. Conversely, what inferences are appropriate when results show very small p-values (e.g., p = .0001) for an endpoint? Is this certain enough? The answer is not always straightforward. Consider a study sized for a survival endpoint, with enrollment of several hundred patients, and with a planned interim analysis of an intermediate endpoint such as time to progression, to be followed later by the final analysis for survival. In such studies, a surprisingly small difference in the results for the intermediate endpoint, sometimes measured in a few days, can often yield very small p-values of high statistical significance but of doubtful clinical significance. Clinicians should not be comforted or coerced by statistical persuasion alone.

**Subgroup and Post Hoc Analyses**

When the statistical testing (primary statistical analysis performed at the prespecified time) indicates that the prespecified primary endpoint differences in the study are not statistically significant, then the entire α (the value allocated to control the false-positive risk) has been expended by that analysis. Any further statistical testing of this or other endpoints poses a high, uncertain, and unacceptable risk of producing false-positive inferences. If secondary endpoints are then assessed as alternative endpoints of importance, the unanswered question is why those endpoints were not considered important enough to be designated as coprimary and to receive prospective statistical planning (including α allocation). Such post hoc examinations produce conclusions that can be called “exploratory” or hypothesis-generating, but should be understood to have a high likelihood of being falsely positive (and must be distinguished from confirmatory [hypothesis-testing] analyses). In contrast, when the differences observed for the planned primary endpoint do demonstrate statistical significance, additional analyses, including subsets based on baseline characteristics, can be informative and valid for inferential purposes, especially to explore interactions of patient characteristics with the outcome. A disinterested statistician should guide this effort. When the differences in the primary endpoint are not statistically significant, and a post hoc “search-and-rescue” analysis is reported to be positive, the follow-on studies to confirm this finding typically prove to be negative (even when the hypotheses appear plausible biologically; these subsequent studies often suffer from publication bias).

This post hoc analysis error arises in part from the “mul-
ticipality trap; that is, performing multiple analyses when the risk for a false-positive error has not been “protected” by a prespecified plan to control that risk (including a allocation) among the multiple endpoint analyses. If this risk has not been carefully protected, such analyses may be reported, but they should carry a prominent caution warning: “statistical inferences should not be drawn from these exploratory results.” To avoid misleading readers, journal editors and peer reviewers should receive and review the formal study protocol and statistical plan for comparison with a manuscript’s assertions. Papers, abstracts, and presentations should prominently distinguish exploratory from confirmatory analyses. A recent discussion of these forms of post hoc statistical roulette was provided by Lagakos [6]. Using the one-sided α risk of .025 (or a two-sided α of .05) as described above, clinicians only need to remember that, if 10 independent subgroup analyses are performed (not so unusual), the likelihood of at least one “positive victory” is 22.4%, almost 10 times the minimally acceptable 2.5% risk. An equally important but more subtle factor contributing to the hazard of overinterpreting post hoc findings is regression to the mean bias, in which outliers (“positive” results) will not be reproducible simply because they are, in fact, outliers.

RESPONDER ANALYSES
“Responders had better outcomes than nonresponders.” Fortunately for readers, this statement signals that a flawed analysis is being presented. In a comparative (two or more study arms) trial, valid inferences are derived from the comparison of the entire groups (or prespecified subgroups) following the randomization and completion of a “fair” trial. Attempts to compare the outcomes of subgroups of patients selected conditionally on some postrandomization event such as responder status are not valid. The reason is that the members of this type of subgroup are selected after some assessment process and time interval and not by randomization. Responders almost always have better outcomes than nonresponders; however, the responder analysis does not establish that a difference in outcome is a result of the intervention (therapy) [7, 8].

CONCLUSIONS
Sound statistical procedures enhance the precision of and allow confidence in the conclusions drawn from clinical trials. However, the clinical significance of a study is far more than a reported “statistically significant” positive outcome. It is a judgment that a well-designed and conducted study, showing prospectively defined substantial statistical significance, has an effect size that is clinically meaningful, in the context of the clinical relevance of the endpoint used in the study. A statistically positive efficacy result also has to be judged in light of the potential toxicities to be incurred. Also, clinical significance requires an assessment of the external validity of the result, outside of the specific study, considering alternative therapy options and the universe of studies asking similar questions in similar settings. If the risk of a false-positive conclusion is not clearly predefined in the study’s analysis plan, a post hoc conclusion about a study result (usually among many results examined) has no statistical meaning and likely is falsely positive. It is uncertain how often such transgressions arise though innocence or by intention. Statistics can enhance clinicians’ knowledge and confidence, but the unwary can be misled. Probability can masquerade as biologic discovery. Statistical significance does not assure truth, but with careful attention to the rules, “we won’t often be wrong” (attributed to Neyman and Pearson in [9]).

ACKNOWLEDGMENT
The views expressed are independent work and should not be considered as an official position of the U.S. FDA.

REFERENCES