

Demystify Statistical Significance—Time to Move on From the *P* Value to Bayesian Analysis

J. Jack Lee

Correspondence to: J. Jack Lee, PhD, Department of Biostatistics, The University of Texas M. D. Anderson Cancer Center, 1400 Pressler St, Unit 1411, Houston, TX 77030 (e-mail: jjlee@mdanderson.org).

Compared to the long histories of mathematics, physics, chemistry, and biology, statistics is a young science. Rooted in mathematics, statistics is a science of quantitative reasoning. It provides a formal framework for assessing the strength of evidence in the midst of uncertainty. More specifically, statistics allows one to quantify the probability of an event to make proper inference. It provides tools that we can use to sift through mountains of information to identify the signal and filter out the noise. Recently, the United Nations declared October 20, 2010, as “World Statistics Day.” The declaration stated “On 20 October 2010, the World will celebrate the first World Statistics Day, to raise awareness of the many achievements of official statistics premised on the core values of service, integrity and professionalism” (1).

Medicine is an ancient profession. Practicing medicine has evolved from empirical based to evidence based. Evidence-based medicine has become a motto for modern medicine, and statistics is an indispensable tool for evaluating the strength of evidence contained in the data (2,3). How much information do the data contain? Formulating the problem in the context of evaluating drug efficacy, one may ask: “Based on the data, does the drug work?” In this issue of the Journal, Ocana and Tannock (4) ask: “When are ‘positive’ clinical trials in oncology truly positive?” They provide a critical assessment of 18 randomized trials that led to the approval of targeted drugs by the United States Food and Drug Administration (FDA) over the past 10 years. They make four major claims: 1) statistical significance does not equate to clinical significance; 2) the *P* value alone is not sufficient to conclude that a drug works; 3) a prespecified magnitude of clinical benefit, δ , that is defined as the difference in primary endpoints between control and experimental groups, is required to gauge whether a drug works or not; and 4) the absolute difference is more relevant than the relative difference in measuring a drug’s efficacy. The excellent commentary by Ocana and Tannock (4) echoes several previous works addressing these issues (5–8). It challenges the current drug approval paradigm and calls out to both the medical and statistical communities to come up with a more robust framework for assessing drug effectiveness to determine more accurately whether a drug really works. The FDA (and European Medicines Agency alike) indeed considers both clinical and statistical evidence in evaluating drug efficacy. The question is how to provide a robust framework for weighing the evidence. I support the general claims of Ocana and Tannock (4) but argue that they do not go far enough. I provide my own assessment on the statistical evaluation of the drug’s efficacy and give my perspective on its future direction.

What is statistical significance? In the mind of many medical researchers, statistical significance means that the *P* value less than or equal to .05. It also translates to a “positive” result; hence, an article can be published in a journal, a grant successfully reviewed, and a drug approved by the FDA. When *P* value is less than or equal to .05, it is assumed that there is sufficient evidence that the drug works, thus it should be approved. This overly simplistic view is, of course, erroneous. But unfortunately, it permeates the medical research community. The commentary by Ocana and Tannock (4) points out that statistical significance does not equate to a clinically meaningful difference. They propose that in addition to a statistically significant *P* value, to declare that a trial is “positive,” the observed difference in a survival outcome must equal or exceed a prespecified clinically important value. In reviewing 18 randomized trials, Ocana and Tannock found that four trials did not report the predefined hazard ratio and another six trials had an observed hazard ratio being less than the predefined hazard ratio [(4), table 1]. Comparing the observed treatment effect with the prespecified effect is certainly a step in the right direction. However, making inference based on a single point estimate is inadequate because the observed treatment effect itself carries uncertainty. Furthermore, requiring the observed treatment effect to be greater than δ can be overly conservative because the trial will have 50% power even if the true treatment effect is equal to δ . A frequentist confidence interval helps in gauging this uncertainty, but a complete solution can only be obtained by taking the Bayesian approach.

In the frequentist hypothesis-testing framework, the *P* value is defined as the probability of observing events as extreme or more extreme than the observed data, given that the null hypothesis (H_0) is true. If the *P* value is small enough (conventionally, $P \leq .05$), the data provide evidence against the null hypothesis, so we reject the null hypothesis. The *P* value is not the probability that the null hypothesis is true. It is an indirect measure to assess whether the null hypothesis is true or not. So, what is the probability that the null hypothesis is true and what is the probability that the alternative hypothesis (H_1) is true? The Bayesian approach addresses these questions directly and provides coherent answers. Bayesian methods treat an unknown parameter (eg, the real treatment effect) as random and the data as fixed and known, which they are. The Bayesian approach calculates the probability of the parameter(s) given the data, whereas the frequentist approach computes the probability of the data given the parameter(s). Since the parameters are unknown and the data have been observed, it makes more sense to formulate a problem using the Bayesian

approach. To provide a clear illustration of how the Bayesian method works in contrast to the frequentist approach, a very simple example is given below. Interested readers are encouraged to consult many superb works on this subject (9–13).

Assume that one is interested in learning the response rate (θ) of a new drug in a single-arm design and wants to test $H_0: \theta = 0.2$ against $H_1: \theta = 0.4$. In addition, suppose that we have two trials. Trial 1 has observed four responses in 10 patients (the estimated response rate, $\hat{\theta} = 0.4$). Based on the binomial one-sided test, Trial 1 has a P value of .033. Does this mean that the chance of the null hypothesis being true is 1 in 30 and the chance of the alternative hypothesis being true is 29 in 30? The answer is definitely “No.” It is well known that P values overstate the evidence against the null hypothesis. The Bayesian approach quantifies such evidence using the Bayes factor (defined as the ratio of the posterior odds of the null hypothesis divided by the prior odds of the null hypothesis). The likelihood of observing four responses in 10 patients is 0.088 and 0.251 when $\theta = 0.2$ and 0.4, respectively. Assuming equal priors for H_0 and H_1 , the Bayes factor, $B_{01} = 0.35$. Therefore, the posterior odds of the null hypothesis against the alternative hypotheses are only 1:3, not 1:30 as when inferring from the frequentist P value. Furthermore, assuming that θ follows a non-informative prior distribution of beta (0.2, 0.8), the Bayesian approach can calculate the posterior probability of a minimum treatment effect. For example, the following efficacy measures can be easily calculated: probability($\theta > 0.2$) = 0.904, probability($\theta > 0.3$) = 0.697, and probability($\theta > 0.4$) = 0.432. Without much prior information, observing four responses in 10 patients suggests that the response rate has a 90% chance of being greater than 0.2, a 70% chance of being greater than 0.3, and only a 43% chance of being greater than 0.4. Thus, upon defining the threshold of a clinically meaningful response rate, calculating the posterior probability that the response rate is greater than the threshold is straightforward.

Now, assume we have another trial. Trial 2 observes 27 responses in 100 patients ($\hat{\theta} = 0.27$) and has a P value of .034. Because Trials 1 and 2 have similar P values, do these two trials provide the same amount of information against the null hypothesis? The answer again is definitely “No.” Following the same calculation, we have $B_{01} = 9.83$ in Trial 2. The posterior odds of the null vs the alternative hypotheses is in fact about 10:1, suggesting that the null hypothesis is more likely to be true than the alternative hypothesis. The posterior probabilities that probability($\theta > 0.2$) = 0.949, probability($\theta > 0.3$) = 0.238, and probability($\theta > 0.4$) = 0.003 indicate that the true response rate is likely to be greater than 0.2 but extremely unlikely to be as high as 0.4.

Statistics in medicine has passed through its infancy and childhood. As it moves into its adolescence, the growing pains of reconciling frequentist and Bayesian views continue. The commentary by Ocana and Tannock (4) and the above examples clearly illustrate the limitation and fallacy of using the P value to gauge a positive treatment effect. Although the frequentist paradigm has

been widely applied and is deeply rooted in medical research, it is time to move on and look for a better solution. The Bayesian approach provides a direct and coherent assessment of the evidence contained in the data. Earlier applications of Bayesian methods were hampered by the notion of subjectivity and computation difficulty. However, these two major roadblocks have been successfully addressed in many contexts. The application of Bayesian methods in clinical trials has also gained ground, particularly in adaptive designs (14–18). The Bayesian approach is complementary to and can provide a superior alternative to the frequentist paradigm (19). I encourage medical researchers to have an open mind, learn more about Bayesian methods, and apply them to provide a more accurate statistical assessment of the results in clinical trials.

References

1. The United Nations Statistics Division. *World Statistics Day 2010*. 2010. <http://unstats.un.org/unsd/wsd/Default.aspx>. Accessed October 26, 2010.
2. Ashby D, Smith AFM. Evidence-based medicine as Bayesian decision-making. *Stat Med*. 2000;19(23):3291–3305.
3. Greenhalgh T. *How to Read a Paper: The Basics of Evidence-Based Medicine*. Chichester, UK: John Wiley & Sons Ltd; 2010.
4. Ocana A, Tannock IF. When are “positive” trials in oncology truly positive? *J Natl Cancer Inst*. 2010;xx(xx):xxx–xxx.
5. Goodman SN. Toward evidence-based medical statistics. 1: The P value fallacy. *Ann Intern Med*. 1999;130(12):995–1004.
6. Goodman SN. Toward evidence-based medical statistics. 2: The Bayes factor. *Ann Intern Med*. 1999;130(12):1005–1013.
7. Wagenmakers E-J. A practical solution to the pervasive problems of p values. *Psychon Bull Rev*. 2007;14(5):779–804.
8. Wijesundera DN, Austin PC, Hux JE, Beattie WS, Laupacis A. Bayesian statistical inference enhances the interpretation of contemporary randomized controlled trials. *J Clin Epidemiol*. 2009;62(1):13–21.
9. Berry DA. *Statistics: A Bayesian Perspective*. Pacific Grove, CA: Duxbury Press; 1996.
10. Spiegelhalter DJ, Abrams KR, Myles JP. *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. West Sussex, UK: John Wiley & Sons Ltd; 2004.
11. Goodman SN. Introduction to Bayesian methods I: measuring the strength of evidence. *Clin Trials*. 2005;2(4):282–290.
12. Louis TA. Introduction to Bayesian methods II: fundamental concepts. *Clin Trials*. 2005;2(4):291–294.
13. Berry DA. Introduction to Bayesian methods III: use and interpretation of Bayesian tools in design and analysis. *Clin Trials*. 2005;2(4):295–300.
14. Berry DA. A guide to drug discovery: Bayesian clinical trials. *Nat Rev Drug Discov*. 2006;5(1):27–36.
15. Biswas S, Liu DD, Lee JJ, Berry DA. Bayesian clinical trials at the University of Texas M.D. Anderson Cancer Center. *Clin Trials*. 2009;6(3):205–216.
16. Berry SM, Carlin BP, Lee JJ, Müller P. *Bayesian Adaptive Methods for Clinical Trials*. Boca Raton, FL: Chapman & Hall; 2010.
17. Zhou X, Liu S, Kim ES, Herbst RS, Lee JJ. Bayesian adaptive design for targeted therapy development in lung cancer—a step toward personalized medicine. *Clin Trials*. 2008;5(3):181–193.
18. Lee JJ, Xuemin Gu, Suyu Liu. Bayesian adaptive randomization designs for targeted agent development. *Clin Trials*. 2010;7(5):584–596.
19. Held L. A nomogram for P values. *BMC Med Res Methodol*. 2010;10:21.

Affiliation of author: Department of Biostatistics, The University of Texas M. D. Anderson Cancer Center, Houston, TX.