Contemporary Clinical Trials 44 (2015) 33–35

Contents lists available at ScienceDirect

Contemporary Clinical Trials

journal homepage: www.elsevier.com/locate/conclintrial

Bayes factor and posterior probability: Complementary statistical evidence to *p*-value

Ruitao Lin, Guosheng Yin *

Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam Road, Hong Kong

ARTICLE INFO

Article history:
Received 22 April 2015
Received in revised form 29 June 2015
Accepted 3 July 2015
Available online 26 July 2015

Keywords:
Bayes factor
Bayesian inference
Frequentist hypothesis testing
Posterior probability
p-value

ABSTRACT

As a convention, a *p*-value is often computed in hypothesis testing and compared with the nominal level of 0.05 to determine whether to reject the null hypothesis. Although the smaller the *p*-value, the more significant the statistical test, it is difficult to perceive the *p*-value in a probability scale and quantify it as the strength of the data against the null hypothesis. In contrast, the Bayesian posterior probability of the null hypothesis has an explicit interpretation of how strong the data support the null. We make a comparison of the *p*-value and the posterior probability by considering a recent clinical trial. The results show that even when we reject the null hypothesis, there is still a substantial probability (around 20%) that the null is true. Not only should we examine whether the data would have rarely occurred under the null hypothesis, but we also need to know whether the data would be rare under the alternative. As a result, the *p*-value only provides one side of the information, for which the Bayes factor and posterior probability may offer complementary evidence.

In Volume 383 of the journal *Lancet* (January, 2014), Fuchs et al. [1] reported an international, randomized, multicenter, and placebo-controlled, phase III clinical trial with patients of advanced gastric cancer to investigate the efficacy of ramucirumab between October, 2009 and January, 2012. Ramucirumab is a monoclonal antibody vascular endothelial growth factor receptor-2 (VEGFR-2) antagonist that can prevent ligand binding and receptor pathway activity. The trial was originally planned to enroll 615 patients for detecting a hazard ratio (HR) of 0.71 in overall survival with 90% power and a two-sided type I error rate of 0.05. Due to slow recruitment, the sample size was modified to 315 based on an amended HR in overall survival of 0.69 and 80% power in November, 2010. Nevertheless, this amended sample size was further changed to a total enrolment of 348 patients with 268 deaths in October, 2011. Finally, a total of 355 patients were enrolled in the trial with 238 randomized to the ramucirumab group and 117 to the placebo group (with an allocation ratio of 2:1).

With respective 179 and 99 deaths in the ramucirumab and placebo groups, the observed HR for overall survival is 0.776 with a 95% confidence interval (CI) of [0.603, 0.998], and a *p*-value of 0.047. The HR estimated from a multivariable adjusted analysis is 0.774 with a 95% CI of [0.605, 0.991], which maintains the statistical significance with a *p*-value of 0.042. By incorporating the predefined stratification factors, the multivariate analysis further yields an estimated HR of 0.767 with a 95% CI of [0.598, 0.984] and a *p*-value of 0.037. As a result, Fuchs et al. [1] concluded that ramucirumab is significantly beneficial to the patients with advanced gastric or gastro-esophageal junction adenocarcinoma progressing after first-line chemotherapy.

The reported HRs in Fuchs et al. [1], including the observed HR, adjusted HR, and stratified HR, are all closer to 1 than those specified in the protocol, and the upper bounds of all the CIs are also close to 1. The Kaplan–Meier estimates of the overall survival (OS) curves are similar between the treatment and control groups, although the progression-free survival (PFS) curves are substantially separated. This observation may cast doubt on whether PFS would be an adequate surrogate for OS [2], as the difference in OS diminishes to the borderline statistical significance. In addition to the HRs for all patients in the trial, Fuchs et al. [1] further classified patients into 33 subgroups, and only 5 out of 33 subgroups show a statistically significant difference in overall survival, as shown in their Fig. 3. This trial can serve as an example that the null hypothesis is rejected based on the conventional significance level, while the observed data show comparable patterns between the experimental and control groups. In addition to *p*-values, we explore more data information to examine the treatment difference comprehensively in evidence-based research.

Hypothesis testing is ubiquitous in modern statistical applications, such as in the fields of biology, medicine, engineering, and economics. The computation of a *p*-value is a critical part of the hypothesis testing procedure. A *p*-value represents the probability of obtaining the data as or more extreme than the observed given that the null hypothesis is true. Conventionally, the statistical significance level is set at 5%, so that a *p*-value below 5% is considered significant leading to rejection of the null hypothesis. Nevertheless, there are misunderstandings on the interpretation of a *p*-value and the debates on misuse of a *p*-value have been

* Corresponding author.
E-mail address: gyin@hku.hk (G. Yin).

http://dx.doi.org/10.1016/j.cct.2015.07.001
1551-7144/

extensive [3–6]. Our goal is to re-evaluate the conclusions based on *p*-values for the trial in Fuchs et al. [1] and further advocate the complementary role of the Bayesian posterior probability in conjunction with the Bayes factor to the *p*-value in the evidence-based clinical research.

Let θ denote the log hazard ratio of OS, $\theta = \log(\text{HR})$, between the ramucirumab and placebo groups. To test whether there are survival differences between the two groups, a two-sided hypothesis is formulated as

$$H_0 : \text{HR} = 1 \quad \text{versus} \quad H_1 : \text{HR} \neq 1,$$

which is equivalent to

$$H_0 : \theta = 0 \quad \text{versus} \quad H_1 : \theta \neq 0.$$

Based on the allocation ratio of 2:1, the log-rank test statistic T asymptotically follows a normal distribution under some regularity conditions

$$T \sim \text{N}\left(\sqrt{2D}\theta/3, 1\right),$$

where D is the total number of deaths [7]. Under the null hypothesis, $H_0 : \theta = 0$, T asymptotically follows the standard normal distribution

$$T|H_0 \sim \text{N}(0, 1).$$

The frequentist approach to such a two-sided test is to reject the null hypothesis if the absolute value of the observed test statistic $|t|$ is greater than the critical constant $z_{\alpha/2}$, where $z_{\alpha/2}$ is the $100(1-\alpha/2)$th percentile of the standard normal distribution and α is the level of statistical significance. In an asymptotic sense, the aforementioned testing procedure is equivalent to testing $H_0 : \theta = 0$, where an independent and identically distributed sample of size $n = 2D/9$ is observed from $\text{N}(\theta,1)$.

Fuchs et al. [1] provided three *p*-values with regard to the observed, adjusted, and stratified hazard ratios, respectively. Based on the frequentist two-sided test of a normal mean, the observed test statistic t can be traced back as

$$|t| = \Phi^{-1}(1-p\text{-value}/2),$$

where $\Phi^{-1}(\cdot)$ is the quantile function of the standard normal distribution and the sign of t depends on whether the observed hazard ratio is greater (+) or less (−) than 1. Let $\phi(\cdot, \mu, \sigma^2)$ denote the density function of a normal random variable with mean μ and variance σ^2, and let $f(\theta|H_i)$ denote the prior density of θ under H_i, $i = 0,1$. Under the null hypothesis H_0, $f(\theta|H_0)$ is a point mass at $\theta = 0$. Based on the observed test statistic t, the Bayes factor in favor of the alternative hypothesis over the null is derived in a similar way as Yuan and Johnson [8]:

$$\text{BF}_{10}(t) = \frac{m_1(t)}{m_0(t)},$$

where $m_0(t)$ and $m_1(t)$ are respectively the marginal densities of the test statistic under H_0 and H_1,

$$m_0(t) = \phi(t, 0, 1) \quad \text{and} \quad m_1(t) = \int_{\infty}^{\infty} \phi\left(t, \sqrt{2D}\theta/3, 1\right) f(\theta|H_1)d\theta.$$

Given an equal prior probability of H_0 and H_1, the posterior probability of H_0 is given by

$$P(H_0|t) = \frac{1}{1 + \text{BF}_{10}(t)}.$$

Both $\text{BF}_{10}(t)$ and $P(H_0|t)$ depend on the specification of $f(\theta|H_1)$. To make a comprehensive comparison, we investigate two classes of $f(\theta|H_1)$. In the first class, we consider simple alternative hypotheses, i.e., $f(\theta|H_1)$ represents a degenerate distribution at a point mass. Under the two-sided simple alternative,

$$m_1(t) = 0.5\phi\left(t, \sqrt{2D}\theta_T/3, 1\right) + 0.5\phi\left(t, -\sqrt{2D}\theta_T/3, 1\right)$$

where θ_T is the point mass specified under the alternative hypothesis. Four cases for θ_T are examined: $\theta_{T_1} = \log(0.69)$ is the log hazard ratio specified in the modified protocol, $\theta_{T_2} = \log(0.776)$ is the observed log hazard ratio, and the last two are specified based on the uniformly most powerful Bayesian test (UMPBT) [9], θ_{T_3} (and θ_{T_4}) $= \sqrt{2\log(2\gamma)/n}$, where γ is the threshold of $\text{BF}_{10}(t)$ in favor of H_1. We take two values for γ: $\gamma = 25$ is considered as a "strong" evidence to support the alternative in Jeffrey's scheme [10], and $\gamma = \exp(z^2_{\alpha/2}/2)/2 \approx 3.41$ results in the same rejection region of UMPBT as that of a two-sided classical test of size α.

In the second class, we consider composite alternatives, for which we specify four continuous prior distributions of θ under H_1. The first one is a standard Cauchy density, $f_1(\theta|H_1) = \text{Cauchy}(\theta)$, which is recommended by Jeffreys [10], and the second is an intrinsic prior [11] with density $f_2(\theta|H_1) = \text{N}(0, 2)$. We also explore two non-local prior densities [12]: $f_3(\theta|H_1) \propto \theta^{-2}\exp(-0.318/\theta^2)$ is the inverse moment prior (IMOM) with the tails behaving like those of the Cauchy prior, and $f_4(\theta|H_1) \propto \theta^2\phi(\theta, 0, 0.159)$ is the moment prior (MOM) that has the same mode as that of $f_3(\theta|H_1)$.

Table 1 presents the reported *p*-values, the Bayes factor, and the posterior probabilities of the null hypothesis, which are computed for the

Table 1
Reported hazard ratios (HRs) and *p*-values, and computed Bayes factors (BF_{10}) and posterior probabilities of H_0 for the ramucirumab trial.

Reported	HR	*p*-value	*t*	BF_{10}	$P(H_0\|t)$	BF_{10}	$P(H_0\|t)$	BF_{10}	$P(H_0\|t)$	BF_{10}	$P(H_0\|t)$
				Simple alternatives							
				$\theta_{T_1} = -0.371$		$\theta_{T_2} = -0.254$		$\theta_{T_3} = -0.336$		$\theta_{T_4} = -0.249$	
Observed	0.776	0.046	−1.995	2.380	0.296	3.647	0.215	2.639	0.275	3.645	0.215
Adjusted	0.774	0.042	−2.033	2.523	0.284	3.797	0.208	2.793	0.264	3.792	0.208
Stratified	0.767	0.037	−2.086	3.110	0.243	4.377	0.186	3.410	0.226	4.362	0.187
				Composite alternatives							
				Cauchy $f_1(\theta)$		Intrinsic $f_2(\theta)$		IMOM $f_3(\theta)$		MOM $f_4(\theta)$	
Observed	0.776	0.046	−1.995	0.688	0.593	0.643	0.609	0.221	0.819	0.785	0.560
Adjusted	0.774	0.042	−2.033	0.715	0.583	0.669	0.599	0.236	0.809	0.826	0.547
Stratified	0.767	0.037	−2.086	0.824	0.548	0.774	0.564	0.294	0.773	0.999	0.500

Note: t is the test statistic, $\theta_{T_1} = -0.371$ is the log hazard ratio specified in the protocol, $\theta_{T_2} = -0.254$ is the observed log hazard ratio, $\theta_{T_3} = -0.336$ and $\theta_{T_4} = -0.249$ are those based on the UMPBT under $\gamma = 25$ and $\gamma = \exp(z^2_{\alpha/2}/2)/2 \approx 3.41$, respectively. Cauchy is the standard Cauchy density function; Intrinsic denotes the intrinsic prior with density $\text{N}(0, 2)$; IMOM is the inverse moment prior with $f_3(\theta|H_1) \propto \theta^{-2}\exp(-0.318/\theta^2)$; and MOM represents the moment prior with $f_4(\theta|H_1) \propto \theta^2\phi(\theta, 0, 0.159)$.

observed, adjusted, and stratified hazard ratios. We can see that the p-values all lead to significant survival differences. If BF_{10} falls inside the interval of [20,150], then there is strong evidence against H_0 [13], which, under an equal prior probability of H_0 and H_1, corresponds to the posterior probability of H_0 between 0.66% and 4.76%. For the simple alternatives, the posterior probability of H_0 varies between 20% and 30%, which indicates that the data do not contain sufficient information to reject the null hypothesis. For $\theta_{T_1} = 0.69$, which is used to determine the sample size, $P(H_0|t)$ attains the largest value of 29.6% based on the observed HR. Most of the Bayes factors BF_{10} for the simple alternatives are around 3, which indicates that the evidence in the data favoring H_1 over H_0 is not worth more than a bare mention [13]. The results based on composite alternatives are more striking: there is at least a 50% posterior probability of H_0 based on the four continuous priors. The Bayesian hypothesis tests under both the simple and composite alternatives show that the posterior probability of the null hypothesis is still moderate even though the p-value is less than 0.05.

Despite the popularity of using the p-value as a measure of evidence against H_0, we show that its role might be exaggerated to some extent [14–16]. Practitioners tend to over-emphasize the importance of a p-value and over-interpret a p-value of 0.05. It is difficult to perceive the p-value in a probability scale and quantify it as the strength of the data against the null hypothesis, while the posterior probability of the null hypothesis has an explicit interpretation of how strong the data support the null. If there exists a treatment difference which although may be small, a p-value of 0.05 could eventually be reached if the sample size of a trial is large enough. Not only should we examine whether the data would have rarely occurred under the null hypothesis, but we also need to know whether the data would be rare under the alternative. As complimentary measures of evidence, the posterior probability of the null hypothesis and Bayes factor provide additional information about the strength of the data, which account for both the alternative hypothesis and sample size.

Acknowledgments

We thank the two referees for their many constructive and insightful comments that have led to significant improvements in the article. The research was supported in part by a grant (17125814) from the Research Grants Council of Hong Kong.

References

[1] C.S. Fuchs, J. Tomasek, C.J. Yong, et al., Ramucirumab monotherapy for previously treated advanced gastric or gastro-oesophageal junction adenocarcinoma (REGARD): an international, randomised, multicentre, placebo-controlled, phase 3 trial, Lancet 383 (2014) 31–39.
[2] K.R. Broglio, D.A. Berry, Detecting an overall survival benefit that is derived from progression-free survival, J. Natl. Cancer Inst. 101 (2009) 1642–1649.
[3] M.J. Bayarri, J.O. Berger, The interplay of Bayesian and frequentist analysis, Stat. Sci. 19 (2004) 58–80.
[4] S. Goodman, Toward evidence-based medical statistics 1: the p-value fallacy, Ann. Intern. Med. 130 (1999) 995–1004.
[5] E.J. Wagenmakers, A practical solution to the pervasive problems of p values, Psychon. Bull. Rev. 14 (2007) 779–804.
[6] R. Hubbard, R.M. Lindsay, Why P values are not a useful measure of evidence in statistical significance testing, Theor. Psychol. 18 (2008) 69–88.
[7] D. Schoenfeld, The asymptotic properties of nonparametric tests for comparing survival distributions, Biometrika 68 (1981) 316–319.
[8] Y. Yuan, V.E. Johnson, Bayesian hypothesis tests using nonparametric statistics, Stat. Sin. 18 (2008) 1185–1200.
[9] V.E. Johnson, Uniformly most powerful Bayesian tests, Ann. Stat. 41 (2013) 1716–1741.
[10] H. Jeffreys, Theory of probability, 3rd edn Oxford University Press, Oxford, 1961.
[11] J.O. Berger, L.R. Pericchi, The intrinsic Bayes factor for linear models, Bayesian Stat. 5 (1996) 25–44.
[12] V.E. Johnson, D. Rossell, On the use of non-local prior densities in Bayesian hypothesis tests, J. R. Stat. Soc. Ser. B 72 (2010) 143–170.
[13] R.E. Kass, A.E. Raftery, Bayes factors, J. Am. Stat. Assoc. 90 (1995) 773–795.
[14] V.E. Johnson, Revised standards for statistical evidence, PNAS 110 (2013) 19313–19317.
[15] J.P. Ioannidis, Why most published research findings are false, PLoS Med. 2 (2005) e124.
[16] R. Nuzzo, Statistical errors: P values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume, Nature 506 (2014) 150–152.