A SIMULATION STUDY OF THE STRENGTH OF EVIDENCE IN THE ENDORSEMENT OF MEDICATIONS BASED ON TWO TRIALS WITH STATISTICALLY SIGNIFICANT RESULTS

Don van Ravenzwaaij
Department of Psychology, University of Groningen
d.van.ravenzwaaij@rug.nl

For the endorsement of new medications, the US Food and Drug Administration typically require two trials, each with p<.05, to demonstrate effectiveness. In this paper, we calculated with simulations what it means to have exactly two trials with p<.05 in terms of the actual strength of evidence quantified by Bayes factors. Our results show that different cases where two trials have a p-value below .05 have wildly differing Bayes factors. In a non-trivial number of cases, evidence actually points to the null hypothesis. We recommend use of Bayes factors as a routine tool to assess endorsement of new medications, because Bayes factors consistently quantify strength of evidence.

INTRODUCTION

Disclaimer: what follows is a short summary of the published paper by van Ravenzwaaij and Ioannidis (2017). I strongly encourage the reader to read instead the full version of this work, which is open-access and available here: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0173184.

Endorsement of medications (drugs and biologics) for clinical use has been under rigorous control by the US Food and Drug Administration (FDA) since 1962. The FDA functions as a gateway for the adoption of new medications by endorsing drugs and biologics through the results of clinical trials. Medications are tested against a placebo condition or an existing alternative and statistical evidence is accumulated to quantify efficacy. By far the most common way to quantify evidence of efficacy is through Null Hypothesis Significance Testing (NHST): a *null hypothesis* of no effect is defined, data is collected, and the probability of observing this data or something more extreme is quantified by a *p-value*. Usually, a *p*-value smaller than .05 is deemed *statistically significant*: it is considered sufficient evidence to *reject* the null hypothesis. Empirical studies have shown that the use of *p*-values is practically ubiquitous in biomedical research and this applies also to clinical trials (Chavalarias, Wallach, Li, & Ioannidis, 2016).

The FDA values the need for rigorous statistical evidence in their policy for drug endorsement, as becomes evident from their guidance for industry (Food and Drug Administration, 1998): "With regard to quantity, it has been FDA's position that Congress generally intended to require at least two adequate and well-controlled studies, each convincing on its own, to establish effectiveness." (p. 3). The typical application in practice is that two independent clinical trials with $p<.05$ are required before a new drug or biologic gets endorsed. Unfortunately, there is no specification of how many trials with $p>.05$ are allowed among the set of trials that contains these two statistically significant trials. Two significant trials out of two attempted trials constitutes a different *strength of evidence* compared to two significant trials out of five attempted trials.

How often does it happen in practice that the FDA endorses medication following two out of five significant trials? An examination of 12 antidepressant agents approved by the FDA between 1987 and 2004 indicates that it is common for antidepressants to be endorsed even though the majority of trials was non-significant (Turner, Matthews, Linardatos, Tell, & Rosenthal, 2008). The FDA acknowledges that this happens as evident from this quote from the approval label of Citalopram hydrobromide (endorsed following 2 out of 5 significant clinical trials): "Highly variable results have been seen in the clinical development of all antidepressant drugs. Furthermore, in those circumstances when the drugs have not been studied in the same controlled clinical trial(s), comparisons among the results of studies evaluating the effectiveness of different antidepressant drug products are inherently unreliable. Because conditions of testing (e.g., patient samples, investigators, doses of the treatments administered and compared, outcome measures, etc.) vary among trials, it is virtually impossible to distinguish a difference in drug effect from a difference due to one of the confounding factors just enumerated." (FDA, 2018).

In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10, July, 2018), Kyoto, Japan.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org

In van Ravenzwaaij and Ioannidis (2017), we examine with simulations the extent to which strength of evidence varies when employing a criterion for drug approval of having two *p*-values lower than .05 for different scenarios. Here, I will focus on the scenario of exactly two statistically significant results out of five attempted clinical trials. Strength of evidence across all attempted trials is quantified using Bayes factors (Jeffreys, 1998; Goodman, 1999). A Bayes factor captures the relative evidence that the data provide for the alternative hypothesis against the null hypothesis in the form of an odds ratio. For example, when $BF = 10$, the data are 10 times more likely to have occurred under the alternative hypothesis than under the null hypothesis. On the other hand, when $BF = 0.1$, the data are 10 times more likely to have occurred under the null hypothesis than under the alternative hypothesis. As for interpreting the strength of evidence as quantified by a Bayes factor, a Bayes factor between 1 and 3 (or, conversely, between 1/3 and 1) is considered 'not worth more than a bare mention', a Bayes factor between 3 and 20 (or, conversely, between 1/20 and 1/3) is considered 'positive', and a Bayes factor between 20 and 150 (or, conversely, between 1/150 and 1/20) is considered 'strong' (Kass & Raftery, 1995).

In the next section, I will describe the part of the simulations that I will present here. Then, I will present the results of our simulations, demonstrating both the range in strength of evidence and the proportion of times the evidence actually points in favor of the null hypothesis. I will conclude with a discussion of the implications of our results for regulatory assessment of new medications.

METHOD

Presented here will be two sets of simulations. Every set consists of 2,500 data sets. All of the data sets were intended to mimic two–condition between–subjects experiments with an experimental group and a control (e.g. placebo) group. A two-tailed *t*-test with a threshold of $p<.05$ combined with selection as "successes" only of those results for which all statistically significant effects are in the direction of the experimental condition being better than the control (e.g. placebo) condition. The two sets of simulations differed on the true population effect size between the two groups. In the first set of simulations, the true population effect size was small (0.2 standard deviations, or 0.2 SD), and in the second set of simulations, the true population effect size was zero (0 SD).

Presented here are simulations for the scenario of 2 significant results out of 5 performed. This was achieved by continuously regenerating data until exactly 2 significant results emerged. We also varied the number of participants per group. We ran five conditions: n=20, n=50, n=100, n=500, and n=1,000.

Thus, to sum up, the simulations presented here varied along the following dimensions:

1. Effect size: small (0.2 SD), and zero (0 SD)
2. Number of participants: 20, 50, 100, 500, and 1,000

This resulted in a total of 10 types of simulations. We replicated each simulation type 500 times. The full range of simulations, as well as further details about the implementation of the Bayes factors and supplementary analyses, may be found in van Ravenzwaaij and Ioannidis (2017).

RESULTS

The selected Bayes factor results are shown in Figure 1. The left column contains results for the small effect size and the right column contains results for the zero effect size. The top row contains the range of Bayes factors obtained for each type of simulation. In the top panels, the y-axis plots the Bayes factor in favor of the alternative hypothesis on a log scale. Within the panels, different columns indicate a different number of participants. The box-plots contain the middle 50% of simulation results, with the tails extending to 100% of the simulation results. The horizontal dashed line represents the case where evidence equally favors the alternative and the null hypothesis, results above the line favor the alternative hypothesis, and results below the line favor the null hypothesis. Note that for the small effect size, the cell for 1000 participants is empty. The reason for this is that these conditions do not realistically occur for this effect size.

The bottom row contains the same results, now binned in terms of Bayes factor size. In the bottom panels, the y-axis plots the percentage of 500 simulations for which the Bayes factor is

lower than a certain cut-off value. Within the panels, different columns indicate a different number of participants, and different colors indicate different cut-off values.

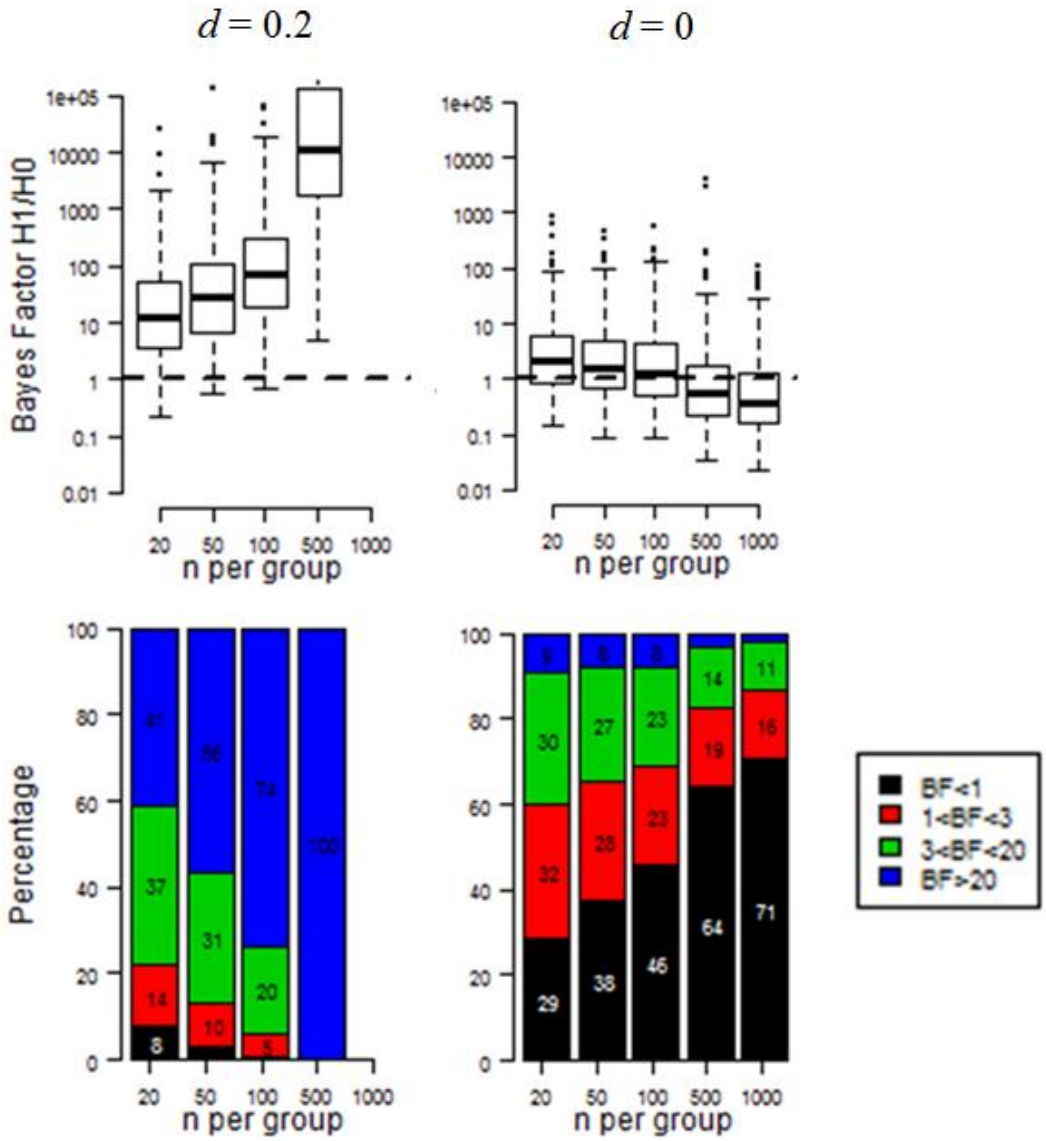

Figure 1. Bayes Factor Results (See text for details)

The results show that there is substantial variability in the evidential strength both across different types of simulations (as reflected by the different heights of the boxes) and within different types of simulations (as reflected by the size of the boxes and the extent of the tails). Summarizing the main trends, increasing sample size leads to higher evidential strength for medications that achieve two trials with statistically significant results out of five attempts if the effect size is small, but increasing sample size leads to lower evidential strength if the effect size is zero. This is how it should be: more data leads to strength of evidence as calculated by the Bayes Factor to point increasingly more strongly towards 'the truth'. The decision criterion based on *p*-values on the other hand invariably leads to an endorsement decision, regardless of the underlying effect size.

DISCUSSION

The result of the presented simulations is simple yet compelling: a criterion of endorsement of two *p*-values lower than .05 leads to a large variety in strength of evidence in favor of a medicine's efficacy and inconsistent decision making. Often, this criterion leads to

endorsement when statistical evidence actually favors the null hypothesis. In van Ravenzwaaij and Ioannidis (2017), we recommend routine consideration of Bayes factors in regulatory assessments and clinical decision-making, and here I will add that such a change can only happen by routinely teaching our (under)grad students about Bayesian methods for statistical inference.

REFERENCES

Chavalarias, D., Wallach, J. D., Li, A. H. T., & Ioannidis, J. P. A. (2016). Evolution of reporting p values in the biomedical literature, 1990-2015. *Journal of the American Medical Association, 315*, 1141-1148.

Food and Drug Administration (1998). *Guidance for industry: providing clinical evidence of effectiveness for human drug and biological products.* Maryland: United States Food and Drug Administration.

Food and Drug Administration (2018). *Celexa Approval Document.* Retrieved from https://www.accessdata.fda.gov/drugsatfda_docs/label/1998/20822lbl.pdf on 23 January 2018.

Goodman, S.N. (1999). Toward evidence-based medical statistics. 2: the Bayes factor. *Annals of Internal Medicine, 130,* 1005-1013.

Jeffreys, H. (1998). *Theory of probability. 3rd ed.* Oxford, UK: Oxford University Press.

Kass, R. E. & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*, 773-795.

Turner, E. H., Matthews, A. M., Linardatos, E., Tell, R. A., & Rosenthal, R. (2008). Selective publication of antidepressant trials and its influence on apparent efficacy. *New England Journal of Medicine, 358,* 252-260.

van Ravenzwaaij, D. & Ioannidis, J. P. A. (2017). A simulation study of the strength of evidence in the recommendation of medications based on two trials with statistically significant results. *PloS one. 2017*: e0173184.