Bayesian reanalysis of null results reported in medicine Strong yet variable evidence for the absence of treatment effects

Hoekstra, R.; Monden, R.; van Ravenzwaaij, D.; Wagenmakers, E.-J.

DOI
10.1371/journal.pone.0195474

Publication date
2018

Document Version
Final published version

Published in
PLoS ONE

PLOS ONE

RESEARCH ARTICLE

Bayesian reanalysis of null results reported in medicine: Strong yet variable evidence for the absence of treatment effects

Rink Hoekstra[1]*, Rei Monden[2☯], Don van Ravenzwaaij[1☯], Eric-Jan Wagenmakers[3☯]

1 University of Groningen, Groningen, The Netherlands, **2** University Medical Center Groningen, Groningen, The Netherlands, **3** University of Amsterdam, Amsterdam, The Netherlands

☯ These authors contributed equally to this work.
* r.hoekstra@rug.nl

Citation: Hoekstra R, Monden R, van Ravenzwaaij D, Wagenmakers E-J (2018) Bayesian reanalysis of null results reported in medicine: Strong yet variable evidence for the absence of treatment effects. PLoS ONE 13(4): e0195474. https://doi.org/10.1371/journal.pone.0195474

Editor: Xiang Li, Janssen Research and Development, UNITED STATES

Received: July 13, 2017

Accepted: March 25, 2018

Published: April 25, 2018

Data Availability Statement: All relevant data are available from the Open Science Framework: https://osf.io/ktyqj/.

Funding: The authors received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Efficient medical progress requires that we know when a treatment effect is absent. We considered all 207 Original Articles published in the 2015 volume of the *New England Journal of Medicine* and found that 45 (21.7%) reported a null result for at least one of the primary outcome measures. Unfortunately, standard statistical analyses are unable to quantify the degree to which these null results actually support the null hypothesis. Such quantification is possible, however, by conducting a Bayesian hypothesis test. Here we reanalyzed a subset of 43 null results from 36 articles using a default Bayesian test for contingency tables. This Bayesian reanalysis revealed that, on average, the reported null results provided strong evidence for the absence of an effect. However, the degree of this evidence is variable and cannot be reliably predicted from the *p*-value. For null results, sample size is a better (albeit imperfect) predictor for the strength of evidence in favor of the null hypothesis. Together, our findings suggest that (a) the reported null results generally correspond to strong evidence in favor of the null hypothesis; (b) a Bayesian hypothesis test can provide additional information to assist the interpretation of null results.

Introduction

Across the medical sciences, null results are of central importance. For both patients and doctors, it is crucial to know that a new treatment does not outperform the current gold standard; that a generic drug is just as effective as an expensive brand-name drug; and that a new surgical procedure does not improve survival rate. Knowing that effects are absent allows the profession to retain existing medical procedures and reallocate its limited resources to the exploration of novel treatments that are potentially effective.

In this manuscript we summarize and reanalyze the null results for primary outcome measures reported in the 2015 volume of the *New England Journal of Medicine* (NEJM). As detailed below, 45 out of 207 Original Articles (21.7%) featured at least one null result. Because of their prominence and impact in the medical literature, null results deserve a detailed and appropriate statistical treatment.

The statistical evaluation of medical hypotheses currently proceeds almost exclusively through the framework of *p*-value null-hypothesis significance testing (NHST) [1,2,3] for a general warning about the use of *p*-values in medical research see [4], and [5], p. 424). This statistical framework is particularly problematic for the assessment of null results, because non-significant *p*-values do not quantify evidence in favor of the null hypothesis (e.g., [6, 7, 8, 9, 10]. This important caveat was recently underscored in a report from the *American Statistical Association*: "a relatively large *p*-value does not imply evidence in favor of the null hypothesis" [11], p. 132.

In order to quantify the evidence in favor of the absence of a treatment effect, we adopt the framework of Bayesian statistics and compute the predictive performance of two competing hypotheses: the null hypothesis that states the effect to be absent and the alternative hypothesis that states the effect to be present (e.g., [12, 13]). The resulting balance of predictive performance is known as the *Bayes factor* (e.g., [2, 14, 15, 16, 17]). In contrast to NHST, the Bayes factor allows researchers to quantify evidence in favor of the null hypothesis. For instance, when the Bayes factor BF = 10, the observed data are 10 times more likely under the null hypothesis than under the alternative hypothesis; when BF = 1 the observed data are equally likely under both hypotheses (i.e., the data are perfectly ambiguous and do not favor one hypothesis over the other); and when BF = 1/10 the observed data are 10 times more likely under the alternative hypothesis than under the null hypothesis. Fig 1 highlights the different interpretations that the two statistical frameworks allow.

Computation of the Bayes factor requires the analyst to quantify the expectation about effect size under the alternative hypothesis. In contrast to a classical power analysis, this

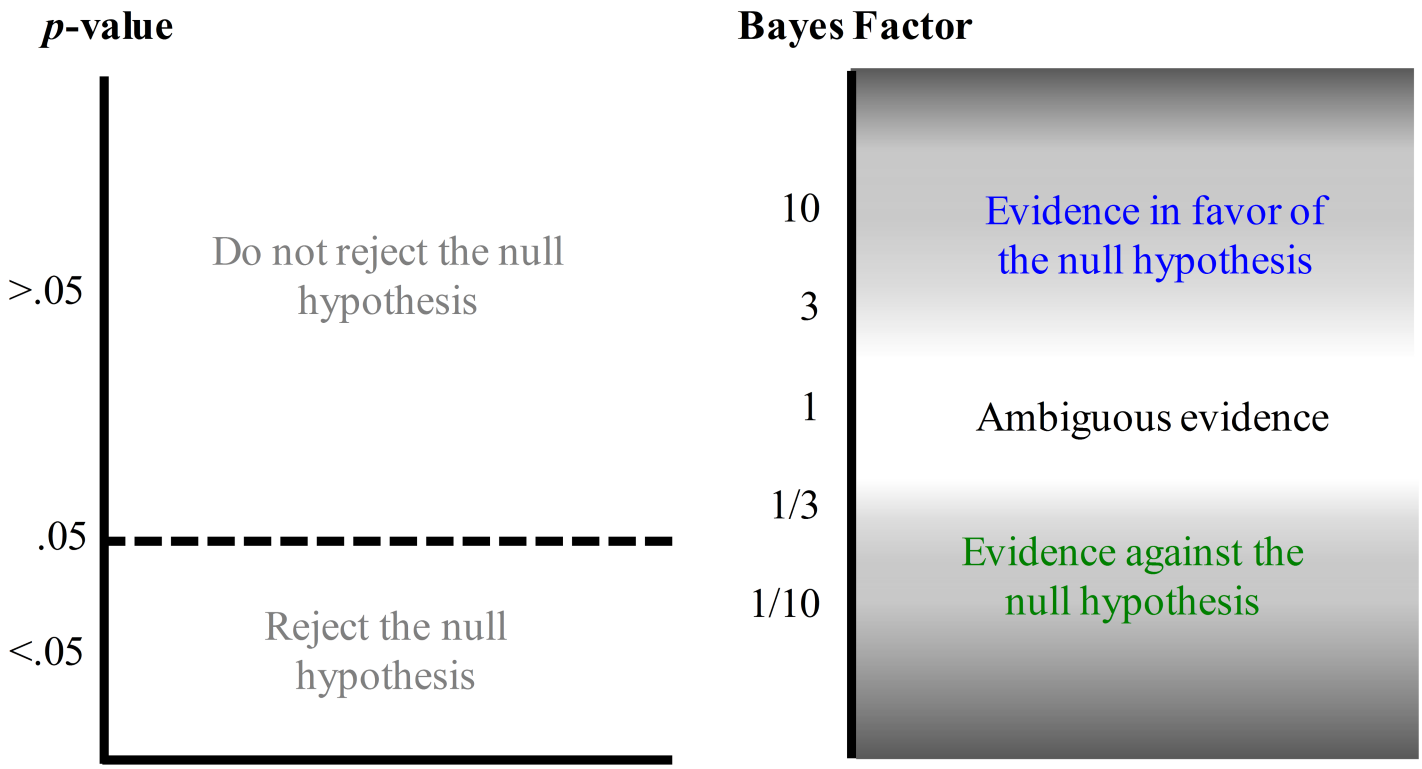

Fig 1. Valid statements based on p-values and Bayes factors. The *p*-value and the Bayes factor allow fundamentally different statements concerning the null hypothesis. The *p*-value can be used to make a discrete decision: reject or retain the null hypothesis. The Bayes factor grades the evidence that the data provide for and against the null hypothesis.

https://doi.org/10.1371/journal.pone.0195474.g001

expectation encompasses a range of different effect sizes, weighted by their prior plausibility. Here we adopt an "objective Bayesian approach" [18] and apply a default test that assigns expectations to effect sizes under the alternative hypothesis such that the test performs well across a wide range of substantively different applications [13].

The conceptual advantage of a Bayesian analysis can be underscored with a simple example. Consider the study by Jolly et al. ([19]; this study is part of the data presented later in this paper) who compared mortality rates of two groups of patients with ST-segment elevation myocardial infarction, one group receiving percutaneous coronary intervention (PCI) and thrombectomy and one group receiving PCI alone. The mortality rate in the PCI-plus-thrombectomy group was 347/5033 (6.89%) and the mortality rate in the PCI-only group was 351/5030 (6.98%). The standard chi-square test yields a p-value of .90 (.87 without application of Yates' continuity correction). Now consider another study by Carrier et al. ([20]; this study is also part of the data presented later in this paper), who examined whether the ability to detect occult cancer is improved by adding computed tomography (CT) to the limited screening practice. The miss rate in the screening-plus-CT group was 5/19 (26.32%) and the miss rate in the screening-only group was 4/14 (28.57%). The standard chi-square test yields a p-value of 1 (.89 without application of Yates' continuity correction).

Both studies fail to reject the null hypothesis and obtain a p-value of similar magnitude. Nevertheless, application of the default Bayes factor hypothesis test reveals that the data from the Carrier et al. study, despite the larger p-value, provide only weak support for the null hypothesis (i.e., BF = 2.7, implying that the data are not much more likely under the null hypothesis than under the alternative hypothesis) whereas the data from the Jolly et al. study provide compelling support in favor of the null hypothesis (i.e., BF = 77.7). This example illustrates how a Bayes factor hypothesis test can differentiate between data that are almost uninformative and data that are highly diagnostic (e.g., [6]).

In sum, the assessment of medical null effects may benefit from an additional Bayesian analysis. Below we explore the extent to which medical null results reported in the 2015 volume of NEJM actually yield compelling evidence in favor of the null hypothesis when assessed by a default Bayes factor hypothesis test.

Method

Sample

We considered all 207 Original Articles published in the 2015 volume of NEJM. This journal was chosen because of its prominence in the field of medicine, and because it publishes papers about a wide range of medical issues. An initial screening identified 45 articles whose abstract contained at least one claim about the absence or non-significance of an effect for a primary outcome measure (21.7%). To facilitate the analysis and the interpretation of the results, we selected a further subset 37 of articles that allowed a simple comparison between proportions, that is, m by k contingency tables. After eliminating one article whose results were significant upon reanalysis, we obtained a final sample of 36 articles. Several of these articles contained more than one claim of no effect, such that the total number of effects available for reanalysis equaled 43. A detailed description of the articles under consideration and the selection procedure is provided in the supplements.

Bayes factor reanalysis

Our Bayesian reanalysis was facilitated by the fact that in order to compute Bayes factors for contingency tables, knowledge of the individual cell counts suffices. Bayes factors were calculated using the default test for a comparison of proportions (e.g., [12, 21, 22, 13];) implemented

in the statistical software package JASP [23]. JASP is an open-source statistical software package with a graphical user interface, supporting both frequentist and Bayesian analyses (for details see jasp-stats.org). The underlying code base is in R ([24]), and our analyses can also be executed in the BayesFactor package [25], available at https://cran.r-project.org/web/packages/BayesFactor/index.html.

The Bayes factors reported here compare the predictive performance of the null-hypothesis (which assumes the absence of association between rows and columns) against predictive performance of the alternative hypothesis (which assumes the presence of an association). The default Bayes factor specifies that under the alternative hypothesis, every combination of values for the proportions is equally likely a priori. For example, in the case of the 2 by 2 table, the alternative hypothesis specifies two independent uniform distributions for the two rate parameters. In specific applications, such a default, reference-style analysis can be supplemented by substantive knowledge based on earlier experience. With a more informative prior distribution, the alternative hypothesis will make different predictions, and a comparison with the null hypothesis will therefore yield a different Bayes factor. The more informed the prior distribution, the more specific the model predictions, and the more risk the analyst is willing to take. Highly informed prior distributions need to be used with care, as they may exert a dominant effect on the posterior distribution, making it difficult to "recover" once the data suggest that the prior was ill-conceived. With informed prior distributions, it is wise to perform a robustness analysis to examine the extent to which different modeling choices lead to qualitatively different outcomes.

In this article, we prefer the default prior, as it is the most common choice and an informed specification would require an elaborate elicitation process from many different experts. We do not, therefore, view the outcomes of this analysis as definitive, although we believe that the qualitative results (i.e., strong but highly variable evidence in favor of the null) hold across a broad range of prior distributions.

Before proceeding it is important to point out that the Bayes factor quantifies the support provided by the data. For any two models, the posterior odds is obtained by multiplying the Bayes factors by the prior odds. In other words, the Bayes factor allows an assessment of the strength of evidence that is independent from the relative prior plausibility of the models.

Results

The main result concerns the default Bayes factors for the 43 null effects reported in the 2015 NEJM volume. Fig 2 shows the Bayes factors (on a logarithmic scale) as a function of the *p*-value. Reassuringly, all Bayes factors are higher than 1, indicating support in favor of the null hypothesis. For ease of interpretation, the right-hand y-axis displays the evidential categories proposed by Jeffreys ([15], Appendix B; the labels have been changed as suggested in [26]). Most Bayes factors provide strong or very strong evidence in favor of the null hypothesis, and only three are weak, anecdotal, or "not worth more than a bare mention" [13].

Fig 2 also shows that the degree of support in favor of the null hypothesis fluctuates considerably from one experiment to the next—the lowest Bayes factors are smaller than 3, whereas the most compelling Bayes factors exceed 100. Moreover, the observed *p*-value is a poor predictor of the Bayes factor. For instance, a *p*-value near 0.9 is almost equally likely to correspond to anecdotal, moderate, strong, or very strong evidence in favor of the null hypothesis. This underscores the usefulness of the Bayes factor over and above the *p*-value.

One possible determinant of the strength of the Bayes factor is sample size. Fig 3 shows the relation between sample size and Bayes factors (both plotted on a logarithmic scale) and confirms that in a small sample, a nonsignificant result is likely to be nondiagnostic (i.e., the

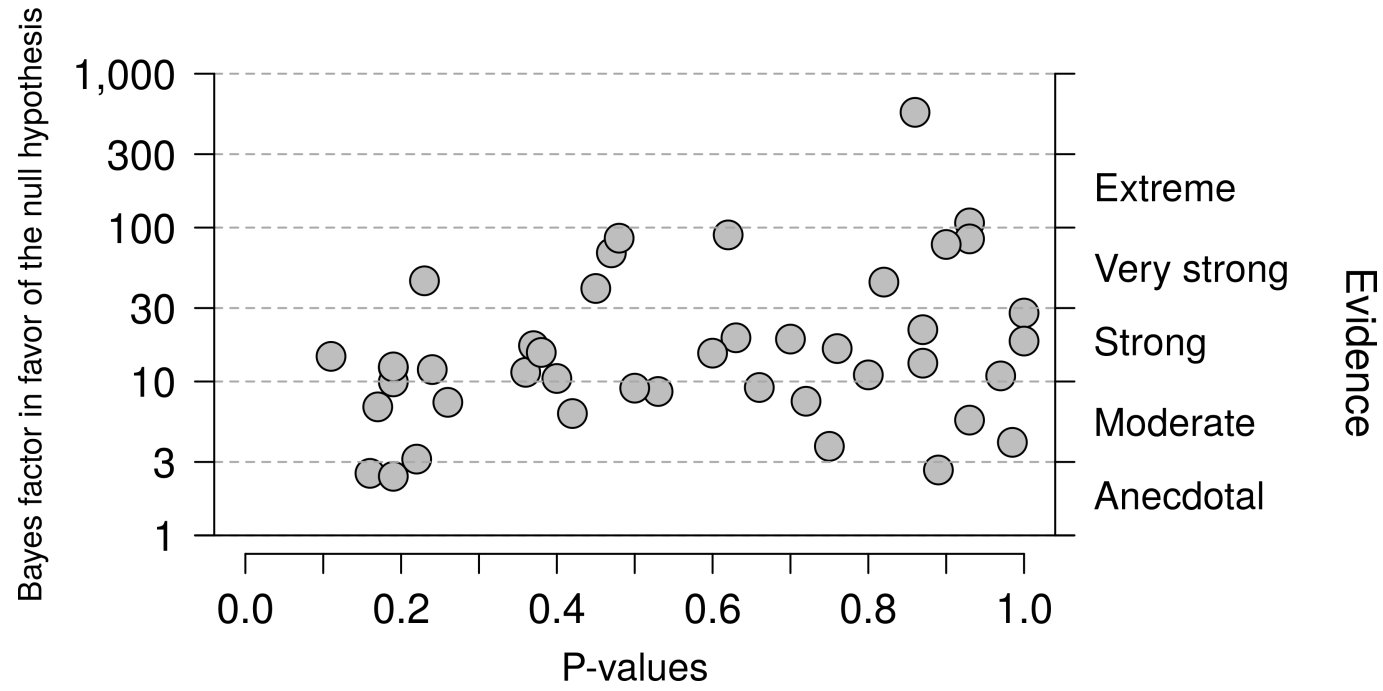

Fig 2. Relation between p-values and Bayes factors. *P*-values and Bayes factors in favor of the null hypothesis for 43 null results from the 2015 volume of NEJM. All Bayes factors indicate support in favor of the null hypothesis, and most Bayes factors do so in a compelling fashion. At the same time, the support in favor of the null hypothesis is highly variable. The *p*-value only explains 8.39% of the variance in the log Bayes factors (r = 0.29).

https://doi.org/10.1371/journal.pone.0195474.g002

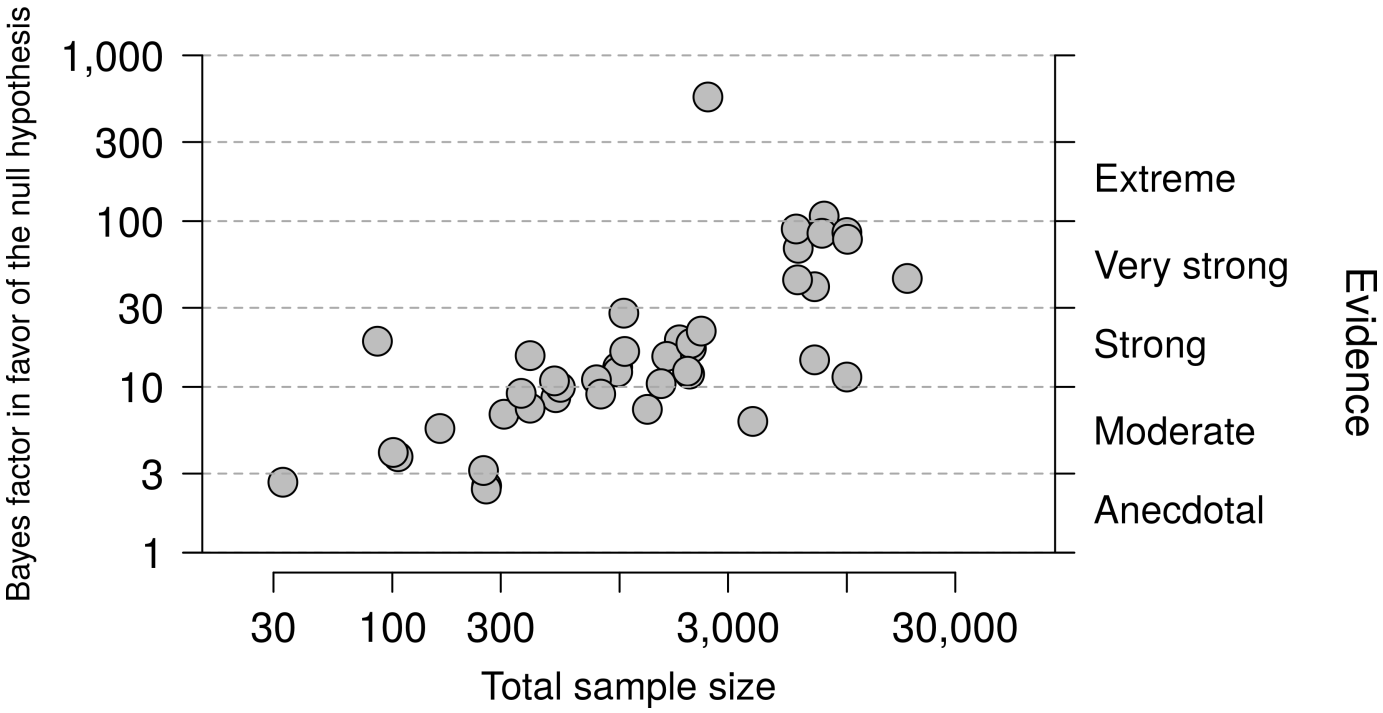

Fig 3. Relation between sample size and Bayes factors. Among the 43 null results from the 2015 volume of NEJM, large samples are more likely to yield compelling evidence in favor of the null hypothesis than small samples (r = 0.72).

https://doi.org/10.1371/journal.pone.0195474.g003

evidence in favor of the null hypothesis is likely to be weak), whereas in a large sample a non-significant result is likely to be diagnostic (i.e., the evidence in favor of the null hypothesis is likely to be strong). Nevertheless, sample size alone is not sufficient to gauge the evidence; for instance, even with a total sample size of 10,000 a nonsignificant result may lead to a Bayes factor slightly higher than 10 or a Bayes factor higher than 100.

Discussion

We applied a default Bayes factor reanalysis to 43 null results published in the 2015 volume of NEJM. Reassuringly, this reanalysis revealed that all null results supported the null hypothesis over the alternative hypothesis, and the overall degree of support was strong. Nevertheless, from experiment to experiment the degree of evidence varied considerably—the smallest Bayes factor was 2.42 ("not worth more than a bare mention", [15]), whereas the largest Bayes factor was 560.9 ("decisive" or "extreme" evidence, [15]; [26]). Our findings also suggest that for non-significant findings, the degree of evidence in favor of the null hypothesis cannot be predicted from the *p*-value, but can be predicted to some extent from sample size: larger samples sizes are more likely to produce compelling evidence.

Several remarks are in order. First, we were pleasantly surprised that as many as 21.7% of the studied papers reported a null result for at least one of the primary outcome measures, considering the strong confirmation bias that is present in the biomedical literature (e.g., [27]). Second, the outcome of our analysis is, naturally, model-dependent. In our comparison of proportions we adopted a default specification of the alternative hypothesis (e.g., [12,13]). We have made our data available online to encourage additional analyses. Third, we limited ourselves here to the assessment of evidence for what is arguably the most popular statistical analysis across the medical sciences: the comparison between two or more proportions. The canonical example is the comparison between the survival rate in a control group versus that in a treatment group. More complicated experimental designs require the application of different statistical models and associated Bayes factors (e.g., [28]).

Fourth, it may be possible to combine the different ingredients of classical statistics and NHST (e.g., power, confidence intervals, *p*-values, sample sizes, effect sizes) to obtain an intuitive and sensible assessment of the support that the data provide in favor of the null hypothesis. Nevertheless, this concept itself is antithetical to classical statistics. The father of modern classical statistics, Ronald Fisher, famously stated, "the null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation. Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis." ([29], p. 19). But in medical research, it is often of crucial importance to be able to quantify the support in favor of the absence of an effect. Rather than combining the various ingredients from classical statistics in order to overcome its inherent limitations, we propose that the desired inference can be attained more simply and more directly through the application of Bayesian statistics.

Fifth, one might argue that the point null hypothesis is never exactly true, and as a result its examination is pointless (i.e., a foregone conclusion [30]), whether from a frequentist of Bayesian standpoint. We disagree with this argument. The null hypothesis represents the idealized position of a skeptic, and claims that contradict this position will have a hard time being accepted by the scientific community when the data fail to discredit the skeptic's position. Moreover, the null hypothesis may often be exactly true. For instance, the effect of homeopathic drugs will be exactly equal to the effect of a placebo alternative, assuming that the study is executed with care. In general, drugs whose active components do not impinge on the relevant biological mechanism will be incapable of outperforming placebo. The most compelling

counterargument, in our opinion, is that the null hypothesis is merely a convenient approximation to the true state of nature, in which effects are so small that they cannot be meaningfully studied with realistic samples sizes. As Cornfield [31] put it, "For finite sized samples the probability of rejecting either H_0 or H_δ [with δ representing a very small effect—HMRW] will be nearly equal, and concern about the high probability of rejecting one is equivalent to concern about rejecting the other" (p. 582).

A final point concerns the difference between testing and estimation. Several researchers and institutions (e.g., [32,33,34]) have promoted the notion that confidence intervals should supplement or replace p-values. In practice, researchers who report confidence intervals often focus exclusively on whether or not the interval contains the value specified by the null hypothesis, thereby executing significance testing by stealth. Also, confidence intervals cannot be used to quantify the strength of evidence, which we expect is something that many researchers are interested in. Moreover, confidence intervals share many of the same interpretational problems with *p*-values (e.g., [35,36]), reducing the appeal of supplementing of replacing *p*-values with confidence intervals. In contrast, Bayes factors do not share the *p*-value problems; for instance, Bayes factors allow researchers to quantify the evidence in favor of the null hypothesis, which may be crucial for medical practitioners who want to know whether or not a particular treatment is effective. But Bayes factors are a tool for testing, and if one prefers estimation (for instance because there is only a single plausible hypothesis or model) the credible interval is a useful Bayesian alternative to the frequentist confidence interval.

In general, null results in medicine can have serious practical ramifications. The importance of medical null results is evident from the fact that in 2015, about 1 in every 5 papers we studied reported a null result for one of its primary outcome measures. For such null results, medical professionals need to be able to gauge the evidence in favor of the absence of an effect. Here we showed how this goal can be accomplished by the application of a default Bayes factor test. For many standard analyses, such default Bayes factor tests are now easy to apply ([23,25]). We recommend that researchers who report a null result consider the conclusions that follow from both a classical and a Bayesian perspective.

In sum, an assessment of all 207 Original Articles in the 2015 volume from NEJM revealed that 21.7% reported a null result for one or more of their primary outcome measures. A standard Bayesian reanalysis of 43 null results revealed that the evidence in favor of the null hypothesis was strong on average, but highly variable. Higher sample sizes generally produced stronger evidence. We suggest that by adopting a statistically inclusive approach, medical researchers confronted with a null result can issue a report that is more informative and more appropriate than the one that is currently the norm.

Supporting information

S1 File. Supplement of "Bayesian reanalysis of null results reported in the New England Journal of Medicine: Strong yet variable evidence for the absence of treatment effects. This is a supplement of "Bayesian Reanalysis of Null Results Reported in the New England Journal of Medicine: Strong yet Variable Evidence for the Absence of Treatment Effects" by Hoekstra R, Monden R, van Ravenzwaaij D and Wagenmaker EJ. This document was written by Rei Monden (November, 2016). These plots were generated based on the 43 test statistics reported in the New England Journal of Medicine 2015.
(PDF)

S2 File. All selected papers from the 2015 volume of the NEJM.
(PDF)

Author Contributions

Conceptualization: Rink Hoekstra, Rei Monden, Don van Ravenzwaaij, Eric-Jan Wagenmakers.

Data curation: Rink Hoekstra, Rei Monden.

Formal analysis: Rei Monden, Don van Ravenzwaaij, Eric-Jan Wagenmakers.

Investigation: Don van Ravenzwaaij, Eric-Jan Wagenmakers.

Methodology: Rink Hoekstra, Rei Monden, Don van Ravenzwaaij, Eric-Jan Wagenmakers.

Project administration: Rink Hoekstra.

Software: Don van Ravenzwaaij, Eric-Jan Wagenmakers.

Supervision: Eric-Jan Wagenmakers.

Validation: Rink Hoekstra, Don van Ravenzwaaij, Eric-Jan Wagenmakers.

Visualization: Rei Monden, Don van Ravenzwaaij.

Writing – original draft: Rink Hoekstra.

Writing – review & editing: Rink Hoekstra, Rei Monden, Don van Ravenzwaaij, Eric-Jan Wagenmakers.

References

1. Chavalarias D, Wallach JD, Li AHT, Ioannidis JPA. Evolution of reporting p values in the biomedical literature, 1990–2015. J Am Med Assoc, 2016: 1141–1148.
2. Goodman SN. Toward evidence-based medical statistics. 1: The P value fallacy. Ann Intern Med. 1999: 130; 995–1004. PMID: 10383371
3. Pocock SJ, Stone GW. The primary outcome fails—what next?. N. Engl. J. Med. 2016; 375: 861–870. https://doi.org/10.1056/NEJMra1510064 PMID: 27579636
4. Rennie D. Vive la difference (p< 0.05). N Engl J Med, 1978; 299: 828–829. https://doi.org/10.1056/NEJM197810122991509 PMID: 692566
5. International Committee of Medical Journal Editors. Uniform requirements for manuscripts submitted to biomedical journals. Reg Anesth Pain Med. 1991:16; 181–185.
6. Dienes Z. Using Bayes to get the most out of non-significant results. Front. Psychol.2016; 5: 781.
7. Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman S N, et al. Statistical tests, p values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol. 2016; 31: 337–350. https://doi.org/10.1007/s10654-016-0149-3 PMID: 27209009
8. Hoekstra R, Finch S, Kiers HAL, Johnson A Probability as certainty: Dichotomous thinking and the misuse of p values. Psychon Bull Rev. 2006; 13: 1033–1037. PMID: 17484431
9. Ioannidis JP. Why most published research findings are false. PLoS Med. 2005; 2: e124. https://doi.org/10.1371/journal.pmed.0020124 PMID: 16060722
10. Wagenmakers EJ. A practical solution to the pervasive problems of p values. Psychonom Bull Rev. 2007; 14: 779–804.
11. Wasserstein RL, Lazar NA. The ASA's statement on p-values: context, process, and purpose. Am Stat. 2016; 70: 129–133.
12. Gunel E, Dickey J. Bayes factors for independence in contingency tables. Biometrika, 1974; 61: 545–557.
13. Jeffreys H. Some tests of significance, treated by the theory of probability. In: Proc Camb Philos Soc. 1935; 31: 203–222.
14. Hobbs BP, Carlin BP. Practical Bayesian design and analysis for drug and device clinical trials. J Biopharm Stat. 2008; 18; 54–80. https://doi.org/10.1080/10543400701668266 PMID: 18161542
15. Jeffreys H. Theory of Probability. 3rd ed. Oxford: Oxford University Press; 1961.
16. Kass RE, Raftery AE. Bayes factors. J. Am. Stat. Assoc. 1995; 90: 773–795.

17. Zaslavsky BG. Bayesian hypothesis testing in two-arm trials with dichotomous outcomes. Biometrics, 2013; 69: 157–163. https://doi.org/10.1111/j.1541-0420.2012.01806.x PMID: 23002906
18. Berger J. The case for objective Bayesian analysis. Bayesian Analysis. 2004; 1: 1–17.
19. Jolly SS, Cairns JA, Yusuf S, Meeks B, Pogue J, Rokoss MJ et al. Randomized trial of primary PCI with or without routine manual thrombectomy. N. Engl. J. Med., 2015; 372: 1389–1398. https://doi.org/10.1056/NEJMoa1415098 PMID: 25853743
20. Carrier M, Lazo-Langner A, Shivakumar S, Tagalakis V, Zarychanski R, Solymoss S et al. Screening for occult cancer in unprovoked venous thromboembolism. N. Engl. J. Med. 2015; 373: 697–704. https://doi.org/10.1056/NEJMoa1506623 PMID: 26095467
21. Jamil T, Ly A, Morey RD, Love J, Marsman M, Wagenmakers E-J. Default "Gunel and Dickey" Bayes factors for contingency tables. Behav Res Methods. 2015: 1–15. https://doi.org/10.3758/s13428-014-0458-y PMID: 24683129
22. Jamil T, Marsman M, Ly A, Morey RD, Wagenmakers E-J. What are the odds? Modern relevance and Bayes factor solutions for MacAlister's problem from the 1881 Educational Times. Educational and Psychological Measurement. 2016, 0013164416667980.
23. JASP Team. JASP (Version 0.7.5.5) [Computer software]. 2016.
24. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2017. URL https://www.R-project.org/.
25. Morey RD and Rouder JN. Bayes Factor (Version 0.9.12–2) [computer software], 2016. URL https://cran.r-project.org/web/packages/BayesFactor/index.html.
26. Lee MD, Wagenmakers E-J. Bayesian Cognitive Modeling: A Practical Course. Cambridge: Cambridge University Press. 2014.
27. Chalmers I, Glasziou P. Avoidable waste in the production and reporting of research evidence. Obstetrics & Gynecology. 2009; 114: 1341–1345.
28. Rouder JN, Morey RD, Speckman PL, Province JM. Default Bayes factors for ANOVA designs. J Math Psychol. 2012; 56: 356–374.
29. Fisher RA. The design of experiments. Edinburgh: Oliver and Boyd. 1935.
30. Berkson J. Tests of significance considered as evidence. J Am Stat Assoc. 1942; 37: 325–335.
31. Cornfield J. A Bayesian test of some classical hypotheses—with applications to sequential clinical trials. J Am Stat Assoc. 1966; 61: 577–594.
32. American Psychological Association. Publication manual of the American Psychological Association (6th ed.). Washington, DC: Author. 2009.
33. Cumming G. Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis. Routledge. 2013.
34. International Committee of Medical Journal Editors. Uniform requirements for manuscripts submitted to biomedical journals. N Engl J Med 1997; 336: 309–316.
35. Hoekstra R, Morey RD, Rouder JN, Wagenmakers E-J. Robust misinterpretation of confidence intervals. Psychon Bull Rev. 2014; 21: 1157–1164. https://doi.org/10.3758/s13423-013-0572-3 PMID: 24420726
36. Morey RD, Hoekstra R, Rouder JN, Lee MD,Wagenmakers E-J. The fallacy of placing confidence in confidence intervals. Psychon Bull Rev. 2016; 23: 103–123. https://doi.org/10.3758/s13423-015-0947-8 PMID: 26450628