2-28-2007

ASSESSING THE UNRELIABILITY OF THE MEDICAL LITERATURE: A RESPONSE TO "WHY MOST PUBLISHED RESEARCH FINDINGS ARE FALSE"

Steven Goodman
The Johns Hopkins University School of Medicine, Department of Biostatistics and Oncology, sgoodman@jhmi.edu

Sander Greenland
University of California, Los Angeles, Departments of Epidemiology and Statistics

Assessing the unreliability of the medical literature:

A response to

"Why Most Published Research Findings are False."

Steven Goodman, MD, PhD
Johns Hopkins University
sgoodman@jhmi.edu

Sander Greenland, MA, MS, DrPH
Departments of Epidemiology and Statistics
University of California, Los Angeles

Word count: ca. 3600

Abstract. A recent article in this journal (Ioannidis JP (2005) Why most published research findings are false. PLoS Med 2: e124) argued that more than half of published research findings in the medical literature are false. In this commentary, we examine the structure of that argument, and show that it has three basic components:

1) An assumption that the prior probability of most hypotheses explored in medical research is below 50%.

2) Dichotomization of *P*-values at the 0.05 level and introduction of a "bias" factor (produced by significance-seeking), the combination of which severely weakens the evidence provided by every design.

3) Use of Bayes theorem to show that, in the face of weak evidence, hypotheses with low prior probabilities cannot have posterior probabilities over 50%.

Thus, the claim is based on a priori assumptions that most tested hypotheses are likely to be false, and then the inferential model used makes it impossible for evidence from any study to overcome this handicap. We focus largely on step (2), explaining how the combination of dichotomization and "bias" dilutes experimental evidence, and showing how this dilution leads inevitably to the stated conclusion. We also demonstrate a fallacy in another important component of the argument –that papers in "hot" fields are more likely to produce false findings.

We agree with the paper's conclusions and recommendations that many medical research findings are less definitive than readers suspect, that *P*-values are widely misinterpreted, that bias of various forms is widespread, that multiple approaches are needed to prevent the literature from being systematically biased and the need for more data on the prevalence of false claims. But calculating the unreliability of the medical

research literature, in whole or in part, requires more empirical evidence and different inferential models than were used. The claim that "most research findings are false for most research designs and for most fields" must be considered as yet unproven.

Introduction

An article recently published in the journal PLoS Medicine encouraging skepticism about biomedical research findings [1] is a welcome reminder that statistical significance is only a small part of the story when attempting to determine the probability that a given finding or claim is true. The prior probability of a hypothesis is a critical determinant of its probability after observing a study result, which the *P*-value does not reflect. This point has been made repeatedly by statisticians, epidemiologists and clinical researchers for at least 60 years [2-9], but is still underappreciated. The new world of gene hunting and biomarkers, in which thousands of relationships are examined at once, is fertile ground for false-positive claims, so prior probabilities play an even more important role today[10-12].

We agree with the qualitative point that there are more false findings in the medical literature than many suspect, even in areas outside genomics. Nonetheless, we will show that the declaration that "it can be proved that most research findings are false" is not confirmed by the argument presented. We will show how the model used to make that claim has a structure that guaranteed conclusions that differed little from its starting assumptions. It did so by treating all "significant" results as though they were mathematically the same - whether P=0.0001 or P=0.04, and then employing a mathematical model for the effect of "bias" – defined overbroadly - that further weakened the evidential value of every study in the literature. Finally, we show a fallacy in the argument that the more scientists publishing in a field, the more likely individual findings are to be false.

Bayes Theorem

We start with the mathematical foundation of the analysis, Bayes theorem. This theorem was presented in terms of the pre-study odds and the "positive predictive value" of a finding (the post-study probability after a "positive" result). We will instead use the standard terminology of the "prior probability" of a hypothesis (its probability before the study), and the "posterior probability" (its probability after the study). The "positive predictive value" language used in [1] is applicable only if a study is deemed to have a "positive" result, like a diagnostic test. The Bayesian approach does not necessarily employ study "positivity" as an input, so the "positive predictive value" language is not generally applicable. We consider the situation in which there is a null hypothesis of no effect, competing against a qualitative alternative hypothesis that there is an effect (perhaps in a particular direction).

Figure 1 displays Bayes theorem in its "odds" form, which shows that the posterior odds of an alternative (non-null) hypothesis equals the prior odds of the hypothesis multiplied by the *Bayes factor*. In addition to being computationally simple, this form is conceptually clear, separating the prior evidence for a hypothesis (reflected in its prior odds) from the evidence provided by the study and its assumptions, represented by the Bayes factor. The Bayes factor compares how well two competing hypotheses predict the observed data, under the study assumptions. It also equals the posterior odds of a hypothesis divided by the prior odds of the hypothesis, thereby showing how strongly the study's evidence affects the plausibility of the hypotheses being tested. In the diagnostic test setting, the Bayes factor corresponds to the likelihood ratio, i.e. how much a given test result changes the odds of disease. Commonly, Bayes factors over 20 are interpreted

strong evidence for a hypothesis, those of 3-5 as weak evidence, and intermediate values as "moderate" evidence.

The role of "bias"

The original paper (1) introduced the term "bias" into the Bayesian calculations. Bias was defined as the probability of reporting of a statistically significant result when the result of the primary, pre-specified analysis was in fact not significant, or should not have been. This definition encompasses poor design, improper analyses, selective misreporting, changing primary endpoints, and outright fraud, all with the effect or aim of producing statistical significance. This is an interesting formulation, but it is important to clearly distinguish it from the traditional statistical definition of bias as a systematic error in an estimate [13], in which statistical significance plays no role.

There is little doubt that "significance questing" practices exist [14]. The question is how pervasive they are, and whether all practices subsumed under the proposed definition have the same detrimental effect. We will explore first the logic of the assumption that all of these forms of significance-seeking are inferentially equivalent. We will then show that the numerical values suggested for this "bias" have a very profound effect on the final claim.

The Global Null Hypothesis

A simple example shows that that the various practices subsumed under the "bias" umbrella are not equal in effect. Suppose that a cardiac drug is tested in an RCT, with a primary endpoint of cardiac mortality and a secondary endpoint of cardiac function. The results are that the mortality endpoint is nonsignificant (P=0.20), but the cardiac function endpoint is highly significant (P=0.001). It is quite possible, and fairly common, that an

investigator might report the trial as "positive" based on the drug's effect on cardiac function, downplaying the mortality result. A second possibility is that the study report mentions only the cardiac function outcome, omitting mention of the mortality result. Finally, a third scenario is that the results are manipulated in such a way as to report the cardiac mortality endpoint as being "positive," P=0.01.

It should be apparent that the last scenario is more pernicious than the first two. It is also probably much rarer, particularly in collaborative studies. These forms of "bias" can be treated identically only if one believes in a single "global null hypothesis," i.e. that the treatment has no effect on *any* outcome, so *any* claimed relationship contravenes the global null. [15] The alternative perspective is that this study is exploring two different hypotheses, i.e. one that the treatment affects mortality, and another that it affects cardiac function.

In the first case of selective emphasis, all the evidence is reported, and a reader could make an independent judgment. The conclusion that the drug improves cardiac function but not mortality could be quite reasonable regardless of which was prespecified as primary. The evidence suppression in the second scenario (also known as selective reporting), would not misrepresent the evidence regarding cardiac function, but it might invite unwarranted extension to an effect on mortality, and it would distort meta-analytic estimates of the mortality effect. In the final scenario, the data manipulation would lead to the erroneous conclusion that there was strong evidence that the drug reduced cardiac death. This example shows that the various practices comprising the proposed definition of "bias" are not equally harmful, and should not be modeled as such. The model used in reference 1 is most appropriate for cases of fraud or manipulation, where empirical

evidence about prevalence is scarce. It should not be applied to selective emphasis or reporting – where empirical evidence is more extensive - unless one subscribes to the global null hypothesis.

One situation not covered by those reporting scenarios is where poor design leads to a distortion in the estimates. This could be similar in effect to fraudulent reporting, since the evidence for the relationship would be misrepresented. But the relationship between significance and distorted estimates is complex, since estimates can be biased in any direction and poor design also affects the variability of those estimates, requiring more complex models.

It is important not to impose an inferential "bias" penalty when the solution may be to adjust the prior probability. For example, in the case of exploratory analyses ("data mining"), one reason for the great concern about multiple testing is that an effect of any particular gene or exposure is unlikely, i.e. each effect has a low prior probability. A Bayesian analysis straightforwardly assigns a very low prior probability to each single predictor, and may even assign a collective low prior probability to finding any predictor. This is exactly what the paper in question [1] does in assigning low probabilities to predictors in exploratory studies, and exceedingly low ones in studies involving genomic scans. To assign "bias" to those studies is to double the penalty for the same inferential transgression, unless some other misdemeanor is being alleged. We will see in the next section how severe the bias penalty is.

The effect of bias

In reference 1, the inferential impacts of different degrees of bias are explored, from 10% to 80%. We will start with the example provided of a study with 30% bias; a

"confirmatory meta-analysis of good quality RCTs" or an "adequately powered, exploratory epidemiologic study" (1) (Table 1). Under the paper's bias definition, if pre-specified analyses were not significant at the 0.05 level (i.e., if $P \geq 0.05$), the researchers would nonetheless find a way to report statistical significance for a claimed primary outcome, or the design would produce it, 30% of the time. Without judging how likely that is for these designs, we will examine the effect of this bias on the evidence, and then on the inference.

The 30% bias increases the false positive error rate from 5% to 34%, an almost 7-fold increase (Appendix 1). This has the effect of decreasing the Bayes factor from 16 to only 2.9 for the meta-analysis and to 2.6 for the epidemiologic study, both exceedingly low numbers. Thus, in Table 1, neither of these designs is able to raise the prior probability very much; from 67% to 85% for the meta-analysis, and from 9% to 20% for the epidemiologic study. Without the 30% bias factor, the probability of the meta-analytic conclusion being true could rise from 67% to a posterior probability of 97%, and the probability of an exploratory hypothesis could rise from a 9% to a posterior of 62%. The assigned bias factor places an extreme impediment on the ability of studies to provide much evidence of anything.

Even the bias of 10%, assigned to "large, well designed RCTs, with minimal bias," plays a large role. The effect of the 10% bias is far from minimal; it translates into a near-tripling of a 5% false-positive rate, and significantly reduces the evidence supplied by $P \leq 0.05$, decreasing the Bayes factor from 16 to 5.7 (Appendix 1).

The above argument is not meant to deny or minimize the effect significance-seeking or significance-producing biases, or to provide alternative claims about the probabilities

of hypotheses examined in these designs. But it should make clear that even seemingly small degrees of this bias modeled in this way have profound effects, and are key determinants of the conclusion of the paper. The two most prevalent forms of significance-seeking bias should not be modeled in this way, and if the falsity of half the published literature is claimed to be "proven," then the specific numbers to assign to the different types of this bias require fairly definitive empirical confirmation.

Treating all significant *P*-values as inferentially equal

The effect of bias on diminishing the impact of experimental evidence, dramatic enough as it is alone, is greatly amplified by a mathematical model that treats all "significant findings" ($P \leq 0.05$) as inferentially equivalent. In the paper's calculations, a result with "$P=0.0001$" has the same effect on the posterior probability as "$P=0.04$"; both are treated as findings significant at $P \leq 0.05$. Leading biomedical journals and reporting guidelines (e.g., the CONSORT statement [16]) require that *P*-values be reported exactly (e.g., as "$P=0.01$," not as "$P \leq 0.05$.") because *P*-values of different magnitudes have different evidential meaning. Although the article says that different thresholds of significance can and should be used, its proof that more than half of published findings are false is based on the $P \leq 0.05$ convention, because that is the convention used in the literature. But the problem we now discuss would actually apply for any significance threshold, in the presence of bias.

Using Bayes theorem, we will illustrate the impact of using "significance" instead of the exact statistical results to calculate the inferential impact of a study; for simplicity of calculation we will ignore the "bias," and calculate the numbers for a study with 80% power.

Table 2 shows the difference in Bayes factors derived from a *P*-values expressed as inequalities (e.g., $P \leq 0.05$) or exactly (e.g., $P=0.05$). A *P*-value of exactly 0.05 increases the odds of this alternative hypothesis only 2.7-fold, whereas "$P \leq 0.05$" increases the odds 16-fold. $P=0.001$ has a 99-fold effect and "$P \leq 0.001$" increases the odds 274 times. As the *P*-value gets smaller the Bayes factor increases without limit (Appendix 2). Yet treating all "findings" as mathematically equivalent caps the Bayes factor at 16. Furthermore, if we include the "bias" factor, even if we lower the significance threshold to near zero (so significance corresponds to extremely strong evidence), the maximum achievable Bayes Factor is still very low (Appendix 1).

Table 2 shows the effect of these Bayes factors on the probability of hypotheses. Consider an unlikely hypothesis, with a prior probability of only 1%. Under the listed assumptions, obtaining a *P*-value of exactly 0.05 would move this 1% prior probability up to only a 2.6% posterior probability, $P \leq 0.05$ would raise it to 14%, $P=0.001$ would raise it to 50%, and "$P<0.001$" would raise it to 73% - a quite respectable level. Even given that these posterior probabilities represent maximum values without consideration of weaknesses in study design, this does show that an implausible alternative hypothesis can be made plausible if the evidence is strong enough. But that cannot happen if all *P*-values below a conventional threshold like 0.05 are treated as evidentially equal.

The above facts show that the claim that half the medical literature is false depends in part on the distribution of *P*-values. Exact *P*-values in the 0.01-0.05 range provide moderate evidence, in the same range as "significance", and are not sufficient to raise hypothesis with prior probabilities less than around 5%-10% to have posterior probabilities over 50%. Conversely, extremely low *P*-values, such as generated in some

genomic explorations, can be highly unstable and hence unreliable. Some empirical evidence has been generated on the distributions of *P*-values in certain fields (e.g. [17]). We agree with Dr. Ioannidis that more empirical evidence of *P*-value distributions and bases for prior probabilities would help address this issue.

The effect of multiple independent groups

A prominent claim in the paper is that "The hotter a scientific field (with more scientific teams involved), the less likely the research findings are to be true." This is described as a "paradoxical" finding. It is indeed paradoxical, since if it were true, the Bayesian argument that preceded it would be invalid; one would have to insert the number of other studies into the formulas to come up with the correct positive predictive value.

The calculations presented to justify the above claim actually do not support it; they show the probability of obtaining one or more positive studies out of a total of N studies, for a given Type I and Type II error. They demonstrate, correctly, that as the total number of studies increases, observing at least one false positive study becomes more likely. The error underlying the claim is that the positive predictive value of a single study (as labeled in the graphs and described in the text) is not what is calculated; rather, it is inferential impact of *all N studies* when we are told only that "one or more were positive". The impact of that data decreases with increasing N because the probability of at least one positive study becomes highly likely whether or not the underlying hypothesis is true. Thus, the curves and equations are for a meta-analytic result, albeit one in which the information in the N studies is reduced to a minimum, i.e. whether any studies were positive.

As the total number of studies increases, there will be more false positive and false negative studies, so it is literally true that there will be "more false findings." But this is not true as a proportion of the total, which is the probability that a given finding is false, i.e. the predictive value. If the number of positive studies is held constant while the total increases, the predictive value of all studies combined decreases, albeit not because the positive predictive value of any positive study decreases, but because the *negative* predictive value of all the *nonsignificant* studies outweighs the positive predictive value of the significant ones. So the positive predictive value of a single significant study, taken alone, with a given prior odds, remains exactly the same no matter how "hot" a field is.

We do not address here those issues mentioned that do not appear in the mathematical models, e.g. publication bias and prejudices of investigators. Although the above explanation refutes the claim that individual positive studies in hot fields are somehow less trustworthy than in other areas, it supports the more important contention in the paper that inferences should never be based on isolated studies, but instead should be based on summaries of all studies in a field addressing the same question.

Discussion

We have shown that the inferential model employed in reference 1 generates the claim made in the title through a model that effectively dilutes evidence in two stages. The first stage does not utilize the evidence represented by the actual *P*-value, but categorizes it as above or below 0.05 (or whatever Type I error is chosen). For a study with 80% power, this reduces the maximum attainable Bayes factor from infinity to 16.

Even if the entire world cooperated in a perfect clinical trial with 100% power, the maximum achievable Bayes factor would be only 20 (= 100%/5%). The introduction of the "bias" term reduces the potential 16-fold change in odds to 5.7 (for 10% bias), 2.6 (for 30% bias) or 1.1 (for 80% bias) (Appendix 1). The definition of bias and the modeling of its effect does not distinguish between misrepresenting the evidence (improper analyses or poor design), suppressing evidence related to different hypotheses (selective reporting) or implicitly misrepresenting the prior probability of a hypothesis (selective emphasis).

In summary, under the conceptual and mathematical model used in reference 1, no result can have much impact on the odds of scientific hypotheses, making it impossible for data of any type from any study to either surprise or to convince us. That is the real claim of the paper. This property explains why the only designs in the paper that give the alternative hypothesis 50% or greater *posterior* probability (meta-analyses of good RCTs and large, well-designed RCTs) are those for which the paper assigns 50% or greater *prior* probability (Table 1). So the conclusion that "most research findings are false for most research designs and for most fields"[18] is based on a combination of *a priori* assumptions that most tested hypotheses are likely to be false with a model that makes it virtually impossible for evidence from any study to overcome this handicap. Finally, the claim that the more scientists publishing research in an area decreases the believability of individual results is unfounded.

The above criticisms of the claimed proof nothwithstanding, we agree with the bulk of the methodologic recommendations that flow from this paper's arguments, and those of many others; the credibility of findings should not be judged from *P*-values alone;

initially implausible hypotheses are not rendered plausible by significance alone; searching to achieve or avoid significance can do a fair bit of damage to the validity of findings; findings in fields looking for small effects (particularly with observational designs) are more tenuous, as are findings from small studies; and we should do all we can to minimize all forms of bias. Finally, when assessing a hypothesis, it is important to examine the entire collection of relevant studies instead of focusing on individual study results.

These issues underscore the importance of efforts and methods, many recommended in the paper, to ensure that investigators do what they planned, and properly report what they did. These efforts include reproducible research and data sharing[19], trial registration[20,21], analyses of publication bias [22,23] and selective reporting [24], conflict of interest [25], and recognition of cognitive biases [26]. Improvement of analysis methods beyond ordinary significance testing is also overdue. Methods for large-scale data exploration [27,28], Bayesian analysis [9,29,30], and bias modeling [31,32] could help us better represent uncertainties and hence properly temper conclusions, particularly from observational studies. Finally, further careful studies of those cases in which seemingly strong findings were later refuted [33,34], an area to which Dr. Ioannidis has contributed, will be critical in informing these kinds of discussions.

We do not believe a valid estimate of the prevalence of false published findings in the entire medical literature can be calculated with the model suggested or from the currently available empirical evidence. That question is best addressed at much smaller-scale levels, which the paper suggests as well; at the level of the field, disease, mechanism, or question. The domains in which the prevalence of false claims are

probably highest are those in which disease mechanisms are poorly understood, data mining is extensive, designs are weak or conflicts of interest rife. More empirical study of the quantitative effects of those factors are badly needed. But we must be very careful to avoid generalizations beyond the limitations of our data and our models, lest in our collective effort to strengthen science through constructive criticism we undermine confidence in the research enterprise, adversely affecting researchers, the public that supports them, and the patients we ultimately serve.

Figure 1: Bayes theorem

$$\underbrace{\text{Odds of Alternative hypothesis after seeing data}}_{\textit{Posterior Odds}} = \underbrace{\left(\text{Odds of Alternative hypothesis before seeing data}\right)}_{\textit{Prior Odds}} \times \underbrace{\left(\frac{\text{Prob(Data | Alternative hypothesis)}}{\text{Prob(Data | Null hypothesis)}}\right)}_{\textit{Bayes Factor}}$$

$$\text{Probability} = \frac{\text{Odds}}{1+\text{Odds}}, \qquad \text{Odds} = \frac{\text{Probability}}{1-\text{Probability}}$$

Null hypothesis: "There is no effect"

Alternative (causal) hypothesis: "There is an effect"

Table 1: Table 4 from Ioannidis [1] showing the effect of being given that $P \leq 0.05$ ("significance") on the probability of a hypothesis, for a study with the given power and "bias" factor. The final column is a new addition, showing the posterior probability if there were no bias factor. (see Appendix 1)

From [1]					*New*
Practical Example	***Power***	***Bias***	***Prior probability***	***Posterior probability with bias***	***Posterior probability without bias***
Adequately powered RCT with little bias and 1:1 pre-study odds	0.80	0.10	50%	85%	94%
Confirmatory meta-analysis of good quality RCTs	0.95	0.30	67%	85%	97%
Meta-analysis of small inconclusive studies	0.80	0.40	25%	41%	84%
Underpowered, but well-performed phase I/II RCT	0.20	0.20	16.7%	23%	45%
Underpowered, poorly performed phase I/II RCT	0.20	0.80	16.7%	17.2%	45%
Adequately powered exploratory epidemiological study	0.80	0.30	9%	20%	66%
Underpowered exploratory epidemiological study	0.20	0.30	9%	12%	33%
Discovery-oriented exploratory research with massive testing	0.20	0.80	0.1%	0.1%	0.4%
As in previous example, but with more limited bias (more standardized)	0.20	0.20	0.1%	0.15%	0.4%

Table 2: Posterior probability of the alternative (non-null) hypothesis, (defined as the hypothesis for which the study has 80% power), as a function of the observed *P*-value, the Bayes factor, and the prior probability (1%, 20% and 50%). (For simplicity of calculation, the hypothesis and *P*-values considered in this table are one-sided, and normality of the test statistic is assumed.)

P-value	Bayes factor	Prior = 1%	Prior = 20%	Prior = 50%
***P*=0.05**	2.7	3%	40%	73%
"*P*≤0.05"	16	14%	80%	94%
***P*=0.01**	15	13%	78%	94%
"*P*≤0.01"	57	36%	93%	98%
***P*=0.001**	99	50%	96%	99%
"*P*≤0.001"	274	73%	98.6%	99.6%

Appendix 1: The effect of bias on the error rates and Bayes factor

Calculation of Bayes factor

The Bayes factor is the ratio of the probability of the data under the alternative hypothesis to that under the null hypothesis. For a result reported as "significant at level α", the probability of obtaining significance under the alternative hypothesis (H_A) is the probability of a true positive (= 1-β) plus the probability of a positive generated by "bias," (= β x bias) i.e. when the original result was not significant. Conversely, the probability of significance under the null hypothesis is the chance of a false positive generated by chance (α) plus the probability of a false positive generated by bias (= bias x (1-α)). This is represented mathematically below.

Equation 1

$$BF(\mathrm{H}_A \text{ vs. } \mathrm{H}_o \mid P \check{S}\ \alpha, \beta, \text{bias}) = \frac{\text{Probability of significance under } \mathrm{H}_A}{\text{Probability of significance under } \mathrm{H}_0}$$

$$= \frac{\Pr(\text{False neg}) \text{ x } bias + \Pr(\text{True pos})}{\Pr(\text{True neg}) \text{ x } bias + \Pr(\text{False pos})} = \frac{\beta \times bias + (1-\beta)}{(1-\alpha) \times bias + \alpha}$$

Where:

BF = Bayes Factor

H_0 = Null hypothesis

H_a = Alternative hypothesis

α = Two-sided Type I error probability

β = One-sided Type II error probability

Power = 1-β

Bias=30%, Power=80% and α=5% (assigned to an exploratory epidemiologic study):

$$\text{Bayes factor} = \frac{0.20 \times 30\% + 0.80}{0.95 \times 30\% + 0.05} = \frac{0.86}{0.335} = 2.6$$

Bias=30%, Power=95% and α=5% (assigned to an confirmatory meta-analysis of RCTs):

$$\text{Bayes factor} = \frac{0.05 \times 30\% + 0.95}{0.95 \times 30\% + 0.05} = \frac{0.965}{0.335} = 2.9$$

Bias=10%, Power=80%, α=5% (assigned to good RCT with minimal bias):

$$\text{Bayes factor} = \frac{0.20 \times 10\% + 0.80}{0.95 \times 10\% + 0.05} = \frac{0.82}{0.145} = 5.7$$

Bias=80%, Power=20%, α=5% (assigned to exploratory study with massive testing):

$$\text{Bayes factor} = \frac{0.80 \times 80\% + 0.20}{0.95 \times 80\% + 0.05} = \frac{0.84}{0.76} = 1.1$$

Bias=80%, Power=80%, α=5%:

$$\text{Bayes factor} = \frac{0.20 \times 80\% + 0.80}{0.95 \times 80\% + 0.05} = \frac{0.96}{0.81} = 1.2$$

Calculation of the maximum Bayes factor as the significance threshold decreases

As $\alpha \rightarrow 0$ in Equation 1**:**

$$\mathbf{BF_{max}} \rightarrow \frac{\beta \times bias + (1-\beta)}{(1-0) \times bias + 0} = \beta + \frac{(1-\beta)}{bias}$$

Using the example of the confirmatory meta-analysis of RCTs (Bias = 30%, and Power=95%), the maximum achievable BF would be:

BF_{max} = 5% + 95%/30% = 3.2

Appendix 2: Bayesian calculations for precise and imprecise *P*-values

The assumptions and calculations that generate the numbers reported in the Table 2 are as follows.

Type I error (α) = 0.05, one-sided

Type II error (β) = 0.20, one-sided

H_o: Null hypothesis, $\mu=0$.

H_A: Alternative hypothesis, $\mu=\mu_{0.80}$,

where μ_{80} is the value of the parameter of interest for which the experiment has 80% power when the one-sided $\alpha=0.05$. The Z-score corresponding to this alternative hypothesis $Z_{\mu 80}= Z_{0.05}+Z_{0.20}=1.644 + 0.842 = 2.49$.

$\Phi(Z)$ is the area under the Gaussian (normal) curve to the left of Z.

Z_α=the Z-score at the α percentile of the Gaussian distribution, i.e. $\Phi(-|Z_\alpha|) = \alpha$.

Bayes factor for imprecise *P*-value, expressed as an inequality $P\leq\alpha$:

If $P\leq\alpha$, the Bayes factor is the area under the alternative distribution curve beyond Z_α, divided by the area under the null distribution beyond the same point (the latter area being α, by definition, for a one-sided test). For one-sided $\alpha = 0.05$, power = 0.80,

$$\text{BF}(H_A \text{ vs. } H_0 \mid P \leq \alpha) = \frac{1-\Phi(Z_\alpha - 2.49)}{\alpha}$$

Bayes factor for a precisely reported *P*-value:

If $P=x$, the Bayes factor is the height of the curve corresponding to the alternative distribution at the point of the observed data, Z_x, divided by the height of the null distribution at the same point. For one-sided $\alpha = 0.05$, power = 0.80,

$$\text{BF}(H_A \text{ vs. } H_0 \mid P=x) = \frac{\phi(2.49 - Z_x)}{\phi(Z_x)} = \frac{\exp(-(2.49 - Z_x)^2 / 2)}{\exp(-Z_x^2 / 2)} = e^{3.1 - 2.49 Z_x}$$

References

1. Ioannidis JP (2005) Why most published research findings are false. PLoS Med 2: e124.

2. Jeffreys, H. (1939) Theory of Probability. Oxford: Oxford University Press.

3. Edwards W, Lindman H, Savage LJ (1963) Bayesian statistical inference for psychological research. Psych Rev 70: 193-242.

4. Diamond GA, Forrester JS (1983) Clinical trials and statistical verdicts: Probable grounds for appeal. Ann Intern Med 98: 385-394.

5. Berger JO, Sellke T (1987) Testing a Point Null Hypothesis: the Irreconcilability of P-values and Evidence. JASA 82: 112-122.

6. Brophy JM, Joseph L (1995) Placing trials in context using Bayesian analysis. GUSTO revisited by Reverend Bayes. JAMA 273: 871-875.

7. Lilford RJ, Braunholtz D (1996) For Debate: The statistical basis of public policy: a paradigm shift is overdue. BMJ 313: 603-607.

8. Goodman SN, Royall R (1988) Evidence and Scientific Research. AJPH 78: 1568-1574.

9. Greenland S (2006) Bayesian perspectives for epidemiological research: I. Foundations and basic methods. Int J Epidemiol 35: 765-775.

10. Todd JA (2006) Statistical false positive or true disease pathway? Nat Genet 38: 731-733.

11. Wang WY, Barratt BJ, Clayton DG, Todd JA (2005) Genome-wide association studies: theoretical and practical concerns. Nat Rev Genet 6: 109-118.

12. Kohane IS, Masys DR, Altman RB (2006) The incidentalome: a threat to genomic medicine. JAMA 296: 212-215.

13. Rothman, K., and Greenland, S. (1998) Modern Epidemiology. Philadelphia: Lippincott-Raven.

14. Rothman K (1986) Significance questing. Ann Int Med 105: 445-447.

15. Rothman KJ (1990) No adjustments are needed for multiple comparisons. Epidemiol 1, 1: 43-46.

16. Moher D, Schulz KF, Altman D (2001) The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. Jama 285: 1987-1991.

17. Pocock SJ, Collier TJ, Dandreo KJ, de Stavola BL, Goldman MB, *0* (2004) Issues in the reporting of epidemiological studies: a survey of recent practice. BMJ 329: 883.

18. Ioannidis JP, Mulrow CD, Goodman SN (2006) Adverse events: the more you search, the more you find. Ann Intern Med 144: 298-300.

19. Peng RD, Dominici F, Zeger SL (2006) Reproducible epidemiologic research. Am J Epidemiol 163: 783-789.

20. Dickersin K, Rennie D (2003) Registering clinical trials. JAMA 290: 516-523.

21. DeAngelis CD, Drazen JM, Frizelle FA, Haug C, Hoey J, *0* (2004) Clinical trial registration: a statement from the International Committee of Medical Journal Editors. JAMA 292: 1363-1364.

22. Ioannidis JP (1998) Effect of the statistical significance of results on the time to completion and publication of randomized efficacy trials. JAMA 279: 281-286.

23. Dickersin K, Min YI (1993) Publication bias: the problem that won't go away. Annals of the New York Academy of Sciences 703: 135-146.

24. Chan AW, Altman DG (2005) Identifying outcome reporting bias in randomised trials on PubMed: review of publications and survey of authors. BMJ 330: 753.

25. Fontanarosa PB, Flanagin A, DeAngelis CD (2005) Reporting conflicts of interest, financial aspects of research, and role of sponsors in funded studies. JAMA 294: 110-111.

26. Lash TL (2007) Heuristic thinking and inference from observational epidemiology. Epidemiology 18: 67-72.

27. Efron B (2004) Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. JASA 99: 96-104.

28. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. J Roy Statist Soc B 1995: 289–300.

29. Goodman SN (1999) Towards Evidence-based Medical Statistics, II: The Bayes Factor. Ann Intern Med 130: 1005-1013.

30. Spiegelhalter, D. J., Abrams, K. R., and Myles, J. P. (2004) Bayesian Approaches to Clinical Trials and Health-Care Evaluation. Chicester, UK: Wiley.

31. Greenland S (2005) Multiple-bias modelling for analysis of observational data. JRSS A 168: 267-306.

32. Greenland S, L. T. L. (2007) Bias Analysis. Ch. 19. In: Rothman KJ, G. S., Lash TL, editor. *Modern Epidemiology*. Philadelphia, PA:: Lippincott-Raven.

33. Ioannidis JP (2005) Contradicted and initially stronger effects in highly cited clinical research. JAMA 294: 218-228.

34. Ioannidis JP, Trikalinos TA (2005) Early extreme contradictory estimates may appear in published research: the Proteus phenomenon in molecular genetics research and randomized trials. J Clin Epidemiol 58: 543-549.