OPEN ACCESS Freely available online

PLoS MEDICINE

Essay

# When Should Potentially False Research Findings Be Considered Acceptable?

Benjamin Djulbegovic*, Iztok Hozo

## Summary

Ioannidis estimated that most published research findings are false [1], but he did not indicate when, if at all, potentially false research results may be considered as acceptable to society. We combined our two previously published models [2,3] to calculate the probability above which research findings may become acceptable. A new model indicates that the probability above which research results should be accepted depends on the expected payback from the research (the benefits) and the inadvertent consequences (the harms). This probability may dramatically change depending on our willingness to tolerate error in accepting false research findings. Our acceptance of research findings changes as a function of what we call "acceptable regret," i.e., our tolerance of making a wrong decision in accepting the research hypothesis. We illustrate our findings by providing a new framework for early stopping rules in clinical research (i.e., when should we accept early findings from a clinical trial indicating the benefits as true?). Obtaining absolute "truth" in research is impossible, and so society has to decide when less-than-perfect results may become acceptable.

As society pours more resources into medical research, it will increasingly realize that the research "payback" always represents a mixture of false and true findings. This tradeoff is similar to the tradeoff seen with other societal investments—for example, economic development can lead to environmental harms while measures to increase national security can erode civil liberties. In most of the enterprises that define modern society, we are willing to accept these tradeoffs. In other words, there is a threshold (or likelihood) at which a particular policy becomes socially acceptable.

In the case of medical research, we can similarly try to define a threshold by asking: "When should potentially false research findings become acceptable to society?" In other words, at what probability are research findings determined to be sufficiently true and when should we be willing to accept the results of this research?

## Defining the "Threshold Probability"

As in most investment strategies, our willingness to accept particular research findings will depend on the expected payback (the benefits) and the inadvertent consequences (the harms) of the research. We begin by defining a "positive" finding in research in the same way that Ioannidis defined it [1]. A positive finding occurs when the claim for an alternative hypothesis (instead of the null hypothesis) can be accepted at a particular, pre-specified statistical significance. The probability that a research result is true (the posterior probability; PPV) depends on: (1) the probability of it being true before the study is undertaken (the prior probability), (2) the statistical power of the study, and (3) the statistical significance of the research result. The PPV may also be influenced by bias [1,4], i.e., by systematic misrepresentation of the research due to inadequacies in the design, conduct, or analysis [1].

However, the calculation of PPV tells us nothing about whether a particular research result is acceptable to researchers or not. Nevertheless, it can be shown that there is some probability (the "threshold probability," $p_t$) above which the results of a study will be sufficient for researchers to accept them as "true" [3]. The threshold probability will depend on the ratio of net benefits/harms (B/H) that is generated by the study [3,5,6]. Mathematically the relationship between $p_t$ and B/H can be expressed as (see Appendix, Equation A1):

$$p_t = \frac{1}{1+\left(\frac{B}{H}\right)} \qquad (1)$$

We define net benefit as the difference between the values of the outcomes of the action taken under the research hypothesis and the null hypothesis, respectively (when in fact the research hypothesis is true). Net harms are defined as the difference between the values of the outcomes of the action taken under the null and the research hypotheses, respectively (when in fact the null hypothesis is true) [3]. It follows that if the PPV is above $p_t$ we can rationally accept the results of the research findings. Similarly, if the PPV is below $p_t$ we should accept the null hypothesis. Note that the research payoffs (the benefits) and the inadvertent consequences (harms)

The Essay section contains opinion pieces on topics of broad interest to a general medical audience.

**Funding:** This work was funded by the Department of Interdisciplinary Oncology of the H. Lee Moffitt Cancer Center and Research Institute at the University of South Florida.

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

**Citation:** Djulbegovic B, Hozo I (2007) When should potentially false research findings be considered acceptable? PLoS Med 4(2): e26. doi:10.1371/journal.pmed.0040026

**Abbreviations:** B/H, net benefits/harms; CI, confidence interval; combined $R_x$, radiotherapy plus chemotherapy; EUT, expected utility theory; PPV, posterior probability; $p_r$, acceptable regret threshold probability; $p_t$, threshold probability; $R_0$, acceptable regret; RT, radiotherapy alone

Benjamin Djulbegovic is in the Department of Interdisciplinary Oncology, H. Lee Moffitt Cancer Center and Research Institute, University of South Florida, Tampa, Florida, United States of America. Iztok Hozo is in the Department of Mathematics, Indiana University Northwest, Gary, Indiana, United States of America.

* To whom correspondence should be addressed: Benjamin.Djulbegovic@moffitt.org

in Equation 1 can be expressed in a variety of units. In clinical research these units would typically be length of life, morbidity or mortality rates, absence of pain, cost, and strength of individual or societal preference for a given outcome [3].

We can now frame the crucial question of interest as: What is the minimum B/H ratio for the given PPV for which the research hypothesis has a greater value than the null hypothesis? Mathematically, this will occur when (see Appendix, Equations A1 and A2):

$$\frac{1}{1+\left(\frac{B}{H}\right)} \le PPV \qquad \text{or} \qquad (2)$$

$$\frac{1}{PPV}-1=\frac{1-PPV}{PPV} \le \left(\frac{B}{H}\right)$$

## Calculation of the Threshold Probability of "Accepted Truth"

Figure 1 shows the threshold probability of "truth" (i.e., the probability above which the research findings may be accepted) as a function of B/H associated with the research results. The graph shows that as long as the probability of "accepted truth" (a horizontal line) is above the threshold probability curve, the research findings may be accepted. The higher the B/H ratio, the less certain we need to be of the truthfulness of the research results in order to accept them.

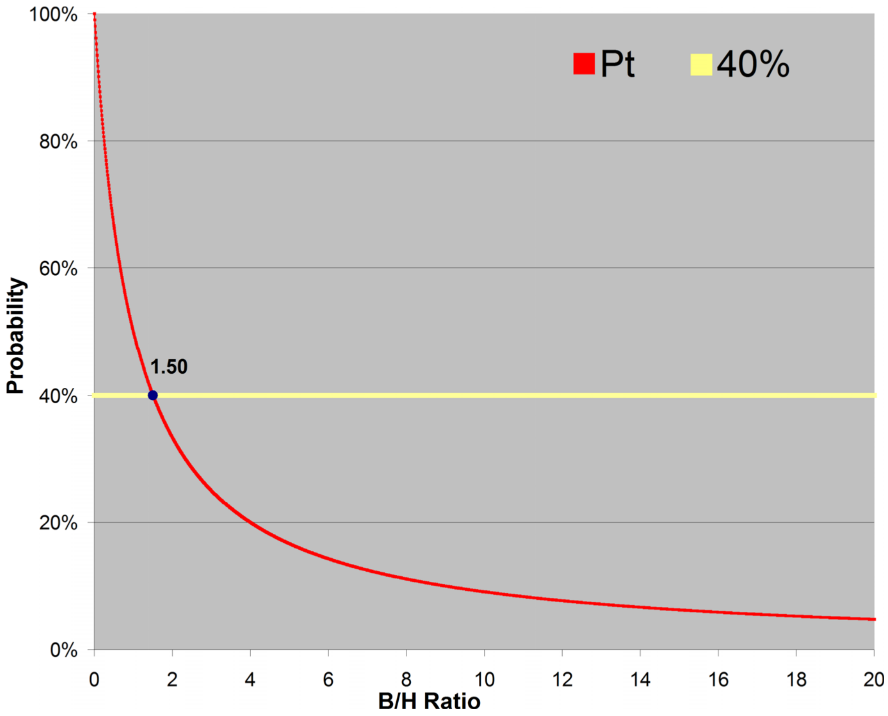
Note that we are following the classic decision theory approach to the results of clinical trials, which states that a rational decision maker should select the research versus the null hypothesis depending on which one maximizes the value of consequences [7–9]. In the parlance of expected utility decision theory, this means that we should choose the option with the higher expected utility [3,5,7–12]. (Expected utility is the average of all possible results weighted by their corresponding probabilities—see Appendix). In other words, the results of the research hypothesis should be accepted when the benefit of the action outweigh its harms.

## A Practical Example: When Should We Stop a Clinical Trial?

Interim analyses of clinical trials are challenging exercises in which researchers and/or data safety monitoring committees have to make a judgment as to whether to accept early promising results and terminate a trial or whether the trial should continue [13,14]. If the interim analysis shows significant benefit in efficacy for the new treatment over the standard treatment, continuing to enroll patients into the trial may mean that many patients will receive the inferior standard treatment [13,14]. The first randomized controlled trial of circumcision for preventing heterosexual transmission of HIV, for example, was terminated early after the interim analysis showed that circumcised men were less likely to be infected with HIV [15]. However, if a study is wrongly terminated for presumed benefits, this could result in adoption of a new therapy of questionable efficacy [13,14].

We now illustrate these issues by considering a clinical research hypothesis: is radiotherapy plus chemotherapy (combined $R_x$) superior to radiotherapy alone (RT) in the management of cancer of the esophagus? (see Box 1). We consider two scenarios: (1) the best-case scenario (B/H = 13.5), and (2) the worst-case scenario (B/H = 1.4). The probability that the research finding is true [16,17] (i.e., that combined treatment is truly better than radiotherapy alone) under the best-case scenario is 95% [95% confidence interval (CI), 89%–99.9%]. Under the worst-case scenario, the probability that combined treatment is better than radiotherapy alone is 80% [95% CI, 61%–99%]. The threshold probability above which these findings should be accepted is 7% [95% CI, 0%–30%] if we assume that B/H = 13.5, or 41% [95% CI, 11%–72%] if we assume B/H = 1.4 (Table 1).

The results indicate that in the best-case scenario, the probability that the research findings are true far exceeds the threshold above which the results should be accepted (i.e., PPV is greater than $p_t$). Therefore, rationally, in this case we should not hesitate to accept the findings from this study as truthful. However, in the worst-case scenario, the lower limit of the PPV's 95% confidence interval intersects with the upper limit of the threshold's 95% confidence interval, indicating that under these circumstances the research hypothesis may not be



doi:10.1371/journal.pmed.0040026.g001

**Figure 1.** The Threshold Probability **Above** ($P_t$ in Red) Which We Should Accept Findings of Research Hypothesis as Being True

The horizontal yellow line indicates the actual conditional probability that the research hypothesis is true in the case of positive findings. This means that for benefit/harm ratios above the threshold (1.5 in this example), the research hypothesis can be accepted.

## Box 1. Is Combined Chemotherapy Plus Radiotherapy Superior To Radiotherapy Alone for Treating Esophageal Cancer?

The Radiation Oncology Cooperative Group conducted a randomized controlled trial to evaluate the effects of combined chemotherapy and radiotherapy versus radiotherapy alone in patients with cancer of the esophagus [28].

A sample size of 150 patients was planned to detect an improvement in the two-year survival rate from 10%–30% in favor of combined $R_x$ (at $\alpha = 0.05$ and $\beta = 0.10$). At the interim analysis, 88% of patients in the control group (RT) had died while only 59% in the experimental arm (combined $R_x$) had died, resulting in a survival advantage of 29% in favor of combined $R_x$ ($p < 0.001$).

For this reason, the trial was terminated prematurely after enrolling 121 patients. Two percent of patients died as a result of treatment in the combined $R_x$ group versus 0% in the RT arm. Thus, the observed net benefit/harm ratio in this trial was [88-59-2]/2 = 13.5 [29] (the *best-case scenario*).

For our *worst-case scenario* we assume that two-thirds of patients who experienced life-threatening toxicities with combined $R_x$ (12%) will have died. This will result in the worst-case net benefit/harms ratio = (88-59-12)/12 = 1.4.

The trial was stopped using classic inferential statistics which indicated that the probability of the observed results, assuming the null hypothesis that combined $R_x$ is equivalent to RT, was extremely small ($p < 0.001$). This, however, tells us nothing about how true the alternative hypothesis is [16,17], i.e., in our case, what is the probability that combined $R_x$ is better than RT? The probability that the research finding is true [16,17] (i.e., that combined $R_x$ is truly better treatment than RT) under the best-case scenario is 95% [95% CI, 89%–99.9%]. Under the worst-case scenario, the probability that combined $R_x$ is better than RT is 80% [95% CI, 61%–99%].

acceptable (since PPV is possibly less than $p_t$). Had the investigators made a mistake when they terminated the trial early?

## Dealing with Unavoidable Erroneous Research Findings

Mistakes are an integral part of research. Positive research findings may subsequently be shown to be false [18]. When we accept that our initially positive research findings were in fact false, we may discover that another alternative (i.e., the null hypothesis) would have been preferable [7,19–21]. When an initially positive research finding turns out to be false, this may bring a sense of loss or regret [19,20,22,23]. However, abundant experience has shown that there are many situations in which we can tolerate wrong decisions, and others in which we cannot [2]. We have previously described the concept of *acceptable regret*, i.e., under certain conditions making a wrong decision will not be particularly burdensome to the decision maker [2].

## Defining Tolerable Limits for Accepting Potentially False Results

We now apply the concept of acceptable regret to address the question of whether potentially false research findings should be tolerated. In other words: which decision (regarding a research hypothesis) should we make if we want to ensure that the regret is less than a predetermined (minimal acceptable) regret, $R_0$ [2]? ($R_0$ denotes acceptable regret and should be expressed in the same units as benefits and harms).

It can easily be shown that we should be willing to accept the results of potentially false research findings as long as the posterior probability of it being true is above the acceptable regret threshold probability, $p_r$ (see Equation 3, Appendix, and Equations A3 and A4):

$$PPV \geq p_r = 1 - \frac{R_o}{H} = 1 - r \cdot \frac{B}{H} \qquad (3)$$

where $r$ is the amount of acceptable regret expressed as a percentage of the benefits that we are willing to lose in case our decision proves to be the wrong one (i.e., $R_o = r \cdot B$).

This equation describes the effect of acceptable regret on the threshold probability (Equation 1) in such a way that the PPV now also needs to be above the threshold defined in Equation 3 for the research results to become acceptable.

Note that actions under expected utility theory (EUT) and acceptable regret may not necessary be identical, but arguably the most rational course of action would be to select those

**Table 1. How True Is the Research Hypothesis that Combined Chemotherapy Is Superior To Radiotherapy Alone in the Management of Esophageal Cancer?**

| Net Benefits (Survival; %)[a] | Net Harms (Treatment-Related Mortality; %)[a] | Benefit/Harms Ratio | Type I ($\alpha$) Error (%) | Type II ($\beta$) Error (%) | The Threshold Probability Above Which Research Hypothesis Should Be Accepted as True Findings (%) | Probability that Research Hypothesis Is True (%) |
|---|---|---|---|---|---|---|
| [88-59-2][b] = 27% | 2 | 13.5 | 5 | 10 | 7% [0%–30%] | 95%<br>[89%–99.9%][d] |
| [88-59-12][c] = 17% | 12 | 1.4 | 5 | 20 | 41% [11%–72%] | 80%[e]<br>[61%–99%] |

Derived from [28].
[a] Calculated as described in [29].
[b] Best-case scenario.
[c] Worst-case scenario.
[d] Assumes 50% prior probability.
[e] Assumes 20% prior probability (see Supplementary Information for details).
doi:10.1371/journal.pmed.0040026.t001

**Table 2.** Probability that Research Findings Are True and Benefit/Harms Ratio Above Which Findings May Become Acceptable

| Type of Research | Probability that Findings Are True[a] (%) | Minimum Benefit/Harms Ratio Above Which Research Hypothesis Can Be Acceptable (No Regret Taken Into Account) | Acceptable Regret[b] for Wrongly Accepting Research Hypothesis (%) | Minimum Benefit/Harms Ratio Above Which Alternative Hypothesis Can Be Acceptable (When Acceptable Regret[b] Is Taken Into Account) |
|---|---|---|---|---|
| Adequately powered RCT with little bias and 1:1: pre-study odds (β = 20%) | 85 | 0.18 | 1 | 15 |
| | | | 20 | 0.75 |
| | | | 30 | 0.5 |
| | | | 40 | 0.38 |
| | | | 60 | 0.25 |
| | | | 80 | 0.19 |
| Confirmatory meta-analysis of good quality RCTs (β = 5%) | 85 | 0.18 | 1 | 15 |
| | | | 20 | 0.75 |
| | | | 30 | 0.5 |
| | | | 40 | 0.38 |
| | | | 60 | 0.25 |
| | | | 80 | 0.19 |
| Meta-analysis of small inconclusive studies (β = 20%) | 41 | 1.44 | 1 | 59 |
| | | | 20 | 2.95 |
| | | | 30 | 1.97 |
| | | | 40 | 1.48 |
| | | | 60 | 0.98 |
| | | | 80 | 0.74 |
| Underpowered, but well-performed phase I/II RCT (β = 80%) | 23 | 3.35 | 1 | 77 |
| | | | 20 | 3.85 |
| | | | 30 | 2.57 |
| | | | 40 | 1.93 |
| | | | 60 | 1.28 |
| | | | 80 | 0.96 |
| Underpowered, poorly performed phase I/II RCT (β = 80%) | 17 | 4.88 | 1 | 83 |
| | | | 20 | 4.15 |
| | | | 30 | 2.77 |
| | | | 40 | 2.08 |
| | | | 60 | 1.38 |
| | | | 80 | 1.04 |
| Adequately powered exploratory epidemiological study (β = 20%) | 20 | 4 | 1 | 80 |
| | | | 20 | 4 |
| | | | 30 | 2.67 |
| | | | 40 | 2 |
| | | | 60 | 1.33 |
| | | | 80 | 1 |
| Underpowered exploratory epidemiological study (β = 80%) | 12 | 7.33 | 1 | 88 |
| | | | 20 | 4.4 |
| | | | 30 | 2.93 |
| | | | 40 | 2.2 |
| | | | 60 | 1.47 |
| | | | 80 | 1.1 |
| Discovery-oriented exploratory research with massive testing (β = 80%) | 0.10 | 999 | 1 | 99 |
| | | | 20 | 5 |
| | | | 30 | 3.33 |
| | | | 40 | 2.5 |
| | | | 60 | 1.67 |
| | | | 80 | 1.25 |

Figures in red: applies only if a decision maker is willing to violate precepts of rational decision making under expected utility theory; otherwise under these circumstances research hypothesis never becomes acceptable. (Research may become acceptable if regret is smaller than the values pre-specified in the table; see text of article (Equation 4) and Text S1 for a longer version of the paper.)

[a] Data from [1].

[b] Expressed as r = 1%, 20%, etc. of benefits (i.e., percentage of benefits that we can tolerate losing in case we wrongly accept research findings).

RCT, randomized controlled trial

doi:10.1371/journal.pmed.0040026.t002

research findings with the highest expected utility while keeping regret below the acceptable levels. The supplementary material (a longer version of the paper and Appendix) show that the maximum possible fraction of benefits that we can forgo (and still be wrong) while at the same time adhering to the precepts of EUT is given by (see Appendix, Equations A3–A6):

$$r \leq \frac{1}{1+\frac{B}{H}} \qquad (4)$$

A practical interpretation of this inequality is that some research findings may never become acceptable unless we are ready to violate the axioms of EUT, i.e., accept value r to be larger than defined in Equation 4 (Table 2).

We return now to the "real life" scenario above, i.e., the dilemma of whether to stop a clinical trial early. In our worst-case analysis (Box 1), we found that the probability that combined $R_x$ is better than radiotherapy alone could potentially be as low as 80% [95% CI, 61%–99%]. This figure overlaps with the probability of the threshold of 41% [95% CI, 11%–72%] above which research findings are acceptable under the worst case scenario (see Table 1) (i.e., PPV is possibly less than $p_t$; see Equations 1 and 2). Thus, it is quite conceivable that the investigators made a mistake when they closed the trial prematurely.

One way to handle situations in which evidence is not solidly established is to explicitly take into account the possibility that one can make a mistake and wrongly accept the results of a research hypothesis. Accepting this possibility can, in turn, help us determine "decision thresholds" that will take into account the amount of error which may or may not be particularly troublesome to us if we wrongly accept research findings.

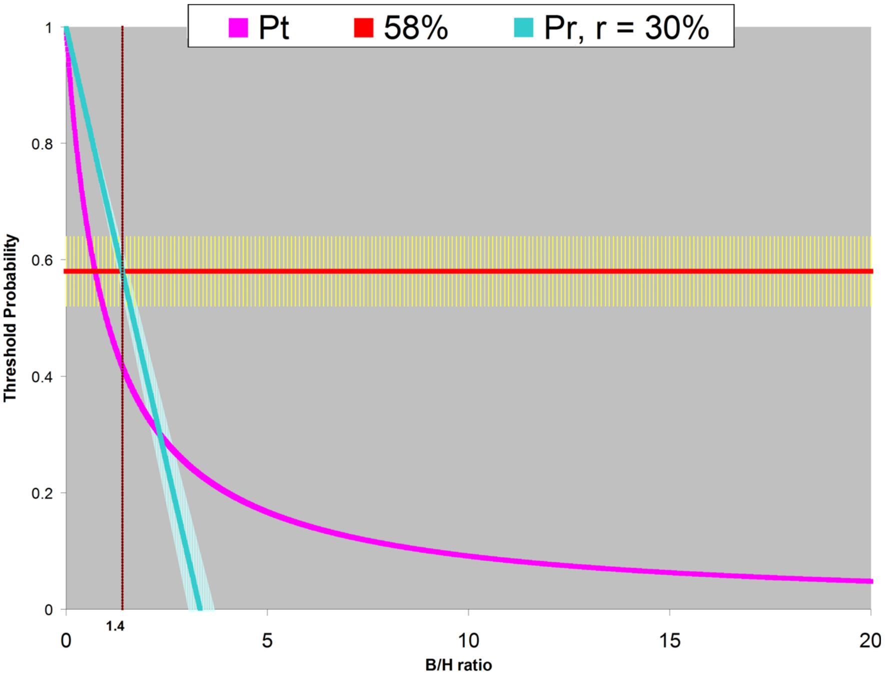
Let us assume that the investigators in the esophageal cancer trial are prepared to accept that they may be wrong and that they were willing to forgo 10%, 30%, or 67% of benefits. Using Equation 3, the calculations in Box 2 and Figure 2 show that for any willingness to tolerate loss of net benefits of greater than 10%, the probability that combined $R_x$ is superior to $R_T$ is above all decision thresholds (since $p_r = 0$ in best-case scenario; Equation 3). Therefore the investigators seemed to have been correct when they terminated the trial earlier than originally anticipated.

## Threshold Probabilities in Various Types of Clinical Research

Table 2 summarizes the results of most types of clinical research showing the probabilities that the research findings are true and the benefit/harms ratio above which the findings become acceptable. For each type of research, the table shows these probabilities with and without acceptable regret being taken into account. What is remarkable is that depending on the amount of acceptable regret, our acceptance of potentially false research findings may dramatically change. For example, in the case of a meta-analysis of small inconclusive studies, we can accept the research hypothesis as true only if B/H > 1.44. However, if we are willing to forgo, say, only 1% of the net benefits in case we prove to be mistaken, the B/H ratio for accepting the findings from the meta-analysis of small inconclusive studies dramatically increases to 59.

## Conclusion

In the final analysis, the answer to the question posed in the title of this paper, "When should potentially false research findings be considered acceptable?" has much to do with our beliefs about what constitutes knowledge itself [24]. The answer depends on the question of how much we are willing to tolerate the research results being wrong. Equation 3 shows an important result: if we are not willing to accept any possibility that our decision to accept a research finding could be wrong (r = 0), that would mean that we can operate only at absolute certainty in the "truth" of a research hypothesis (i.e., PPV = 100%). This is clearly not an attainable goal [1]. Therefore, our acceptability of "truth" depends on how much we care about being wrong. In our attempts to balance these tradeoffs, the value that we place on benefits, harms, and degrees of errors that we can tolerate becomes crucial.



doi:10.1371/journal.pmed.0040026.g002

**Figure 2.** The Threshold Probability ($P_t$) **Above** Which We Should Accept Findings of Research Hypothesis as Being True (Pink Line) as a Function of Benefit/Harm Ratio

The calculated (acceptable regret) threshold above which we should accept research findings is shown for the worst-case scenario (B/H = 1.4; see text for details) with a (hypothetical) assumption that we are willing to forgo 30% of the benefits (slanted line). The calculated threshold probability (acceptable regret threshold) has a value of 58% when B/H = 1.4 (the horizontal line). This means that as long as the probability that research findings are true is above this acceptable regret threshold, these research findings could be accepted with tolerable amount of regret in case the research hypothesis proves to be wrong (for didactic purposes only one acceptable regret threshold is shown). See Box 2 and text for details.

## Box 2. Determining the Threshold Above Which Research Findings Are Acceptable When Acceptable Regret Is Taken Into Account

You will recall (in Box 1) that the Radiation Oncology Cooperative Group investigators hoped to detect an absolute difference of 10%–30% in survival in favor of combined $R_x$. By finding that combined $R_x$ improved survival by 29%, they appeared to have realized their most optimistic expectations [28]. This implies that the investigators would consider their trial a success even if the survival was improved by 10% instead, i.e., less than 67% of the realized, but most optimistic outcome [1-(.10/.30) × 100% = 67%].

Therefore, we assume that the investigators in the esophageal cancer trial are prepared to accept that they may be wrong and that they were willing to forgo 10%, 30%, or 67% of benefits.

We applied Equation 3 to calculate acceptable regret thresholds above which we can accept research findings as true (i.e., when PPV > $p_r$).

**Best-case scenario (benefit/harm ratio: 13.5).** The calculated thresholds above which we should accept the findings are zero, regardless of whether our tolerable loss of benefits was 10%, 30%, or 67%. Note that these thresholds ($p_r = 0$) are well below calculated probability that the research hypothesis is true [PPV = 95% (88%–99.9%)] ( i.e., PPV > $p_r = 0$ for all acceptable regret assumptions; Equation 3, Table 1) and hence the research hypothesis should be accepted.

**Worst-case scenario (benefit/harm ratio: 1.4).** The calculated threshold above which we should accept the findings from this study is 86% [95% CI, 84%–88%] for a loss of 10% of benefits, 58% [95% CI, 52%–64%] for a loss of 30% of net benefits, and 6% [95% CI, 0%–19%] if we are willing to tolerate a loss of 67% of net benefits. This means that, except in the case when acceptable regret is 10% or less, the probability that combined $R_x$ is superior to RT [80% (61%–99%)] is above all other decision thresholds and its "truthfulness" can be accepted (because PPV [= 80% (61%–99%)] > acceptable regret threshold [= 58% (52%–64%)] and PPV > acceptable regret threshold [= 6% (0%– 19%)]). Note that in case of our willingness to tolerate loss of 30% of benefits for being wrong, the upper limit of the acceptable regret CI (=64%) still overlaps with the lower limit of PPV's CI (=61%), but that is not the case if we are willing to forgo 67% of treatment benefits. See Equation 3, Table 1.

However, because a typical clinical research hypothesis is formulated to test for benefits, we have here postulated a relationship between *acceptable regret* and the fraction of benefits that we are willing to forgo in the case of false research findings. Unfortunately, when we move outside the realm of medical treatments and interventions, the immediate and long-term harms and benefits are very difficult to quantify. On occasion, wrongly adopting some false positive findings may lead to the adoption of other false findings, thus creating fields replete with spurious claims. One typical example is the use of stem cell transplant for breast cancer, which resulted in tens of thousands of women getting aggressive, toxic, and very expensive treatment based on strong beliefs obtained in early phase I/II trials until controlled, randomized trials demonstrated no benefits but increased harms of stem cell transplants compared with conventional chemotherapy [25]. Therefore, even for clinical medicine, where benefits and harms are more typically measured, we should acknowledge that often the quality of the information on harms is suboptimal [26]. There is no guarantee that the "benefits" will exceed the "harms." Although (as noted in Text S1) there is nothing to prevent us from relating $R_0$ to harms, or both benefits and harms, one must acknowledge that there is much more uncertainty, often total ignorance, about harms (since data on harms is often limited). As a consequence, under these circumstances research may become acceptable only if we relax our criteria for acceptable regret, i.e., accept value r to be larger than defined in Equation 4. In other words, unless we are ready to violate the precepts of rational decision making (see the figures in red in Table 2), a research finding with low PPV (the majority of research findings) should not be accepted [1].

We conclude that since obtaining the absolute "truth" in research is impossible, society has to decide when less-than-perfect results may become acceptable. The approach presented here, advocating that the research hypothesis should be accepted when it is coherent with beliefs "upon which a man is prepared to act" [27], may facilitate decision making in scientific research. ■

## Supporting Information

**Text S1.** Longer version of the paper.

Found at doi:10.1371/journal.pmed.0040026.sd001 (657 KB DOC).

**Text S2.** Appendix

Found at doi:10.1371/journal.pmed.0040026.sd002 (163 KB DOC).

## Acknowledgments

We thank Drs. Stela Pudar-Hozo, Heloisa Soares, Ambuj Kumar, and Madhu Behrera for critical reading of the paper and their useful comments. We also wish to thank Dr. John Ioannidis for important and constructive insights particularly related to the issue of quality of data on harms and overall context of this work.

### References

1. Ioannidis JP (2005) Why most published research findings are false. PLoS Med 2: e124. doi:10.1371/journal.pmed.0020124
2. Djulbegovic B, Hozo I, Schwartz A, McMasters K (1999) Acceptable regret in medical decision making. Med Hypotheses 53: 253–259.
3. Djulbegovic B, Hozo I (2002) At what degree of belief in a research hypothesis is a trial in humans justified? J Eval Clin Practice 8: 269–276.
4. Wacholoder S, Chanock S, Garcia-Closas M, El Ghormli L, Rothman N (2004) Assessing the probability that a positive report is false: An approach for molecular epidemiology studies. J National Cancer Inst 96: 432–442.
5. Pauker S, Kassirer J (1975) Therapeutic decision making: A cost benefit analysis. N Engl J Med 293: 229–234.
6. Djulbegovic B, Desoky AH (1996) Equation and nomogram for calculation of testing and treatment thresholds. Med Decis Making 16: 198–199.
7. Bell DE, Raiffa H, Tversky A (1988) Decision making: Descriptive, normative, and prescriptive interactions. Cambridge: Cambridge University Press.
8. Hastie R, Dawes RM (2001) Rational choice in an uncertain world. London: Sage Publications.
9. Ciampi A, Till JE (1980) Null results in clinical trials: The need for a decision-theory approach. Br J Cancer 41: 618–629.
10. Browner WS, Newman TB (1987) Are all significant p values created equal? The analogy between diagnostic tests and clinical research. JAMA 257: 2459–2463.
11. Hulley SB, Cummings SR (1992) Designing clinical research. Baltimore (MD): Williams and Wilkins.
12. Pater JL, Willan AR (1984) Clinical trials as diagnostic tests. Controlled Clin Trials 5: 107–113.
13. DAMOCLES Study Group (2005) A proposed charter for clinical trial data monitoring committees: Helping them to do their job well. Lancet 365: 721–722.
14. Pocock SJ (2005) When (not) to stop a clinical trial for benefit. JAMA 294: 2228–2230.

15. Auvert B, Taljaard D, Lagarde E, Sobngwi-Tambekou J, Sitta R, et al. (2005) Randomized, controlled intervention trial of male circumcision for reduction of HIV infection risk: The ANRS 1265 trial. PLoS Med 2 :e298. doi:10.1371/journal.pmed.0020298
16. Goodman SN (1999) Toward evidence-based medical statistics. 1: The p value fallacy. Ann Intern Med 130: 995–1004.
17. Goodman SN (1999) Toward evidence-based medical statistics. 2: The Bayes factor. Ann Intern Med 130: 1005–1013.
18. Ioannidis JPA (2005) Contradicted and initially stronger effects in highly cited clinical research. JAMA 294: 218–228.
19. Bell DE (1982) Regret in decision making under uncertainty. Oper Res 30: 961–981.
20. Loomes G, Sugden R (1982) Regret theory: An alternative theory of rational choice. Economic J 92: 805–824.
21. Loomes G (1987) Testing for regret and disappointment in choice under uncertainty. Economic J 92: 805–824.
22. Allais M (1953) Le compartment de l'homme rationnel devant le risque. Critque des postulates et axiomes de l'ecole Americaine. Econometrica 21: 503–546.
23. Hilden J, Glasziou P (1996) Regret graphs, diagnostic uncertainty and Youden's index. Stat Med 15: 969–986.
24. Ashcroft R (1999) Equipoise, knowledge and ethics in clinical research and practice. Bioethics 13: 314–326.
25. Welch HG, Mogielnicki J (2002) Presumed benefit: Lessons from the American experience with marrow transplantation for breast cancer. BMJ 324: 1088–1092.
26. Ioannidis JPA, Evans SJW, Gotzsche PC, O'Neill RT, Altman DG, et al. (2004) Better reporting of harms in randomized trials: An extension of the CONSORT statement. Ann Intern Med 141: 781–788.
27. deWaal C (2005) On pragmatism. Belmont (CA): Wadsworth.
28. Herskovic A, Martz K, Al-Sarraf M, Leichman L, Brindle J, et al. (1992) Combined chemotherapy and radiotherapy compared with radiotherapy alone in patients with cancer of the esophagus. N Engl J Med 326: 1593–1598.
29. Djulbegovic B, Hozo I, Lyman G (2000) Linking evidence-based medicine therapeutic summary measures to clinical decision analysis. MedGenMed 2: E6. Available: http://www.medscape.com/viewarticle/403613_1. Accessed 21 November 2006.

# Appendix: When should potentially false research findings be acceptable?

Benjamin Djulbegovic and Iztok Hozo

(see PLoS Medicine 2007;4(2):e26.doi:10.1371/journal.pmed.0040026)

## Analogies between Hypothesis Testing and Clinical Management

In a typical **Hypothesis Testing** scenario we have to accept one of the two hypotheses, the null hypothesis ($H_o$) and the alternative, research hypothesis ($H_a$). The possible results of our testing are two research findings: *RF*+ = "research finding positive", and *RF*- = "research finding negative". The true state of reality can be described as either $H_a$- = "null hypothesis is true (research hypothesis is false)", and $H_a$+ = "null hypothesis is false (research hypothesis is true)".

The probability of false negative research finding (type II error) is $\beta = P(RF- \mid H_a +)$ and is usually set at $\beta = 0.20$. The probability of a false positive research finding (type I error) is $\alpha = P(RF+ \mid H_a -)$; typically α is assumed to be $\alpha = 0.05$. The positive and negative likelihood ratios are $LR+ = \dfrac{1-\beta}{\alpha}$ and $LR- = \dfrac{\beta}{1-\alpha}$.

In a typical **Clinical Testing** scenario we conduct a clinical test in order to confirm one of the two hypotheses, the absence of disease ($H_o$) or the alternative, the presence of the disease ($H_a$). The possible results of our testing are two test results: *RF*+ = "positive test result" (indicates presence of the disease), and *RF*- = "negative test result". The true state of reality can be described as either $H_a$- = "the patient does not have the disease.", and $H_a$+ = "the patient has the disease".

The probability of false negative test result is $\beta = P(RF- \mid H_a +)$ (the opposite of sensitivity, $\beta = 1 - S$). The probability of a false positive test result is $\alpha = P(RF+ \mid H_a -)$ (the opposite of specificity, $\alpha = 1 - S_p$). The positive and negative likelihood ratios are $LR+ = \dfrac{S}{1-S_p} = \dfrac{1-\beta}{\alpha}$ and $LR- = \dfrac{1-S}{S_p} = \dfrac{\beta}{1-\alpha}$.

The posterior probability that the positive research finding (presence of the disease) is accurate is given by

$$PPV = P(H_a + \mid RF +) = \frac{P(RF+ \mid H_a +)P(H_a +)}{P(RF+ \mid H_a +)P(H_a +) + P(RF+ \mid H_a -)P(H_a -)}.$$

If we denote by $\pi = P(H_a +)$ the prior probability of positive research finding (presence of the disease), we can rewrite the equation above as

$$PPV = \frac{(1-\beta)\pi}{(1-\beta)\pi + \alpha(1-\pi)}.$$

Interested readers can found further details in the references(1-4).

## Decision Theory and Expected Utilities

Net benefit, B, is the difference between the utilities of outcomes of the action taken under research hypothesis and the null hypothesis, respectively (when in fact research hypothesis is true), $B = U_1 - U_3$ (see Figure A.1). Net harms, H, are defined as the difference between the utilities of outcomes of the action taken under the null and research hypothesis, respectively (when in fact null hypothesis is true), $H = U_4 - U_2$ (see Figure A.1).

In the context of classical decision theory (see Figure A1), we select the hypothesis with higher expected utility (EU) involving Benefits and Harms associated with our decision. Expected utility is the average of all possible results weighted by their corresponding probabilities. In case of the positive Research Finding $(RF+)$, the expected utility of accepting the alternate hypothesis is $E(H_a) = PPV \cdot U_1 + (1 - PPV) \cdot U_2$, and the expected utility of accepting the null hypothesis is $E(H_0) = PPV \cdot U_3 + (1 - PPV) \cdot U_4$. Setting $\mathrm{E}(\mathrm{H_a}) = \mathrm{E}(\mathrm{H_o})$ and solving for p, we find the probability at which either decision results in the same expected utility, the threshold probability (*p<sub>t</sub>*)(5):

$$p_t = \frac{1}{1 + \left(\frac{B}{H}\right)} \qquad \text{(A.1)}$$

In terms of prior probability, *π* (the probability that the research hypothesis is true before we perform the hypothesis testing), this threshold splits into two threshold probabilities: a) the treatment threshold probability above which we will accept the results of research hypothesis (reject a null hypothesis) ($p_{tRH}$) and b) the no-treatment threshold probability below which we will accept the results of null hypothesis ($p_{tNH}$). For unbiased results, mathematically these two thresholds can be expressed as(6, 7):

$$p_t = \frac{1}{1 + LR \cdot \frac{B}{H}} \qquad \text{(A.1')}$$

where $p_t = p_{tRH}$ if $\mathrm{LR} = \mathrm{LR-}$, and $p_t = p_{tNH}$ if $\mathrm{LR} = \mathrm{LR+}$.

Ioannidis(8) defines the effect of bias (u) on research findings in such a way that a fraction of false negative results can be classified as true positives (u*β) and a fraction of true negatives can be classified as false positives [u*(1-α)] because of bias. When incorporating bias, expressions for LR become(9):

$$LR+ = \frac{[(1-\beta) + u * \beta]}{[\alpha + u * (1-\alpha)]} = \frac{[(1-\beta) * (1-u)]}{[\alpha * (1-u) + u]} \text{ and } LR- = \frac{[\beta * (1-u)]}{[(1-\alpha) * (1-u)]}$$

Note that under this definition, bias effects cancel out in the expression for LR-. This indicate that it would be more desirable to assume separate effects of bias on true positive results (u1) and true negative results (u2), respectively.

The minimum B/H ratio for the given posterior probability for which the research hypothesis has a greater expected utility than the null hypothesis will occur when:

$$\frac{1}{1+\left(\frac{B}{H}\right)} \le PPV \quad \text{or} \quad \frac{1}{PPV}-1=\frac{1-PPV}{PPV} \le \left(\frac{B}{H}\right) \qquad \text{(A.2)}$$

In terms of prior probability *π*, we can rewrite the formula (2) as

$$\frac{1-\pi}{\pi}\cdot\frac{1}{(LR-)} \le \left(\frac{B}{H}\right) \qquad \text{(A.2')}$$

## Regret and Acceptable Regret

A reader is referred to reference (10) for details. Briefly, regret (Rg) is the difference between the utility of the outcome of the action taken and the utility of the outcome of another action we should have taken, in retrospect. For example, regret associated with acceptance of the research hypothesis (treatment Rx1) when in fact the null hypothesis is true is given by

$$Rg(Rx1,H_a -) = \max[U(Rx1,H_a -), U(Rx2,H_a -)] - U(Rx1,H_a -) = U_4 - U_2 = H$$

The regret associated with acceptance of the null hypothesis (treatment Rx2) when in fact the null hypothesis is true is given by

$$Rg(Rx2,H_a -) = \max[U(Rx1,H_a -), U(Rx2,H_a -)] - U(Rx2,H_a -) = U_4 - U_4 = 0$$

The regret associated with acceptance of the null hypothesis (treatment Rx2) when in fact the research hypothesis is true is given by

$$Rg(Rx2,H_a +) = \max[U(Rx1,H_a +), U(Rx2,H_a +)] - U(Rx2,H_a +) = U_1 - U_3 = B$$

The regret associated with acceptance of the research hypothesis (treatment Rx1) when in fact the research hypothesis is true is given by

$$Rg(Rx1,H_a +) = \max[U(Rx1,H_a +), U(Rx2,H_a +)] - U(Rx1,H_a +) = U_1 - U_1 = 0$$

Repeating the expected utilities procedure described above, we can define expected regret associated with selection of the research and the null hypothesis, respectively. A solution of these two equations will produce the same equation (A1) as the one defined under classic expected utility theory. However, at the intersection where the decisions of selecting the research hypothesis vs. the null are the same, the expected regret ($ER[.]$) is maximal:

$$ER\left[Rx1\right]=\left(1-p_t\right)H=ER\left[Rx2\right]=p_t\cdot B=\frac{B}{1+\frac{B}{H}}$$

Unlike the threshold probability, the maximal level of expected regret does not depend on the benefit/harm ratio only but also on the absolute magnitude of the net benefit.

The acceptable regret, $R_0$, is the utility we find acceptable of losing. We are interested in finding out at which probability *ER[Rx1]≤ R₀.* Solving this inequality, it follows that we should be willing to accept results of potentially false research findings as long as probability (p) of it being true is above the threshold probability, $p_r$

$$p\geq p_r=1-\frac{R_0}{H} \qquad \text{(A.3)}$$

Since regret among individuals does differ and is typically related to the magnitude of perceived benefits or harms, we will also assume the amount of acceptable regret is equal to the percentage of the benefits (*r*% of benefits) that we are willing to lose in case our decision proves to be the wrong one.

Therefore, if $R_o=r\cdot B$ (the percentage of benefit),

$$p\geq p_r=1-\frac{R_o}{H}=1-r\cdot\frac{B}{H} \qquad \text{(A.4)}$$

When acceptable regret is taken into account, a decision-maker may choose to violate expected-utility precepts in order to minimize his sense of loss. For example, if the probability PPV is between the thresholds $\frac{1}{1+\frac{B}{H}}\leq PPV\leq 1-r\cdot\frac{B}{H}$, the expected utility theory would prescribe acceptance of the research hypothesis (Rx1), but the decision maker would be exposed to an uncomfortable level of expected regret.

Nevertheless, under some circumstances the threshold probability based on the classic decision-theoretic approach to research findings (equation A.1) will be equal to the threshold probability derived from the acceptable regret approach. That is, these two thresholds will intersect when: $\frac{1}{1+\frac{B}{H}}=1-r\cdot\frac{B}{H}$. Solving for benefit/harms ratio, we have: $\frac{B}{H}=\frac{1}{r}-1$ (A.5)

or, solving for *r*: $r=\frac{1}{1+\frac{B}{H}}$ (A.6)

Equation A.6 indicated maximum possible loss (as a percent of the benefit) that we are willing to forgo (and be wrong) while at the same time adhering to the precepts of expected utility theory according to classic decision-theory.

## Confidence intervals

Using Taylor's expansion we can approximate the variance of a multivariable function of independent variables $w = f\left(x_1, x_2, x_3, ..., x_k\right)$ using the formula $\sigma_w{}^2 = \sum_{i=1}^{k}\left(\frac{\partial f\left(x_1, x_2, x_3, ..., x_k\right)}{\partial x_i}\right)^2 \sigma_{x_i}{}^2$ . Using this formula we can estimate the confidence intervals for the variables in formulas (A.1) – (A.6). For details, see the reference(11).

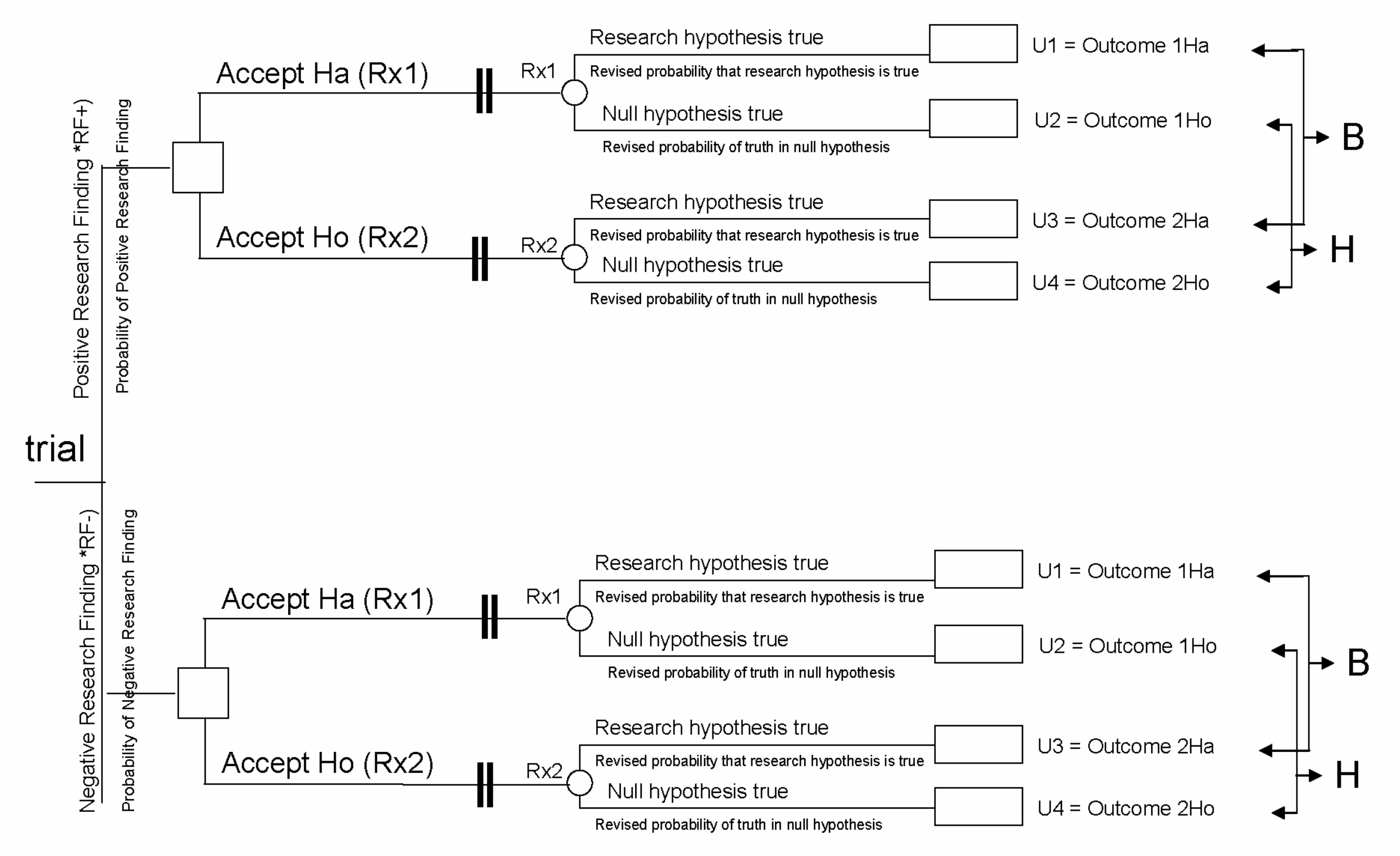

Positive Research Finding *RF+)
Probability of Positive Research Finding
Accept Ha (Rx1)
Rx1
Research hypothesis true
Revised probability that research hypothesis is true
Null hypothesis true
Revised probability of truth in null hypothesis
U1 = Outcome 1Ha
U2 = Outcome 1Ho
B
Accept Ho (Rx2)
Rx2
Research hypothesis true
Revised probability that research hypothesis is true
Null hypothesis true
Revised probability of truth in null hypothesis
U3 = Outcome 2Ha
U4 = Outcome 2Ho
H
trial
Negative Research Finding *RF-)
Probability of Negative Research Finding
Accept Ha (Rx1)
Rx1
Research hypothesis true
Revised probability that research hypothesis is true
Null hypothesis true
Revised probability of truth in null hypothesis
U1 = Outcome 1Ha
U2 = Outcome 1Ho
B
Accept Ho (Rx2)
Rx2
Research hypothesis true
Revised probability that research hypothesis is true
Null hypothesis true
Revised probability of truth in null hypothesis
U3 = Outcome 2Ha
U4 = Outcome 2Ho
H

**Figure A1:** Decision tree outlining the choice in a typical clinical research setting between accepting research hypothesis ($H_a$: Treatment Rx1 is superior) vs. null hypothesis ($H_o$: Rx2 is superior).

## References:

1. Browner WS, Newman TB. Are all significant p values created equal. The analogy between diagnostic tests and clinical research. JAMA. 1987;257:2459-63.
2. Goodman SN. Toward evidence-based medical statitistics. 1: the p value fallacy. Ann Intern Med. 1999;130:995-1004.
3. Goodman SN. Toward evidence-based medical statistics. 2: the Bayes factor. Ann Intern Med. 1999;130:1005-13.
4. Pianadosi S. Clinical Trials. A Methodologic perspective. Hoboken, New Jersey: Wiley-Interscience; 2005.
5. Pauker S, Kassirer J. Therapeutic decision making: a cost benefit analysis. N Engl J Med. 1975;293:229 -34.
6. Djulbegovic B, Desoky AH. Equation and nomogram for calculation of testing and treatment thresholds. Med Decis Making. 1996 Apr-Jun;16(2):198-9.
7. Djulbegovic B, Hozo I. At what degree of belief in a research hypothesis is a trial in humans justified? J Eval Clin Practice. 2002;8:269-76.
8. Ioannidis JP. Why most published research findings are false. PLoS Med. 2005 Aug;2(8):e124.
9. Pauker SG. The clinical interpretation of research. PLoS Medicine. 2005;2:e395.
10. Djulbegovic B, Hozo I, Schwartz A, McMasters K. Acceptable regret in medical decision making. Med Hypotheses. 1999;53:253-9.
11. Hozo I, Djulbegovic B. Calculating confidence intervals for posttest and threshold probabilities. MD Computing. 1997;15:110-5.