When do Scientists Change their Minds?

Week 7 – Science and reproducibility II

EGMT-1520 Monday, Feb 28, 2022
Bill Pearson wrp@virginia.edu

Overview of this session:

- Are most scientific papers wrong?
  - Testing: sensitivity / specificity
  - Prior probabilities – Bayes
  - False positives for Covid/ for HIV
  - False positives in scientific publications
  - Measuring the False Discovery Rate for papers
- Science – results vs process
- Final project work– time (final presentations due Weds. Mar 2)
- Also DUE March 2 – Peer Evaluation II

1

Final project products (Weds Mar 2):

1. A 5 minute presentation (Powerpoint/Google slides) with 5 – 6 slides
   1. 2 slides explaining why the incorrect explanation is correct – please try to make a convincing case for the wrong explanation that a 10 year old would believe
   2. 1-2 slides describing the change of perspective – focus on the perspective – what is being "seen" differently (not just equations)
   3. 2 slides explaining how the change of perspective explains the phenomena, highlighting the contrast between the "intuitive" perspective and the "correct" perspective
2. A 750–1000 word paper making the arguments in text. Arguments should be developed in paragraphs with topic sentences and complete sentences.
3. Each slide in the presentation or section of the paper should be attributed to at least one member of the group. Each member of the group should have an attributed contribution. Slides should not overlap with other slides; like wise paragraphs in the paper should have minimal overlap.
4. At least one person from each group should look at the presentation as a whole, to make certain that statements in one part do not contradict statements in another part.

2

Ioannidis, J. P. A. *PLoS Med.* **2,** e124 (2005).

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true. Moreover, for many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias. In this essay, I discuss the implications of these problems for the conduct and interpretation of research.

Published research findings are

factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a *p*-value less than 0.05. Research is not most appropriately represented and summarized by *p*-values, but, unfortunately, there is a widespread notion that medical research articles

It can be proven that most claimed research findings are false.

should be interpreted based only on *p*-values. Research findings are defined here as any relationship reaching formal statistical significance, e.g., effective interventions, informative predictors, risk factors, or associations. "Negative" research is also very useful. "Negative" is actually a misnomer, and the misinterpretation is widespread.

is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R + 1)$. The probability of a study finding a true relationship reflects the power $1 - \beta$ (one minus the Type II error rate). The probability of claiming a relationship when none truly exists reflects the Type I error rate, α. Assuming that c relationships are being probed in the field, the expected values of the 2×2 table are given in Table 1. After a research finding has been claimed based on achieving formal statistical significance, the post-study probability that it is true is the positive predictive value, PPV. The PPV is also the complementary probability of what Wacholder et al. have called the false positive report probability [10]. According to the 2×2 table, one gets $PPV = (1 - \beta)R/(R - \beta R + \alpha)$. A research finding is thus

Sensitivity, specificity true-positives, true-negatives

real state / truth	meas True	meas False
real True infected / loaded	True Positive TP	False Negative FN Type II error
real False healthy / fair	False Positive FP Type I error	True Negative TN

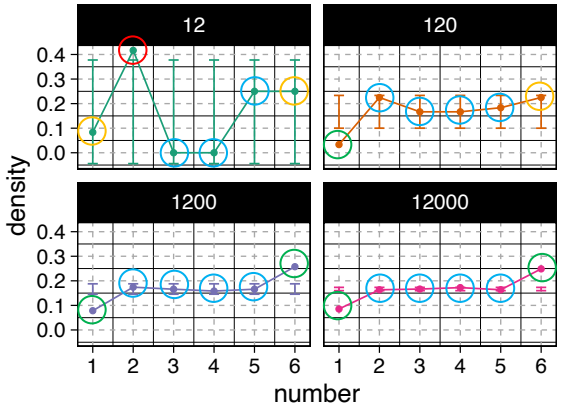
Sensitivity: TP / (TP + FN)
Specificity: TN / (TN + FP)

False Discovery Rate (FDR): FP / (TP + FP)
Positive predictive value: TP / (TP + FP)

In general, false positives are considered more harmful than false-negatives (except for infectious diseases)

sensitivity and specificity: non-random die

P(1)=1/12
P(2,3,4,5)=1/6
P(6)=3/12



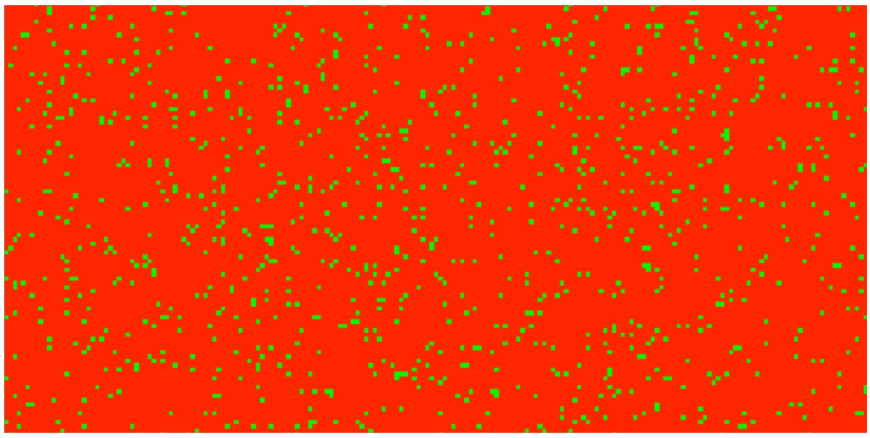
False Negative
False Positive
True positive
True negative

ice_nr.r 1

Sens: TP/(TP+FN)		Spec: TN/(TN+FP)
12:	sens: 0/2	spec: 3/4
120:	sens: 1/2	spec: 4/5
1200:	sens: 2/2	spec: 4/4

5

How many tests?



20,000 simultaneous t-tests on random normal data from the same distribution.
There are 1,009 green points (false positives), making up 0.05 of the comparisons (at α = 0.05).

6

Accuracy vs Prior Probability Covid19 testing

- ?Question – if you get a positive Covid PCR test, what are the chances you have Covid?
  - How accurate is the test (sensitivity/specificity)?
  - How likely is it that you have Covid?
- Covid19 RT-PCR tests are about 85% sensitive (15% false negatives), but 99% specific (1% false positives)
  www.idsociety.org/covid-19-real-time-learning-network/diagnostics/RT-pcr-testing/

7

If you have a positive Covid19 tests, and Covid tests are 99% specific, how likely are you to have Covid19?

- The simple (wrong) answer:
  - Tests are 85% sensitive (misses 15%) and 99% specific (1% false positives)
  - a positive test means a 99% chance of Covid19
- What else matters?
  - How many people have Covid when you do the test
- What if only 1,000 people have Covid19, but you test 1,000,000 per week?
  - From the 1 million tests, there will be 10,000 false positives (1%)
  - If only 1,000 people have Covid19, the odds of having the disease after a positive test is 1,000/10,000 or 10%.

Sensitivity and specificity are important, but so is the background (prior) probability of having Covid19

8

Accuracy vs Prior Probability
Bayes Theorum

A little notation:

$p(A)$ - probability of event A (must be ≤ 1.0)
coin came up heads

$p(B)$ - probability of event B (must be ≤ 1.0)
second toss came up heads
opposite side came up tails

$p(AB)$ - probability of event A and event B (≤1.0)
heads first toss, heads second p()=0.25
heads first toss, tails on opposite side p()=0.5

$p(A|B)$ - probability of event A given event B
p(H on second | H on first) = 0.5
p(T on bottom | H on top) = 1.0
p(having covid | +covid test)

9

Accuracy vs Prior Probability
Bayes Theorum

- Conditional probability:
  - If $p(A) = 0.5$ and $p(B) = 0.5$, what is the probability of both A and B $p(AB)$?
    - If A and B are independent, $p(AB) = p(A) \bullet p(B)$
      (consecutive coin tosses, $p(HH) = p(TT) = 0.25$
    - But if A and B are not independent, then
      $p(AB) = p(A|B) \bullet p(B) = p(B|A) \bullet p(A)$
      (one coin toss, p(H on bottom|T on top) = 1.0) $p(Ttop\ |Hbottom) = p(Ttop|Hbottom) \bullet p(T) = 0.5$
      Getting a Covid test is probably NOT independent of having Covid
- Bayes rule:
  - Since $p(AB) = p(A|B) \bullet p(B) = p(B|A) \bullet p(A)$
  - $p(A|B) \bullet p(B) = p(B|A) \bullet p(A)$ so
  - $p(A|B) = p(B|A) \bullet p(A) / p(B)$ (divide by $p(B)$)
  - $p(covid \mid +test) = p(+test \mid covid) \bullet p(covid) / p(+test)$

10

Accuracy and Prior Probability: Bayes Theorum

To estimate p(covid) given a positive test:

$$p(covid \mid +test) = p(+test \mid covid) \bullet p(covid) \,/\, p(+test)$$

1. What is $p(+test \mid covid)$ given a test based on sensitivity: $p(+test \mid covid) = 0.85$
2. What is $p(covid)$ in the population?
   - In the US in January, 2022, there were about 675,000 new cases/week • 3 weeks of illness= 2 million active cases
   - Cases are 2-4X under reported, so perhaps 6 million active
   - US population Jan 2022, 2020: 332 million
   - $p(covid) = 6/332 = 0.018\ (1.8\%)$
3. What is $p(+test)$?

Accuracy and Prior Probability: Bayes Theorum

To estimate p(covid) given a positive test:

$$p(covid \mid +test) = p(+test \mid covid) \bullet p(covid) \,/\, p(+test)$$

3. What is $p(+test)$?

$p(+test)$
$= p(+test|covid) \bullet p(covid) + p(+test|nocovid) \bullet p(nocovid)$
sens: 0.85 0.018 1–spec: 0.01 1-0.018=0.982
= 0.85 • 0.018 + 0.01 • 0.982
= 0.0153 + 0.0098 = 0.025

So: $p(covid \mid +test) =$

$$p(+test \mid covid) \bullet p(covid) \,/\, p(+test)$$
$$= 0.85 \bullet 0.018 \,/\, 0.025 = \boxed{0.61}$$

Accuracy and Prior Probability: Bayes Theorum

What if there is 5X as much Covid19??

From: $p(covid \mid +test) =$

$$p(+test \mid covid) \bullet p(covid) \,/\, p(+test)$$

$$= 0.85 \bullet 0.018 \,/\, 0.025 = \boxed{0.61}$$

To (5x as much):

$$p(+test \mid covid) \bullet p(covid) \,/\, p(+test)$$

$$= 0.85 \bullet 0.09 \,/\, p(+test)$$

$$p(+test) = 0.85 \bullet 0.09 + 0.01 \bullet 0.91$$

$$= 0.077 + 0.0091 = 0.0856$$

$$= 0.85 \bullet 0.09 \,/\, 0.0856 = \boxed{0.89}$$

What if you have symptoms? (and symptoms are from Covid 20%):

$$p(+test \mid covid) \bullet p(covid) \,/\, p(+test)$$

$$= 0.85 \bullet 0.20 / p(+test)$$

$$p(+test) = 0.85 \bullet 0.02 + 0.01 \bullet 0.8$$

$$= 0.17 + 0.008 = 0.18$$

$$= 0.85 \bullet 0.2 \,/\, 0.18 = \boxed{0.95}$$

13

Accuracy and Prior Probability: Bayes Theorum

To estimate p(covid) given a *negative* test:

$$p(covid \mid -test)$$

$$= p(-test \mid covid) \bullet p(covid) \,/\, p(-test)$$

$$= 1 - 0.85 \bullet 0.018 \,/ p(-test)$$

$p(-test)$

$= p(-test|covid) \bullet p(covid) + p(-test|nocovid) \bullet p(nocovid)$

1–sens: 0.15 0.018 spec: 0.99 1−0.018=0.982

$= 0.15 \bullet 0.018 + 0.99 \bullet 0.982$

$= 0.0027 + 0.972 = 0.975$

$$p(covid \mid -test) = 0.15 * 0.018/0.975 = 0.003$$

14

Sensitivity, specificity, and prior probability

- Covid19 PCR tests are 85% sensitive (15% of infections missed, False Negatives), but 99% specific (1% False Positives)
- At current infection levels (1.8% of population)
  - Positive test means 61% chance of infection
  - Negative test means 0.3% chance of infection
- If Covid19 incidence were 5x higher (9%):
  - Positive: 89% chance
- If you have symptoms (20% of Covid, not 1.8%):
  - Positive: 95% chance

The same test, with the same sensitivity and specificity, means different things depending on how many people are sick

15

Accuracy – HIV testing

To estimate p(HIV) given a positive test:

$$p(HIV \mid +test)$$
$$= p(+test \mid HIV) \bullet p(HIV) \,/\, p(+test)$$

$$= 0.98 \bullet 0.01 \,/\, (0.01 \cdot 0.98 + 0.99 \cdot 0.06)$$
$$= \boxed{0.142}$$

sens. 1.0-spec.

dlsun.github.io/probability/bayes.html

16

Ioannidis, J. P. A. *PLoS Med.* **2,** e124 (2005).

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true. Moreover, for many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias. In this essay, I discuss the implications of these problems for the conduct and interpretation of research.

Published research findings are

factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a *p*-value less than 0.05. Research is not most appropriately represented and summarized by *p*-values, but, unfortunately, there is a widespread notion that medical research articles

It can be proven that most claimed research findings are false.

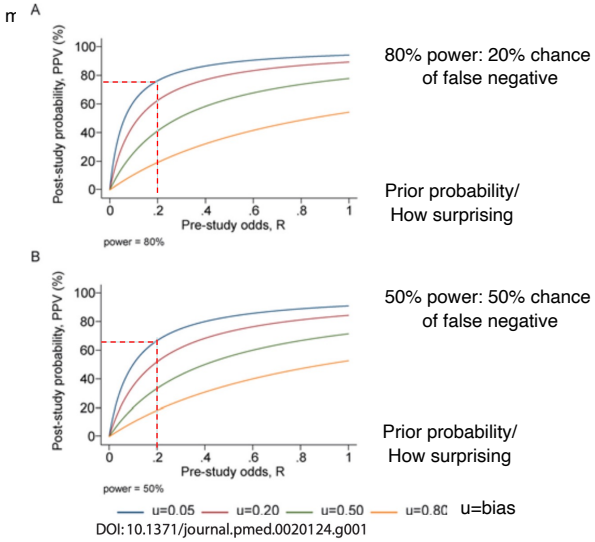
should be interpreted based only on *p*-values. Research findings are defined here as any relationship reaching formal statistical significance, e.g., effective interventions, informative predictors, risk factors, or associations. "Negative" research is also very useful. "Negative" is actually a misnomer, and the misinterpretation is widespread.

is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R + 1)$. The probability of a study finding a true relationship reflects the power $1 - \beta$ (one minus the Type II error rate). The probability of claiming a relationship when none truly exists reflects the Type I error rate, α. Assuming that c relationships are being probed in the field, the expected values of the 2×2 table are given in Table 1. After a research finding has been claimed based on achieving formal statistical significance, the post-study probability that it is true is the positive predictive value, PPV. The PPV is also the complementary probability of what Wacholder et al. have called the false positive report probability [10]. According to the 2×2 table, one gets $PPV = (1 - \beta)R/(R - \beta R + \alpha)$. A research finding is thus

17

"unexpectedness" vs "accuracy"

Just as low prior probability and high false positive rates misdiagnose diseases, surprising results with p<0.05 (5% false positives) are often wrong.



80% power: 20% chance of false negative

Prior probability/ How surprising

50% power: 50% chance of false negative

Prior probability/ How surprising

DOI: 10.1371/journal.pmed.0020124.g001

PLoS Med. **2,** e124 (2005)

18

Contradicted and initially stronger effects in highly cited clinical research

Results: Of 49 highly cited original clinical research studies, 45 claimed that the intervention was effective. Of these, 7 (16%) were contradicted by subsequent studies, 7 others (16%) had found effects that were stronger than those of subsequent studies, 20 (44%) were replicated, and 11 (24%) remained largely unchallenged. Five of 6 highly-cited non-randomized studies had been contradicted or had found stronger effects vs 9 of 39 randomized controlled trials (P = .008). Among randomized trials, studies with contradicted or stronger effects were smaller (P = .009) than replicated or unchallenged studies although there was no statistically significant difference in their early or overall citation impact. Matched control studies did not have a significantly different share of refuted results than highly cited studies, but they included more studies with "negative" results.

Conclusions: Contradiction and initially stronger effects are not unusual in highly cited research of clinical interventions and their outcomes. … Controversies are most common with highly cited nonrandomized studies, but even the most highly cited randomized trials may be challenged and refuted over time, especially small ones.

Ioaniddis (2005) *JAMA* **294,** 218–228

More than 50% (7+20)/(7+20+7)= 80% replicated

19

An estimate of the science-wise false discovery rate and application to the top medical literature

The accuracy of published medical research is critical for scientists, physicians and patients who rely on these results. However, the fundamental belief in the medical literature was called into serious question by a paper suggesting that most published medical research is false. … We then collect *P*-values from the abstracts of all 77430 papers published in *The Lancet*, *JAMA*, *The New England Journal of Medicine*, *The British Medical Journal*, and *The American Journal of Epidemiology* between 2000 and 2010. Among these papers, we found 5322 reported *P*-values. We estimate that the overall rate of false discoveries among reported results is 14% (s.d. 1%), contrary to previous claims. … Statistical analysis must allow for false discoveries in order to make claims on the basis of noisy data. But our analysis suggests that the medical literature remains a reliable record of scientific progress.

Jager & Leek (2014). *Biostatistics* **15:**1–12

20

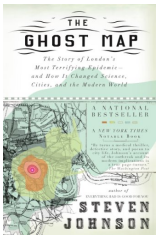
Science as a process and results

- The discovery of DNA as the genetic material
  - TMV paper "showed" protein transferred in infection (False positive)
  - Avery (1944) paper showed DNA is the transforming material
  - others failed to confirm Avery in other transformation systems (False negative)
  - Hershey and Chase (1952) phage transfer of DNA
  - Watson and Crick (1953) structure of DNA

21

Science as a process and results

- Bacterial basis of Cholera: John Snow, 1849
- Darwin, 1859
  - No mechanism (Mendel, DNA)
- The origin of continents and oceans (Wegener, 1925)
  - No mechanism
- HIV and AIDS
- Global warming



22

Are most scientific findings false?

- Many results cannot be replicated
  - technically challenging experiments are harder to replicate
- Journals strongly prefer to publish "positive" or "significant" results, which increases the odds of "false" discoveries
- Statisticians have reliable methods for correcting for multiple tests
  - these methods work best for controlling false-positives
- "Absence of evidence" is not "Evidence of absence" – false-negatives are common at the "bleeding edge"
- 70 - 90% of findings are probably correct (but <50% for very surprising results)

Final project products (Weds March 2):

1. A 5 minute presentation (Powerpoint/Google slides) with 5 – 6 slides
   1. 2 slides explaining why the incorrect explanation is correct – please try to make a convincing case for the wrong explanation that a 10 year old would believe
   2. 1-2 slides describing the change of perspective – focus on the perspective – what is being "seen" differently (not just equations)
   3. 2 slides explaining how the change of perspective explains the phenomena, highlighting the contrast between the "intuitive" perspective and the "correct" perspective
2. A 750–1000 word paper making the arguments in text. Arguments should be developed in paragraphs with topic sentences and complete sentences.
3. Each slide in the presentation or section of the paper should be attributed to at least one member of the group. Each member of the group should have an attributed contribution. Slides should not overlap with other slides; like wise paragraphs in the paper should have minimal overlap.
4. At least one person from each group should look at the presentation as a whole, to make certain that statements in one part do not contradict statements in another part.

Also DUE March 2 – Peer evaluation II