Is most published research false?

An exploration of false discovery and missed discovery in biomedical research

David Sidebotham

Auckland City Hospital

In this talk I address the issue of false discovery and missed discovery in the medical literature and attempt to answer the question: is most published research false? The issue of false discovery is based on the work of two investigators, David Colquhoun[1,2], a statistician from University College, London, and John Ioannidis,[3-7] a contrarian epidemiologist from Stanford University, California. The ideas on missed discovery are my own.

The title of the talk is taken from an essay written by John Ioannidis, and published in *PLoS Medicine* in 2005: Why most published research is false.[3]

Introduction: a randomised controlled trial, from hypothesis to publication

A researcher wishes to investigate whether a nadir haematocrit less than 0.26 is associated with increased mortality during cardiopulmonary bypass. She decides to perform a randomised controlled trial (RCT). Her null hypothesis (Devil's advocate position) is that a haematocrit less than 0.26 does not influence mortality. Her alternative hypothesis (what she actually suspects may be true) is that a haematocrit less than 0.26 does influence mortality. She then performs a power calculation to determine the sample size she will need. She chooses a power of 0.8 (beta of 0.2) and an alpha of 0.05. Statistical power is the probability of correctly rejecting a false null hypothesis (i.e., the probability of obtaining a true positive result). Beta is the probability of obtaining a false negative result, also known as a type II error. Power is one minus beta. Alpha (the critical value) is the probability of obtaining a false positive result, also known as a type I error. These values for power (0.8), beta (0.2), and alpha (0.05) are virtually ubiquitous in medical research.

Our researcher goes to an online sample size calculator to perform a 'power' calculation. She plugs in his chosen values for power, beta, and alpha. She then estimates the 'treatment effect' (proportional reduction in mortality with a haematocrit less than 0.26) and the 'event rate' (mortality in the control group). She chooses an expected mortality in the control group of 30% and in the treatment group of 20%. The online calculator tells him he needs a total of 293 patients in each group.

Then, over several years, and with much effort, our researcher performs the study. Following which, she measures a test statistic on the data (in this case a Fisher exact test). The test statistic produces a p-value of 0.04. The p-value tells her the probability that the observed (or a more extreme) outcome would occur if the null hypothesis was true. Since the p-value is less than the critical value (alpha) – and since mortality was lower in the high haematocrit group than the low haematocrit group – the researcher declares the result to be ‘statistically significant’, and concludes that a nadir less than 0.26 is associated with increased mortality during cardiopulmonary bypass. The study is written up and accepted for publication in in a prestigious journal.

False discovery

Given an alpha of 0.05 is ubiquitous in medical research, you might think that the proportion of findings that are false in the published medical literature was around 5%; however, the proportion is actually much higher than 5%. In some areas of medical research, the proportion of published studies that are false likely exceeds 50%.

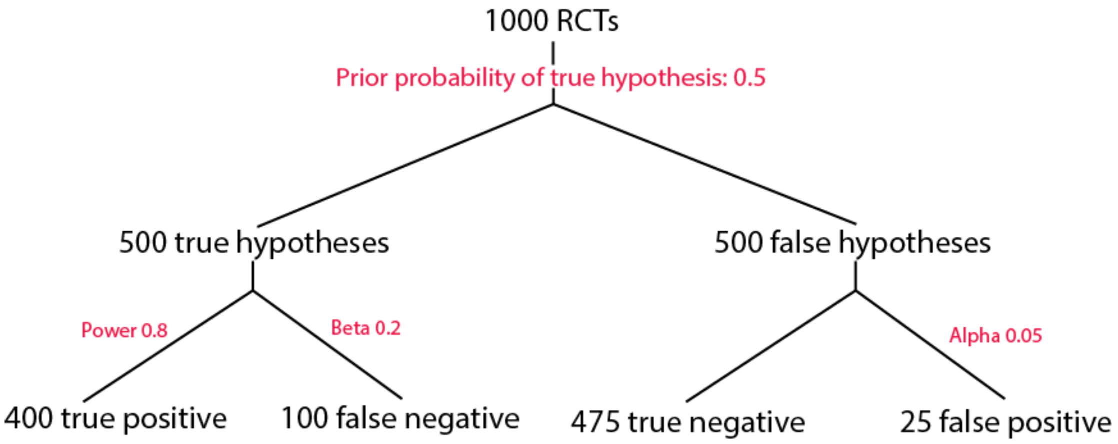
To explain why this is the case, I will use the metric of the false discovery rate (FDR), which is the proportion of published studies that are false:

$$FDR = \frac{FP}{FP+TP},$$

where FP is the number of false positive results and TP is the number of true positive results.

Imagine a 1000 RCTs investigating variety of hypotheses. Assume that the prior probability that each hypothesis is true is 0.5 (50%). We will call this parameter P(real). There are 500 true hypotheses and 500 false hypotheses. Using a power of 0.8 (beta 0.2) and alpha of 0.05, the expected FDR is 6%, as shown in Figure 1.

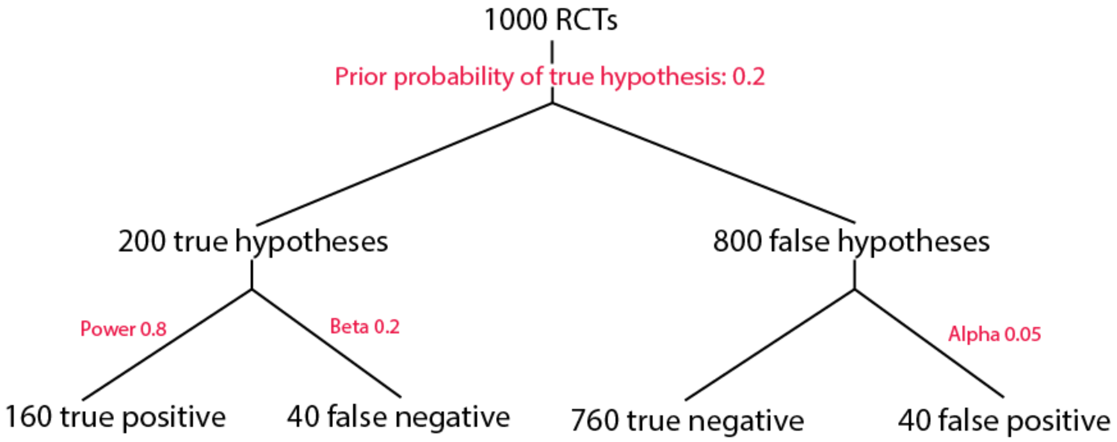
Figure 1: The False Discovery Rate



$$\text{FDR} = \frac{25}{25+400} = 0.06$$

When P(real) is 0.5 and power is 0.8, the FDR of 6% is similar to the ‘expected’ figure of 5%. However, when P(real) and power are reduced from these values, the FDR increases. For instance, if P(real) is 0.2, the FDR increases to 0.2 (20%), as shown in Figure 2.

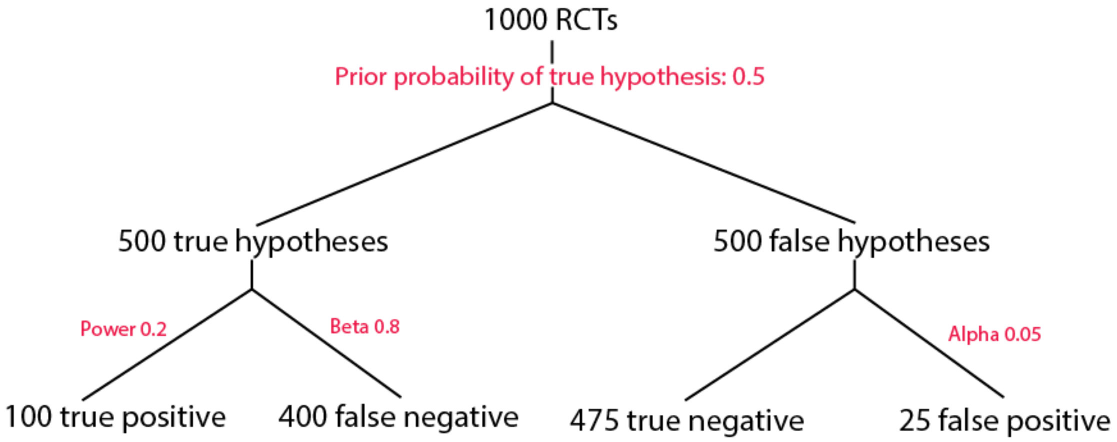
Figure 2: The effect of a low prior probability of a true hypothesis on the FDR



$$\text{False discovery rate} = \frac{FP}{FP+TP} = \frac{40}{40+160} = 0.2$$

Low power is widely appreciated to increase the likelihood of a type II error (false negative result); indeed, this is how a type II error is defined. What is less appreciated, however, is that low power also increases the FDR. This effect occurs because when a study has a reduced power, the number of true positive results decreases, and therefore the proportion of positive results that are false increases. Figure 3 shows the effect on FDR when power is 0.2 (and P(real) is 0.5). In this scenario the FDR is also 20%.

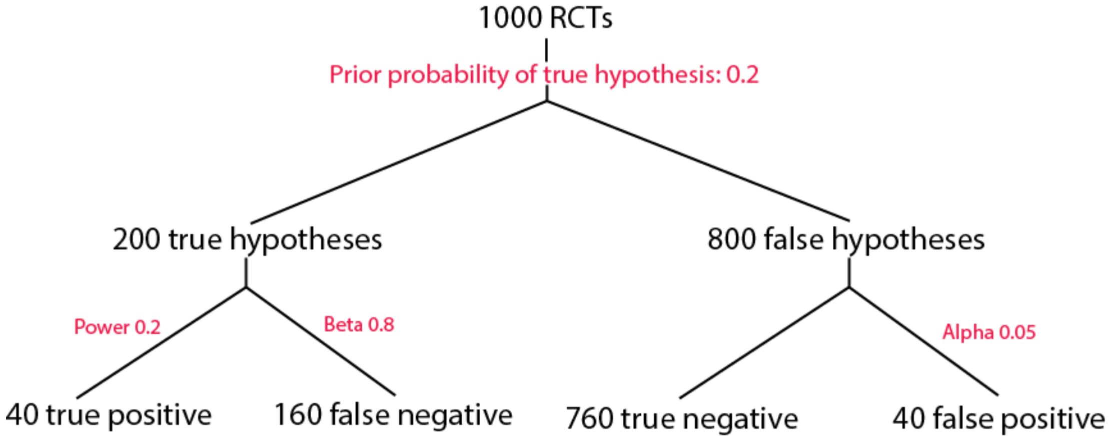
Figure 3: The effect of low power on the FDR



$$\text{False discovery rate} = \frac{FP}{FP+TP} = \frac{25}{25+100} = 0.2$$

When both power and P(real) are low, the FDR increases substantially. When power is 0.2 and P(real) is 0.2, the FDR is 50%, as shown in Figure 4.

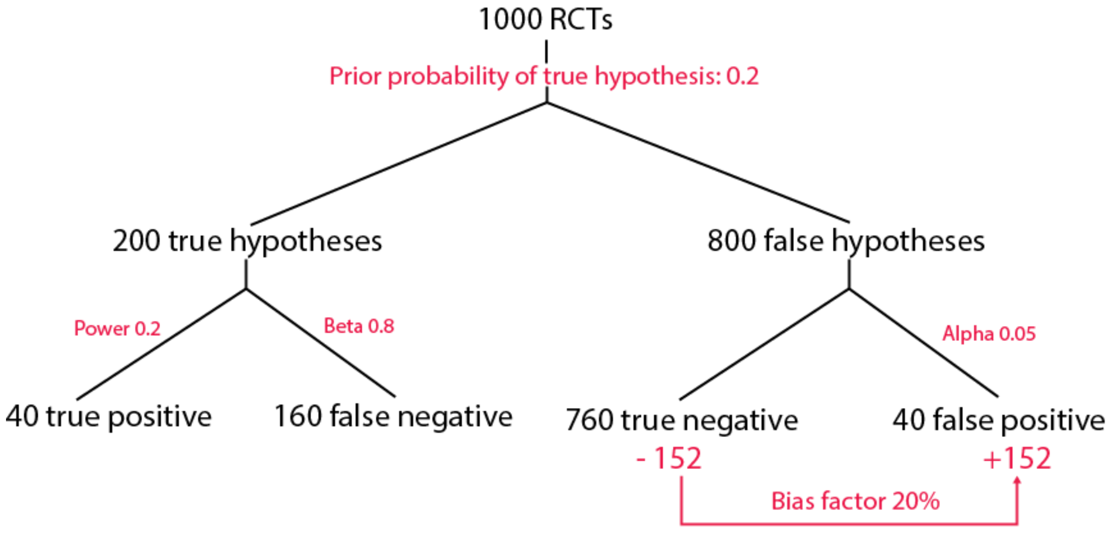
Figure 4: The effect of low prior probability and low power on the FDR



$$\text{False discovery rate} = \frac{\text{FP}}{\text{FP+TP}} = \frac{40}{40+40} = 0.5$$

Note the above examples all relate to well-conducted RCTs *in the absence of bias*. When the effect of bias, such as selective reporting or multiple testing (p-hacking), is considered the situation deteriorates further. In Figure 5, a factor of 0.2 has been introduced, in which 20% of true negative results become false positive results as a result of bias. Now, with P(real) of 0.2, power of 0.2, and bias of 0.2, the FDR is 83%.

Figure 5: The effect of low prior probability, low power, and bias on the FDR



$$\text{False discovery rate} = \frac{\text{FP}}{\text{FP+TP}} = \frac{192}{192+40} = 0.83$$

So, in the presence of a low prior probability, low power, *and* bias, the FDR can easily exceed 50%, a situation in which the great majority of published findings are false.

A Bayesian approach to false discovery[1]

In the same way as Bayes' formula can be used to evaluate the relationship between a positive test for a disease and the likelihood of having the disease (see my earlier talk: Should I get tested?"), a similar approach can be used for evaluating the relationship between the probability of a positive statistical test (statistically significant result) and the probability that the hypothesis is true. Here goes...

We are concerned with two probabilities: (1) the probability of a true hypothesis ($P(H_1)$) and (2) the probability of a statistically significant test result ($P(T^+)$). We can express $P(H_1)$ as $1 – P(H_0)$ and $P(T^-)$ as $1 – P(T^+)$.

We are interested in $P(H_0|T^+)$, the probability that the null hypothesis (H_0) is true, given a statistically significant test result. (Which is the same as saying, the probability that the alternative (H_1) hypothesis is false given a statistically significant test result.) $P(H_0|T^+)$ is equivalent to the FDR.

Applying Bayes' formula (see details from my first talk, we have:

$$P(H_0|T^+) = FDR = \frac{P(H_0)P(T^+|H_0)}{P(T^+)}$$

A statistically significant test result (T^+) can arise in the context of a true hypothesis (true positive finding) or of a false hypothesis (false positive finding), so (and again using the principles outlined in my first talk) we have:

$$P(T^+) = \ P(T^+|H_0)P(H_0) + \ P(T^+|H_1)P(H_1)$$

Which gives:

$$P(H_0|T^+) = FDR = \frac{P(H_0)P(T^+|H_0)}{(T^+|H_0)P(H_0) + (T^+|H_1)P(H_1)}$$

We have met the probabilities in the above equation earlier in this essay:

- $P(H_1)$ (and $1 – P(H_0)$) is equivalent to the prior probability of the hypothesis being true, termed P(real) above.
- $P(T^+|H_0)$ is the probability that the test is statistically significant given the null is true (probability of a false positive result). This probability is equivalent to alpha.
- $P(T^+|H_1)$ is the probability that the test is positive given the null is false (i.e., the alternative is true). That is, the probability of a true positive result, which is equivalent to the power (or 1-beta).

The Bayes' formulation provides the same information as obtained by the tree diagrams used earlier. Thus, for a P(real) of 0.2, power of 0.8 and alpha of 0.05, we have:

[1] Author note: this section can be daunting on first exposure (and even on subsequent exposures). I have been through this material numerous times and I still feel disorientated and confused each time I review it. Hopefully I have all the details correct. It is hard to be sure. Feel free to skip to the next section.

$$P(H_0|T^+) = \ FDR = \frac{(1-0.2)0.05}{(1-0.2)0.05+0.8\times 0.2} = \frac{0.04}{0.2} = 0.20.$$

Which is the same result as obtained above using the tree diagram in Figure 2.

The Bayesian approach gives a particular insight into the meaning of p-values and helps explain why the FDR is different from the alpha, the probability of a false positive result. The ideal probability for deciding whether to reject a null hypothesis in favour of the alternative hypothesis is $P(H_0|T^+)$, the probability that the null hypothesis is true given the test is statistically significant. This is after all, what we are interested in. However, the actual probability that is used for deciding to reject the null hypothesis is $P(T^+|H_0)$, the probability that the test is statistically significant given the null hypothesis is true. Another way of expressing $P(T^+|H_0)$ is the probability that the calculated p-value is less than or equal to the critical value (alpha) given the null hypothesis is true.

The two probabilities are not the same; that is, $P(H_0|T^+)$ is not equivalent to $P(T^+|H_0)$. The former is the FDR and the latter is the false positive rate for the statistical test. As outlined above, when $P(H_1)$ is low, $P(H_0|T^+)$ greatly exceeds $P(T^+|H_0)$. This is why the FDR is typically greater than the 'expected' value of 5% when P(real) is low.

The situation is analogous to the difference between $P(D^+|T^+)$, the probability you have a disease given a positive test (what you want to know) and $P(T^+|D^+)$, the probability a test is positive given you have a disease (the sensitivity of the test), outlined in my previous talk.

Phew!

Is it really this bad?

Are assumptions used in the above calculations reflective of real-world research? The first point to make is that there is strong (but indirect) evidence from replication studies, that false discovery *is* a major problem in biomedical research. The most detailed and impressive replicability study comes from the Open Science Collaboration, who attempted to *exactly* replicate 100 articles from three psychology journals. Published in *Science* in 2015, the results were sobering.[8] The mean effect size in the replicated studies was half that of in the original studies. While 97% of the original studies reported statistically significant results, only 36% of the replicated studies demonstrated statistically significant findings.

Closter to home, in the field critical care, concordance between studies investigating the same treatments is poor. (Here we are dealing with similar studies, not identical studies as in the *Science* article.) Niven and colleagues looked at 158 critical care practices that had been the subject of RCTs published in *The New England Journal of Medicine (N Engl J Med)*, *The Lancet*, and *The Journal of the American Medical Association (JAMA)*.[9] Replication attempts had been made for 66 practices. Overall, larger effect sizes were reported for the original studies than the replicated studies and more than half of the replication studies demonstrated a clinical effect that was inconsistent with the original study. Similar lack of concordance between studies examining the same intervention have been demonstrated across biomedical research. Many examples immediately spring to mind within the fields of anaesthesia and critical care: bispectral index monitoring for awareness; lignocaine for neuroprotection; tight perioperative glucose control, corticosteroids for acute respiratory distress syndrome, etcetera. The list is long.

As demonstrated above, a low P(real) and a low power predict a high FDR. However, are a low P(real) and low power common in published RCTs? Almost by definition, P(real) is unknown prior to performing a series studies testing a single hypothesis. Furthermore, the metric likely varies dramatically according to the type of research that is conducted. However, based on first principles and evidence from replication studies, an educated guess of P(real) in different scenarios can be made. Ioannidis[3] has estimated that for large confirmatory meta-analyses and RCTs, P(real) is around 50% and for small discovery-orientated RCTs, around 20%. For speculative or ‘long shot’ studies, P(real) is likely much less than 20%. Thus, ‘significant’ p-values need to be interpreted in light of the likelihood that the hypothesis being tested is true. The presence or lack of confirmatory evidence from other studies is crucial. If the result is dramatic or unsupported by other evidence, in all likelihood the finding is false. A low p-value (< 0.01) is somewhat reassuring but still leaves room for doubt. As Carl Sagan said, “Extraordinary claims require extraordinary evidence.”

While many investigators strive for (and report) a power of 0.8, in practice this is rarely the case. Problems frequently arise during the design phase of a study, due to the authors overestimating the event rate and potential efficacy of the intervention, which in turn leads to the sample size being underestimated. Alternatively, recruitment may be difficult or the study terminated early (due to lack of funding or for ‘ethical’ reasons, either efficacy or futility), resulting in a sample size that is less than planned. A number of authors have pointed out that the *actual* power of published studies is far less (often < 0.2) than the *claimed* power. Fundamentally, more patients are usually required for studies than funding or infrastructure allows. For instance, in the scenario outlined at the beginning of this article, nearly 600 patients are required to achieve a power of 0.8. Six-hundred patients is formidable target for any single-centre study. However, the estimated event rate (30%) and treatment effect (a 66% relative mortality reduction) are wildly optimistic. If mortality were more realistically estimated to be 5% in the control group (low haematocrit) and 3% in the treatment group (high haematocrit), a total of 3012 patients would be required to achieve a power of 0.8. Thus, rather than 0.8, the *actual* power achieved in the study, assuming a more realistic event rate of 5% and a treatment effect of 2% (absolute mortality reduction) is 0.23! Overestimation of the event rate in during the design phase is probably the main reason for the low power of published studies. This point highlights the importance of carefully reading the methods section of study before accepting the findings.

As a rough guide, Ioannidis[3] estimates the following rates of FDR according to the type of study:

Confirmatory meta-analysis of good quality RCTs	15%
Large well-conducted RCTs	15%
Meta-analysis of small inconclusive RCTs	60%
Underpowered well-conducted RCTs	75%
Underpowered, poorly-conducted RCTs	85%

(Note, Ioannidis uses the positive predictive value, which is 1-FDR)

A final point is worth mentioning. The p-value is an important determinant of the FDR, with values closer to 0.05 increasing the likelihood of a false discovery. The increased risk of false discovery with p-values close to 0.05 has led Ioannidis, and others, to advocate using an alpha of 0.01 for declaring 'statistical significance'.[6]

To summarise, risk factors for false discovery include:

1. P-values close to 0.05
2. Small, underpowered RCTs
3. Results that are unexpected or dramatic (low prior probability)
4. A lack of corroborating studies
5. Unrealistic event rate or treatment effect in the power calculation
6. Low-quality journal (however, false discoveries can be reported in good quality journals)
7. Low-quality investigators (however, false discoveries can be reported by high-quality investigators)

Missed discovery

Given the high risk for false discovery with small RCTs, the answer would appear to be to perform large RCTs that are adequately powered to answer hypotheses with a high P(real)? Indeed, large, well-designed, multi-centre RCTs that attempt to definitively answer hypotheses generated from smaller, less conclusive studies have been much embraced by the medical community over the last two decades. However, in my view – and at least in the area of critical care, the domain with which I am most familiar – there is a problem. Too many of these large, multi-centre RCTs are negative with respect to their primary outcome variable, particularly when the outcome is mortality. So, do we have a problem with missed discovery in large trials?

Using the same approach as for FDR, and assuming a P(real) of 0.5, a power of 0.8, and alpha of 0.05, the proportion of studies that would be expected to be positive (true positive + false positive/total studies) is 42%. This is a conservative estimate, as ideally a large RCT attempting to confirm data from smaller studies should have a P(real) of *at least* 50% and a power of *at least* 0.8. Indeed, it would seem unreasonable to spend millions of dollars of public funds on an inadequately powered fool's errand. However, the proportion of positive RCTs in critical mortality is the primary outcome is substantially less than 42%.

Harhay and colleagues reviewed 146 RCTs published in the critical care literature and found that overall 37% of trials were positive but only 10% were positive when mortality was the primary outcome variable.[10] Fifty-four studies were single-centre and 92 were multi-centre studies. Multi-centre trials were significantly *less* likely to be positive than single-centre trials.

Taking that most prestigious of medical journals, the *N Engl J Med*, there have been (by my count) 33 large RCTs or meta-analyses of RCTs published in the field of adult critical care in the last 10 years in which mortality was the primary outcome variable. Excluding non-inferiority trials, only six (18%) were positive with respect to the primary outcome. Of the 6 positive trials, one had a p-value of 0.03 and two had p-values of 0.04 with respect to mortality. One study was stopped early for harm in the treatment group (increased mortality in patients receiving high-frequency oscillation). An additional trial had a p-value of 0.05 for the primary outcome; however, the authors had chosen a predetermined p-value of 0.044 as the cut-off for statistical significance. Only one of the 33 trials can be considered to demonstrate a convincing benefit for the study treatment, prone positioning for treating acute respiratory distress syndrome.[11]

Perhaps the world's leading trials group in critical care is the Australia and New Zealand Intensive Care Society (ANZICS). This organisation performs well-conducted large multi-centre RCTs, the results of which results are published in prestigious medical journals, predominantly the *N Engl J Med*. Of the 10 multi-centre RCTs run by or endorsed in by ANZICS in which mortality was the primary end-point, none were positive.

What is the explanation for the low rate of positive large RCTs in the critical care literature? I believe there are several reasons, all of which relate to the fact that the studies are effectively underpowered, increasing the risk of a type II error (high false negative rate).

First, it is hard to demonstrate a difference in mortality by altering a single element of patient care, even if that intervention *is* beneficial. For instance, consider a patient admitted to ICU with pneumonia. The patient's risk of dying is primarily determined by the severity of his or her admission diagnosis and comorbidities. While in hospital, the patient is subject to multiple therapeutic interventions (e.g., mechanical ventilation, renal replacement therapy, nutrition, surveillance for nosocomial infection, prophylaxis for venous thromboembolism, fluid therapy, blood component therapy, skin cars, antimicrobial treatment, etcetera), all of which are modifiable. It is implausible that modifying a *single* intervention – for instance, the mode of sedation, or whether resuscitation fluid is given as normal saline or a balanced crystalloid solution – will have a large effect on mortality. Particularly, if the ‘dose’ of the intervention is low. Randomising such a patient to a bundle of care, involving changes to several interventions, could have a large (or at least measurable) impact on mortality. However, bundles of care are rarely evaluated; most trials test a single intervention.

Second, for pragmatic reasons, large multi-centre RCTs typically recruit ‘all comers’, which for a critical care trial is essentially all patients admitted to an ICU who meet certain broad criteria (e.g., ventilation for at least two days, requirement for a blood transfusion, evidence of sepsis, etcetera). Patients admitted to ICU encompass is a broad range of illness severity, and most patients recruited into a trial testing a single intervention will fall into one of two categories: a (larger) group who are ‘hard to harm” and a (smaller) group who are ‘hard to cure. Only a fraction of recruited patients will have an illness severity such that their mortality risk is substantially modifiable by the study intervention. Furthermore, the intervention may be beneficial for some patients but harmful for others. Imagine a RCT trial in which patients admitted to ICU received intravenous frusemide or placebo. For most patients their outcome will be unchanged. Some patients (e.g., those with heart failure) in the frusemide arm may benefit and others (e.g., those with hypovolaemia or bleeding) may be harmed. Such a trial is likely to be negative with respect to a ‘hard’ outcome such as mortality. A negative trial. However, it would be wrong to conclude that frusemide has no effect on mortality.

Third, power calculations for large RCTs are often flawed. A typical large RCT in critical care might recruit 2,500-5,000 patients. Given that mortality in critical care trials is consistently around 25-30%, a trial involving 3500 patients at a power of 80% and an alpha of 0.05, will detect an absolute mortality difference of about 4%. A mortality effect less than that will result in a non-significant p-value. For a most trials investigating a single intervention, say, two different sedation regimes, expecting a 4% absolute mortality effect is, in my view, unrealistic. A more realistic treatment effect of, say, a 1% on a baseline event rate of 25% would require more than 50,000 patients to achieve 80% power at an of alpha of 0.05. Bear in mind, if the results from small studies – which are prone to random effects and publication bias – are used in the power calculation for a subsequent large RCT, the treatment effect may be significantly overestimated in the power calculation.

As an example, take the TRANSFUSE study, published in 2017 in the *N Engl J Med*.[12] In TRANSFUSE, critically ill patients were randomised to receive either the ‘freshest’ or ‘oldest’ blood. Inclusion criteria included patients over 18-years with a predicted ICU stay longer than 24-hours who required one or more units of blood to be transfused. A total of 4919 patients in 59 centres were recruited over four years. The estimated baseline mortality was 25%. The median number of units transfused was 2 (interquartile range [IQR] 1-4). Observed mortality was 24.8% in the 'freshest' group compared to 24.1% in the 'oldest' group (p = 0.57, absolute mortality difference 0.7% [confidence interval -1.7-3.1%]). The treatment effect used to power the study was an absolute mortality difference of 4.2%. Thus, for patients admitted to an ICU for whatever reason, the relative mortality attributable to fresh versus old blood was predicted to be 16.8% (4.2/25). Further, the 'dose' of the intervention was low (median 2 units transfused). While it *is* plausible that for a trauma patient receiving a massive transfusion of, say, 50 units of blood, that a significant outcome difference would be found depending on whether they received old or new blood, for most patients it is not. Most patients recruited into the study, receiving a modest transfusion of 1-2 units, in the context of an admission diagnosis unrelated to blood loss – for example traumatic brain injury, subarachnoid haemorrhage, intraabdominal sepsis, etcetera – it is highly implausible that the intervention would have a major impact on their outcome.

For these reasons, the *effective* number of participants is far less than the actual number. As such, many of these trials are underpowered, and therefore subject to an increased risk of a type II error. The FDR for these trials is low, as P(real) is high and there is typically minimal bias. Low power does not affect the FDR in this circumstance, as low power can only affect the FDR if P(real) is low. So while small trials are skewed toward false discovery, large trials testing a single intervention with mortality as the primary outcome are skewed towards missed discovery.

Conclusion

Returning to the question posed at the start, 'Is most published research false?' While I hesitate to claim that most published research is false, in certain domains, I think a compelling argument can be made that a substantial proportion of the published literature is either falsely positive or falsely negative. My advice for consumers of the biomedical literature is simple: be sceptical, read the methods section first, and be mindful that a lot of what you read is probability not true.

Oh, and finally, don't assume that because a paper is published in a prestigious journal it is free from error, bias, (or fraud). After all, it was the *N Engl J Med* that published the initial papers by Mangano[13] and Poldermans[14] on perioperative betablockade; two papers that at best can be considered discredited.

References

1. Colquhoun D. An investigation of the false discovery rate and the misinterpretation of p-values. *Royal Society open science.* 2014;1(3):140216.
2. Colquhoun D. The reproducibility of research and the misinterpretation of p-values. *Royal Society open science.* 2017;4(12):171085.
3. Ioannidis JP. Why most published research findings are false. *PLoS medicine.* 2005;2(8):e124.
4. Ioannidis JP. Failure to Replicate: Sound the Alarm. *Cerebrum : the Dana forum on brain science.* 2015;2015.
5. Ioannidis JP. Evidence-based medicine has been hijacked: a report to David Sackett. *Journal of clinical epidemiology.* 2016;73:82-86.
6. Ioannidis JPA. Lowering the P Value Threshold-Reply. *Jama.* 2018;320(9):937-938.
7. Ioannidis JP. Discussion: Why "An estimate of the science-wise false discovery rate and application to the top medical literature" is false. *Biostatistics (Oxford, England).* 2014;15(1):28-36; discussion 39-45.
8. PSYCHOLOGY. Estimating the reproducibility of psychological science. *Science (New York, NY).* 2015;349(6251):aac4716.
9. Niven DJ, McCormick TJ, Straus SE, et al. Reproducibility of clinical research in critical care: a scoping review. *BMC medicine.* 2018;16(1):26.
10. Harhay MO, Wagner J, Ratcliffe SJ, et al. Outcomes and statistical power in adult critical care randomized trials. *Am J Respir Crit Care Med.* 2014;189(12):1469-1478.
11. Guerin C, Reignier J, Richard JC, et al. Prone positioning in severe acute respiratory distress syndrome. *N Engl J Med.* 2013;368(23):2159-2168.
12. Cooper DJ, McQuilten ZK, Nichol A, et al. Age of Red Cells for Transfusion and Outcomes in Critically Ill Adults. *N Engl J Med.* 2017;377(19):1858-1867.
13. Mangano DT, Layug EL, Wallace A, Tateo I. Effect of atenolol on mortality and cardiovascular morbidity after noncardiac surgery. Multicenter Study of Perioperative Ischemia Research Group. *N Engl J Med.* 1996;335(23):1713-1720.
14. Poldermans D, Boersma E, Bax JJ, et al. The effect of bisoprolol on perioperative mortality and myocardial infarction in high-risk patients undergoing vascular surgery. Dutch Echocardiographic Cardiac Risk Evaluation Applying Stress Echocardiography Study Group. *N Engl J Med.* 1999;341(24):1789-1794.