A patient called Medical Research*

Tomas Furst[a], Jan Strojil[b]

We report the case of a patient called Medical Research who presents with multiple life threatening symptoms, including a plethora of false positive results. This paper describes the course of the disease, discusses possible etiologies and offers options for future management to ensure the survival of the patient and that of our civilization.

Key words: reproducibility crisis, perverse incentives, *P* value, publication bias, frequentist, Bayesian

Received: January 26, 2017; Accepted: February 23, 2017; Available online: March 22, 2017
https://doi.org/10.5507/bp.2017.005

[a]Department of Mathematical Analysis and Applications of Mathematics, Faculty of Science, Palacky University Olomouc, Czech Republic
[b]Department of Pharmacology, Faculty of Medicine and Dentistry, Palacky University Olomouc, Czech Republic
**Invited article*
Corresponding author: Tomas Furst, e-mail: Tomas.Furst@seznam.cz

The scientific method has shaped our world, allowed us to understand nature and ourselves in ways that have made it possible for us to control the environment we live in and gain an upper hand over a number of diseases, extend the lifespan and improve the quality of life[1]. It would be difficult to overstate the importance of Science for our modern civilization. However, the authors believe that Science, especially Medical Research, suffers from problems that affect its ability to drive further progress.

A patient called Medical Research comes to us, presenting with a number of symptoms including fatigue and the inability to achieve reproduction. Diagnostic tests seem to be returning mostly false positives.

Recent years have seen the start of a conversation that has gradually occupied the pages of some of the most prestigious journals in the world, including Nature, The Economist and others. Among the first papers to raise red flags were two publications by large pharmaceutical companies showing that only 11-20% of landmark biomedical studies could be replicated[2,3]. Since then, poor reproducibility of basic research has been discussed in several other high profile papers. One example is the complete failure of potential amyotrophic lateral sclerosis drugs to reproduce their alleged effect in the very same mouse model on which they have been reported to be effective[4]. More recently, a publication brought to light the high rate of false positive results in functional magnetic resonance imaging (fMRI) studies caused by flawed analysis software that led to a 14-fold overestimation of positive results, potentially affecting tens of thousands of published papers[5]. The title of the most downloaded paper in PLoS Medicine of 2015 speaks for itself: Why most published research findings are false[6].

This crisis of reproducibility has vital implications for the ability of the medical industry to use the results of basic research in the development of new diagnostic and therapeutic methods. With false positives and biased results muddying the biomedicine waters, relying on published results for further research and development and in particular for making informed decisions about patient care becomes more and more problematic.

A self-proclaimed Medical Consilium (composed of the authors of this letter) has carefully examined the patient and assessed the state of the disease and its course. Vis á vis the etiology of the disease, it came to the following conclusion:

The precarious state of Medical Research is caused by the compounding of two factors: Perverse incentives and inappropriate statistical methods.

Perverse incentives in Science are widely understood to be a problem but are seldom discussed openly. In recent decades, almost all aspects of the evaluation of academic research have been narrowed down to a single metric – the quantity of published results. In the Czech Republic, the quantity of results is directly linked to the funding of research institutions (via institutional support), awarding of academic titles, hiring and firing of academic staff, assessing grant applications and scholarships, and evaluation of PhD students. Indirectly but increasingly often, it is also used to assess study programs (via accreditation standards), and to fund educational institutions. The excessive use of this metric creates a perverse incentive to publish results of any quality and origin. The sheer number of articles of dubious quality submitted for publication has necessarily compromised the peer-review process. Failures of the system have been reported multiple times not sparing reputable journals[7,8]. This makes it even more tempting to submit low-quality papers. Thus, the vicious cycle completes itself.

However strong these perverse incentives may be, they alone would not ruin Medical Research. Sadly, they have found a uniquely unfortunate companion in the classical statistical toolbox of hypotheses testing. Let us present a simple example to explain.

Let us randomly choose a location in the human genome and study a single base pair mutation in a pop-

Table 1. Example of two binary variables that have the same values but lead to different conclusions.

	CR +	CR –
mutation +	60	40
mutation –	40	60

	CR +	CR –
mass low	60	40
mass high	40	60

ulation of 200 patients with lymphoma, 100 of whom achieved complete remission (CR) after treatment and the other 100 did not. The mutation is binary (present or not) and so is CR. Let us assume our study produced the results shown in Table 1 on the left. Then we look at the relationship between CR and tumor mass (classified as low and high), shown in the right side of Table 1.

The asymptotic Chi-squared test of independence returns *P*=0.0047. The hypothesis of independence is therefore rejected in both cases. The test result is of course the same because the numbers are the same in both tables. But common sense tells us that the genetic result is most likely due to chance since we tested for a random position in a random gene, while the association between tumor mass and CR seems very plausible. Still, the *P*-values are the same.

What is going on? Is the test appropriate in one case and not in the other? No, the test is in fact used correctly in both cases but the meaning of its *P*-value needs to be interpreted correctly. It tells us the probability of obtaining the data in Table 1 (or data where the difference is even larger) **provided** there is no association between the mutation and CR (or tumor mass and CR in the latter case). The *P*-value is therefore a probability conditional on our null hypothesis being true. But in real life we are rarely interested in the probability of observing specific data provided our hypothesis is true. What we are typically interested in is the probability of our hypothesis being true given the results of our experiment. Let us ponder the fundamental difference between these questions and their answers. The *P*-value in no way reflects how likely our null hypothesis was to start with (i.e. its prior probability). In the first example, the association between CR and a randomly chosen base pair is of course negligible – there are billions of base pairs and an overwhelming majority of them will have no relationship to tumor response to treatment. A low p-value therefore means that the probability of an association between this mutation and CR increased from "ridiculously tiny" to "utterly negligible". The probability of a true association remains incredibly low despite a *P*-value of 0.0047.

The latter case is quite different. There are many valid reasons to think that the response to treatment is related to tumor size. The probability of an association is therefore quite high to begin with and the low *P*-value increases it further. Identical *P*-values can therefore lead to different conclusions regarding the validity of our hypothesis.

We fear that many false positive (irreproducible) results in medical research come into being this way. A profusion of hypotheses with very low prior probability get tested and every rejected null hypothesis (i.e. a statistically significant effect) is mistaken for a demonstrated fact and published. To make things worse, null hypotheses that failed to be rejected (i.e. non-significant effects) often never see the light of publication further distorting the evidence base we so fervently invoke in modern medicine. Most false positives get published but only very few true negatives do.

It is interesting to note that in medicine there is one context where the above misunderstanding does not happen – screening tests. To illustrate, let us take a disease with prevalence of 0.1% (one in thousand). Imagine a diagnostic test with sensitivity and specificity both equal to 99%. What is the probability that a randomly selected person who tests positive does in fact have this disease? In

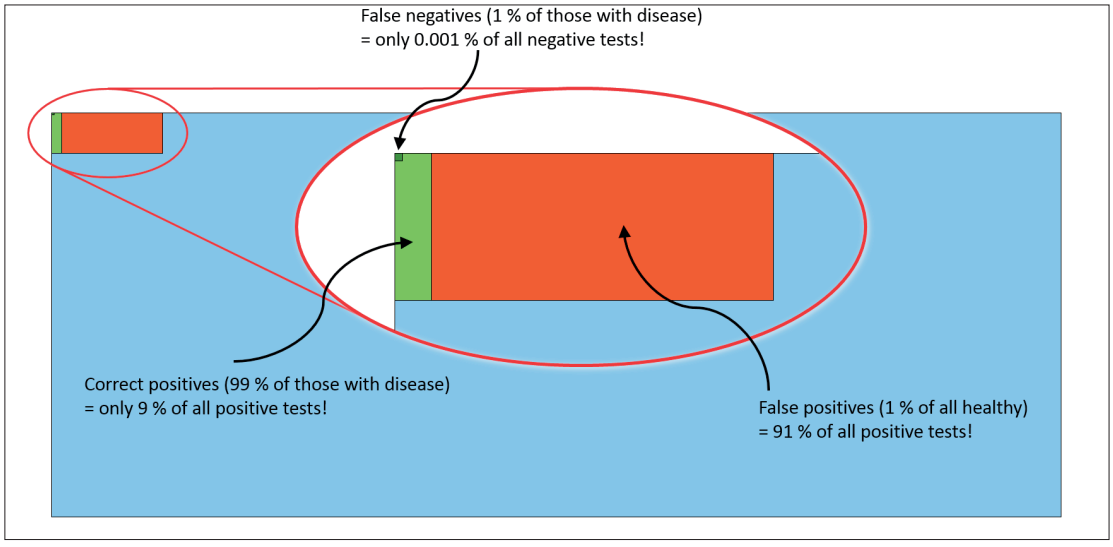

Fig. 1. Visualization of the rate of false positives when both sensitivity and specificity are extremely high. This is the best case scenario – real life performance of diagnostic or statistical tests will be much worse. (Specificity of a diagnostic test corresponds to [1 - type I error] of a hypothesis test and sensitivity corresponds to [1 - type II error]). Ignoring the prior probability leads to misleading conclusions. Key: Blue = true negatives, red = false positives, light green = true positives, dark green = false negatives.

this case, the *P*-value is the probability of a positive test in a healthy person (i.e. 1 – specificity = 0.01). Still, very few people would argue that we can reject the null hypothesis that the person is healthy based on this test alone.

In a population of one million, the prevalence of 0.1% means there are 1,000 people with the disease. False negatives (due to sensitivity < 100%) apply only to these 1,000 persons, therefore there will be only 10 false negatives and 990 (99%) will be correctly identified as having the disease. In a similar fashion, this test with a specificity of 99% will falsely flag 1% of healthy subjects as positive. One percent of 999,000 translates into 9,990 false positive results. This means that of all who tested positive (9,990+990), less than 10% will in fact have the disease. Over 90% of all positive results will be false positives. It is worrying that this kind of scrutiny which seems natural in the context of screening tests, is not applied in the field of hypotheses testing.

We have shown that the current method of hypotheses testing combined with an arbitrary threshold for rejecting them serves as a powerful generator of false positive results when prior probability gets neglected. This torrent of false positives combines in a very unfortunate synergy with the extreme pressure to publish. The result is the above described crisis of reproducibility in medical research.

Having examined the patient and assessed the disease we conclude that the condition is grave with little chance of spontaneous improvement should it be left untreated. A shift in culture, motivation and statistical tools is required for recovery. We offer the following treatment options:

First, let us start by removing the perverse incentives. When evaluating and funding research institutions, reward quality over quantity. We are glad to report that the proposed reform of institutional support in the Czech Republic seems to be headed in this direction. Restore the responsibility for hiring staff to department heads whose judgement has been replaced by automated quantity-based metrics. When evaluating teaching institutions, let us assess the quality of teaching, not the quantity of publications. Let us abolish the obsolete system of academic titles and opt for the more standard system of academic positions.

We should encourage replication of results. In a culture where only original, positive and striking results are valued, false positives are promoted and amplified. Reproducibility is one of the core tenets of science. It is lost when results are not challenged and rigorously retested. A relentless march forward without ever looking back to check the way can take us off the correct path into expensive dead ends that delay true progress. For example, the careful replication of a landmark result would be a valuable contribution of PhD students to their field and serve as a useful teaching tool. The resulting vast pool of replication studies would provide needed clarity to the scientific record.

All results should be published. We should encourage publication of well interpreted negative results. Overwhelmingly, these tend to be correct (remember the screening test above – there were virtually no false negatives!) and reduce wasting of resources due to re-testing the same failed hypothesis over and over (typically until a false positive result is obtained and published). Better yet, let us abolish the arbitrary threshold between "significant" and "non-significant".

These interventions will not ensure that the patient recovers unless we take another leap: let us adopt the Bayesian approach towards statistics which is in fact older than the more common frequentist doctrine. **We must realize that the frequentist toolbox attempts to answer the question "What is the probability that I obtain this data (or worse) given my hypothesis is true" while, more often than not, the question in real life stands "What is the probability of my hypothesis being true given my data".** This is exactly what the Bayesian method enables us to quantify. Imagine a radiologist examining a chest X-ray. She wants to know the probability that the patient has lung cancer. So why do we stick to statistical methods that rather tell her the probability of seeing this X-ray (or worse) in a healthy patient?

Two generations ago, computational power was not available to run the algorithms needed for Bayesian inference. Hypothesis testing was the only option for obtaining quantitative answers. This is no longer the case. The Bayesian toolbox is mature and ready for prime time. Some research areas have embraced it from the beginning – most notably the fascinating field of machine learning – and achieved spectacular success while doing so[9]. Many others are switching to Bayesian inference as we speak. It is of crucial importance that medical research embraces the correct inference as soon as possible because its results directly inform decisions about life and death.

Bayesian inference requires assumptions (and thus an open discussion) about the prior probability of the examined hypothesis. It then systematically updates this *prior* knowledge with all the new information that emanated from data to produce the *posterior* probability of the examined hypothesis. A simple case of Bayesian inference was presented in the screening example above – note that it produced the correct prediction, in contrast to hypothesis testing which produced a false positive. In the lymphoma example mentioned above, the very low prior probability of the mutation affecting CR is somewhat increased by the information from the data. However, the posterior probability remains very low. Despite the small *P*-value, the hypothesis is still very likely wrong. On the other hand, the prior probability of tumor mass affecting CR is reasonably high and it is further increased by the evidence. A Bayesian treatment enables correct inference, in stark contrast to hypothesis testing which (due to identical *P*-values) classifies both predictors of CR as "significant".

We implore the patient to start treatment as soon as possible, to waste no more time. We recognize that adherence to treatment recommendations may be an issue and that life-style modification are among the most difficult to follow. To help the patient on their way to recovery, we strongly

recommend a healthy dose of ASA. Not aspirin in this case but rather the American Statistical Association and their "Statement on p-values"[10].

The American Statistical Association has made a statement regarding the use of *P*-values. The statement explicitly says that:

- "Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold."
- "A *P*-value, or statistical significance, does not measure the size of an effect or the importance of a result."
- "By itself, a *P*-value does not provide a good measure of evidence regarding a model or hypothesis."

The ASA did not go as far as rejecting the entire toolbox of hypothesis testing. It rather advocated for multiple other approaches, including Bayesian treatment. Nevertheless, some journals, e.g. Basic and Applied Social Psychology, have stopped accepting articles containing *P*-values at all[11].

We wish the patient a speedy recovery and many more joyful years amongst us – her devoted and caring relatives.

Acknowledgement: Both authors wish to express their gratitude towards the Fulbright Commission for offering them a year of sheer intellectual joy in the USA which they spent reading, learning, and thinking without the obligation to publish as many papers as possible.

Conflict of interest statement: The authors declare that they wrote the article on behalf of *4BIN* – a movement for correct, i.e. Bayesian inference – which they are currently founding together with Halina Šimková, Jan Zimmer, Jiří Drábek, and Ondřej Vencálek. The authors declare that correct, i.e. Bayesian inference is in the interest of everyone, so that no conflict can arise.

REFERENCES

1. Ferguson N. Civilization: The West and the Rest. Penguin, 2011.
2. Begley CG, Ellis LM. Drug development: Raise standards for preclinical cancer research. Nature 2012;483(7391):531-3. doi: 10.1038/483531a
3. Prinz F, Schlange T, Asadullah K. Believe it or not: how much can we rely on published data on potential drug targets? Nat Rev Drug Discov 2011;10(9):712. doi: 10.1038/nrd3439-c1
4. Perrin S. Preclinical research: Make mouse studies work. Nature 2014;507(7493):423-5.
5. Eklund A, Nichols TE, Knutsson H. Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. Proc Natl Acad Sci U S A 2016;113(28):7900-5. doi: 10.1073/pnas.1602413113
6. Ioannidis JPA. Why most published research findings are false? PLoS Med 2005;2(8):e124.
7. Schroter S, Black N, Evans S, Godlee F, Osorio L, Smith R. What errors do peer reviewers detect, and does training improve their ability to detect them? J R Soc Med 2008;101(10):507-14. doi: 10.1258/jrsm.2008.080062.
8. Bohannon J. Who's afraid of peer review? Science 2013;342(6154):60-5. doi: 10.1126/science.342.6154.60
9. Bishop C. Pattern Recognition and Machine Learning. New York: Springer, 2007.
10. Wasserstein RL, Lazar NA. The ASA's Statement on p-Values: Context, Process, and Purpose. Am Stat 2016;70(2):129-33. doi: 10.1080/00031305.2016.1154108
11. Woolston C. Psychology journal bans P values. Nat News 2015;519(7541):9.