# The Truth Wears Off

## Fundamental limitation of statistical studies of living organisms and other complex systems

Yitzhak Rabin
Department of Physics
Bar-Ilan University

THE NEW YORKER

**Annals of Science**

# The Truth Wears Off

## Is there something wrong with the scientific method?

**by Jonah Lehrer December 13, 2010**

On September 18, 2007, a few dozen neuroscientists, psychiatrists, and drug-company executives gathered in a hotel conference room in Brussels to hear some startling news. It had to do with a class of drugs known as atypical or second-generation antipsychotics, which came on the market in the early nineties. The drugs, sold under brand names such as Abilify, Seroquel, and Zyprexa, had been tested on schizophrenics in several large clinical trials, all of which had demonstrated a dramatic decrease in the subjects' psychiatric symptoms. As a result, second-generation antipsychotics had become one of the fastest-growing and most profitable pharmaceutical classes. By 2001, Eli Lilly's Zyprexa was generating more revenue than Prozac. It remains the company's top-selling drug.

But the data presented at the Brussels meeting made it clear that something strange was happening: the therapeutic power of the drugs appeared to be steadily waning. A recent study showed an effect that was less than half of that documented in the first trials, in the early nineteen-nineties. Many researchers began to argue that the expensive pharmaceuticals weren't any better than first-generation antipsychotics, which have been in use since the fifties. "In fact, sometimes they now look even worse," John Davis, a professor of psychiatry at the University of Illinois at Chicago, told me.

the Atlantic

November 2010

Print | Close

# Lies, Damned Lies, and Medical Science

MUCH OF WHAT MEDICAL RESEARCHERS CONCLUDE IN THEIR STUDIES IS MISLEADING, EXAGGERATED, OR FLAT-OUT WRONG. SO WHY ARE DOCTORS—TO A STRIKING EXTENT—STILL DRAWING UPON MISINFORMATION IN THEIR EVERYDAY PRACTICE? DR. JOHN IOANNIDIS HAS SPENT HIS CAREER CHALLENGING HIS PEERS BY EXPOSING THEIR BAD SCIENCE.

*By David H. Freedman*

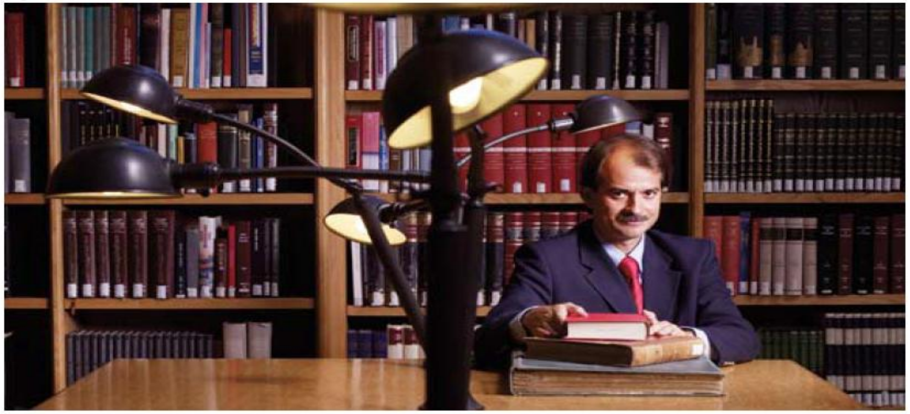

IMAGE CREDIT: ROBYN TWOMEY/REDUX

IN 2001, RUMORS were circulating in Greek hospitals that surgery residents, eager to rack up scalpel time, were falsely diagnosing hapless Albanian immigrants with appendicitis. At the University of Ioannina medical school's teaching hospital, a newly minted doctor named Athina Tatsioni was discussing the rumors with colleagues when a professor who had overheard asked her if she'd like to try to prove whether they were true—he seemed to be almost daring her. She accepted the challenge and, with the professor's and other colleagues' help, eventually produced a formal study showing that, for whatever reason, the appendices removed from patients with Albanian names in six Greek hospitals were more than three times as likely to be perfectly healthy as those removed from patients with Greek names. "It was hard to find a journal willing to publish it, but we did," recalls Tatsioni. "I also

WORLD VIEW *A personal take on events*

# Unpublished results hide the decline effect

*Some effects diminish when tests are repeated.* **Jonathan Schooler** *says being open about findings that don't make the scientific record could reveal why.*

Many scientifically discovered effects published in the literature seem to diminish with time. Dubbed the decline effect, this puzzling anomaly was first discovered in the 1930s in research into parapsychology, in which the statistical significance of purported evidence for psychic ability declined as studies were repeated. It has since been reported in a string of fields — both in individual labs (including my own) and in meta-analyses of findings in biology and medicine. The issue has been recognized in some circles within the scientific community, but rose to wider prominence last December when it was discussed in an article in the magazine *The New Yorker*.

Some scientists attribute the decline effect to statistical self-correction of initially exaggerated outcomes, also known as regression to the mean. But we cannot be sure of this interpretation, or even test it, because we do not generally have access to 'negative results': experimental outcomes that were not noteworthy or consistent enough to pass peer review and be published.

WE NEED A BETTER **RECORD** TO LEARN HOW WELL SCIENCE DISTINGUISHES **TRUTH** FROM **FALLACY.**

How could the availability of unpublished results be improved? I suggest an open-access repository for all research findings, which would let scientists log their hypotheses and methodologies before an experiment, and their results afterwards, regardless of outcome. Such a database would reveal how published studies fit into the larger set of conducted studies, and would help to answer many questions about the decline effect.

Availability of unpublished findings could also address other shortcomings of the current scientific process, including the regular failure of scientists to report experiments, conditions or observations that are inconsistent with hypotheses; the addition or removal of participants and variables to generate statistical significance; and the probable existence of numerous published findings whose non-replicability is shrouded because it is difficult to report null results.

To address the decline effect, such a database could pinpoint whether the phenomenon reflects how scientists design experiments, how they write them up or how journals decide what to publish. It could be used to explore whether genuine changes in studied phenomena could stem from conventional mechanisms; for example, in social sciences, decline effects could be the result of participants no longer being naive about the effect under investigation. Less likely, but not inconceivable, is an effect stemming from some unconventional process. Perhaps, just as the act of observation has been suggested to affect quantum measurements, scientific observation could subtly change some scientific effects. Although the laws of reality are usually understood to be immutable, some physicists, including Paul Davies, director of the BEYOND: Center for Fundamental Concepts in Science at Arizona State University in Tempe, have observed that this should be considered an assumption, not a foregone conclusion.

➲ NATURE.COM
Discuss this article online at:
go.nature.com/xbvqlt

More prosaic explanations for the decline effect include the previously mentioned regression to the mean. If early results are most likely to be reported when errors combine to magnify the apparent effect, then published studies will show systematic bias towards initially exaggerated findings, which are subsequently statistically self-corrected (although this would not account for the typically linear nature of the decline).

Publication bias could also be responsible. Researchers might only be able to publish initial findings on an effect when it is especially large, whereas follow-up studies might be more able to report smaller effects. Other potential answers include unreported aspects of methods, exclusive reporting of findings consistent with hypotheses, changes in researcher enthusiasm, more rigorous methodologies used in later studies, measurement error resulting from experimenter bias and the general difficulty of publishing failures of replication.

An open-access database of research methods and published and unpublished findings would go a long way towards testing these ideas. For example, both the regression to the mean and degradation of procedure explanations assume that early published studies benefit from being at one statistical end of a larger body of (unpublished) findings. Publication bias and selective reporting of data are similarly difficult to investigate without knowing about unpublished data.

An open-access repository of findings would be difficult to introduce. It would need an automated protocol to enable study methods and results to be entered and retrieved. Some way to assess the quality of the work would be required — perhaps through open-access commentaries moderated in a manner similar to Wikipedia. We would need to assure the qualifications of researchers who use it, and maintain a blackout period to protect hypotheses and findings prior to publication. Reluctant scientists would need incentives — and perhaps new rules from funders — to take part.

Such challenges would not be insurmountable. Similar, if more narrowly defined, databases have already been set up for clinical trials (http://clinicaltrials.gov) and educational research (http://pslcdatashop.web.cmu.edu). A good starting point might be to develop a host of subject-specific repositories. However it is implemented, we need a better record of unpublished research before we can know how well the current scientific process, based on peer review and experimental replication, succeeds in distinguishing grounded truth from unwarranted fallacy. ■

**Jonathan Schooler** *is a professor of psychology at the University of California, Santa Barbara.*
*e-mail: schooler@psych.ucsb.edu*

# 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 sometimes refuted by subsequent evidence, with ensuing confusion and disappointment. Refutation and controversy is seen across the range of research designs, from clinical trials and traditional epidemiological studies [1–3] to the most modern molecular research [4,5]. There is increasing concern that in modern research, false findings may be the majority or even the vast majority of published research claims [6–8]. However, this should not be surprising. It can be proven that most claimed research findings are false. Here I will examine the key 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 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. However, here we will target relationships that investigators claim exist, rather than null findings.

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

As has been shown previously, the probability that a research finding is indeed true depends on the prior probability of it being true (before doing the study), the statistical power of the study, and the level of statistical significance [10,11]. Consider a 2 × 2 table in which research findings are compared against the gold standard of true relationships in a scientific field. In a research field both true and false hypotheses can be made about the presence of relationships. Let $R$ be the ratio of the number of "true relationships" to "no relationships" among those tested in the field. $R$ 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, $\alpha$. 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

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

**Citation:** Ioannidis JPA (2005) Why most published research findings are false. PLoS Med 2(8): e124.

**Abbreviation:** PPV, positive predictive value

John P. A. Ioannidis is in the Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, Greece, and Institute for Clinical Research and Health Policy Studies, Department of Medicine, Tufts-New England Medical Center, Tufts University School of Medicine, Boston, Massachusetts, United States of America. E-mail: jioannid@cc.uoi.gr

**Competing Interests:** The author has declared that no competing interests exist.

**DOI:** 10.1371/journal.pmed.0020124

# Statistical ensemble of complex systems

- Lets assume that we have $N \gg 1$ "identical" systems ( mice, people, etc.), each of which possesses a measurable attribute (variable) $X$. This variable represents a complex property of the system such as the ability to recognize a face, the beneficial effect of a drug, etc. and therefore a measurement of $X$ does not, in general, yield a yes or no (binary) result.

- In order to describe this property we assume that $X$ can take a large number ($M_x \gg 1$) of values in the range $[x_{min}, x_{max}]$.

- We perform an experiment in which we measure the property $X$ for each of the systems and obtain a set of $N$ results $\{x_i\}_{i=1...N}$. In order to generate a good statistical sample, the number of systems has to be much larger than the number of possible $x$ values, i.e., $N \gg M_x$.

- We construct a histogram of these results and obtain the distribution $P(x)$. Consider the two simplest limiting cases:
- (a) Flat distribution $P(x)=const$. (b) Peaked distribution (around $x=a$).

- If the measured distribution is flat, we will conclude that $X$ is not an interesting/relevant property of these systems and forget about it. However, if the distribution is peaked, we will conclude that we discovered an important (i.e., publishable) property of our systems.

- Both limits can be described by e.g., a Gaussian distribution of width $\sigma$:

$$P(x)=A\exp\{-[(x-a)^2)/(2\sigma^2)]\}$$

- Each of the systems is complex in the sense that it is not completely defined by the property $X$ ( if it was not the case, observation of different values of $x$ would imply that the systems are not even approximately identical) and that are a huge number $\aleph$ of hidden variables $\{Y, Z, W, \ldots\}$ etc., such that $N \ll \aleph$. In the following, we refer to all the hidden variables as $Y$. The unknown variables $Y$ can take $M_y$ values where $M_y \gg N$.

- In general, unless we are either lucky or have deep knowledge/intuition about the complex system, the variables $X$ and $Y$ will not be statistically independent.

- Each time we perform an experiment that measures a property $X$, we sample a particular subset $\{y_i\}_{i=1..N}$ of all $M_y$ possible values of $Y$. Some of these subsets will yield very broad distributions of $X$ but it is only when a particular subset yields a peaked distribution,

$$P_{apparent}(X) = P(X|\{y\}) = A\exp\{-[x-a(\{y\})]^2/2\sigma^2(\{y\})\}$$

that we conclude that the experiment yielded a significant (publishable) result.

- The next time we perform an experiment on another group of "identical" individuals, we will sample another subset $\{y'\}$ of the set of possible values of the unknown variable $Y$ and will obtain different values $a'$ and $\sigma'$ which, in general, will no longer correspond to a narrow distribution.

- Since we already decided that X is a "good" variable, we are no longer free to dismiss the new results and we will conclude that "the truth wears off" with time. The extent to which this happens will depend on the correlation between the measured X and the hidden Y variables: weak correlations will yield relatively robust results which will change only weakly in repeated studies. This would happen, for example, for logarithmic dependence of the width on Y. Strong correlations will lead to irreproducible results and loss of significance once the experiment is repeated (e.g., when the dependence of the width on Y is of power law or exponential type).

- The use of control groups will not help to overcome this problem since such groups consist of similarly complex individuals and therefore introduce a baseline that may shift with time.

- Although the time scale for the truth to wear off may be affected by the intrinsic dynamics of the system (e.g., due to mutations or environmental changes that affect the population), in many cases such changes are too slow to be observed and the relevant time scale is the response time of the scientific community in question (the time it takes new results to be disseminated through publication and seminars, and the time it takes to design and carry out a new study) - usually, several years.

# What about Physics of Complex Systems?

- In equilibrium statistical physics we consider relatively simple fundamental elements for which the number of internal states is rather small (just 2 in the case of electron spin) and one can always generate a large enough ensemble of identical elements in order to obtain reliable statistics. It is only when one considers a large number of interacting elements that one observes complex phenomena such as phase transitions.

- Question: Why is the thermodynamic state of arbitrarily complex systems completely characterized by a small number of thermodynamic variables such as number of particles, volume, pressure, temperature, etc. ?

  Answer: we use symmetries and conservation laws (conservation of mass, energy, momentum, electric charge, etc.) to identify the thermodynamics variables and deduce the relationships between them, without getting into the largely unknown microscopic details of the system under consideration (e.g., a steam engine).

- No prescription for identifying the "good" variables that capture the behavior of complex physical systems far from equilibrium. In some cases (e.g., hydrodynamics ), we are sufficiently familiar with the system studied, to be able to identify the slowly changing variables and represent the cumulative effect of all the other (fast) degrees of freedom as dissipation and random noise.

Does truth wear off in Biophysics?

Lets take a vote!