emagnification:

a tool for estimating effect size magnification and performing design calculations in epidemiological studies

David J. Miller,[1] James T. Nguyen,[1] and Matteo Bottai [2]

[1] Health Effects Division
Office of Pesticide Programs
U.S. Environmental Protection Agency
Washington, DC,USA

[2] Unit of Biostatistics
Institute of Environmental Medicine
Karolinska Institute
Stockholm, Sweden

2019 Nordic and Baltic Stata Users Group Meeting

**Karolinska Institute
Stockholm**

30 August 2019

EPA
United States
Environmental Protection
Agency

Outline

- Background
- Reproducibility and Reliability… continuing interest
- Effect Size Magnification (ESM): understanding what it is
- Why ESM is of regulatory interest
- Stata’s `-emagnification-` command : An epidemiological example
- ESM as “Type M Error” (Gelman and Carlin, 2014)
- Other Stata code of interest

Background (or where this began)

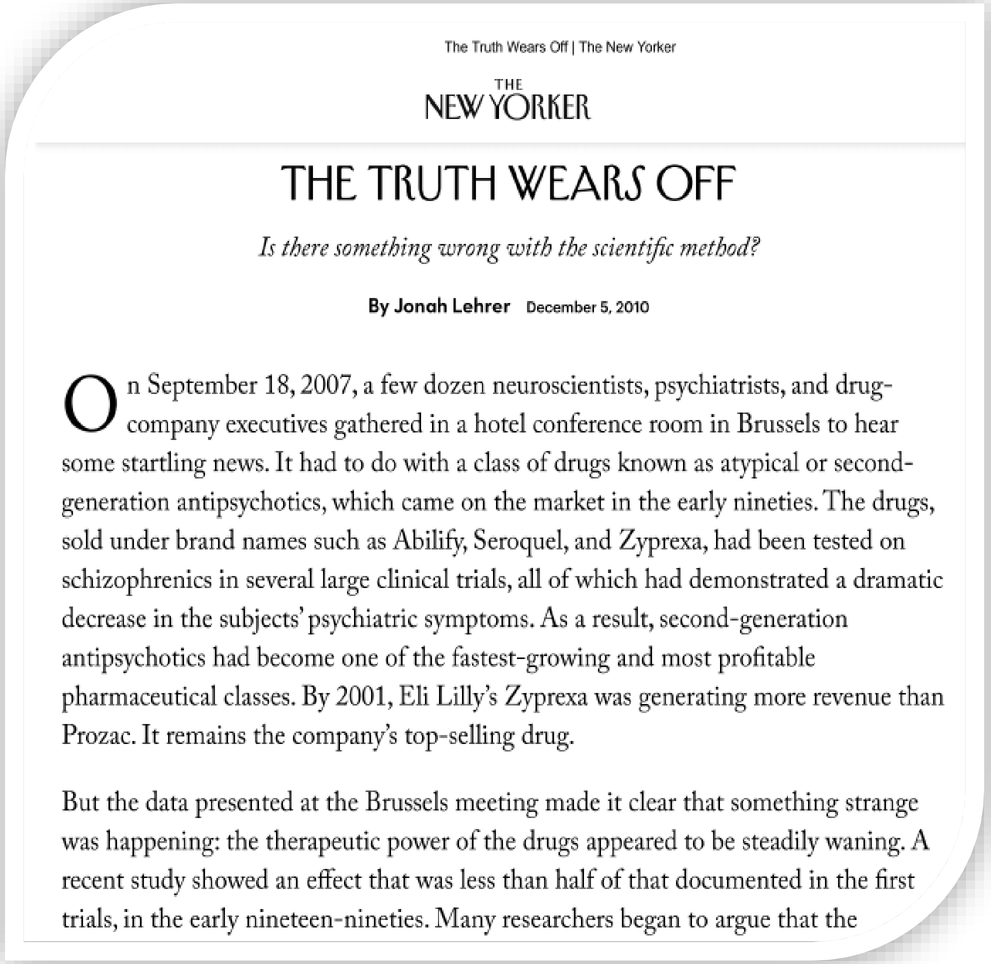
- There is increasing interest and concern in the scientific community in recent years on the “replication crisis” in science.
  - Specifically, scientists are finding that the result from scientific experiments can be difficult to reliably replicate on subsequent investigations.
    - Some have gone so far as to assert and provide support for a contention that most published research findings are false (Ioannidis, 2005).
    - Others have pointed out that even the more modest goal of reproducing previous research – demonstrating that others can calculate using the same data and methods – is frequently difficult or impossible (ASA 2017).

- Several ideas have been advanced with respect to the reasons for this increased difficulty in replicating scientific results
  - “vibrational effects”, which develop from the multitude of choices in the way the data are analyzed;
  - increased pressures to publish;
  - publication bias;
  - small power and the prevalence of and emphasis in research on null-hypothesis-significance-testing.

Background (or where this began)

the prelude

- New Yorker article “The Truth Wears Off… Is there something wrong with the Scientific Method?”
  - published in 2010
- Discusses declining effect sizes over time
  - Psychiatric Drugs (2^{nd} generation antipsychotics)
  - Psychological Testing (verbal overshadowing, ESP)
  - Evolutionary Biology/Ecology (fluctuating asymmetry)
- Referred to as “Decline Effect”
  - “Cosmic Habituation”



The Truth Wears Off | The New Yorker

THE NEW YORKER

THE TRUTH WEARS OFF

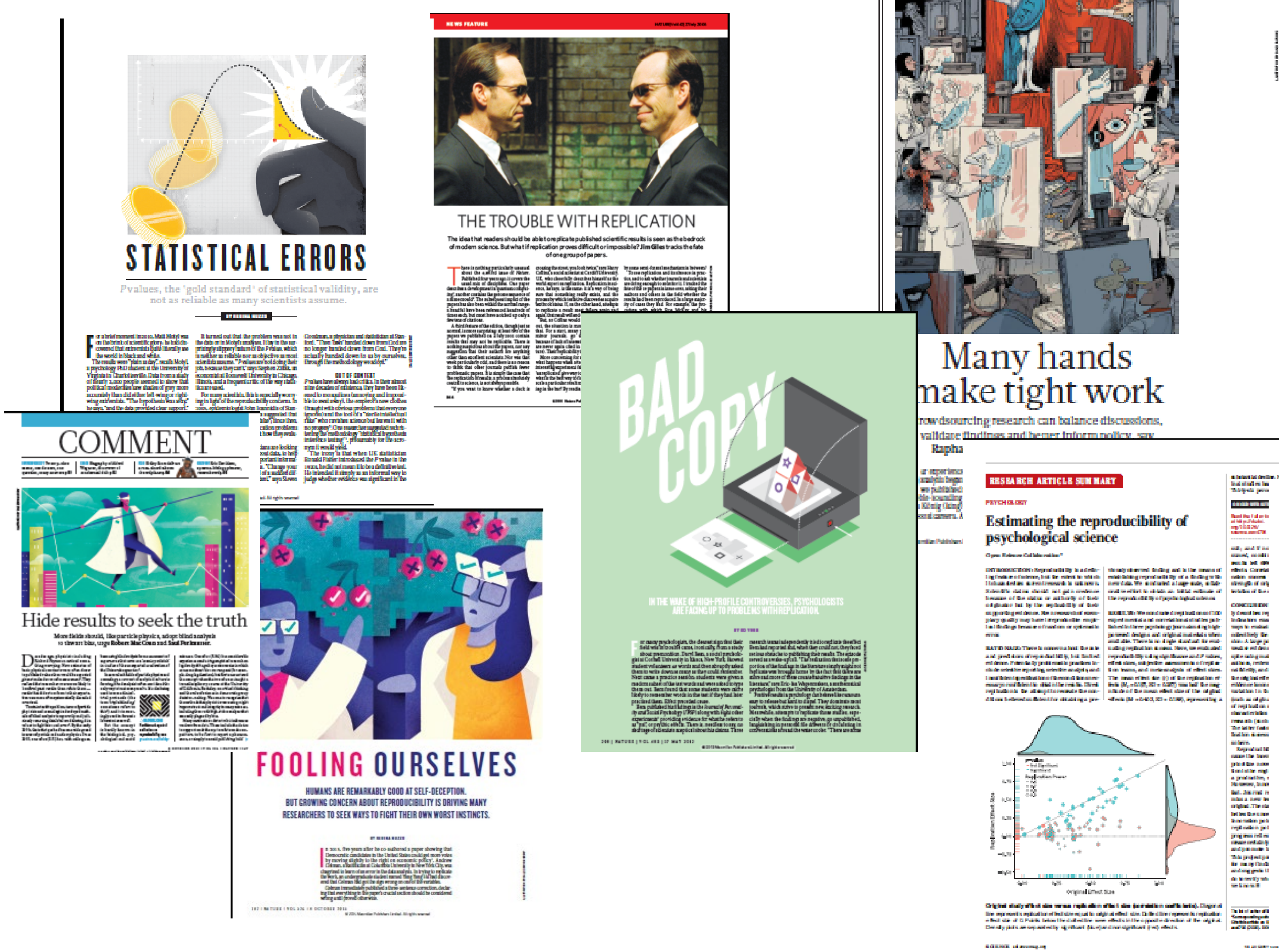
Is there something wrong with the scientific method?

By Jonah Lehrer December 5, 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

Reproducibility and Reliability… continuing interest



STATISTICAL ERRORS

P values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume.

NEWS FEATURE

THE TROUBLE WITH REPLICATION

The idea that readers should be able to replicate published scientific results is seen as the bedrock of modern science. But what if replication proves difficult or impossible? Jim Giles tracks the fate of one group of papers.

COMMENT

Hide results to seek the truth

More fields should, like particle physics, adopt blind analysis to thwart bias, urge Robert MacCoun and Saul Perlmutter.

BAD COPY

IN THE WAKE OF HIGH-PROFILE CONTROVERSIES, PSYCHOLOGISTS ARE FACING UP TO PROBLEMS WITH REPLICATION.

FOOLING OURSELVES

HUMANS ARE REMARKABLY GOOD AT SELF-DECEPTION. BUT GROWING CONCERN ABOUT REPRODUCIBILITY IS DRIVING MANY RESEARCHERS TO SEEK WAYS TO FIGHT THEIR OWN WORST INSTINCTS.

COMMENT

Many hands make tight work

Crowdsourcing research can balance discussions, validate findings and better inform policy, say Rapha

RESEARCH ARTICLE SUMMARY

PSYCHOLOGY

Estimating the reproducibility of psychological science

Open Science Collaboration*

Power failure: why small sample size undermines the reliability of neuroscience

Katherine S. Button, John P. A. Ioannidis, Claire Mokrysz, Brian A. Nosek, Jonathan Flint, Emma S. J. Robinson and Marcus R. Munafò

Abstract | A study with low statistical power has a reduced chance of detecting a true effect, but it is less well appreciated that low power also reduces the likelihood that a statistically significant result reflects a true effect. Here, we show that the average statistical power of studies in the neurosciences is very low. The consequences of this include overestimates of effect size and low reproducibility of results. There are also ethical dimensions to this problem, as unreliable research is inefficient and wasteful. Improving reproducibility in neuroscience is a key priority and requires attention to well-established but often ignored methodological principles.

COMMENT

A tragedy of errors

Mistakes in peer-reviewed papers are easy to find but hard to fix, report David B. Allison and colleagues.

Essay

Most Published Research Findings Are False—But a Little Replication Goes a Long Way

ToxSci Advance Access published March 19, 2015

SOT Society of Toxicology

Raising the Bar for Reproducible Science at the U.S. Environmental Protection Agency Office of Research and Development

Neuroscience Sifting the evidence

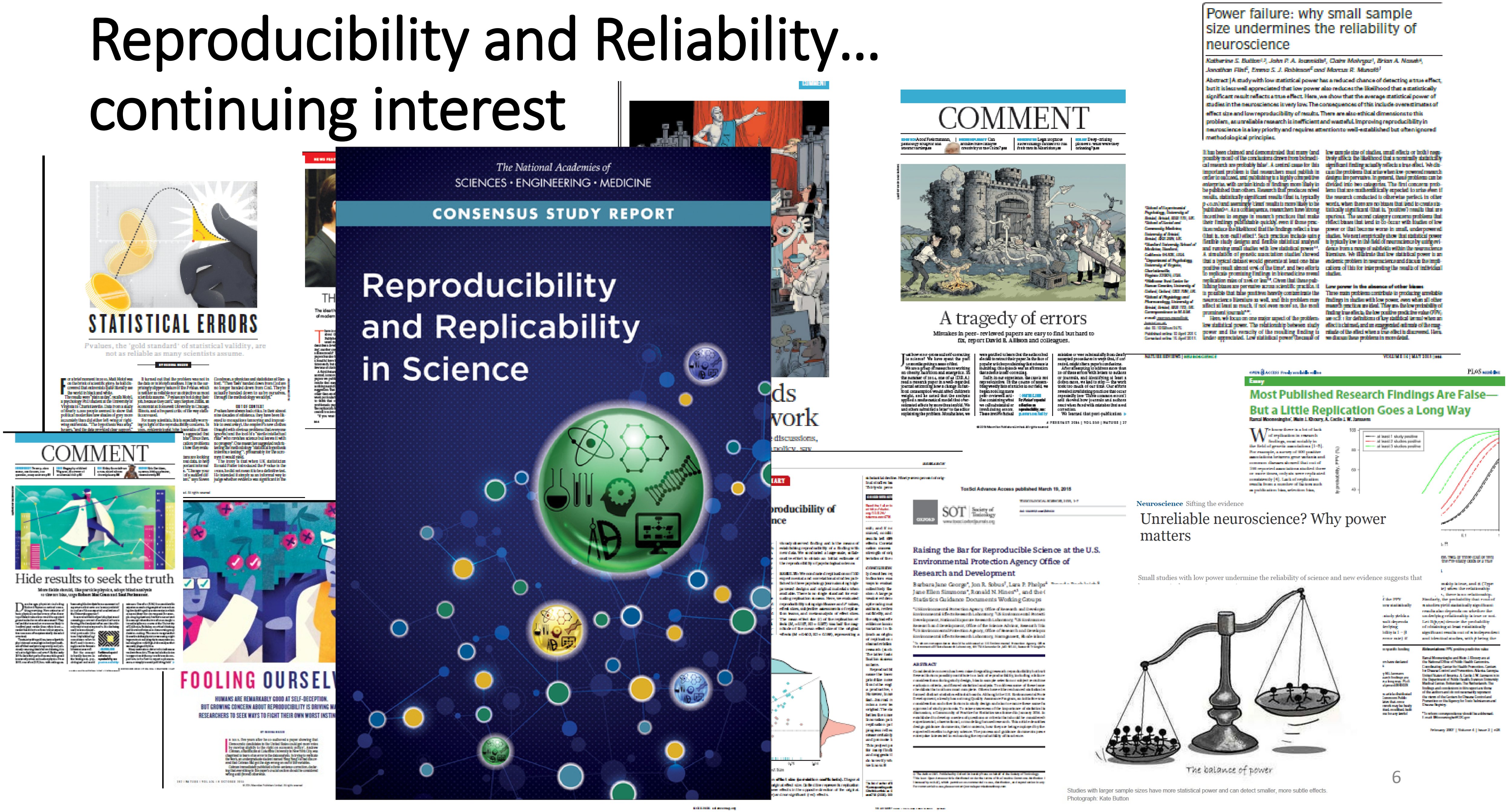
Unreliable neuroscience? Why power matters

Small studies with low power undermine the reliability of science and new evidence suggests that

The balance of power

Studies with larger sample sizes have more statistical power and can detect smaller, more subtle effects.
Photograph: Kate Button

Reproducibility and Reliability… continuing interest



The National Academies of SCIENCES · ENGINEERING · MEDICINE

CONSENSUS STUDY REPORT

Reproducibility and Replicability in Science

STATISTICAL ERRORS

P values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume.

COMMENT

Hide results to seek the truth

More fields should, like particle physics, adopt blind analysis to thwart bias, urge Robert MacCoun and Saul Perlmutter.

FOOLING OURSELVES

HUMANS ARE REMARKABLY GOOD AT SELF-DECEPTION. BUT GROWING CONCERN ABOUT REPRODUCIBILITY IS DRIVING MANY RESEARCHERS TO SEEK WAYS TO FIGHT THEIR OWN WORST INSTINCTS.

COMMENT

A tragedy of errors

Mistakes in peer-reviewed papers are easy to find but hard to fix, report David B. Allison and colleagues.

Power failure: why small sample size undermines the reliability of neuroscience

Katherine S. Button, John P. A. Ioannidis, Claire Mokrysz, Brian A. Nosek, Jonathan Flint, Emma S. J. Robinson and Marcus R. Munafò

Abstract | A study with low statistical power has a reduced chance of detecting a true effect, but it is less well appreciated that low power also reduces the likelihood that a statistically significant result reflects a true effect. Here, we show that the average statistical power of studies in the neurosciences is very low. The consequences of this include overestimates of effect size and low reproducibility of results. There are also ethical dimensions to this problem, as unreliable research is inefficient and wasteful. Improving reproducibility in neuroscience is a key priority and requires attention to well-established but often ignored methodological principles.

Essay

Most Published Research Findings Are False—But a Little Replication Goes a Long Way

ToxSci Advance Access published March 19, 2015

SOT Society of Toxicology

Raising the Bar for Reproducible Science at the U.S. Environmental Protection Agency Office of Research and Development

Barbara Jane George, Jon R. Sobus, Lara P. Phelps, Jane Ellen Simmons, Ronald N. Hines, and the Statistics Guidance Documents Working Groups

Neuroscience Sifting the evidence

Unreliable neuroscience? Why power matters

Small studies with low power undermine the reliability of science and new evidence suggests that

The balance of power

Studies with larger sample sizes have more statistical power and can detect smaller, more subtle effects. Photograph: Kate Button

Reproducibility and Reliability, continuing interest

STATISTICAL ERRORS

P values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume.

COMMENT

Hide results to seek the truth

FOOLING OURSELVES

HUMANS ARE REMARKABLY GOOD AT SELF-DECEPTION. BUT GROWING CONCERN ABOUT REPRODUCIBILITY IS DRIVING MANY RESEARCHERS TO SEEK WAYS TO FIGHT THEIR OWN WORST INSTINCTS.

The National Academies of SCIENCES · ENGINEERING · MEDICINE

CONSENSUS STUDY REPORT

Reproducibility and Replicability in Science

Public Symposium: Reproducibility and Replicability in Science

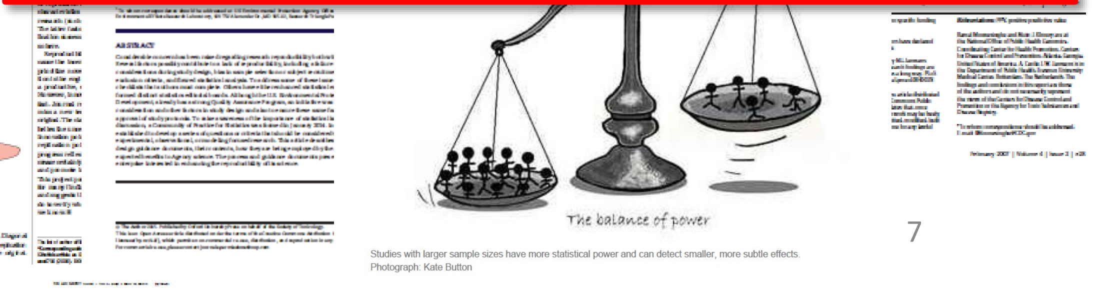
September 24, 2019

National Academy of Sciences, Engineering, and Medicine
Lecture Room
2101 Constitution Avenue NW
Washington, DC
Available by webinar.

See http://sites.nationalacademies.org/sites/reproducibility-in-science/index.htm

Agenda available at http://sites.nationalacademies.org/cs/groups/sitessite/documents/webpage/sites_194816.pdf

Download free PDF of report from https://www.nap.edu/catalog/25303/reproducibility-and-replicability-in-science



Studies with larger sample sizes have more statistical power and can detect smaller, more subtle effects. Photograph: Kate Button

Background (or where this began)

Open access, freely available online

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 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, 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

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.

Copyright: © 2005 John P. A. Ioannidis. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

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

Background (or where this began)

Open access, freely available online

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 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.

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

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, 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 $\mathrm{PPV} = (1 - \beta)R/(R - \beta R + \alpha)$. A research finding is thus

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

Copyright: © 2005 John P. A. Ioannidis. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

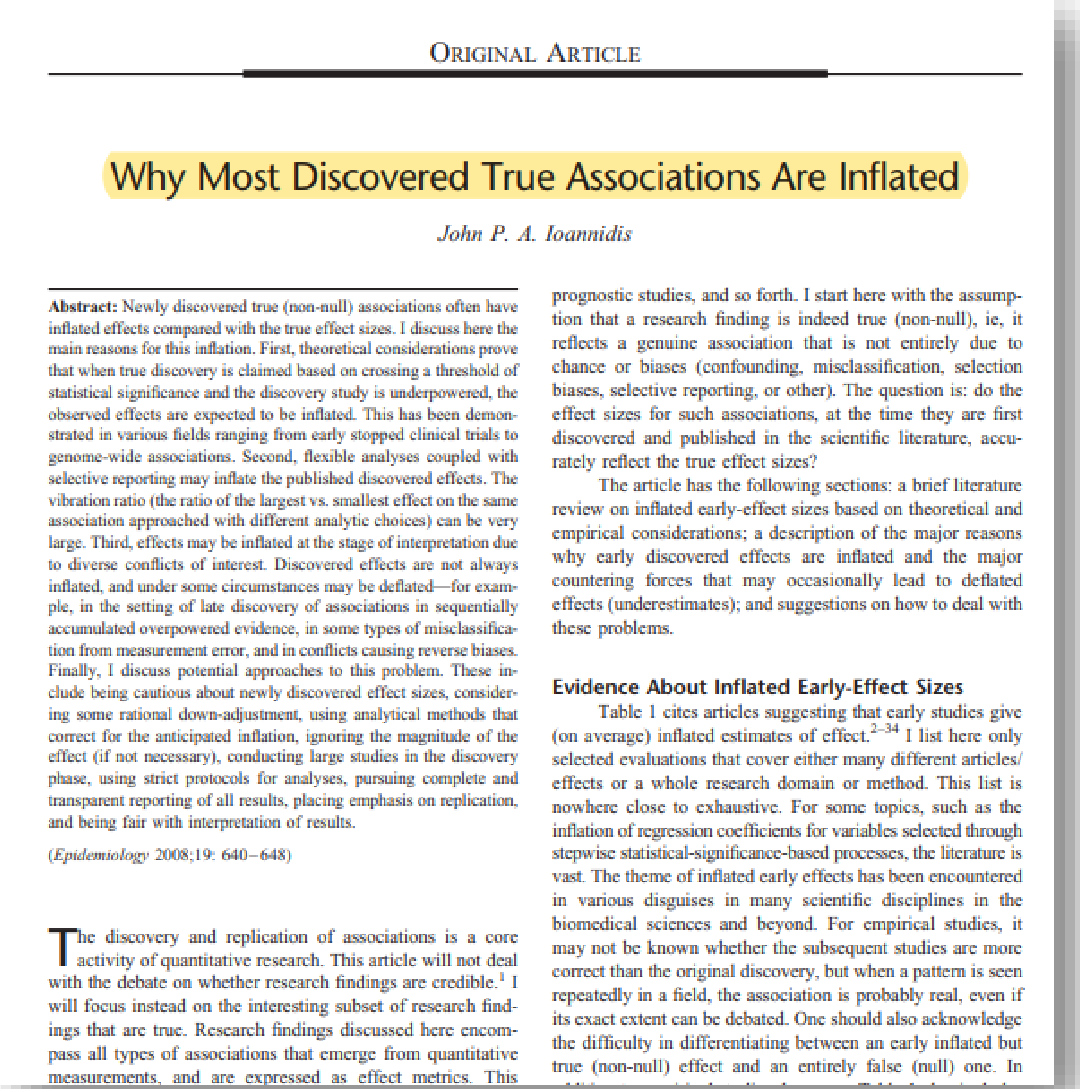
Abbreviation: PPV, positive predictive value

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

PLoS Medicine | www.plosmedicine.org 0696 August 2005 | Volume 2 | Issue 8 | e124



ORIGINAL ARTICLE

Why Most Discovered True Associations Are Inflated

John P. A. Ioannidis

Abstract: Newly discovered true (non-null) associations often have inflated effects compared with the true effect sizes. I discuss here the main reasons for this inflation. First, theoretical considerations prove that when true discovery is claimed based on crossing a threshold of statistical significance and the discovery study is underpowered, the observed effects are expected to be inflated. This has been demonstrated in various fields ranging from early stopped clinical trials to genome-wide associations. Second, flexible analyses coupled with selective reporting may inflate the published discovered effects. The vibration ratio (the ratio of the largest vs. smallest effect on the same association approached with different analytic choices) can be very large. Third, effects may be inflated at the stage of interpretation due to diverse conflicts of interest. Discovered effects are not always inflated, and under some circumstances may be deflated—for example, in the setting of late discovery of associations in sequentially accumulated overpowered evidence, in some types of misclassification from measurement error, and in conflicts causing reverse biases. Finally, I discuss potential approaches to this problem. These include being cautious about newly discovered effect sizes, considering some rational down-adjustment, using analytical methods that correct for the anticipated inflation, ignoring the magnitude of the effect (if not necessary), conducting large studies in the discovery phase, using strict protocols for analyses, pursuing complete and transparent reporting of all results, placing emphasis on replication, and being fair with interpretation of results.

(*Epidemiology* 2008;19: 640–648)

The discovery and replication of associations is a core activity of quantitative research. This article will not deal with the debate on whether research findings are credible.[1] I will focus instead on the interesting subset of research findings that are true. Research findings discussed here encompass all types of associations that emerge from quantitative measurements, and are expressed as effect metrics. This

prognostic studies, and so forth. I start here with the assumption that a research finding is indeed true (non-null), ie, it reflects a genuine association that is not entirely due to chance or biases (confounding, misclassification, selection biases, selective reporting, or other). The question is: do the effect sizes for such associations, at the time they are first discovered and published in the scientific literature, accurately reflect the true effect sizes?

The article has the following sections: a brief literature review on inflated early-effect sizes based on theoretical and empirical considerations; a description of the major reasons why early discovered effects are inflated and the major countering forces that may occasionally lead to deflated effects (underestimates); and suggestions on how to deal with these problems.

Evidence About Inflated Early-Effect Sizes

Table 1 cites articles suggesting that early studies give (on average) inflated estimates of effect.[2–34] I list here only selected evaluations that cover either many different articles/effects or a whole research domain or method. This list is nowhere close to exhaustive. For some topics, such as the inflation of regression coefficients for variables selected through stepwise statistical-significance-based processes, the literature is vast. The theme of inflated early effects has been encountered in various disguises in many scientific disciplines in the biomedical sciences and beyond. For empirical studies, it may not be known whether the subsequent studies are more correct than the original discovery, but when a pattern is seen repeatedly in a field, the association is probably real, even if its exact extent can be debated. One should also acknowledge the difficulty in differentiating between an early inflated but true (non-null) effect and an entirely false (null) one. In

Effect Size Magnification: *What it is.*

- Effect size magnification (ESM) refers to the phenomenon that low-powered studies that find evidence of an effect often provide inflated estimates of the size of that effect

Effect Size Magnification: *What it is.*

Conduct experiment/observational study today

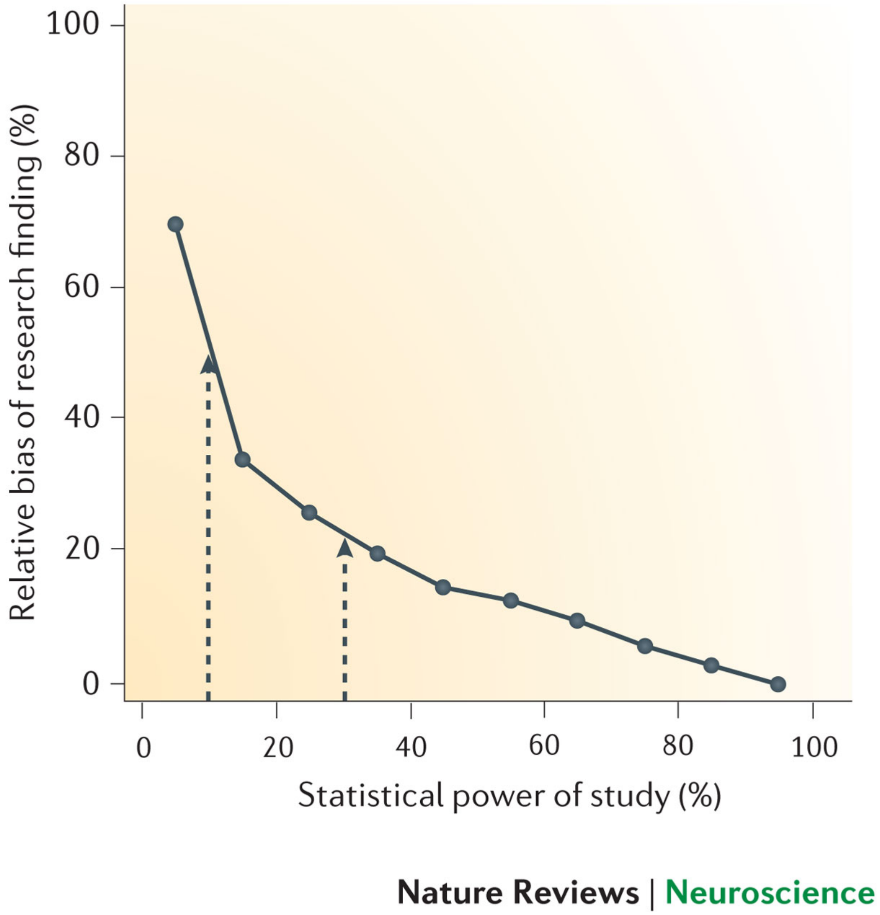
Discover a statistically significant effect size of importance

Repeat the study again tomorrow because you discovered an statistically significant effect size of interest and … effect size diminishes

- Effect size magnification (ESM) refers to the phenomenon that low-powered studies that find evidence of an effect often provide inflated estimates of the size of that effect

… so that when that study is repeated (US NAS term: “replicated”)**, the observed effect size is likely to decline**

Effect Size Magnification: *What it is.*



- Effect size magnification (ESM) refers to the phenomenon that low-powered studies that find evidence of an effect often provide inflated estimates of the size of that effect

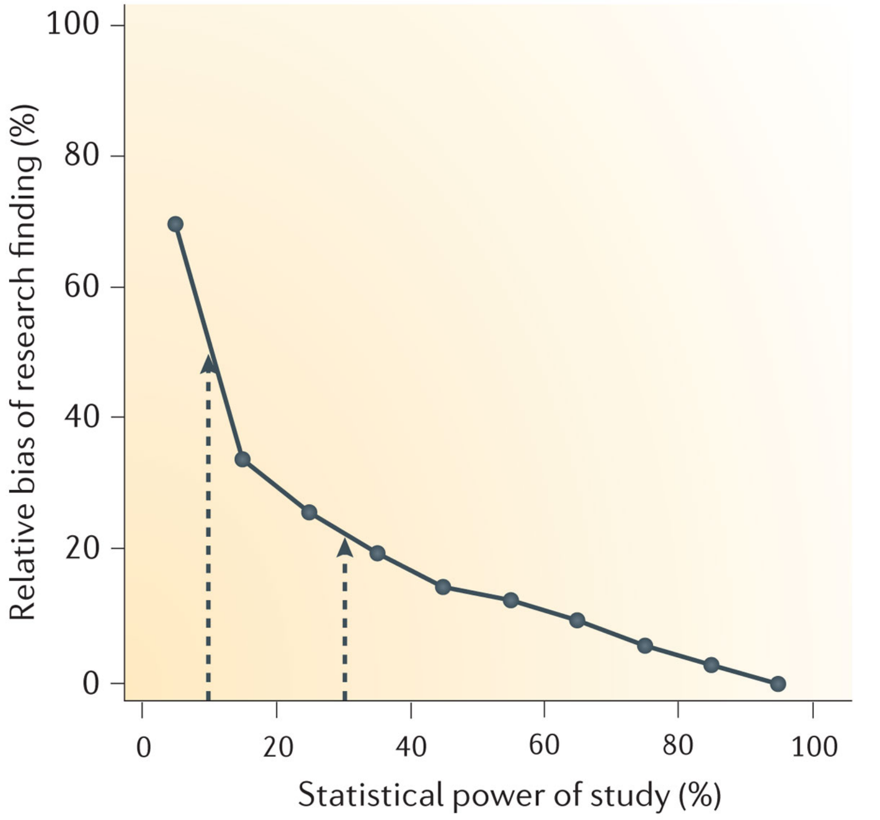
… so that when that study is repeated (US NAS term: "replicated")**, the observed effect size is likely to decline**

…degree of decline (amount of ESM) is inversely related to power

- Sample size
- True Effect Size
- Background or Control Rate

From: http://www.nature.com/nrn/journal/v14/n5/fig_tab/nrn3475_F5.html

Effect Size Magnification: *What it is.*



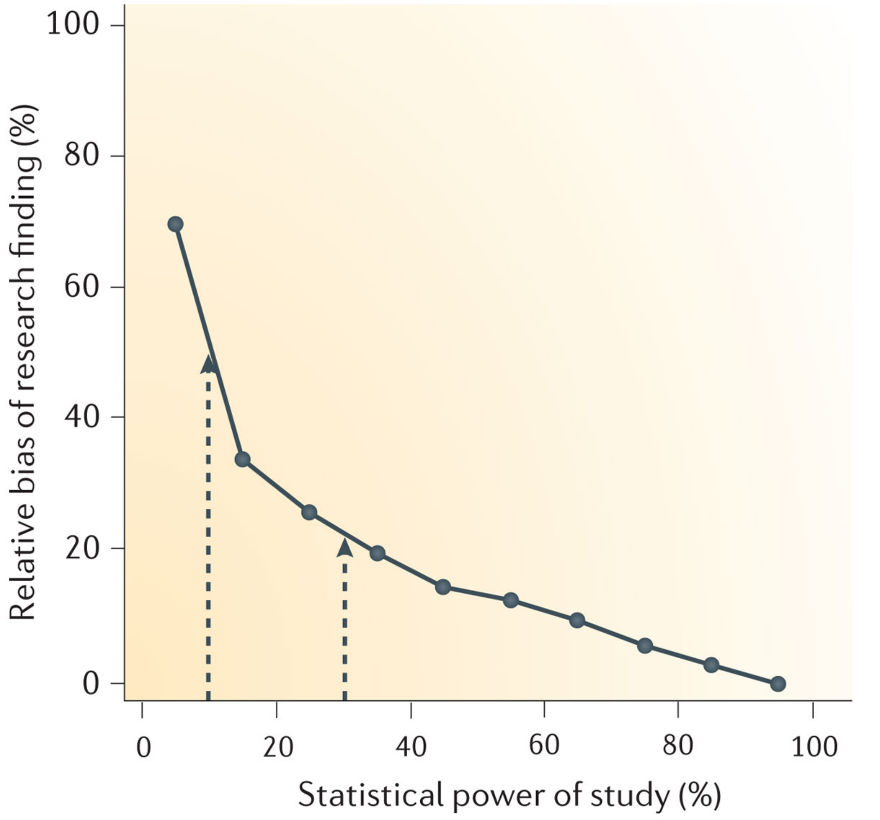
Nature Reviews | Neuroscience

Key Points

- ESM is expected when an effect has to pass a certain threshold — such as reaching statistical significance — in order for it to have been 'discovered'.
- ESM is worst for small, low-powered studies, which can only detect effects that happen to be large.
  - In practice, this means that research findings of small studies are biased in favor of finding inflated effects.
- While most researchers recognize issues associated with small/low powered studies *vis-a-vis* the failure to detect true effects, fewer recognize issues associated with small/low powered studies and their tendency to produce inflated estimates.

From: http://www.nature.com/nrn/journal/v14/n5/fig_tab/nrn3475_F5.html

Effect Size Magnification: *What it is.*



Nature Reviews | Neuroscience

Key Points

- ESM is expected when an effect has to pass a certain threshold — such as reaching statistical significance — in order for it to have been 'discovered’.
- ESM is worst for small, low-powered studies, which can only detect effects that happen to be large.
    - In practice, this means that research findings of small studies are biased in favor of finding inflated effects.
- While most researchers recognize issues associated with small/low powered studies *vis-a-vis* the failure to detect true effects, fewer recognize issues associated with small/low powered studies and their tendency to produce inflated estimates.

From: http://www.nature.com/nrn/journal/v14/n5/fig_tab/nrn3475_F5.html

A simulated numerical illustration of ESM…

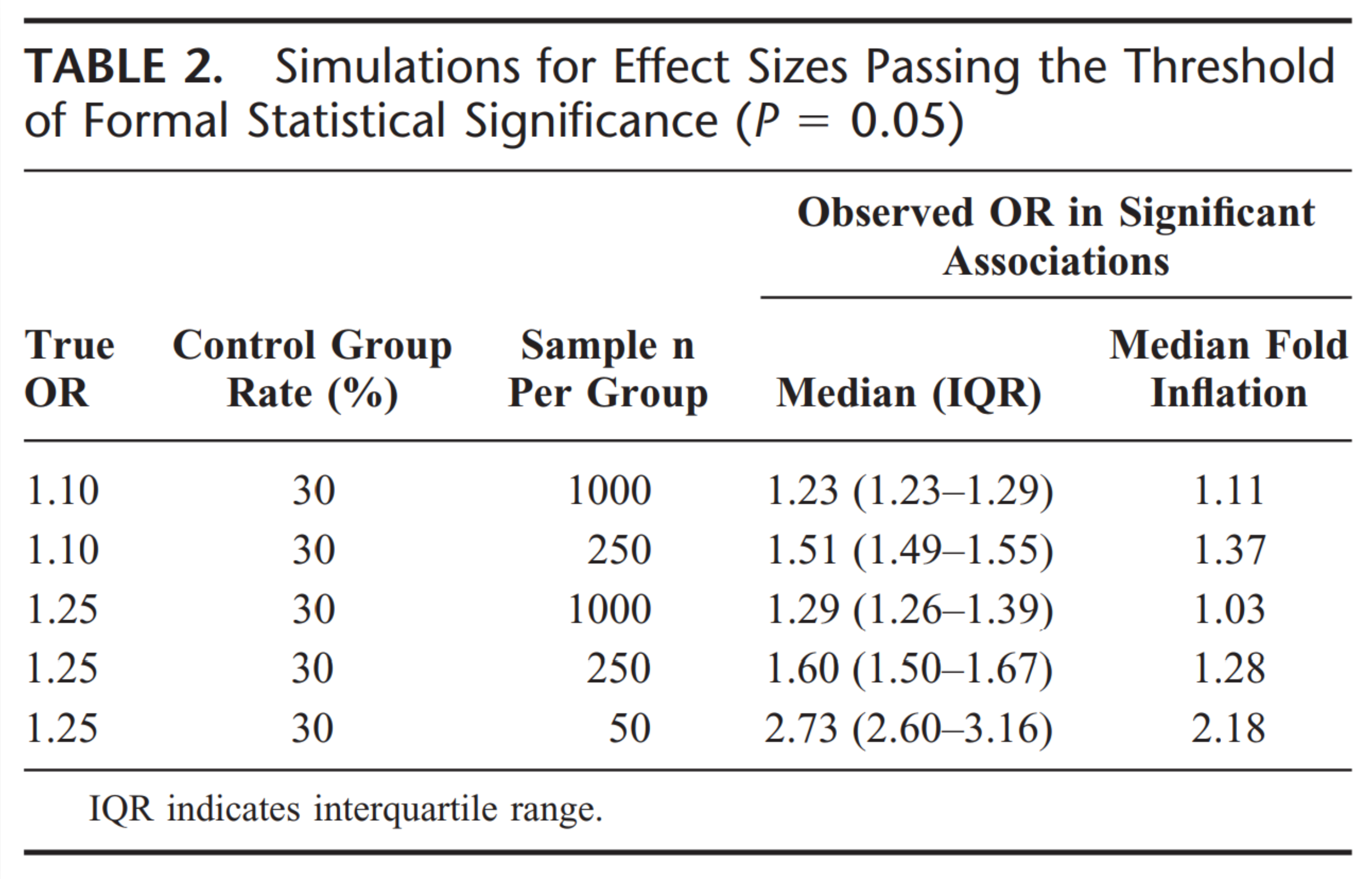
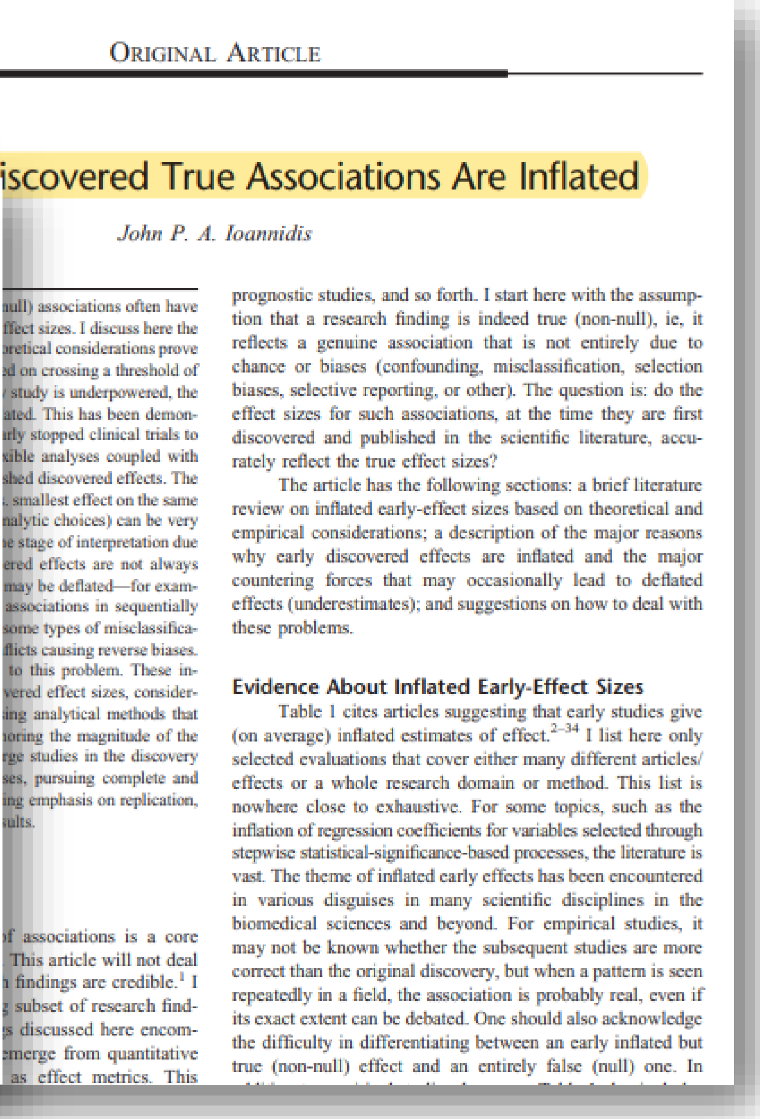


TABLE 2. Simulations for Effect Sizes Passing the Threshold of Formal Statistical Significance (*P* = 0.05)

			Observed OR in Significant Associations	
True OR	Control Group Rate (%)	Sample n Per Group	Median (IQR)	Median Fold Inflation
1.10	30	1000	1.23 (1.23–1.29)	1.11
1.10	30	250	1.51 (1.49–1.55)	1.37
1.25	30	1000	1.29 (1.26–1.39)	1.03
1.25	30	250	1.60 (1.50–1.67)	1.28
1.25	30	50	2.73 (2.60–3.16)	2.18

IQR indicates interquartile range.



ORIGINAL ARTICLE

Why Most Discovered True Associations Are Inflated

John P. A. Ioannidis

null) associations often have
ffect sizes. I discuss here the
oretical considerations prove
ed on crossing a threshold of
study is underpowered, the
ated. This has been demon-
arly stopped clinical trials to
xible analyses coupled with
shed discovered effects. The
. smallest effect on the same
nalytic choices) can be very
he stage of interpretation due
ered effects are not always
may be deflated—for exam-
associations in sequentially
some types of misclassifica-
flicts causing reverse biases.
to this problem. These in-
vered effect sizes, consider-
ing analytical methods that
noring the magnitude of the
rge studies in the discovery
ses, pursuing complete and
ing emphasis on replication,
sults.

of associations is a core
. This article will not deal
h findings are credible.[1] I
g subset of research find-
gs discussed here encom-
emerge from quantitative
as effect metrics. This

prognostic studies, and so forth. I start here with the assumption that a research finding is indeed true (non-null), ie, it reflects a genuine association that is not entirely due to chance or biases (confounding, misclassification, selection biases, selective reporting, or other). The question is: do the effect sizes for such associations, at the time they are first discovered and published in the scientific literature, accurately reflect the true effect sizes?

The article has the following sections: a brief literature review on inflated early-effect sizes based on theoretical and empirical considerations; a description of the major reasons why early discovered effects are inflated and the major countering forces that may occasionally lead to deflated effects (underestimates); and suggestions on how to deal with these problems.

Evidence About Inflated Early-Effect Sizes

Table 1 cites articles suggesting that early studies give (on average) inflated estimates of effect.[2–34] I list here only selected evaluations that cover either many different articles/effects or a whole research domain or method. This list is nowhere close to exhaustive. For some topics, such as the inflation of regression coefficients for variables selected through stepwise statistical-significance-based processes, the literature is vast. The theme of inflated early effects has been encountered in various disguises in many scientific disciplines in the biomedical sciences and beyond. For empirical studies, it may not be known whether the subsequent studies are more correct than the original discovery, but when a pattern is seen repeatedly in a field, the association is probably real, even if its exact extent can be debated. One should also acknowledge the difficulty in differentiating between an early inflated but true (non-null) effect and an entirely false (null) one. In

An simulated numerical illustration of ESM…

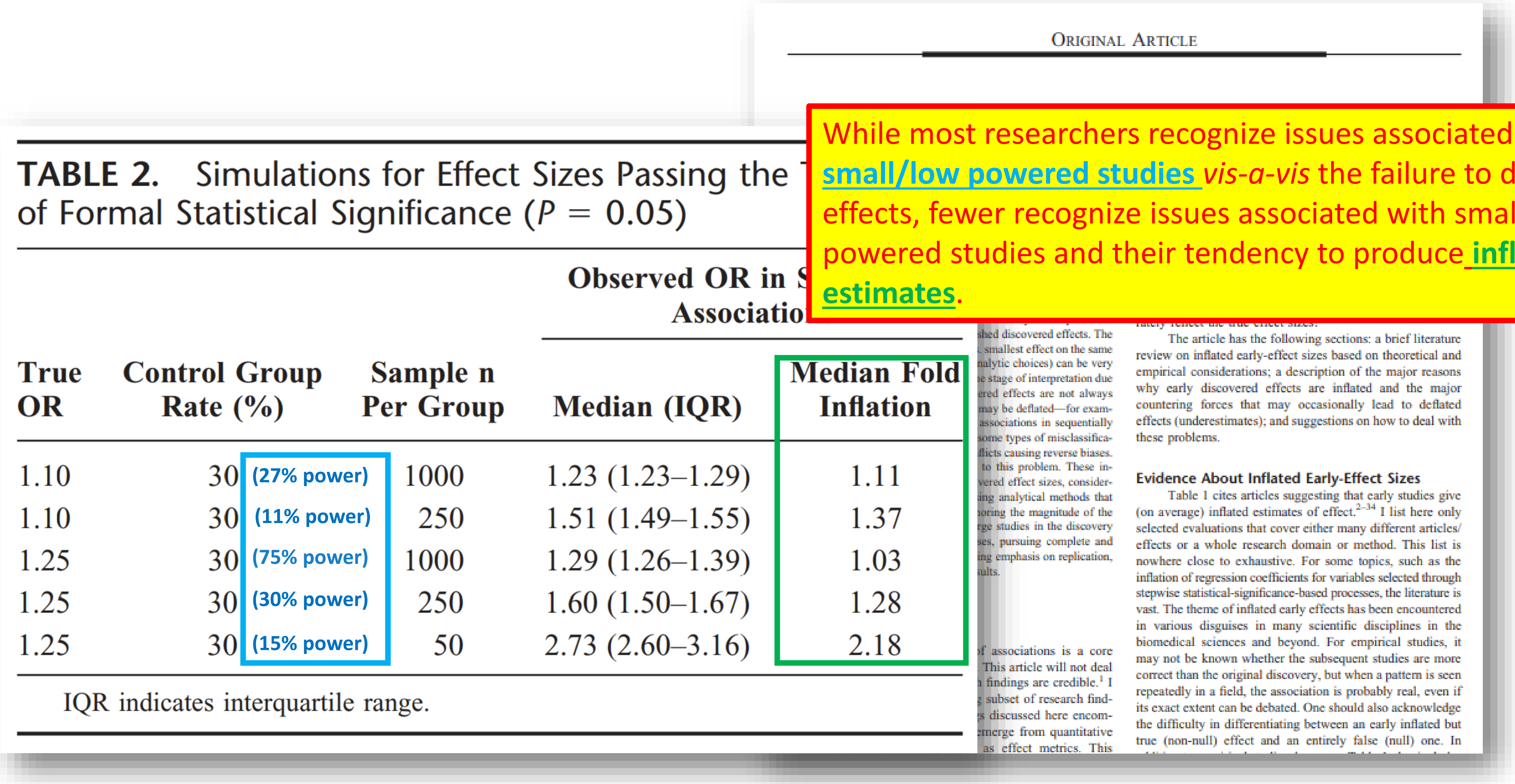


TABLE 2. Simulations for Effect Sizes Passing the T
of Formal Statistical Significance ($P = 0.05$)

				Observed OR in S Associatio	
True OR	Control Group Rate (%)		Sample n Per Group	Median (IQR)	Median Fold Inflation
1.10	30	(27% power)	1000	1.23 (1.23–1.29)	1.11
1.10	30	(11% power)	250	1.51 (1.49–1.55)	1.37
1.25	30	(75% power)	1000	1.29 (1.26–1.39)	1.03
1.25	30	(30% power)	250	1.60 (1.50–1.67)	1.28
1.25	30	(15% power)	50	2.73 (2.60–3.16)	2.18

IQR indicates interquartile range.

ORIGINAL ARTICLE

shed discovered effects. The . smallest effect on the same nalytic choices) can be very ne stage of interpretation due ered effects are not always may be deflated—for example associations in sequentially some types of misclassification flicts causing reverse biases. to this problem. These invered effect sizes, considering analytical methods that noring the magnitude of the rge studies in the discovery ses, pursuing complete and ing emphasis on replication, sults.

of associations is a core . This article will not deal h findings are credible.[1] I g subset of research findgs discussed here encomemerge from quantitative as effect metrics. This

rately reflect the true effect sizes.

The article has the following sections: a brief literature review on inflated early-effect sizes based on theoretical and empirical considerations; a description of the major reasons why early discovered effects are inflated and the major countering forces that may occasionally lead to deflated effects (underestimates); and suggestions on how to deal with these problems.

Evidence About Inflated Early-Effect Sizes

Table 1 cites articles suggesting that early studies give (on average) inflated estimates of effect.[2–34] I list here only selected evaluations that cover either many different articles/effects or a whole research domain or method. This list is nowhere close to exhaustive. For some topics, such as the inflation of regression coefficients for variables selected through stepwise statistical-significance-based processes, the literature is vast. The theme of inflated early effects has been encountered in various disguises in many scientific disciplines in the biomedical sciences and beyond. For empirical studies, it may not be known whether the subsequent studies are more correct than the original discovery, but when a pattern is seen repeatedly in a field, the association is probably real, even if its exact extent can be debated. One should also acknowledge the difficulty in differentiating between an early inflated but true (non-null) effect and an entirely false (null) one. In

While most researchers recognize issues associated with small/low powered studies *vis-a-vis* the failure to detect true effects, fewer recognize issues associated with small/low powered studies and their tendency to produce inflated estimates.

A simulated numerical illustration of ESM…

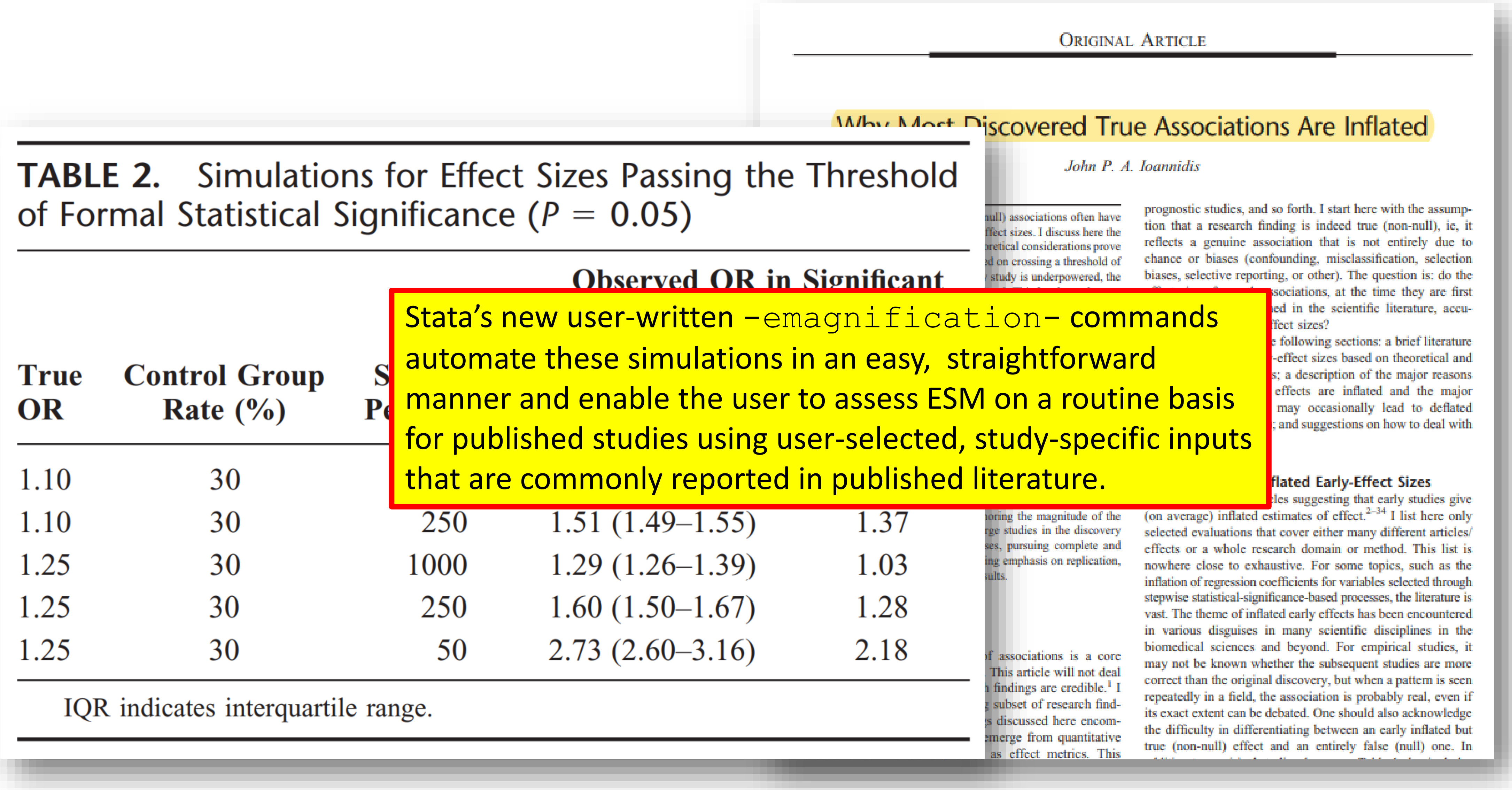


TABLE 2. Simulations for Effect Sizes Passing the Threshold of Formal Statistical Significance ($P = 0.05$)

True OR	Control Group Rate (%)	S Pe	Observed OR in Significant	
1.10	30			
1.10	30	250	1.51 (1.49–1.55)	1.37
1.25	30	1000	1.29 (1.26–1.39)	1.03
1.25	30	250	1.60 (1.50–1.67)	1.28
1.25	30	50	2.73 (2.60–3.16)	2.18

IQR indicates interquartile range.

Stata's new user-written –`emagnification`– commands automate these simulations in an easy, straightforward manner and enable the user to assess ESM on a routine basis for published studies using user-selected, study-specific inputs that are commonly reported in published literature.

ORIGINAL ARTICLE

Why Most Discovered True Associations Are Inflated

John P. A. Ioannidis

null) associations often have
ffect sizes. I discuss here the
oretical considerations prove
ed on crossing a threshold of
study is underpowered, the

oring the magnitude of the
rge studies in the discovery
ses, pursuing complete and
ing emphasis on replication,
sults.

of associations is a core
. This article will not deal
h findings are credible.[1] I
g subset of research find-
s discussed here encom-
emerge from quantitative
as effect metrics. This

prognostic studies, and so forth. I start here with the assumption that a research finding is indeed true (non-null), ie, it reflects a genuine association that is not entirely due to chance or biases (confounding, misclassification, selection biases, selective reporting, or other). The question is: do the
ssociations, at the time they are first
ed in the scientific literature, accu-
ffect sizes?
e following sections: a brief literature
-effect sizes based on theoretical and
s; a description of the major reasons
effects are inflated and the major
may occasionally lead to deflated
; and suggestions on how to deal with

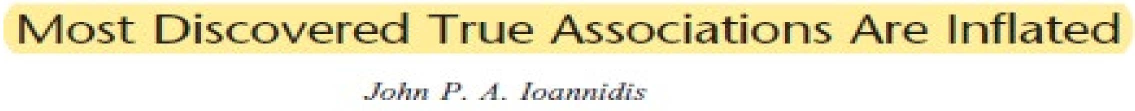
flated Early-Effect Sizes

les suggesting that early studies give (on average) inflated estimates of effect.[2–34] I list here only selected evaluations that cover either many different articles/effects or a whole research domain or method. This list is nowhere close to exhaustive. For some topics, such as the inflation of regression coefficients for variables selected through stepwise statistical-significance-based processes, the literature is vast. The theme of inflated early effects has been encountered in various disguises in many scientific disciplines in the biomedical sciences and beyond. For empirical studies, it may not be known whether the subsequent studies are more correct than the original discovery, but when a pattern is seen repeatedly in a field, the association is probably real, even if its exact extent can be debated. One should also acknowledge the difficulty in differentiating between an early inflated but true (non-null) effect and an entirely false (null) one. In

Why is ESM of regulatory interest?

- If the results of a study or studies of interest cannot -- in theory or practice -- be reliably replicated and might reflect systematically inflated effect sizes, how much confidence can we have in regulatory decisions that rely upon them?

- Statistical significance can play an important role in “eliminating chance as a potential explanation for study results”.
  - “Statistical *significance* testing (via the p-value) is the first-line defense against being fooled by randomness” [Y. Benjamini, 2017]

- If 

  Most Discovered True Associations Are Inflated

  John P. A. Ioannidis

  …. under what circumstances does this occur (why and when)?

  …and how do regulators know when this is happening, evaluate/consider it, and incorporate it into decision-making?

  e.g., “a statistically significant doubling of the lung cancer risk”

  “what is an adequate sample size”

  “how big is big [enough]?”

- Might inflated effect sizes from small studies be in part a reason for the reproducibility issues (“crisis”) being increasingly discussed in science?

Why is ESM of regulatory interest?

Can we - as regulators - ***understand***, ***reproduce***, and finally ***apply*** the ESM work to better understand (epidemiological) studies that are of potential regulatory interest?

Why is ESM of regulatory interest?

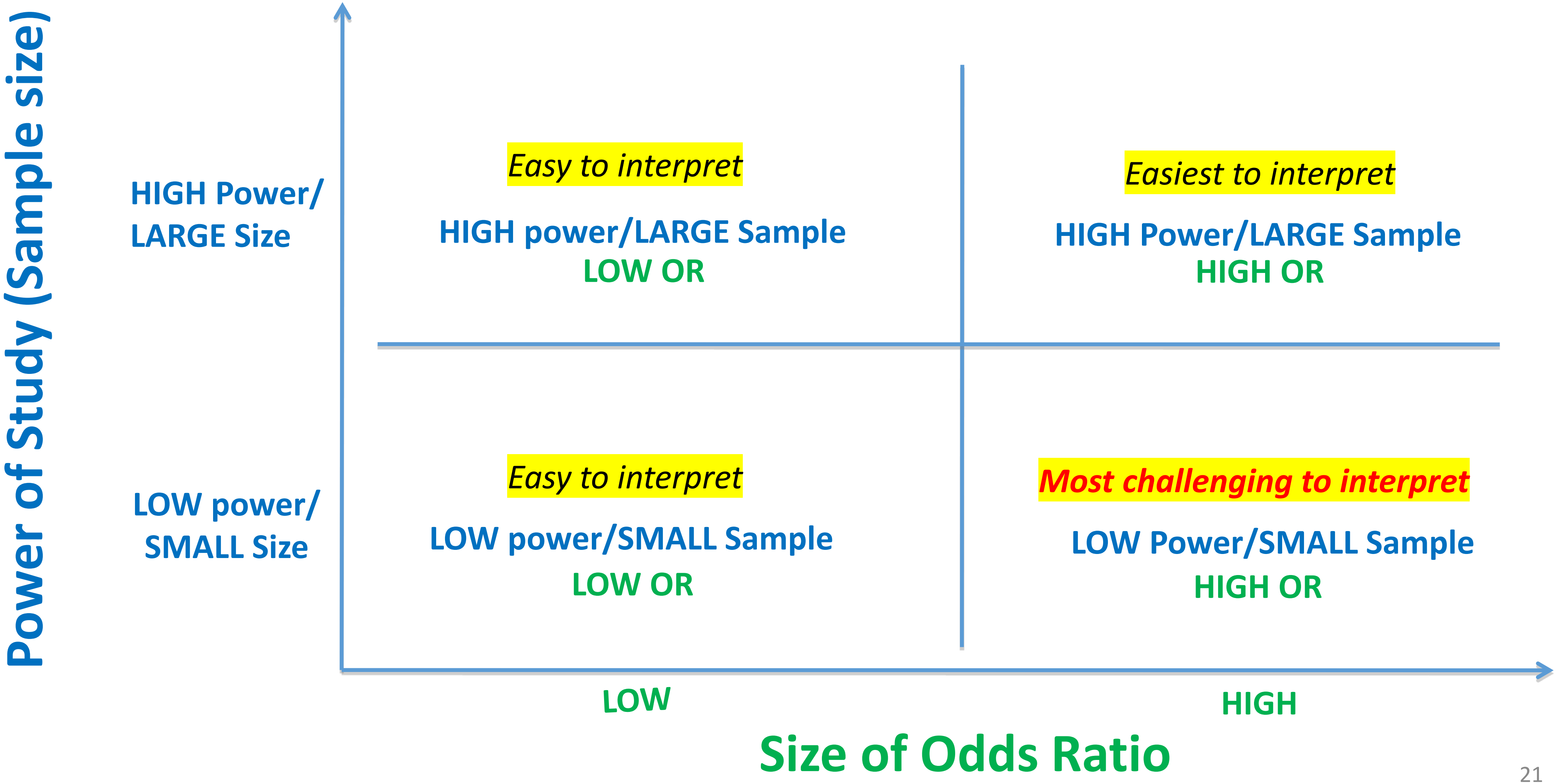
Can we - as regulators - ***understand***, ***reproduce***, and finally ***apply*** the ESM work to better understand (epidemiological) studies that are of potential regulatory interest?

-AND-

Can we use this to better evaluate the reliability of reported (statistically significant) effect sizes and put these into a fuller context with respect to potential implications for epidemiological study conclusions?

Why is ESM of regulatory interest?

Statistical Significant Results from High Quality Study:



Power of Study (Sample size)

	LOW	HIGH
HIGH Power/ LARGE Size	*Easy to interpret* HIGH power/LARGE Sample LOW OR	*Easiest to interpret* HIGH Power/LARGE Sample HIGH OR
LOW power/ SMALL Size	*Easy to interpret* LOW power/SMALL Sample LOW OR	***Most challenging to interpret*** LOW Power/SMALL Sample HIGH OR

Size of Odds Ratio

An Epidemiological Example

- An epidemiological example uses a case study example published by Greenland (1994)[1]
  - relevant to **case-control studies** using **odds ratios**[2]
- Greenland studied the rates of lung cancer deaths among cases and controls from occupational exposure to resins in a facility that assembled transformers.
  - **45 exposed cases**; **94 unexposed cases**; **257 exposed controls**; and **945 unexposed controls**.
    - Odds Ratio$_{\text{crude}}$ = 1.76; 95% CI: 1.20, 2.5

[1] The data is also provided in Rothman *et al.*'s *Modern Epidemiology*. See Table 19-1 (p. 349) in the third edition. It is used here by Rothman *et al.* to illustrate quantitative sensitivity analyses, **not** effect size inflation. Adjusted OR from original article is 1.72 (95% CI: 1.17, 2.52)

[2] Stata's –`emagnification`– command can also perform ESM simulations for cohort studies using Rate Ratios (see Working Paper at http://www.imm.ki.se/biostatistics/emagnification/ for an example)

An Epidemiological Example:

Setting this up in Stata

```
cci 45 94 257 945, woolf
```

```
                                                         Proportion
                 |   Exposed   Unexposed  |      Total     Exposed
-----------------+------------------------+------------------------
           Cases |        45          94  |        139       0.3237
        Controls |       257         945  |       1202       0.2138
-----------------+------------------------+------------------------
           Total |       302        1039  |       1341       0.2252
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
      Odds ratio |         1.760286       |    1.202457    2.576898 (Woolf)
 Attr. frac. ex. |         .4319106       |    .1683693    .6119365 (Woolf)
 Attr. frac. pop |         .1398272       |
                 +-------------------------------------------------
                               chi2(1) =     8.63  Pr>chi2 = 0.0033
```

QUESTION: To what extent might effect size inflation be important here if one were looking for a statistically significant result?

Sample size
True Effect Size
Background or Control Rate

Effect Size Magnification – essential inputs

- In order to determine the potential degree of effect size magnification for any given study, the reviewer needs to perform various “design effect” calculations. This, in turn, requires that we know four values:
  1. the number of subjects in the *reference* (or control) group
  2. the number of subjects in the *comparison* group
  3. the proportion of interest in the *reference* group;

     e.g., the proportion of **exposed** subjects in the control group for case-control studies
  4. a target value of interest to detect a difference of a given (pre-determined) size in a comparison of two groups (e.g., exposed vs. not exposed)

The first three listed values are provided in or must be obtained from the publication while the target value of interest (typically an OR or RR in epidemiology studies) is selected by the risk managers (and is ultimately a policy decision).

An Example

Resin Exposure and Lung Cancer

Here, we have:

i. the number of subjects in the (reference) control group = **1202**

945 non-exposed controls + 257 resin-exposed controls

ii. the number of subjects in the case group = **139**

94 non-exposed cases + 45 resin- exposed cases

iii. the number of resin exposed subjects in the (reference) control group = **257**

```
                                                    Proportion
                 |   Exposed   Unexposed  |      Total     Exposed
-----------------+------------------------+------------------------
           Cases |        45          94  |        139      0.3237
        Controls |       257         945  |       1202      0.2138
-----------------+------------------------+------------------------
           Total |       302        1039  |       1341      0.2252
```

```
. emagnification proportion, p0(`=257/1202') or(1.1 1.2 1.5 2.0 3.0) n0(1202) n1(139) pctile(25 50 75)
        ifactor(50) nsim(1000) level(0.05) onesided seed(123) log

Scenario 1: p0 = .21381032, or = 1.1, n0 = 1202, n1 = 139
Completed: 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%
Scenario 2: p0 = .21381032, or = 1.2, n0 = 1202, n1 = 139
Completed: 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%
Scenario 3: p0 = .21381032, or = 1.5, n0 = 1202, n1 = 139
Completed: 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%
Scena
Compl
Scena
Compl

The t
```

```
. emagnification proportion, p0(`=257/1202') or(1.1 1.2 1.5 2.0 3.0) n0(1202)
n1(139) pctile(10 50 90)ifactor(50) nsim(1000) level(0.05) onesided seed(123)
log
```

```
p0              p1    true_or     n0     n1    valid    power    p25     p50     p75      if_p50

.2138103     .230268        1.1   1202    139     1000     .147  1.450   1.508   1.593    1.371
.2138103    .2460507        1.2   1202    139     1000     .223  1.461   1.547   1.698    1.289
.2138103    .2897407        1.5   1202    139     1000     .658  1.508   1.653   1.847    1.102
.2138103    .3522961          2   1202    139     1000     .967  1.760   2.015   2.289    1.007
.2138103    .4493007          3   1202    139     1000        1  2.648   3.003   3.436    1.001
```

```
. emagnification proportion, p0(`=257/1202') or(1.1 1.2 1.5 2.0 3.0) n0(1202) n1(139) pctile(25 50 75)
        ifactor(50) nsim(1000) level(0.05) onesided seed(123) log

Scenario 1: p0 = .21381032, or = 1.1, n0 = 1202, n1 = 139
Completed: 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%
Scenario 2: p0 = .21381032, or = 1.2, n0 = 1202, n1 = 139
Completed: 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%
Scenario 3: p0 = .21381032, or = 1.5, n0 = 1202, n1 = 139
Completed: 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%
Scenario 4: p0 = .21381032, or = 2, n0 = 1202, n1 = 139
Completed: 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%
Scenario 5: p0 = .21381032, or = 3, n0 = 1202, n1 = 139
Completed: 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

The tests are one-sided with level = .05


p0              p1   true_or     n0     n1    valid    power    p25     p50     p75      if_p50

.2138103    .230268        1.1   1202    139     1000     .147  1.450   1.508   1.593    1.371
.2138103   .2460507        1.2   1202    139     1000     .223  1.461   1.547   1.698    1.289
.2138103   .2897407        1.5   1202    139     1000     .658  1.508   1.653   1.847    1.102
.2138103   .3522961          2   1202    139     1000     .967  1.760   2.015   2.289    1.007
.2138103   .4493007          3   1202    139     1000        1  2.648   3.003   3.436    1.001
```

Simulations for Effect Sizes Passing a Threshold of Formal Statistical Significance (p = 0.05) for Greenland *et al.* (1994) Epidemiology Study

			Observed OR in Significant Associations		
True OR	Control Group Rate, p_0 (%)	Sample n Per Group (n_0/n_1)	Median (10th-90th)[a]	Median Fold Inflation	
1.1	21.4	1202/139	1.508 (1.417– 1.684)	1.371	14% power
1.2	21.4	1202/139	1.547 (1.415– 1.833)	1.289	22% power
1.5	21.4	1202/139	1.653 (1.440– 2.044)	1.102	66% power
2	21.4	1202/139	2.015 (1.584– 2.560)	1.007	97% power
3	21.4	1202/139	3.003 (2.347– 3.810)	1.001	>99% power

[a] 10th to 90th indicates the 10th and 90th percentiles of the statistically significant results.

```
emagnification proportion, p0(`=257/1202') or(1.1 1.2 1.5 2.0 3.0) n0(1202) n1(139) pctile(25 50 75) ifactor(50) nsim(1000) level(0.05)
onesided seed(123) log
```

Simulations for Effect Sizes Passing a Threshold of Formal Statistical Significance ($p = 0.05$) for Greenland *et al.* (1994) Epidemiology Study

			Observed OR in Significant Associations		
True OR	Control Group Rate, p_0 (%)	Sample n Per Group (n_0/n_1)	Median (10th-90th)[a]	Median Fold Inflation	
1.1	21.4	1202/139	1.508 (1.417– 1.684)	1.371	14% power
1.2	21.4	1202/139	1.547 (1.415– 1.833)	1.289	22% power
1.5	21.4	1202/139	1.653 (1.440– 2.044)	1.102	66% power
			2.015 (1.584– 2.560)	1.007	97% power
			3.003 (2.347– 3.810)	1.001	>99% power

ificant results.

```
ema                                     75) ifactor(50) nsim(1000) level(0.05) onesided seed(123)
log
```

What does this mean?

Here, the authors "discovered" an odds ratio of 1.76 for an association between resin exposure and lung cancer.

…which the (low) power of the study suggests could be attributable to effect size inflation at a true OR of as low as 1.2 and for which power is only 22%

Simulations for Effect Sizes Passing a Threshold of Formal Statistical Significance (p = 0.05) for Greenland *et al.* (1994) Epidemiology Study

True OR	Control Group Rate, p_0 (%)	Sample n Per Group (n_0/n_1)	Observed OR in Significant Associations: Median (10^{th}-90^{th})[a]	Observed OR in Significant Associations: Median Fold Inflation	
1.1	21.4	1202/139	1.508 (1.417– 1.684)	1.371	14% power
1.2	21.4	1202/139	1.547 (1.415– 1.833)	1.289	22% power
1.5	21.4	1202/139	1.653 (1.440– 2.044)	1.102	66% power
			2.015 (		% power
			3.003 (		% power

ificant resul

ema

log

75) ifactor(50) nsim(1000) level(0.05) onesided seed(123)

<u>What does this mean?</u>

Here, the authors "discovered" an odds ratio of 1.76 for an association between resin exposure and lung cancer.

…which the (low) power of the study suggests could be attributable to effect size inflation at a true OR of as low as 1.2 and for which power is only 22%

<u>Thus:</u> Given the size (power) of the study, the "discovered" odds ratio of 1.76 would not be unexpected if the true odds ratio were in fact as low as 1.2.

Where else has this ESM approach appeared?

Design Calculations

(aka "Post-hoc design analysis" methods to evaluate effect magnification)

- Introduced conceptually by Gelman and Carlin (2014) as **Type M(agnitude)** and Type S(ign) errors but for *continuous* (not categorical) data. Recently expanded upon by Lu et al (2019)
  - ESM calculations introduced here can be considered "sister" calculations to these
- Gelman and Carlin's design calculations can inform a statistical data summary and are recommended when apparently strong (statistically significant) evidence for non-null effects has been found.
  - not 'What is the power of a test?', but instead the more relevant *post-hoc* 'What might be expected to happen in studies of this size?'.
- Further informs if interpretation of a statistically significant result can change drastically depending on the plausible size of the underlying effect
- **NOT** post-hoc power
  - See "**Yes, it makes sense to do design analysis ('power calculations') after the data have been collected"** at https://statmodeling.stat.columbia.edu/2017/03/03/yes-makes-sense-design-analysis-power-calculations-data-collected/ 3 March 2017

How can I download the **`-emagnification-`** command from Stata?

```
net install emagnification,
from(http://www.imm.ki.se/biostatistics/stata)
```

Where can I get additional information?

See KI working paper at: http://www.imm.ki.se/biostatistics/emagnification/

More Stata Code of potential interest for epidemiological studies:

- Klein, D. (2019). **RDESIGNI**: Stata module to perform design analysis. Statistical Software Components, Boston College Department of Economics. https://ideas.repec.org/c/boc/bocode/s458423.html
- Linden A. (2019). **RETRODESIGN**: Stata module for computing type-S (Sign) and type-M (Magnitude) errors. Statistical Software Components, Boston College Department of Economics. http://ideas.repec.org/c/boc/bocode/s458631.html
- Linden A, Mathur M. B., VanderWeele, T. J. (2018). **EVALUE**: Stata module for conducting sensitivity analyses for unmeasured confounding in observational studies. Statistical Software Components, Boston College Department of Economics. https://ideas.repec.org/c/boc/bocode/s458592.html
- Orsini, N., Bellocco, R., Bottai, M. and Greenland S. (2006). **EPISENS**: Stata module for Deterministic and probabilistic sensitivity analysis. Statistical Software Components, Boston College Department of Economics. Revised 14 March 2013. https://ideas.repec.org/c/boc/bocode/s456792.html

More Stata Code of potential interest for epidemiological studies:

- Klein, D. (2019). **RDESIGNI**: Stata module to perform design analysis. Statistical Software Components, Boston College Department of Economics. https://ideas.repec.org/c/boc/bocode/s45[illegible]2.html
- Linden, A. (2019). **RETRODESIGN**: Stata module for computing type-S (Sign) and type-M (Magnitude) errors. Statistical Software Components, Boston College Department of Economics. http://ideas.repec.org/c/boc/bocode/s458631.html

RDESIGNI and **RETRODESIGN** both perform *post-hoc* design analysis for *continuous* variables

- Linden, A., Mathur, M. B., VanderWeele, T. J. (2018). **EVALUE**: Stata module for conducting sensitivity analyses for unmeasured confounding in observational studies. Statistical Software Components, Boston College Department of Economics. https://ideas.repec.org/c/boc/bocode/s458592.html

EVALUE evaluates sensitivity of results to unmeasured confounding

- Orsini, N., Bellocco, R., Bottai, M. and Greenland S. (2006). **EPISENS**: Stata module for Deterministic and probabilistic sensitivity analysis. Statistical Software Components, Boston College Department of Economics. Revised 14 March 2015. https://ideas.repec.org/c/boc/bocode/s456792.html

EPISENS performs Quantitative Bias Analysis (QBA)

Take Home Messages –

1. **Effect Size Magnification refers to the phenomenon that studies that find evidence of an effect often provide inflated estimates of the size of that effect**
   - Occurs when studies have low power
   - Such magnification is expected when an effect has to pass a certain threshold — such as reaching statistical significance — in order for it to have been 'discovered'
2. **Many epi studies are under-powered to find low to moderate effects –**
   - Can lead to exaggerated or inflated effect size estimates if primary interest is in "discovered" effects
3. **If an epi study has low power, we must be suspect of 'large' or 'significant' ORs, since these values may be inflated**
   - Don't rely just on p-values, as these may only be meaningful/reliable in adequately powered studies
4. **If an epi study does have low power and a 'large' discovered odds ratio, then perform a *post-hoc* design calculation to assist in quantitatively evaluating how reliable the odds ratio estimate may be**
   - Such calculations can help calibrate (simultaneous) thinking around sample size and reported odds ratios in published research

Summing it up

What is of critical importance is to recognize that adequately powered studies are necessary to be able to have at least some minimal degree of confidence in the estimate of the effect size, particularly in "discovery" phases with effect sizes that are statistically significant

…and…

Design calculations (such as done by `-emagnification-`) can assist in determining if effect size magnification may be present and the extent to which it may be an issue or should be accounted for in interpretation of results.

Thank you !

UNITED STATES ENVIRONMENTAL PROTECTION AGENCY

Contact information:

David J. Miller CAPT|USPHS

Acting Chief, Toxicology and Epidemiology Branch

Health Effects Division

Office of Pesticide Programs

Email: miller.davidj@epa.gov

Additional Slides

emagnification:

a tool for estimating effect size magnification and performing design calculations in epidemiological studies

Abstract. Artificial effect size magnification (ESM) may occur in underpowered studies where effects are only reported because they or their associated p-value have passed some threshold. Ioannidis (2008) and Gelman and Carlin (2014) have suggested that the plausibility of findings for a specific study can be evaluated by computation of ESM, which requires statistical simulation. In this talk, we present a new Stata package called -emagnification- that allows straightforward implementation of such simulations in Stata. The commands automate these simulations for epidemiological studies and enable the user to assess ESM on a routine basis for published studies using user-selected, study-specific inputs that are commonly-reported in published literature. The intention of the package is to allow a wider community to use ESMs as a tool for evaluating the reliability of reported effect sizes and to put an observed statistically significant effect size into a fuller context with respect to potential implications for study conclusions.

Select References

- American Statistical Association. The ASA's Statement on p-Values: Context, Process, and Purpose. https://www.tandfonline.com/doi/full/10.1080/00031305.2016.1154108
- Button, K., J. P. A. Ioannidis, C. Mokrysz, B. A. Nosek, J. Flink, Emma S.J. Robinson, and M.R. Munafo. 2013. Power failure: why small sample size undermines the reliability of neuroscience. Nature Reviews Neuroscience 14: 365-376. [accessed 6 September 2017 at http://www.nature.com/nrn/journal/v14/n5/full/nrn3475.html]
- Button, K. 2013. "Unreliable neuroscience? Why power matters". The Guardian newspaper (UK). 10 April [Accessed 6 September 2017 at https://www.theguardian.com/science/sifting-the-evidence/2013/apr/10/unreliable-neuroscience-power-matters]
- Gelman, A. and J. Carlin. 2014. Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. Perspectives on Psychological Science. Vol 9(6): 641-651. [accessed 05 May 2018 at http://www.stat.columbia.edu/~gelman/research/published/retropower_final.pdf]
- Greenland, S., A. Salvan, D. H. Wegman, M. F. Hallock, and T. J. Smith. 1994. A case–control study of cancer mortality at a transformer-assembly facility. *International Archives of Occupational and Environmental Health* 66: 49–54.
- Halsey, Lewis g., Douglas Curan-Everett, Sarah L. Vowler, and Gordon B. Drummond (2015). The fickle P value generates irreproducible results. Nature Methods. 12(3): 179-185 [accessed 06 September 2018 at https://www.mathworks.com/matlabcentral/answers/uploaded_files/55204/The%20fickle%20P%20value%20generates%20irreproducible%20results.pdf]
- Ioannidis, J. P. A. 2005. Why most published research findings are false. PLoS Medicine 2(8). E124.doi:10.1371.pmed.0020124. [accessed 6 September 2017 at http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0020124]
- Ioannidis, J. P. A. 2008. Why most discovered true associations are inflated. Epidemiology 19: 640-648. [accessed 24 February 2018 at http://www.dcscience.net/ioannidis-associations-2008.pdf]
- Lash, Timothy, Lindsay J. Collin, and Miriam E. Van Dyke. The Replication Crisis in Epidemiology: Snowball, Snow Job, or Winter Solstice? *Current Epidemiology Reports* (published online 12 April 2018)
- Lehrer, J. 2010. "The Truth Wears Off: Is there something wrong with the scientific method". New Yorker. 13 December. [Accessed 6 September 2017 at http://www.newyorker.com/magazine/2010/12/13/the-truth-wears-off
- Rothman, KJ, Greenland, S and Lash, TL. Modern Epidemiology. 2008. 3rd ed. Lippincot, Williams, and Wilkins. Philadelphia.

It’s a recognized issue… by some

(but not necessarily well-publicized)

“**It is not sufficiently well understood that ‘significant’ findings from studies that are underpowered (with respect to the true effect size) are likely to produce wrong answers, both in terms of the direction and magnitude of the effect**. ..There is a range of evidence to demonstrate that it remains the case that too many small studies are done and preferentially published when “significant”. We suggest that one reason for the continuing lack of real movement on this problem is the historic focus on power as a lever for ensuring statistical significance, **with inadequate attention being paid to the difficulties of interpreting statistical significance in underpowered studies**.

Because insufficient attention has been paid to these issues, we believe that too many small studies are done and preferentially published when ‘significant’. **There is a common misconception that if you happen to obtain statistical significance with low power, then you have achieved a particularly impressive feat, obtaining scientific success under difficult conditions.”**

Gelman, Andrew and John Carlin (2014) Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. *Perspectives in Psychol. Sci.* 9(6): 641-651.

It’s a recognized issue… by some

(but not necessarily well-publicized)

“Focusing on the P value during statistical analysis is an entrenched culture. **The P value is often used without the realization that in most cases the statistical power of a study is too low for P to assist the interpretation of the data**. Among the many and varied reasons for a fearful and hidebound approach to statistical practice, a lack of understanding is prominent. A better understanding of why P is so unhelpful should encourage scientists to reduce their reliance on this misleading concept….

Although statistical power is a central element in reliability, it is often considered only when a test fails to demonstrate a real effect (such as a difference between groups): a ‘false negative’ result. Many scientists who are not statisticians do not realize that the power of a test is equally relevant when considering statistically significant results, that is, when the null hypothesis appears to be untenable. This is because the statistical power of the test dramatically affects our capacity to interpret the P value and thus the test result. **It may surprise many scientists to discover that interpreting a study result from its P value alone is spurious in all but the most highly powered designs. The reason for this is that unless statistical power is very high, the P value exhibits wide sample-to sample variability and thus does not reliably indicate the strength of evidence against the null hypothesis.**”

Halsey, Lewis g., Douglas Curan-Everett, Sarah L. Vowler, and Gordon B. Drummond (2015). The fickle P value generates irreproducible results. *Nature Methods. 12*(3): 179-185.

It’s a recognized issue… by some

(but not necessarily well-publicized)

“**In a scientific culture that focuses on statistically significant results [67], effects are more likely to be overestimated than underestimated whenever power is less than 100%,** as seen in one of the replication projects [48]… In that project, 82 of 99 studies showed a stronger effect size in the original study than in the replication study. This pattern is what should be expected if the original studies were selected because their results were statistically significant. On average, these studies’ results should be overestimates. … **By focusing on results that are statistically significant, null hypothesis significance testing has built a machine to overestimate the truth.** These pressures cause early studies to have inflated estimates, and then subsequent studies may use the inflated results as the target estimates when designing a replication study, leading to underpowered replication studies that falsely fail to demonstrate reproducibility. **One cannot rationally label the resulting poor reproducibility as a crisis; the accumulation of evidence is behaving exactly as expected.”**

Lash, Timothy, Lindsay J. Collin, and Miriam E. Van Dyke. The Replication Crisis in Epidemiology: Snowball, Snow Job, or Winter Solstice? *Current Epidemiology Reports* (published online 12 April 2018)

It’s a recognized issue… by some

(but not necessarily well-publicized)

- **John Ioannidis** on *Statistical Significance, Economics, and Replication*.

  http://www.econtalk.org/john-ioannidis-on-statistical-significance-economics-and-replication/

  Jan 22 2018 podcast

- **Andrew Gelman** on *Social Science, Small Samples, and the Garden of the Forking Paths*.

  http://www.econtalk.org/andrew-gelman-on-social-science-small-samples-and-the-garden-of-the-forking-paths/

  Mar 20 2017 podcast

- **Geoff Cumming** on Dance of the p-values

  https://www.bing.com/videos/search?q=dance+of+the+p+values&view=detail&mid=6D48A4D9F8A653BA10496D48A4D9F8A653BA1049&FORM=VIRE

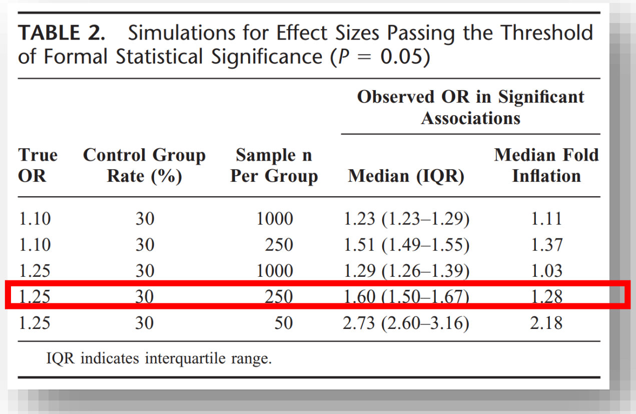
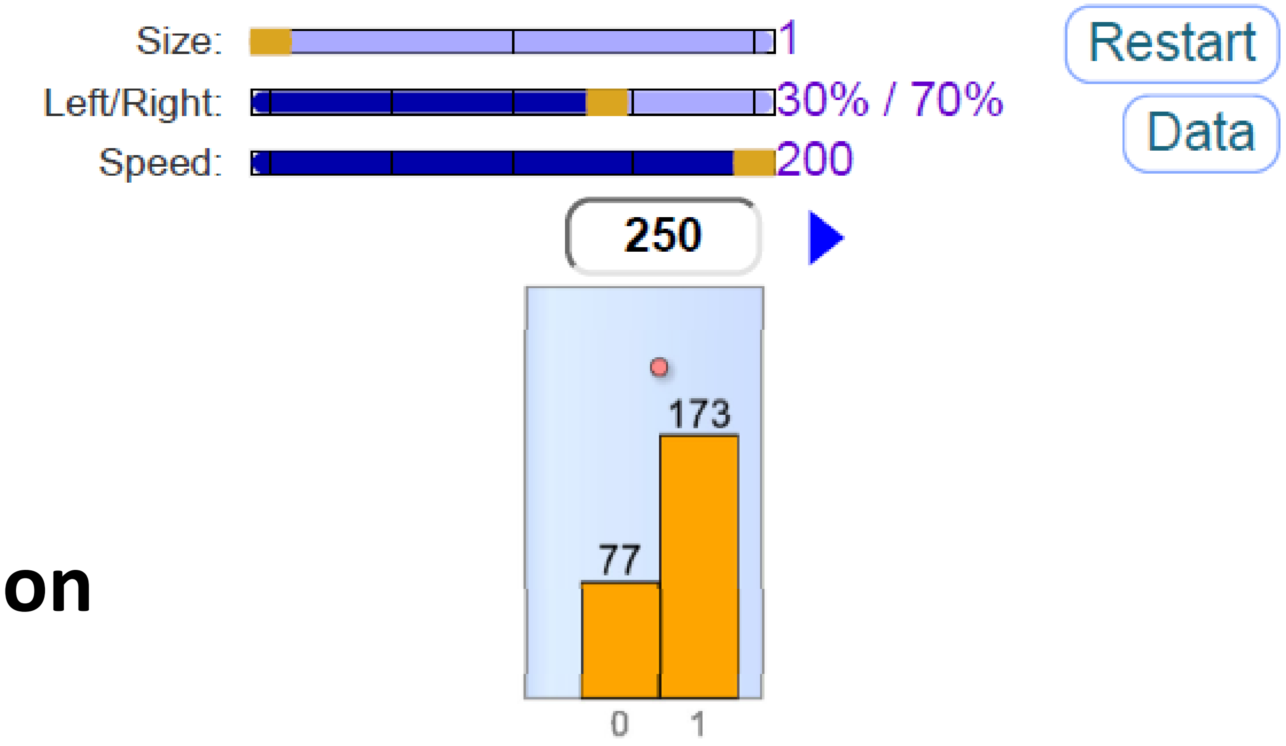


TABLE 2. Simulations for Effect Sizes Passing the Threshold of Formal Statistical Significance ($P = 0.05$)

			Observed OR in Significant Associations	
True OR	Control Group Rate (%)	Sample n Per Group	Median (IQR)	Median Fold Inflation
1.10	30	1000	1.23 (1.23–1.29)	1.11
1.10	30	250	1.51 (1.49–1.55)	1.37
1.25	30	1000	1.29 (1.26–1.39)	1.03
1.25	30	250	1.60 (1.50–1.67)	1.28
1.25	30	50	2.73 (2.60–3.16)	2.18

IQR indicates interquartile range.

30% of the controls are exposed, 70% are not



Effect Size Magnification:
the mechanics of the simulation

For this iteration:

- 77 of 250 controls are exposed (30.8%)
- 173 of 250 controls are not exposed

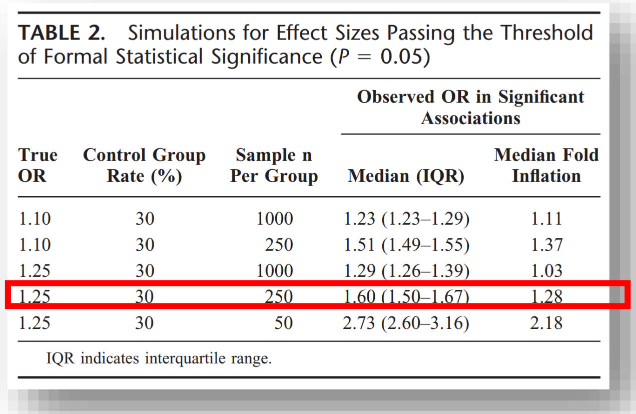


TABLE 2. Simulations for Effect Sizes Passing the Threshold of Formal Statistical Significance ($P = 0.05$)

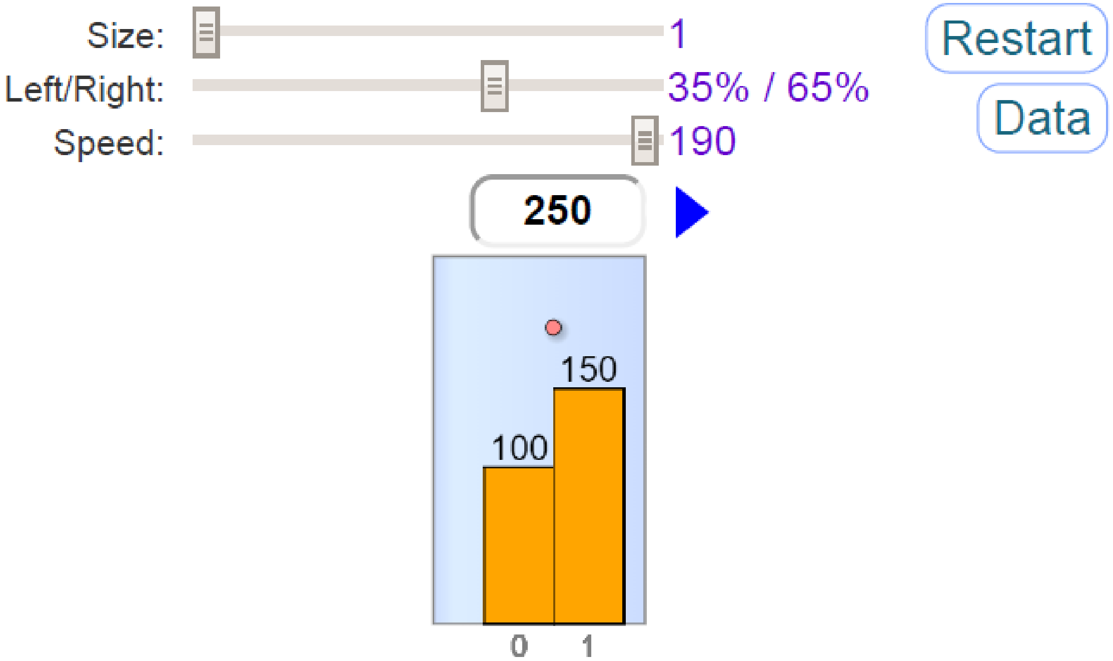
True OR	Control Group Rate (%)	Sample n Per Group	Observed OR in Significant Associations Median (IQR)	Median Fold Inflation
1.10	30	1000	1.23 (1.23–1.29)	1.11
1.10	30	250	1.51 (1.49–1.55)	1.37
1.25	30	1000	1.29 (1.26–1.39)	1.03
1.25	30	250	1.60 (1.50–1.67)	1.28
1.25	30	50	2.73 (2.60–3.16)	2.18

IQR indicates interquartile range.

For an odds ratio of 1.25, need 35% of the controls to be exposed, 65% not

P1 = (P0 x OR) / [(1 – P0) + (P0 x OR)]

```
. display (0.30  * 1.25) / ((1-0.30) + (0.30 * 1.25))
.34883721
```



Size: 1

Left/Right: 35% / 65%

Speed: 190

Effect Size Magnification: the mechanics of the simulation

https://www.mathsisfun.com/data/quincunx.html

For this iteration:

- 100 of 250 controls are exposed (40%)
- 150 of 250 controls are not exposed

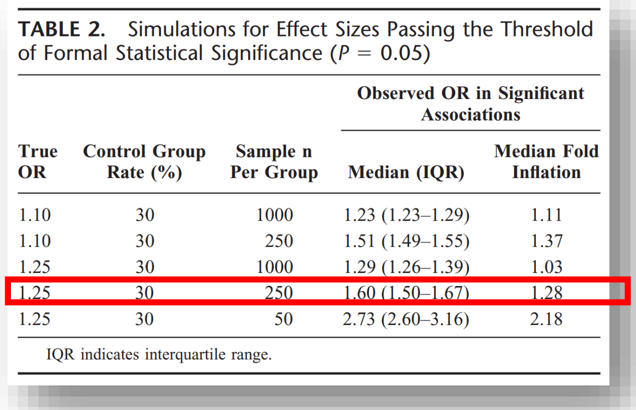
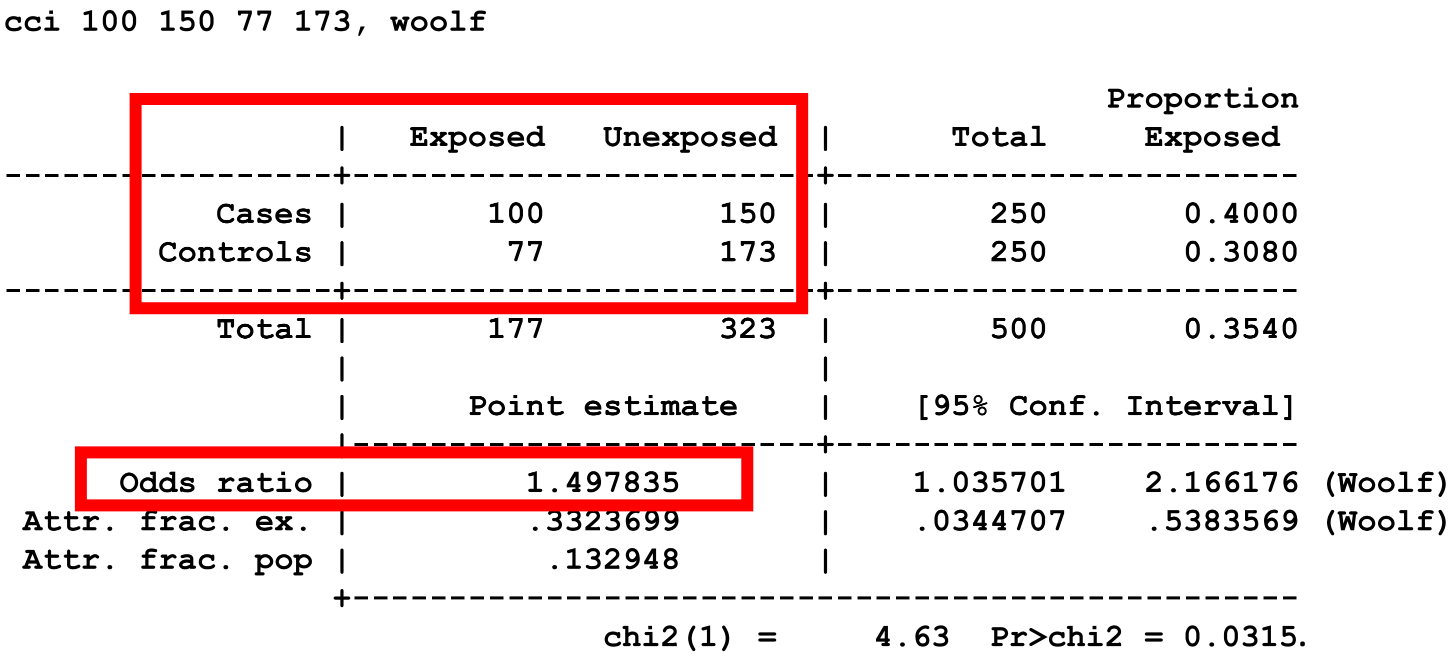


TABLE 2. Simulations for Effect Sizes Passing the Threshold of Formal Statistical Significance (P = 0.05)

True OR	Control Group Rate (%)	Sample n Per Group	Observed OR in Significant Associations Median (IQR)	Median Fold Inflation
1.10	30	1000	1.23 (1.23–1.29)	1.11
1.10	30	250	1.51 (1.49–1.55)	1.37
1.25	30	1000	1.29 (1.26–1.39)	1.03
1.25	30	250	1.60 (1.50–1.67)	1.28
1.25	30	50	2.73 (2.60–3.16)	2.18

IQR indicates interquartile range.

Effect Size Magnification: the mechanics of the simulation



```
cci 100 150 77 173, woolf

                                                         Proportion
                 |   Exposed   Unexposed  |      Total     Exposed
-----------------+------------------------+------------------------
           Cases |       100         150  |        250      0.4000
        Controls |        77         173  |        250      0.3080
-----------------+------------------------+------------------------
           Total |       177         323  |        500      0.3540
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
      Odds ratio |         1.497835       |   1.035701    2.166176 (Woolf)
 Attr. frac. ex. |         .3323699       |   .0344707    .5383569 (Woolf)
 Attr. frac. pop |          .132948       |
                 +-------------------------------------------------
                               chi2(1) =     4.63  Pr>chi2 = 0.0315.
```

Then repeat 999 more times…

What to do… ?

Original Article

Why Most Discovered True Associations Are Inflated

John P. A. Ioannidis

[illegible]

Evidence About Inflated Early-Effect Sizes

[illegible]

640

Epidemiology • Volume 19, Number 5, September 2008

TABLE 3. Avoiding Being Misled on Effect Sizes of True Associations in Early Discovery

Be cautious about effect sizes (and even about the mere presence of any effect in new discoveries)
Consider rational down-adjustment of effect sizes
Consider analytical methods that correct for anticipated inflation
Ignore effect sizes arising from discovery research
Conduct large studies in discovery phase
Use strict protocols for analyses
Adopt complete and transparent reporting of all results
Use methodologically rigorous, unbiased replication (potentially ad infinitum)
Be fair with interpretation

Ioannidis, John P.A. (2008). Why Most True Associations Are Inflated. *Epidemiology. 18*(5): 640-648.

What to do… ?

"At the time of the first postulated discovery, we usually cannot tell whether an association exists at all, let alone judge its effect size. As a starting principle, one should be cautious about effect sizes. Uncertainty is not conveyed simply by CIs (no matter if these are 95%, 99% or 99.9%)"

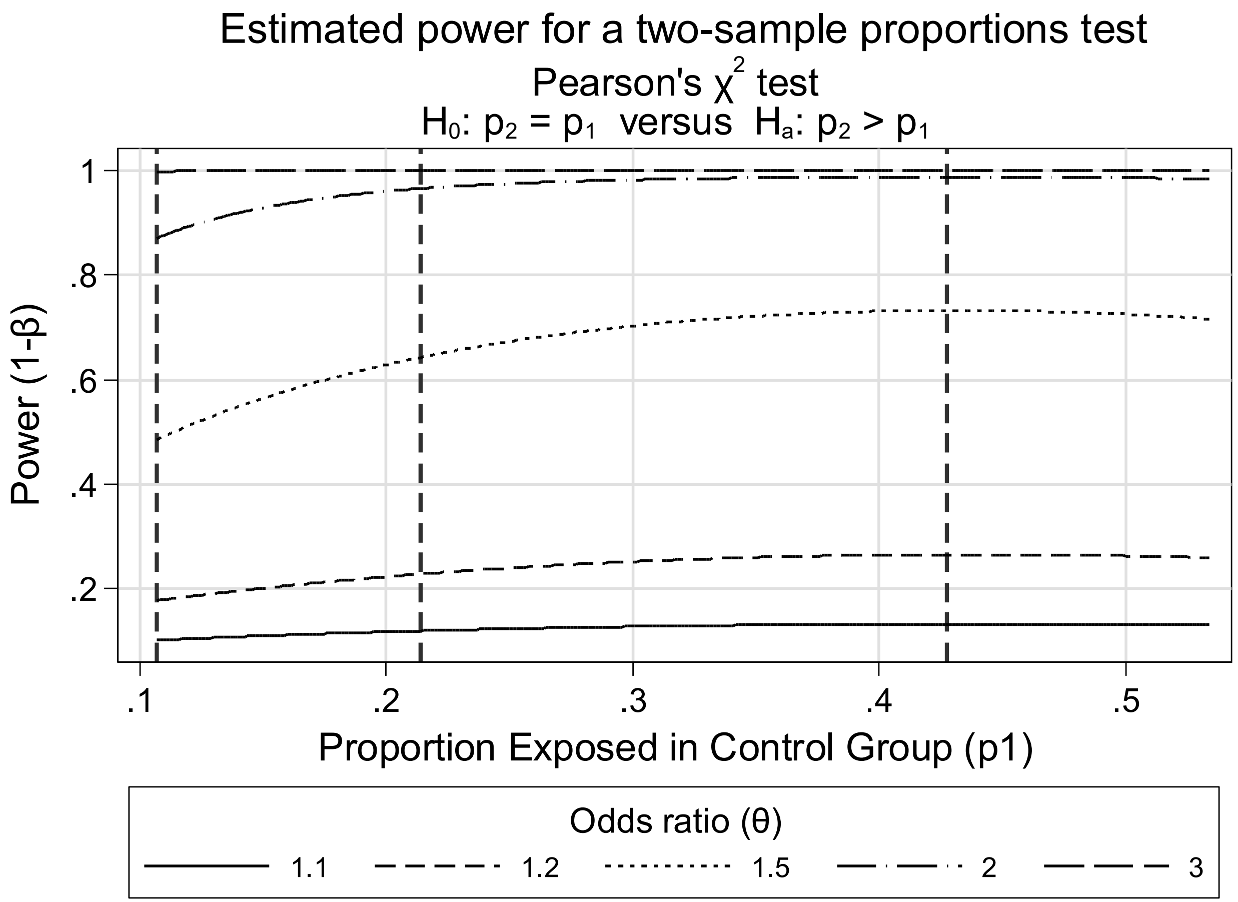
"For a new proposed association, credibility and accuracy of the proposed effect varies depending on the case. One may ask the following questions:

- Does the research community in the field adopt widely statistical significance or similar selection thresholds for claiming research findings?
- Did the discovery arise from a small study?
- Is there room for large flexibility in the analyses?
- Are we unprotected from selective reporting (e.g., was the protocol not fully available upfront?)
- Are there people or organizations interested in finding and promoting specific 'positive' results?
- Finally, are the counteracting forces that would deflate effects minimal?"

Ioannidis, John P.A. (2008). Why Most True Associations Are Inflated. *Epidemiology. 18*(5): 640-648.

Sensitivity Analysis on Control Group Proportion, Greenland *et al.* (1994) Example

- “Proportion Exposed in Control Group” can be an important parameter in sensitivity analysis
- It is useful to vary this to determine how sensitive power is to this (observed) quantity
  - ½ x-, 1x-, and 2x- variations (heavy vertical dashed lines) on observed proportion of **257/1202** illustrated here
  - Results suggest that conclusion that observed OR of 1.76 could be attributable to effect size inflation at a true OR of as low as 1.2 is not sensitive to observed proportion exposed in control group



Vertical dash lines represent 1/2x, 1x, and 2x observed Proportion Observed in Control Group

```
powertwoproportions (`=0.5* 257/1202'(0.001) `=2.5 * 257/1202'), test(chi2) oratio(1.1 1.2 1.5 2.0 3.0) n1(1202) n2(139)graph(recast(line)
xline(`=0.5* 257/1202' `= 257/1202' `=2*257/1202',lpattern(dash)lwidth(medthick))legend(rows(1)size(small) position(6)) ylabel(0.2(0.2)1.0)
xtitle("Proportion Exposed in Control Group (p1)") note("Vertical dash lines represent 1/2x, 1x, and 2x observed Proportion Observed in Control
Group", size(vsmall)) scheme(s1manual)) onesided
```