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 Bottai²

¹Health 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



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 (2nd generation antipsychotics)
 - Psychological Testing (verbal overshadowing, ESP)
 - Evolutionary Biology/Ecology (fluctuating asymmetry)
- Referred to as "Decline Effect"
 - "Cosmic Habituation"



Is there something wrong with the scientific method?

By Jonah Lehrer December 5, 2010

O n September 18, 2007, a few dozen neuroscientists, psychiatrists, and drugcompany 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 secondgeneration 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









Without Children Reproducts by to a defarestantion to the product of parameter product of parameters in the product of the second of graduate a choice parameters is not served. A second of the se epitodellity of here is necessariled second REGENERATION We could add drept had enser (100 Deproduction ensert do not convert hand studies pub-random or optionally. School in first producting journalised og hightage is made

solared hereby the share A large p total a cell de and be and a set for any or and for each or the only for entities are the for the



ter of the state of an in original officer state. On the data way before the data within some officer is the separate data of with significant data of an data significant (see) official

A tra Mistakes in pee	agedy of en r- reviewed papers are easy to port David B. Allison and colle	101 find burb agues.
Justices store person and self-connecting is a cleaned? We have spear the part as monthly which are spear that part with same prior dimension that working on observe for an end of an IDB A1 reads are summing to the same for an IDB A1 person descent paper in a web-oppoint person descent person descent person descent person person descent person descent person descent person descent person descent person descent person descent person descent person person descent person descent p	were profiled in here that the authorshol electific to extend here paper. In the bost of papelse to this proceeding up whether all tension in hereings that papers were all tension data defore interformations, the cause is more representations. If the causes of suscen- bling working into interfaces is no refueld, we have notify the new re-	minister or accepted po- rected, migh After alle or joarnals dobut more track too m



COMMENT





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. Phelps⁴ Jane Ellen Simmons^{*}, Ronald N. Hines^{*1}, and the (Statistics Guidance Documents Working Groups

ovicemental Resection Agency, Office of Research and Developm connectal Effects Research i abcorpore NK Conference invisionmental Effects Research Laboratory "US Environmental Protecti Reselpances, National Exposure Research Laboratory "US Environment Research in al Development, Office of Invisione Advisor, Research I vid VIS Revisionmental Effects Research Laboratory, Million of Research and Developm Revisionmental Effects Research Laboratory, Nationagement, it is do is laborat The sub-concentration of the solid control of 100 features repeated in the state size of the solid control of the

USTRAC

Children and a later



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

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

Katherine S. Button 12, John P. A. Joannidbil, Claire Mohrgazi, Brian A. Nasehl onathan Flint^E, Emma S. J. Robinson^E and Marcus R. Munalö

bstract (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 gnificant result reflects a true effect. Here, we show that the average statistical power of udies in the neurosciences is very low. The consequences of this include overest imates 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 seuroscience is a key priority and requires attention to well-established but often ignored chodological principles

	It has been claimed and demonstrated that many (and possibly most) of the conclusions drawn from bidmedi-	low sample size of shalles, small effects or both) nega- itedy affects the likelihood that a nominally statistically
	cal meterch are probably blas'. A central cause for bia	significant finding schedly reflects a true effect. We da-
	important problem is that researchers must publish in	case the problems that arise when low-powered research
	order to succeed, and publishing is a highly chaped the	designs for pervasive. In general, these problems can be
	enterprise, with carbon kinds of findings more likely to	dvided into two categories. The first concerns prob-
	be published than others. Research that produces novel	lens that are mathematically expected to arise even if
	results, statistically significant results (that is, typically	the research conducted is otherwise periods in other
	the count of the second s	words, when there are no biases that lend to create siz-
-	published-4. As a collisionerus, researchers have strong	tatically significant (that is, 'positive') results that are
	monthes in energy in research practices that make	meriosa. The second calegory concerns problems that
	their findings publishable quickly, even if these proc-	reflect binare that lend in Co-foccar with English of low
	the reduce the Hadhood that the Indiag reflect is true	power or that become worse in small, underpowered
£	(that is, non-real) effect?, Such practices include using	studies. We next empirically show that statistical power
	forthis study designs and forthis statistical analyses	is typically low in the field of memorization by uning evi-
	and remains small studies with low statistical power**.	dence from a more of addicide within the neuroscience.
204	A simulation of genetic association studies showed	Herature. We illustrate that low obtinitical power is an
and the second	that a typical dataset would generate at least one take	endemic problem in mercacience and discus the impli-
	positive result simpsi www.of the time", and two efforts	cations of this for interpreting the results of individual
	to replicate promising findings in biomedicine reveal	studies.
in le	replication rates of 2006 or line". Often that these pub-	
in givening of	lithing biase are pervisive across scientific practice, it	Low power in the absence of other bieses
1200.00	is possible that have positive heavily contaminate the	Three main problems contribute to producing anyibitie
	neuroscience literature as well, and this problem may	findings in studies with low power, even when all other
100.00	affect at least as much, if not even more so, the most	mean's practices are ideal. They are the low probability of
M II M	prominent journals**.	finding the effects, the low positive predictive value GWs,
	Here, we incus on one major aspect of the problem.	are sufficient of the statistical in real when an
	low statistical power. The relationship between study	effect is claimed, and an emprystated estimate of the mag-
And Persons	power and the veracity of the moulting finding in	minute of the effect when a true effect is discreted. Here,
April 2011	lander appreciated. Low statistical power (because of	we discuss these problems in more detail.

eren 🗟 Access fredy weihide eilber

Most Published Research Findings Are False-

PLOS and the

But a Little Replication Goes a Long Way

	for a second pro, see 1 second pro-	COLUMN A			
Neuroscience Sifting the evidence	We have have a to be of hele of explanation is essential the field of generic manifolds (1). For example, a narroy of 600 parallel manufations interaction (1), 2000. The symposic anarroy of 600 parallel manufations interaction and 600 to manufations interaction and 600 to manufations interaction and 600 to manufations in the field states of the manufation of the field states of the field states remained by (4). Lack for regularization and the field states of the field states of the field states are publication bias, arises in bias,	It protecting, PPV (%)	100 - 100 - 40 -	al logis 1 duck profile al logis 2 ducks profile al logis 3 ducks profile	
Unreliable neur	oscience? Wł	ıy j	00	wer	- AND
matters			-		E.1
					tes, mes, or more cost or rem the me-study costs or a mus
Small studies with low power undern	nine the reliability of science	and n	ew ev	ridence suggests that	L making in processed in 17 years

Λ	f the PPV reputationary analysisted as and depends belog again for a set remove sale of the remove sale of the	1, Berry Xandariy, De prob n sholine yield stat renada alan depres underlying relation of obtaining at her eignificant renafia and timestical stud.
ALL A	in sparific lending in here declared it 901, Lensan and being our as leng use, Robi down 2008 NEW in a star being our here and the star of the star of the star in and the star being being the star of the star in a star of the star of the star of the star in a star of the star of the star of the star in a star of the star of the star of the star of the star in a star of the star in a star of the star in a star of the star of th	Relevantations (1% pr Band Manarchine and Relational Control of the and Relational Control of Homes Inter Vision of Homes Inter Vision of Homes Relation of Homes Relation of Homes Relational Control Relational Control

The balance of power

Galand of Coperating Sector (2014), Data Sector (2014), Data Sector (2014), Constructing of Data Sector (2014), Constructing of Data Sector (2014), Construction (2014), Construction (2014), Construction (2014), Construction (2014), Construction (2014), Sector (2014), Construction (2014), Sector (2014), Construction (2014),

Boast Boast

de L'Oth

Reproducibility and Reliability... continuing interest



The National Academies of SCIENCES · ENGINEERING · MEDICINE **CONSENSUS STUDY REPORT**

Reproducibility and Replicability in Science





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

a stor-prime and self-converting new Yes have speech the part this pointing a man collisit, and point in an annual converting the speech store section work and yes a start in a web regarded interesting laws are changed in the same paper in a web regarded interesting laws are changed in the same of a second that the same place of a second that the same place and the same place and the same place and the same place and the same place and the same place and the same place and the same place and the same place and the same place and the same place and	were particle in have the the selection to obtain a restoration of the selection of papeler article pool along the statements in the selection of the selection of the interface of the selection of the selection of the selection of the selection of the selection of the selection of the selection of the selection of the selection of the selection	minister or very a schedul die yben deels anzeige is proseder in a wegt das 11 deel anzeig de generatie in wegt das 11 deel Alter zie specifie bestehen zure ein best op isoaratie, nod diese alter generatie wegt de schedung de schedulingen als leist doch notes, wie als alter die schedulingen werdet aller die schedulingen als sekt doch de schedulingen die schedulingen auf dereich bewei journalist auf schedulingen auf die schedulingen journalist auf schedulingen auf schedulingen bestehen als schedulingen aller schedulingen als schedulingen auf die schedulingen als schedulingen auf die schedulingen als schedulingen auf die schedulingen als schedulingen auf die schedulingen als schedulingen als schedulingen als schedulingen als schedulingen wiel schedulingen als schedulingen als schedulingen wiel schedulingen als schedulingen als schedulingen wiel schedulingen als schedulingen als schedulingen als schedulingen wiel schedulingen als schedulingen als schedulingen als schedulingen als schedulingen als schedulingen wiel schedulingen als schedulingen also schedulingen als schedulingen als schedulingen als schedulingen als schedulingen also sched

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

sk om nor ogen dære skon id he selfræsed at UK forden mental Briterian Agenry (Br

and and the



Photograph: Kate Buttor

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

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

Ratherine S. Button R. John P. A. Joannidtel, Claire Mohrauel, Brian A. Navel nathan Flint⁶, Emma S. J. Robinson⁶ and Marcus R. Munalö

ostract (A study with low statistical power has a reduced chance of detecting a true effect It it is less well appreciated that low power also reduces the likelihood that a statistically nificant result reflects a true effect. Here, we show that the average statistical power of dist in the neurosciences is very low. The consequences of this include overestimates of flect size and low reproducibility of results. There are also ethical dimensions to this oblem, as unreliable research is inefficient and wasteful, improving reproducibility in proscience is a key priority and requires attention to well-established but often ignored

ndraied that many (and low accepts due of shalles, small effects or both) no to drawn from bidraedi- birdy affects the likelihood that a nominally statistic al meanch are probably blac'. A control cause for this — significant finding sciently reflects a bras effect. We do eprise, with certain kinds of its obtainally similarit really that is twicely . Be resurch conducted is observing periods in other about that the findings reflect a true rower or that became worse in small, under now studies with low statistical power tion of eventic space taken studies' show rais that low statistical power is a but a braical chizard would emergie at least one blue real almost were of the time", and two efforts o replicate promiting findings in biométicane revea splication miss of 20% or inst¹⁰. Olven that these pub-thing biase are pervasive across scientific practice, it a possible that have positives heavily contaminate the Three main problem neuroscience literature as well, and this problem may affect at least as much, if not even more so, the most that Brain Bill 1 prominent jour mak¹⁴. Inding the effects, the low positive predictive value (70). Here, we incur on one major aspect of the problem, we will 1 for definitions of any statistical inread when a less statistical power. The relationship between study effect is chimed, and an ex-power and he wencity of the resulting finding is inder appreciated. Los statistical power (because of we discuss here proteined)

eren Statester realementale aller

Most Published Research Findings Are False-

PLOS

But a Little Replication Goes a Long Way

a second gen, said 2 hand 5 h char		
When the start is the dist of experiments of the start is the start of the start is the start is the start of the start is the start is the start is the start is the start is the start is the start is the start is the start is the start is the start is the start is the start is the start is	power	
		the mark of more coal of the the me staticy coals or a more
the reliability of esience and	now ovidence suggests that	ь.
the reliability of science and	i new evidence suggests that	



Small studies with low power undermine

CONTRACTOR OF BRIDE

vork fiscussions





Unreliable neuros matters

max for row of

Reproducibility and Reliability continuing interest



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 <u>http://sites.nationalacademies.org/sites/reproducibility-in-</u> <u>science/index.htm</u>

Agenda available at <u>http://sites.nationalacademies.org/cs/groups/sitessite/documents/</u> webpage/sites 194816.pdf

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

A manufacture of the second se







Minute 207 | Marcel | Jacob 2 | 428

The balance of powe

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

No. of Surgery of

Background (or where this began)

Open access, freely available online

is characteristic of the field and can

vary a lot depending on whether the field targets highly likely relationships

Essay

Why Most Published Research Findings Are False

Summary

Interests inclusiong concern truct more current published research indings 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 nalationships 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 greater number and lesser preselection of tested relationships, when there is greater fluxibility in designs, definitions, outcomes, and analytical modes, when there is greater fluxibility in designs, definitions, outcomes, and analytical modes, when there is greater fluxibility in tansa are involved in a scientific field in chase of statistical significance. Simulations show that for more stauly 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 to simply accurate measures of the prevailing blas. In this essay, i discuss the implications of these problems for the

ublahed 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 chims [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

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

PLoS Medicine | www.plosmedicine.org

some corollaries thereof. Modeling the Framework for False

factors that influence this problem and

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 lif-founded strategy of chaiming conclusive research findings solely on the hasis of a single study assessed by formal statistical significance, typically for a *p*-splue less than 0.05. Research is not most appropriately represented and summarized by *p*-sulues, but, unfortunately, there is a widespread notion that medical research articles

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

should be interpreted based only on a prature. Research findings are defined there as any relationship reaching formal statistical significance, e.g., fifticitive interventions, informative predictors, risk factors, or associations. "Negative" research is also very useful. "Negative" a schully a minomer, and Negative "a schully a minomer, and the minimterpretation is widespread. However, here we will arget

relationships that investigators claim exist, rather than null findings. As has been shown previously, the probability that a research finding is indeed true depends on the prior probability of it being true (before of the study, and the level of statistical againticance [10,11]. Consider a 2 × 2 table in which research findings are compared against the gold standard for 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

0696

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 linding a true relationship for a study linding a true relationship filects the power 1 – β (one minus

the Type II error rate). The probability of claiming a relationship when none truly exists reflects the Type I error rate, 0. 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 = $(1 - \beta)R/(R$ - βR + α). A research finding is thus

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

Copyright: 0 2005 John P. A. Joannidis, This Is an open-access article distributed under the terms of the Crossive Common Attribution Likensa, which permits unrewhicked use, distribution, and reproduction In any medium, provided the original work is properly click.

Abbreviation: PTV, positive predictive value

John F.A. Kannoldin Isi The Department of Hygiane and Epidemiology University of Kanonina School of Medicines, kanonina, Genera, and Institute for Clinical Bearanch and Health Policy School. Department of Medicines, Turthe Heart England Medical Content, Turth University School of Medicine, Readon, Manachusette University School of Medical Content, Turths th

Competing Interests: The author has declared that no competing interests exist. DOI: 10.1271/journal.comed.0000124

August 2005 | Volume 2 | Issue 8 | e124

August 2005 | Volu

8

Background (or where this began)

Open access, freely available online

Why Most Published Research Findings Are False John P.A. Joannidis

ummary

There is increasing concern that most alse. The probability that a research cla s true may depend on study power and bias the number of other studies on the lationships among the robed in each scientific nework, a research find reludice: and when more tions show that for most study ed research findings may nply accurate measures of th plas. In this essay, I discuss the lications of these problems for the duct and interpretation of research.

ublahed 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 chims [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

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

3. PLoS Medicine | www.plosmedicine.org

factors that influence this problem and is characteristic of the field and can ome corollaries thereof. vary a lot depending on whether the field targets highly likely relationships Modeling the Framework for False or searches for only one or a few Positive Findings true relationships among thousands and millions of hypotheses that may Several methodologists have pointed out [9-11] that the high be postulated. Let us also consider.

rate of nonreplication (lack of for computational simplicity, circumscribed fields where either there confirmation) of research discoveries is only one true relationship (among is a consequence of the convenient, yet ill-founded strategy of claiming many that can be hypothesized) or the power is similar to find any of the conclusive research findings solely on the basis of a single study assessed by several existing true relationships. The formal statistical significance, typically pre-study probability of a relationship being true is R/(R+1). The probability for a twalne less than 0.05. Research is not most appropriately represented of a study finding a true relationship reflects the power 1 - B (one minus and summarized by evalues, but, unfortunately, there is a widespread the Type II error rate). The probability notion that medical research articles of claiming a relationship when none truly exists reflects the Type I error rate, 0. Assuming that c relationships

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

achieving formal statistical significance, should be interpreted based only on evalues. 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 mininterpretation is widespread. However, here we will target

relationships that investigators claim exist, rather than null findings. 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 relationshins in a scientific field. In a research field both true and false hypotheses can be made about the presence of relationships. Let Rbe the ratio of the number of "true relationships" to "no relationships"

among those tested in the field. R

0696

the post-study probability that it is true is the positive predictive value, PPV. The PPV is also the complementar probability of what Wacholder et al. have called the false positive report prohability [10]. According to the 2 \times 2 table, one gets PPV = $(1 - \beta)R/(R$ $-\beta R + \alpha$). A research finding is thus

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

are being probed in the field, the

given in Table 1. After a research

finding has been claimed based on

expected values of the 2 × 2 table are

Copyright: 0 2005 John P.A. Joannidis 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 clied.

Abbreviation: PTV, positive predictive value

John F.A. loannids is in the Department of Hygiene and Epidemiology, University of loanning School of Medicine, loanning, 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: Joannid@cc.uol.gr

August 2005 | Volume 2 | Issue 8 | e124

Competing Interests: The author has declared that no competing interests exist. DOI: 10.1371/journal.pmed.0020124

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 (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



Conduct experiment/observational study today



Discover a statistically significant effect size of importance

 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



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



Nature Reviews | Neuroscience

 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





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



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...

ORIGINAL ARTICLE

Why Most Discovered True Associations Are Inflated

John P. A. Ioannidis

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.

ull) associations often have ect sizes. I discuss here the etical considerations prove d on crossing a threshold of study is underpowered, the ted. This has been demondy stopped clinical trials to ible analyses coupled with hed discovered effects. The mallest effect on the same lytic choices) can be very stage of interpretation due ed effects are not always v be deflated-for examsociations in sequentially me types of misclassificaets causing reverse biases. this problem. These inred effect sizes, consideranalytical methods that ing the magnitude of the studies in the discovery pursuing complete and emphasis on replication,

f associations is a core This article will not deal findings are credible.¹ I subset of research finds discussed here encommerge 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...

					Original	ARTICLE		
TABL of For	E 2. Simulations rmal Statistical Sign	for Effect hificance	Sizes Passing the $(P = 0.05)$	While most small/low effects, fev	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			
True	Control Group S	ample n	Observed OR in Associat	io <u>estimates</u> . Median Fold	shed discovered effects. The smallest effect on the same nalytic choices) can be very se stage of interpretation due ered effects are not always	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		
1.10	30 (27% power)	1000	1.23 (1.23–1.29)	1.11	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	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		
1.10 1.25	30 (11% power) 30 (75% power)	250 1000	1.51 (1.49–1.55) 1.29 (1.26–1.39)	1.37 1.03	toring the magnitude of the rge studies in the discovery ses, pursuing complete and ing emphasis on replication, sults.	(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		
1.25 1.25	30 (30% power) 30 (15% power)	250 50	1.60 (1.50–1.67) 2.73 (2.60–3.16)	1.28 2.18	of associations is a core This article will not deal	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		
IOI	R indicates interquartile ra	nge.			subset of research find-	repeatedly in a field, the association is probably real, even if its exact extent can be debated. One should also acknowledge		

discussed here encom-

merge from quantitative

is effect metrics. This

the difficulty in differentiating between an early inflated but

true (non-null) effect and an entirely false (null) one. In

IQR indicates interquartile range.

A simulated numerical illustration of ESM...



- 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?



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



Can we - as regulators - <u>understand</u>, <u>reproduce</u>, and finally <u>apply</u> 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?



Statistical Significant Results from High Quality Study:





An Epidemiological Example

- An epidemiological example uses a case study example published by Greenland (1994)¹
 - relevant to case-control studies using odds ratios²
- 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_{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, <u>not</u> 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 <u>Rate Ratios</u> (see Working Paper at <u>http://www.imm.ki.se/biostatistics/emagnification/</u> for an example)



An Epidemiological Example:

Setting this up in Stata

cci 45 94 257 945, woolf



<u>QUESTION</u>: 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 <u>number of subjects</u> in the *reference* (or control) group
 - 2. the <u>number of subjects</u> in the *comparison* group
 - 3. the <u>proportion of interest</u> in the *reference* group;
 - e.g., the proportion of **exposed** subjects in the control group for case-control studies
 - 4. a <u>target value</u> 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 <u>first three listed values</u> are provided in or must be obtained from the publication while <u>the</u> <u>target value of interest</u> (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:

- the number of subjects in the (reference) control group = <u>1202</u>
 945 non-exposed controls + 257 resin-exposed controls
- ii. the number of subjects in the case group = <u>139</u>
 <u>94 non-exposed cases</u> + 45 resin- exposed cases
- iii. the number of resin exposed subjects in the (reference) control group = $\frac{257}{257}$

	Exposed	Unexposed	Total	Proportion Exposed
Cases Controls	45 257	<mark>94</mark> 945	139 1202	0.3237 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%
```

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

The

Scena

p0	p1	true_or	n 0	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 Signif	icant Associations	
True OR	Control Group Rate, p ₀ (%)	Sample n Per Group (n ₀ /n ₁)	Median (10 th -90 th) ^a	Median Fold Inflation	
1.1	21.4	1202/139	1.508 (1.417– 1.684)	1.371 149	% 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 >9	9% power

^a10th 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 Significa	nt Associatio	ons
	True OR	Control Group Rate, p ₀ (%)	Sample n Per Group (n ₀ /n ₁)	Median (10 th -90 th) ^a	Median Fol Inflation	d
	1.1	21.4	1202/139	1.508 (1.417– 1.684)	1.371	14% power
	1.2	21.4	1202/139	1.547 (<mark>1.415– 1.833</mark>)	1.289	22% power
	4 5	Nhat doos this r	1202/120 noan2	1.653 (1.440– 2.044)	1.102	66% power
				2.015 (1.584– 2.560)	1.007	97% power
He an	re, the author association b	rs "discovered" an etween resin expo	odds ratio of 1.76 for sure and lung cancer.	3.003 (2.347– 3.810)	1.001	>99% power
be	which the (l attributable t	ow) power of the store of the s	study suggests could ion at a true OR of as	ificant results. 75) ifactor(50) nsim(1000) level(0.05) onesided se	ed (123)	
	v as 1.2 and n	or which power is (29

ema 100

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

			_	Observed OR in Signi	ficant Associations	
	True OR	Control Group Rate, p ₀ (%)	Sample n Per Group (n ₀ /n ₁)	Median (10 th -90 th) ^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 (<mark>1.415– 1.833</mark>)	1.289 22%	power
	4 E	Nhat doos this r	1202/120 noan2	1.653 (1.440– 2.044)	1.102 66%	power
				2.015 (Thus: Given the size the "discovered" of the the discovered of the	e (power) of the study,	power
He an	re, the author association b	rs "discovered" an <mark>etween resin expo</mark>	odds ratio of 1.76 for sure and lung cancer.	3.003 would not be unex	pected if the true odds	% power
be	which the (l attributable t	low) power of the s to effect size inflati	study suggests could ion at a true OR of as	75) ifactor(50) nsim(1000) level(0.05) ones	sided seed(123)	_

ema 100

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 <u>Type M(agnitude)</u> and <u>Type S(ign) errors</u> 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
- **<u>NOT</u>** post-hoc power
 - See "Yes, it makes sense to do design analysis ('power calculations') after the data have been collected" at <u>https://statmodeling.stat.columbia.edu/2017/03/03/yes-makes-sensedesign-analysis-power-calculations-data-collected/</u> 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: <u>http://www.imm.ki.se/biostatistics/emagnification/</u>



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. <u>https://ideas.repec.org/c/boc/bocode/s458423.html</u>
- Linden A. (2019). RETRODESIGN: Stata module for computing type-S (Sign) and type-M (Magnitude) errors. Statistical Software Components, Boston College Department of Economics. <u>http://ideas.repec.org/c/boc/bocode/s458631.html</u>
- 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. <u>https://ideas.repec.org/c/boc/bocode/s458592.html</u>
- 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. <u>https://ideas.repec.org/c/boc/bocode/s456792.html</u>



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://rdeSign1 and RETRODESIGN both perform post-hoc
- Linden design analysis for <u>continuous</u> variablesting type-S (Sign) and type-M (Magn tude) errors. Statistical Software components, Boston College Department of Economics. <u>http://ideas.repec.org/c/boc/bocode/s458631.html</u>
- Linden EVALUE evaluates sensitivity of results to State modules. Statistical sensiti ry analyses for conducting confounding partment of Economics.
 Softwa unmeasured confounding partment of Economics.
- Orsini, N., Bellocco, R., Bottai, M. and Greenland S. (2006). EPISENS: Stata module for Detern Boston EPISENS performs Quantitative Bias Analysis (QBA)
 https://ideas.repec.org/c/boc/bocode/s456792.html



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 <code>-emagnification-</code>) 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 !



Contact information:

David J. Miller CAPTIUSPHS 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, studyspecific 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. <u>https://www.tandfonline.com/doi/full/10.1080/00031305.2016.1154108</u>
- 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 <u>http://www.stat.columbia.edu/~gelman/research/published/retropower_final.pdf]</u>
- 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 <u>https://www.mathworks.com/matlabcentral/answers/uploaded_files/55204/The%20fickle%20P%20value%20generates%20irreproducible%20results.pdf]</u>
- 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 <u>http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0020124</u>]
- Ioannidis, J. P. A. 2008. Why most discovered true associations are inflated. Epidemiology 19: 640-648. [accessed 24 February 2018 at <u>http://www.dcscience.net/ioannidis-associations-2008.pdf]</u>
- 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. 41

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."

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=6D48A4D9F8A6 53BA10496D48A4D9F8A653BA1049&FORM=VIRE

			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

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

			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
IOI	indicates interquartil	e range		

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

 $P1 = (P0 \times OR) / [(1 - P0) + (P0 \times OR)]$

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



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

cci 100 150 77 173, woolf



			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	
IQI	R indicates interquartil	e range.			

Effect Size Magnification:

Then repeat 999 more times...

What to do...?

Original Article

Why Most Discovered True Associations Are Inflated

😑 John P. A. Ioannidiz

ABETRE: Newly discovered into (converd) associations of its have inflated effects compared with the tear effect sizes. I discuss here the main mason for the inflation. Next, there the I can identican prove that when two discovery is claimed based on crossing a threshold of excitational seguritaneous and the discovery study is undergonversed, the observed effects are reported to be indicated. This has been demonstrastrated in various fields ranging from early stopped clinical tricle to genume with manufacture. Second, flexible analyses mapped with admine sparing may offer the published discovered effects. The vibution ratio (for ratio of the largest on analysis during) on he care association approached with different analysis during) on he very large. Third, effects may be inflated at the stage of interpretation due to downs could at a street. Discoveral effects are not always inflated, and under some circumstances may be deflated - for example, in the setting of late discovery of associations in sequentially accordated overpresent evidence, in same types of minimalities tion from measurement owner, and in conflicts making reverse biases. Finally, I discuss potential approaches to this problem. These inshale being mations about needly discovered effect sizes, considering none rational descendjustment, using analytical methods that merest for the articipated collation, ignoring the magnitude of the effect (if not consump), conducting large studies in the discovery phase, using strict protocols for analyses, possing complete and transported reporting of all results, placing mephasis on replication, and being fair with interpretation of results.

(Reidenicing 2008,19-640-648)

The decovery and replication of accordings is a conscibility of quarifiative research. This article will not doit with the defaults on wholler research findings are cradible.¹ I will focus instand on the interesting subset of research indings that area iras. Knownch findings documed here seemipose all types of zanochistons that sensing them quarifiative measurements, and area expressed as effect metrics. This includes tradinent effects from clinical triat, measures of tak for documentional tak factors, progenitie effects for

Salamited 17 March 2008, sampled 17 May 2008, peaked 16 May. From the Dipartment of Hyperson and Hydroxidegy. University of Francisco Soluted of Multision, Survey, and Caparison of Multision, Tables 1 intervently Soluted of Multision, Standardson, Miller' and Academic Solution of Multision, Standardson, Miller's Associated Solids approximated of Hypers and Spith. Solidson, Discourse March 20, A Solid Solid Solidson, Statistics, Comparison of Solid Y-Lephanet Balance of Multision, Statistics, Comparison of Solid Solid Solid Solid Solidson, Statistics, Communication, Statistics, Statistics, Statistics, Statistics, Communication, Statistics, Statistics, Statistics, Statistics, Statistics, Discourse Solid Solid Solid Solidson, Statistics, St prognostic sindian, and so lorih. I start hare with the assumpion that a reasonth finding is indeed true (non-nill), is, it reflects a genuine association that is not entirely due to chance or bitsom (continuinting, micclosoffication, soluciton bases, solicitive reporting, or other). The quantion is: do the effect state for such associations, all the time they are first discovered and published in the scientific likewises, scientific inder states.

The article has the following socions: a brief literature raview on influid early-offici sizus hand on theoretical and empirical considerations; a description of the major reasons why early discovered efficit are influid and the major continuing forces that may occusionally lead to deflated efficite (indexemination); and magnetions on how in deal with these problems.

Evidence About inflated Early-Effect Stree

Table 1 cites articles suggesting that early studies give (on average) inflated estimates of effect.2.14 I list here only adacted evaluations that cover either many different articles effects or a whole research domain or method. This list is nowhere close to exhaustive. For some lopics, such as the infairs of memories coefficients for variables selected foreach stepwise statistical-significance-based processes, the literature is veri. The thome of initial early effects has been encountered in various doputous in many scientific disciplings in the biomodical 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 need repeatedly in a field, the association is prohably real, even if its start extent can be deltated. One should also addrewind se the difficulty in differentiating between an early inflated but true (non-null) effect and an entirely false (null) one. In addition to empirical station, however, Table I also includes. theretical work that proves why inflation is anticipated; some of these arguments are discussed in the next section.

I metion here a low stample to demonstrate the arrisonant of the problem. The progenitic significance of a 70-gene expression significant for lymph-node-negative brand cancer to accepted heyend death.²⁴ However, while the first study published in Nature showed almost perficit somitivity and specificity, even in an independent replication exercise of 10 published in Nature showed almost perficit somitivity waves showed assumivity of 60% and specificity of only 40% (AUC: for survival 0.648).²⁵ Prognantic ability in

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 <u>257/1202</u> 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(slmanual)) onesided