Εκτίμηση της αξιοπιστίας της επιστημονικής βιβλιογραφίας

ΕΦΕ 11/2016

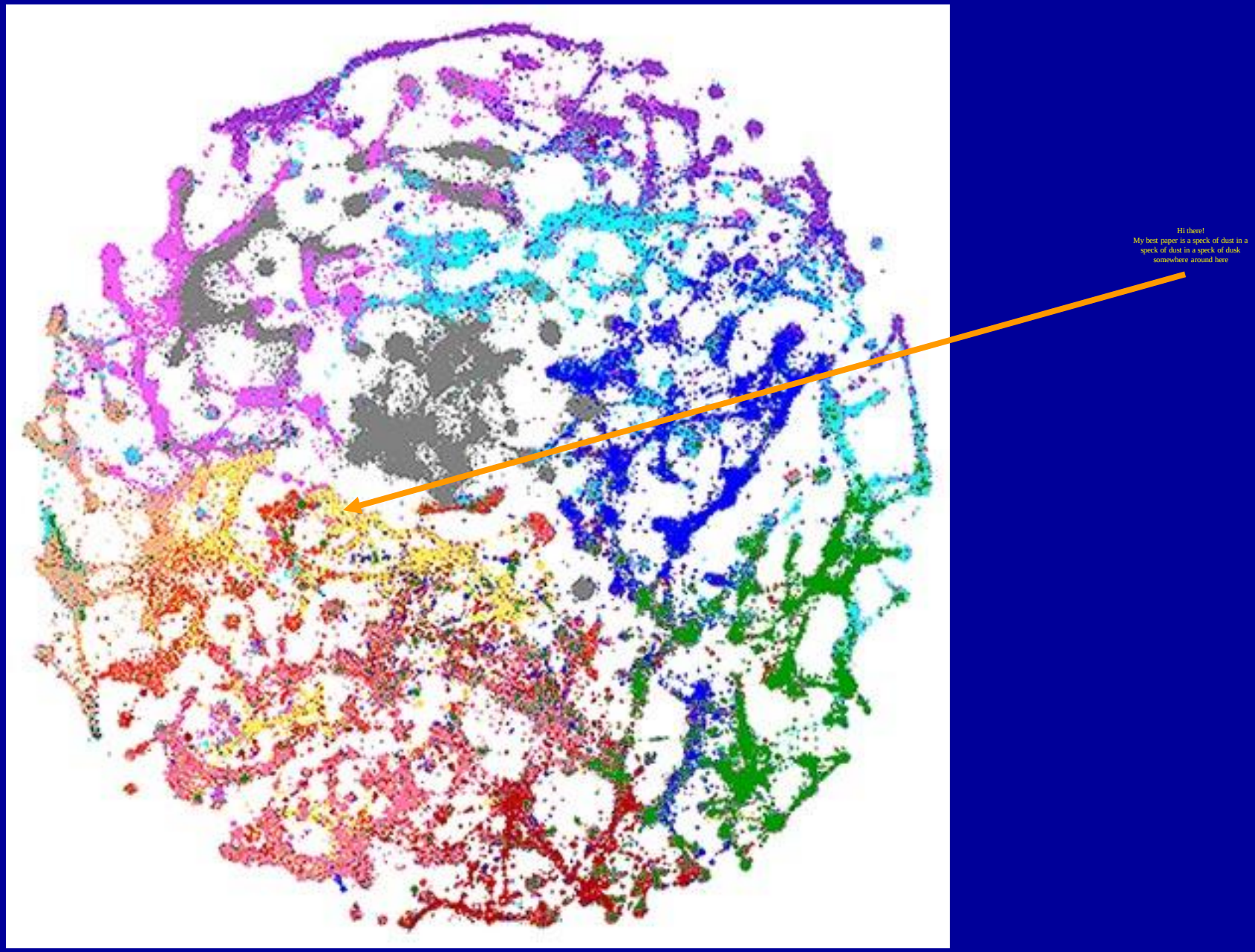
Ιωάννης Π.Α. Ιωαννίδης

C.F. Rehnborg Chair in Disease Prevention
Professor of Medicine, of Health Research and Policy, and of Statistics
Stanford University
Co-Director, Meta-Research Innovation Center at Stanford (METRICS)

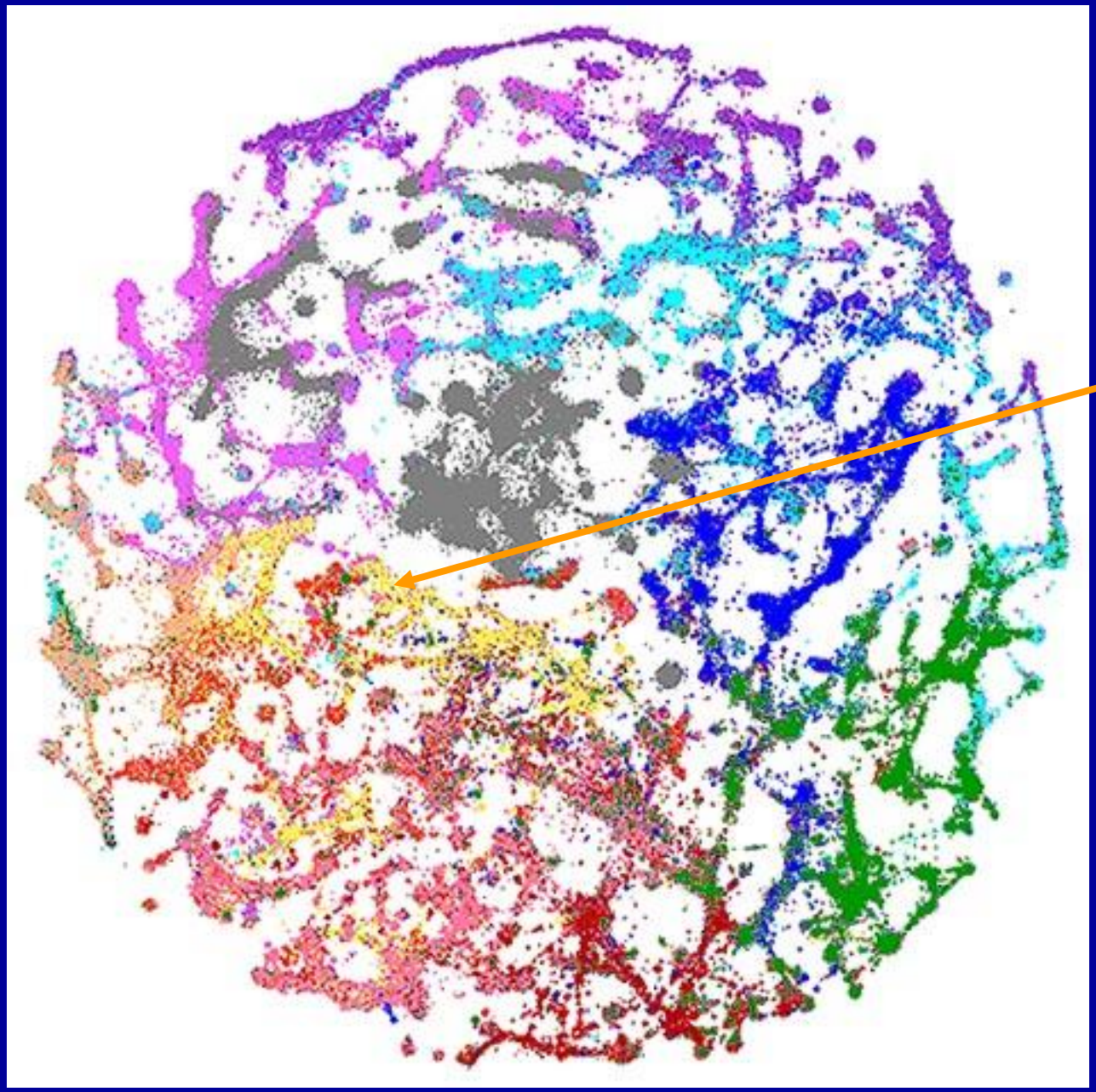
Science: a bird's eye view

- 15,153,100 different scientists publishing papers in major scientific journals in 1996-2011 (Scopus)
- An estimated 160 million scholarly documents in Google Scholar
- Each empirical paper can include anywhere from a few up to many millions of results

Ioannidis, et al. EJCI 2013



A map of recent science: 20 million papers, 2 million patents, 200000 clusters lasting 2-16 years each



Hi there!
My best paper is a speck of dust in a speck of dust in a speck of dust somewhere around here

A map of recent science: 20 million papers, 2 million patents, 200000 clusters lasting 2-16 years each

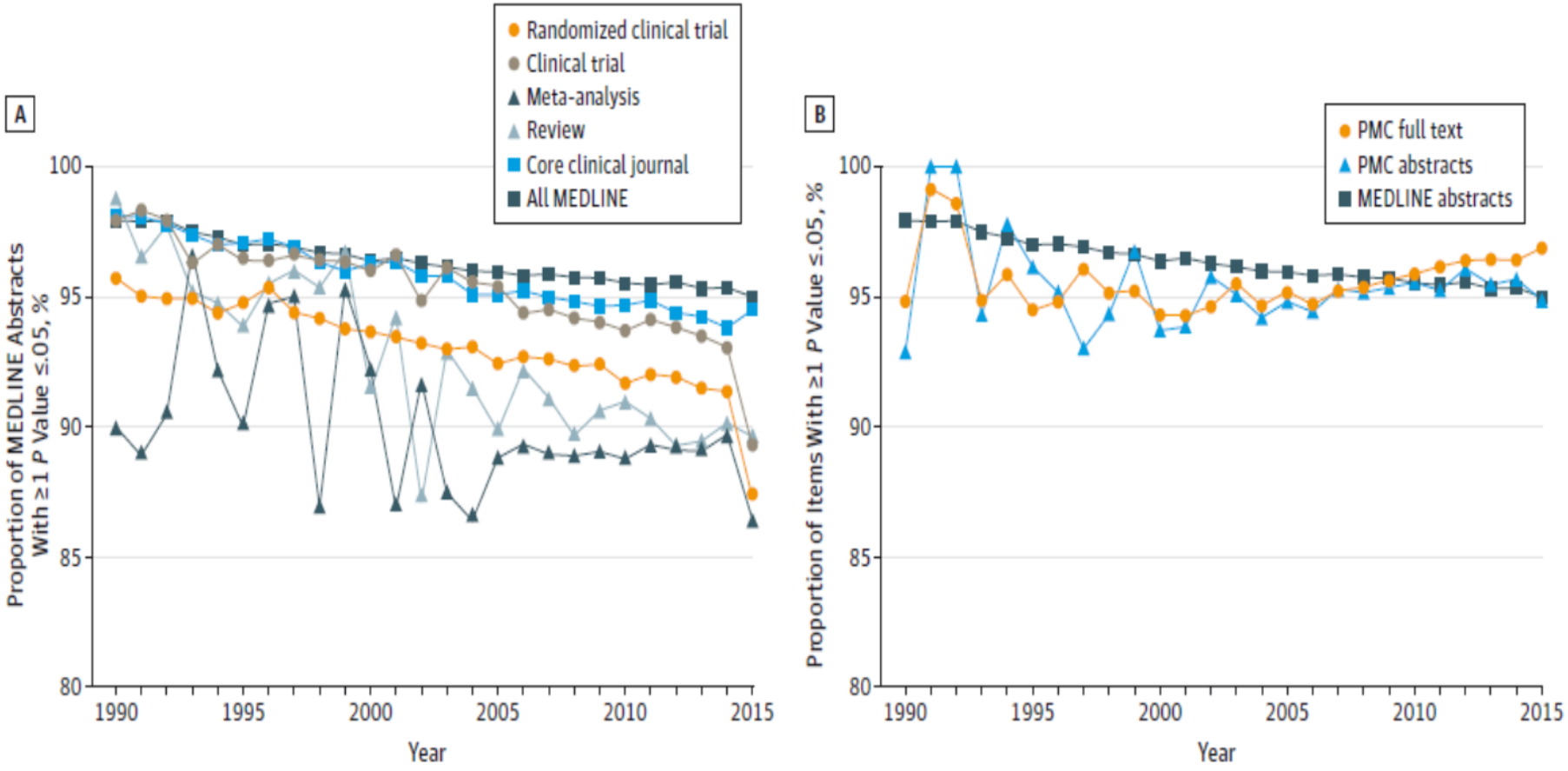
Definitions

- Credibility=how likely a research finding is to be true
- Significance=how likely a research finding is to attract attention
- Statistical significance=a key criterion for attracting attention

“Credible” has little to do with “statistically significant”

- Peer review improves credibility but not necessarily impressively so

Scientific discovery has become a boring nuisance: 96% of the biomedical literature claims significant results



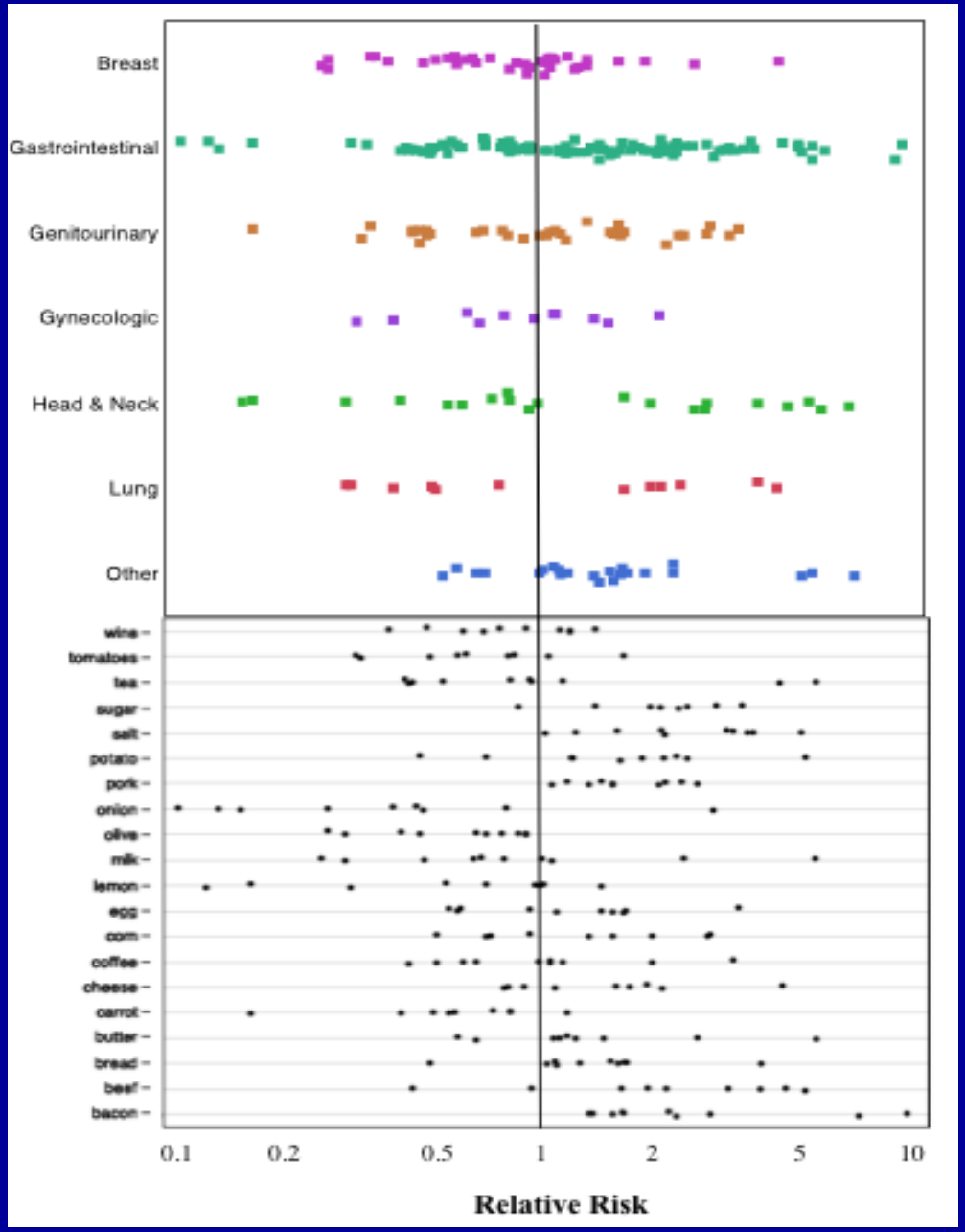
Chavalarias, Wallach, Li, Ioannidis, JAMA 2016

Diet causes cancer

- Open a popular cookbook
- Randomly check 50 ingredients
- How many of those are associated with significantly increased or significantly decreased cancer risk in the scientific literature?

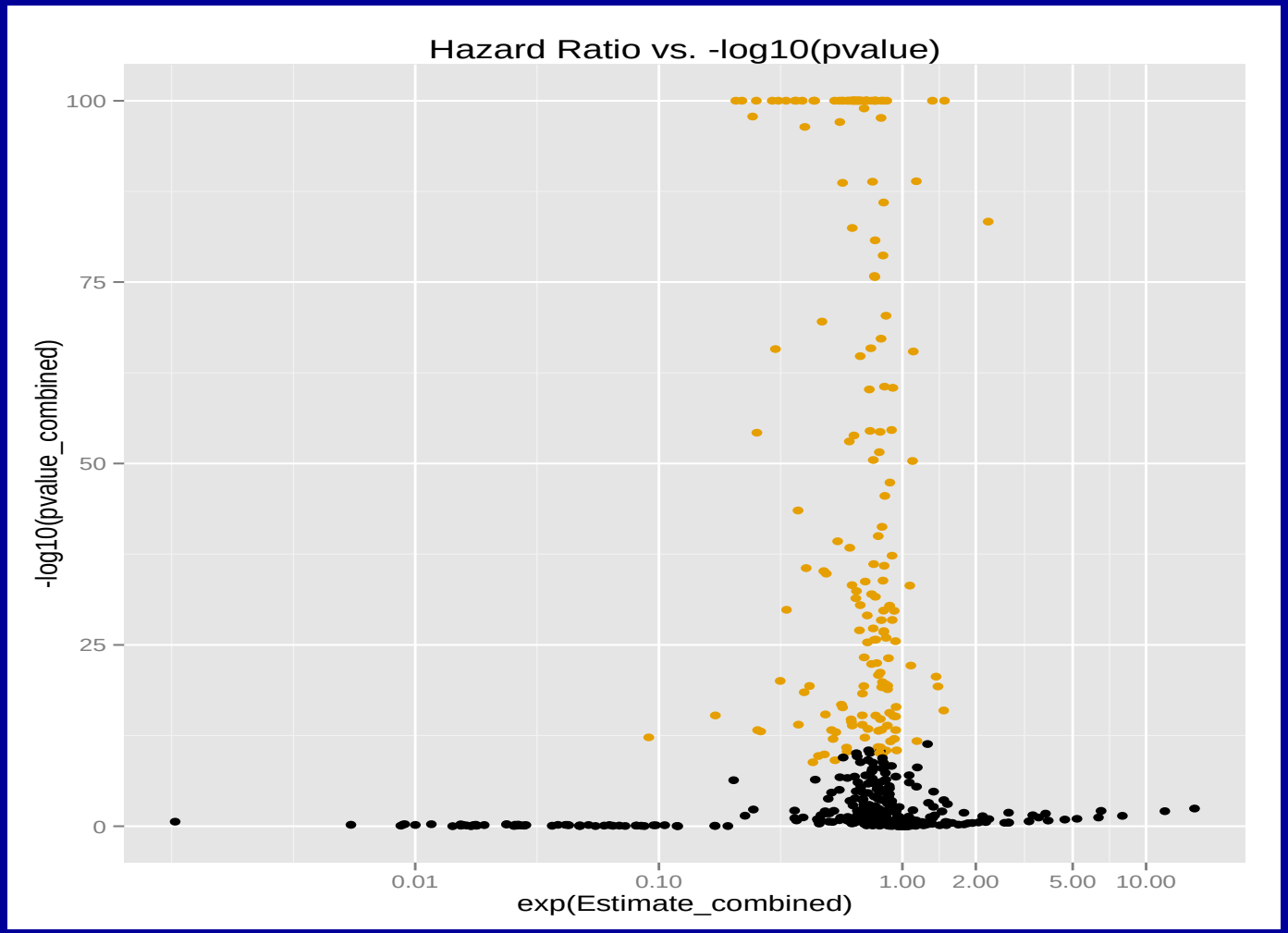
Associated with cancer risk

- veal, salt, pepper spice, flour, egg, bread, pork, butter, tomato, lemon, duck, onion, celery, carrot, parsley, mace, sherry, olive, mushroom, tripe, milk, cheese, coffee, bacon, sugar, lobster, potato, beef, lamb, mustard, nuts, wine, peas, corn, cinnamon, cayenne, orange, tea, rum, raisin



Schoenfeld and Ioannidis AJCN 2013

One third of known medications may affect cancer risk (!?)



Patel et al, Sci Reports 2016

Why research findings may not be credible?

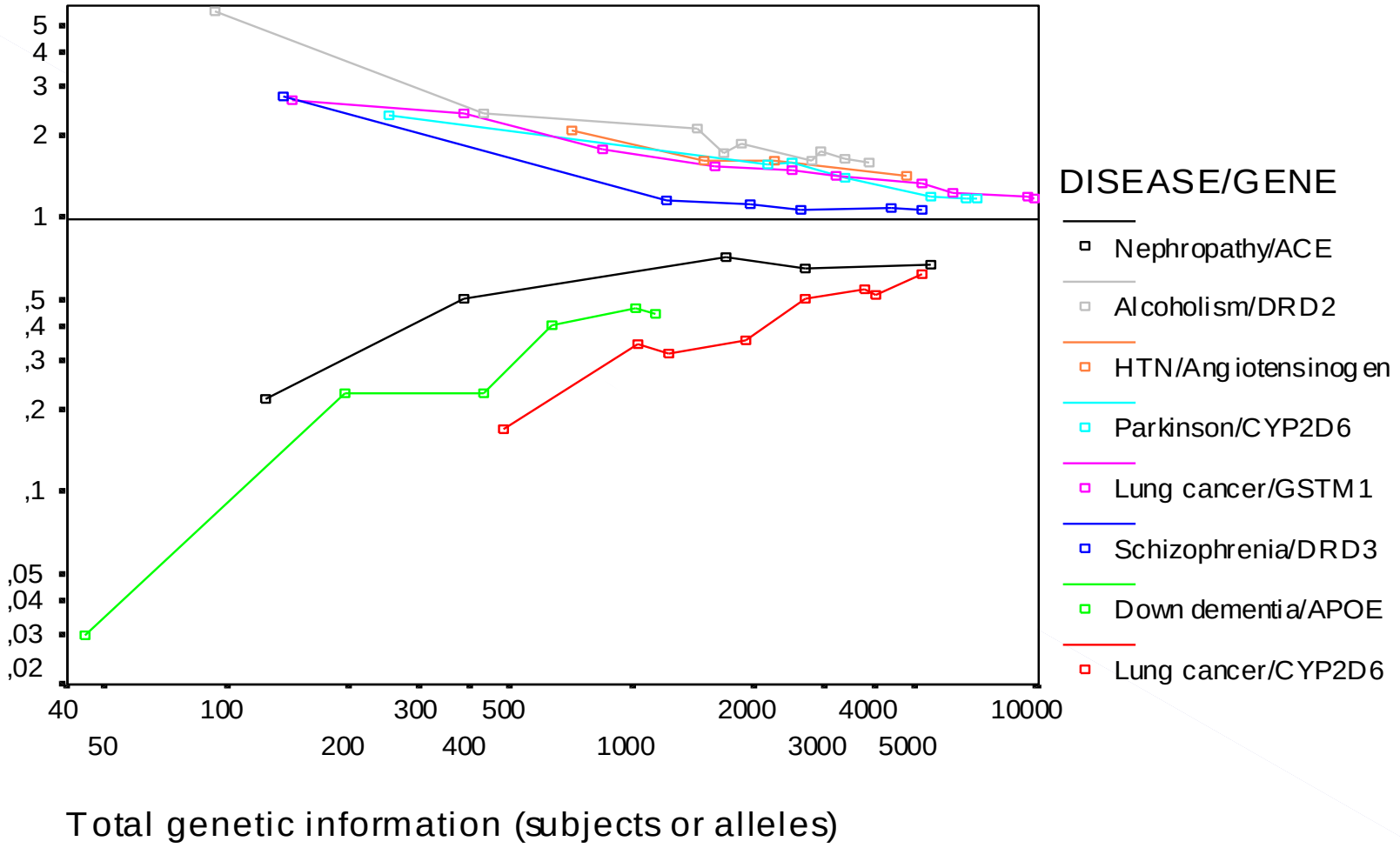
- There is bias
- There is random error (see multiple comparisons)

- Usually there is plenty of both

Bias

- Any deviation from the truth beyond chance error
- Conscious, subconscious, or unconscious
- One may create theory (or theories) about bias or may study its consequences
- The former seem more robust, but it is the latter that we measure, witness, and eventually suffer

Non-replicated diminishing effects



Ioannidis et al, Nature Genetics 2001

Candidate genes replicated through GWAS: replication rate = 1.2%

Table. Large-scale efforts to massively replicate reported candidate gene associations

First author	Disease/phenotype	Gene loci tested	Sample size (design)	Replicated gene loci
Bosker (16)	Major depressive disorder	57	3540 (Case-control)	1
Caporaso (17)	Smoking (7 phenotypes)	359	4611 (Cohort)	1
Morgan (18)	Acute coronary syndrome	70	1461 (Case-control)	0
Richards (19)	Osteoporosis (2 phenotypes)	150	19,195 (Cohort)	9
Samani (20)	Coronary artery disease	55	4864; 2519 (Case-control)	1
Scuteri (21)	Obesity (3 phenotypes)	74	6148 (Cohort)	0
Soeber (22)	Blood pressure	149	1644; 8023 (Cohort)	0
Wu (23)	Childhood asthma	237	1476 (Triads)	1

Ioannidis, Tarone, McLaughlin, Epidemiology 2011

Contradicted and Initially Stronger Effects in Highly Cited Clinical Research

John P. A. Ioannidis, MD

Context Controversy and uncertainty ensue when the results of clinical research on the effectiveness of interventions are subsequently contradicted. Controversies are most prominent when high-impact research is involved.

Objectives To understand how frequently highly cited studies are contradicted or find effects that are stronger than in other similar studies and to discern whether specific characteristics are associated with such refutation over time.

Design All original clinical research studies published in 3 major general clinical journals or high-impact-factor specialty journals in 1990-2003 and cited more than 1000 times in the literature were examined.

Main Outcome Measure The results of highly cited articles were compared against subsequent studies of comparable or larger sample size and similar or better controlled designs. The same analysis was also performed comparatively for matched studies that were not so highly cited.

Results Of 49 highly cited original clinical research studies, 45 claimed that the intervention was effective. Of these, 7 (16%) were contradicted by subsequent studies, 7 others (16%) had found effects that were stronger than those of subsequent studies, 20 (44%) were replicated, and 11 (24%) remained largely unchallenged. Five of 6 highly-cited nonrandomized studies had been contradicted or had found stronger effects vs 9 of 39 randomized controlled trials ($P=.008$). Among randomized trials, studies with contradicted or stronger effects were smaller ($P=.009$) than replicated or unchallenged studies although there was no statistically significant difference in their early or overall citation impact. Matched control studies did not have a significantly different share of refuted results than highly cited studies, but they included more studies with "negative" results.

Conclusions Contradiction and initially stronger effects are not unusual in highly cited research of clinical interventions and their outcomes. The extent to which high citations may provoke contradictions and vice versa needs more study. Controversies are most common with highly cited nonrandomized studies, but even the most highly cited randomized trials may be challenged and refuted over time, especially small ones.

JAMA. 2005;294:218-228 www.jama.com

CLINICAL RESEARCH ON IMPORtant questions about the efficacy of medical interventions is sometimes followed by subsequent studies that either reach opposite conclusions or suggest that the original claims were too strong. Such disagreements may upset clinical practice and acquire publicity in both scientific circles and in the lay press. Several empirical investigations have tried to address whether specific types of studies are more likely to be contradicted and to explain observed controversies. For example, evidence exists that small studies may sometimes be refuted by larger ones.[1,2]

Similarly, there is some evidence on disagreements between epidemiological studies and randomized trials.[3-5] Prior investigations have focused on a variety of studies without any particular attention to their relative importance and scientific impact. Yet, most research publications have little impact while a small minority receives most attention and dominates scien-

Highly-cited contradicted findings

- Vitamin E and cardiovascular mortality (two large prospective cohorts and one trial of 2,002 subjects claimed large decreases in mortality)
- Hormone replacement therapy and coronary artery disease (major benefits claimed by the Nurses' Health Study)
- Nitric oxide found initially to markedly improve outcomes in adult respiratory distress syndrome

Some other major refuted claims

- Flavonoids decrease cardiovascular mortality by 80%
- Low-fat diet dramatically decreases colorectal cancer, heart disease, stroke, and breast cancer
- Aspirin is highly protective against heart disease in both men and women
- Beta-carotene is highly effective in preventing against cancer and heart disease
- Fruit intake diminishes breast cancer risk by up to 90%

Different types of reproducibility

- Reproducibility of methods: the ability to understand or repeat as exactly as possible the experimental and computational procedures.
- Reproducibility of results: the ability to produce corroborating results in a new study, having followed the same experimental methods.
- Reproducibility of inferences: the making of knowledge claims of similar strength from a study replication.

Goodman, Fanelli, Ioannidis. Science Transl Med 2016

Overall credibility

- Depends on the pre-evidence odds (multiplicity of comparisons against true associations)
- Depends on the data (the study at hand)
- Depends on bias
- Depends on the field
- All of these may depend on each other

Simple model: no bias, one team of researchers

Table 1. Research Findings and True Relationships

Research Finding	True Relationship		
	Yes	**No**	**Total**
Yes	$c(1-\beta)R/(R+1)$	$c\alpha/(R+1)$	$c(R+\alpha-\beta R)/(R+1)$
No	$c\beta R/(R+1)$	$c(1-\alpha)/(R+1)$	$c(1-\alpha+\beta R)/(R+1)$
Total	$cR/(R+1)$	$c/(R+1)$	c

Bias present

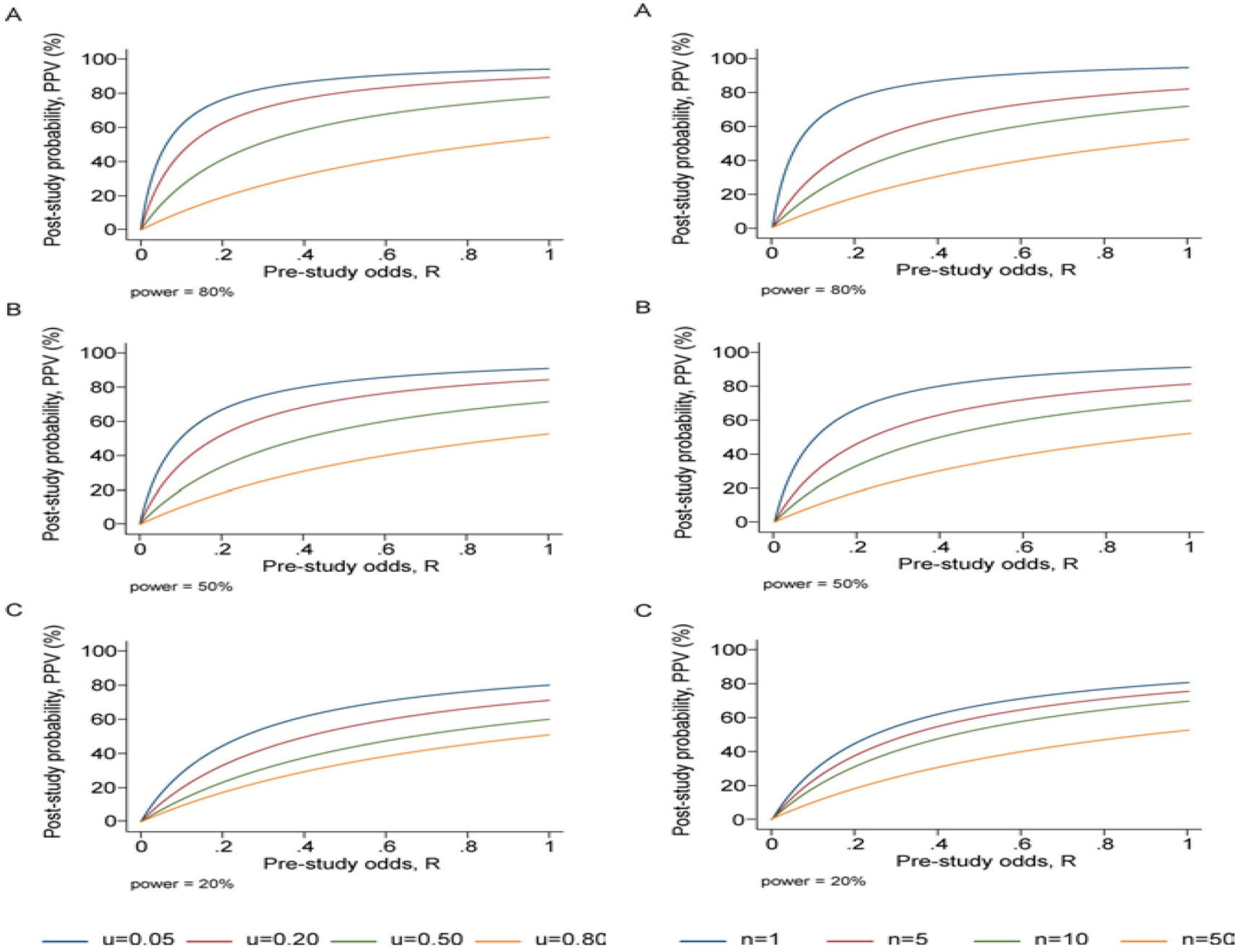
Table 2. Research Findings and True Relationships in the Presence of Bias

Research Finding	True Relationship		
	Yes	No	Total
Yes	$(c[1-\beta]R + uc\beta R)/(R+1)$	$c\alpha + uc(1-\alpha)/(R+1)$	$c(R+\alpha-\beta R+u-u\alpha+u\beta R)/(R+1)$
No	$(1-u)c\beta R/(R+1)$	$(1-u)c(1-\alpha)/(R+1)$	$c(1-u)(1-\alpha+\beta R)/(R+1)$
Total	$cR/(R+1)$	$c/(R+1)$	c

Many teams of researchers

Table 3. Research Findings and True Relationships in the Presence of Multiple Studies

Research Finding	True Relationship		
	Yes	No	Total
Yes	$cR(1-\beta^n)/(R+1)$	$c(1-[1-\alpha]^n)/(R+1)$	$c(R+1-[1-\alpha]^n-R\beta^n)/(R+1)$
No	$cR\beta^n/(R+1)$	$c(1-\alpha)^n/(R+1)$	$c([1-\alpha]^n+R\beta^n)/(R+1)$
Total	$cR/(R+1)$	$c/(R+1)$	c


A
Post-study probability, PPV (%)
100
80
60
40
20
0
0
.2
.4
.6
.8
1
Pre-study odds, R
power = 80%
B
power = 50%
C
power = 20%
u=0.05
u=0.20
u=0.50
u=0.80
n=1
n=5
n=10
n=50

Science at low pre-study odds of true findings

Ioannidis. Why most published research findings are false? PLoS Medicine 2005

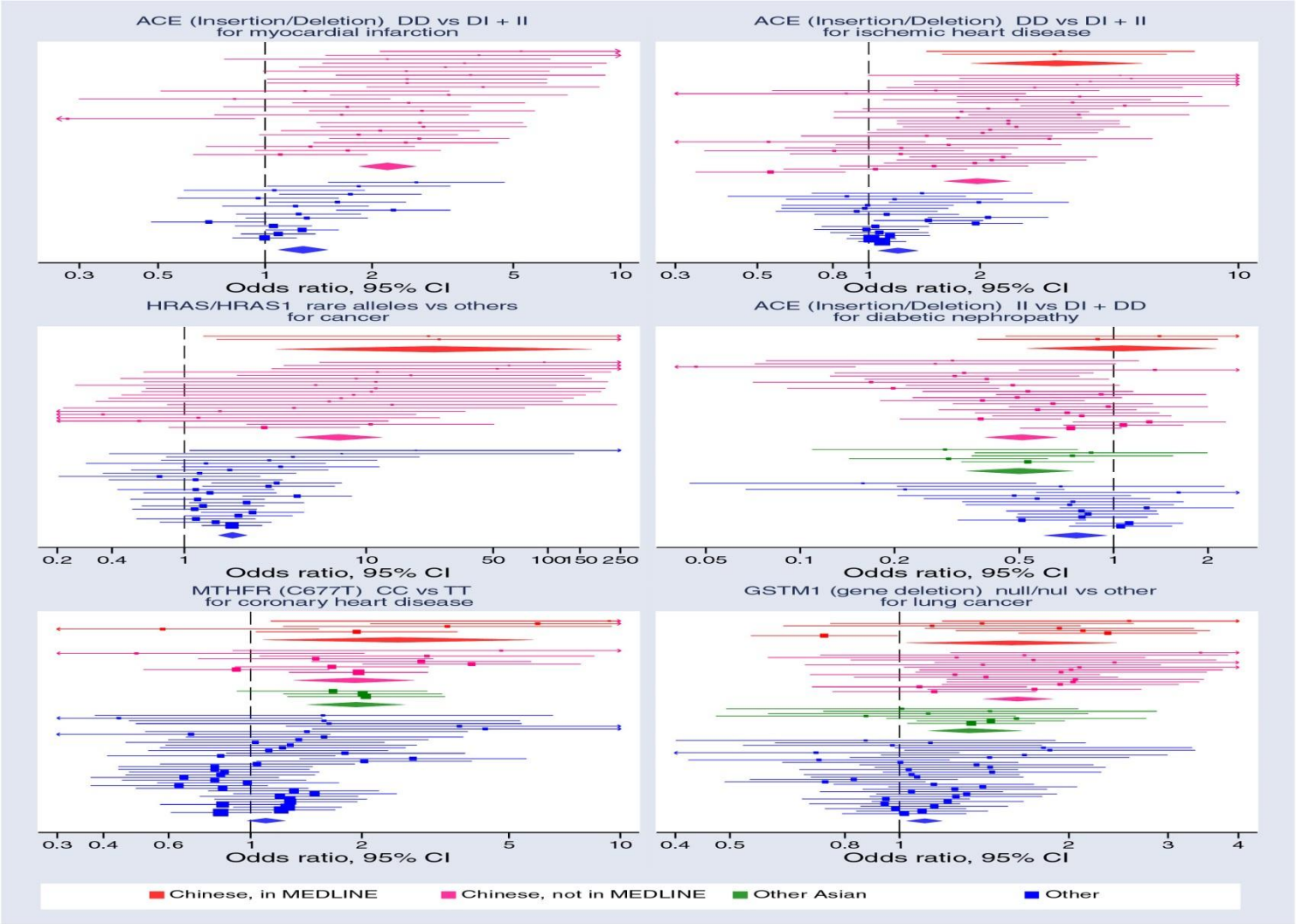
Positive predictive value (PPV) of research findings for various combinations of power (1-β), ratio of true to no relationships (R) and bias (u)

1-β	R	u	Practical example	PPV
0.80	1:1	0.10	Adequately powered RCT with little bias and 1:1 pre-study odds	.85
0.95	2:1	0.30	Confirmatory meta-analysis of good quality RCTs	.85
0.80	1:3	0.40	Meta-analysis of small inconclusive studies	.41
0.20	1:5	0.20	Underpowered, phase I/II well-performed RCT	.23
0.20	1:5	0.80	Underpowered, phase I/II poorly performed RCT	.17
0.80	1:10	0.30	Adequately powered, exploratory epidemiological study	.20
0.20	1:10	0.30	Underpowered, exploratory epidemiological study	.12
0.20	1:1000	0.80	Discovery-oriented exploratory research with massive testing	.0010
0.20	1:1000	0.20	As above, but with more limited bias (more standardized)	.0015

Effect size = bias

- In several scientific disciplines, the effect sizes observed in different studies are, on average, accurate estimates of the extent of net bias operating in the field
- Thus, disciplines that find larger effect sizes (=are scientifically considered more successful) are simply more biased than others that find smaller effect sizes
- In the same scientific discipline, the most successful and appreciated studies are simply the ones that suffer the worst net bias

Effect size = bias
A Chinese language lesson



Pan et al. PLoS Med 2005

Post-study odds of a true finding are small

- When effect sizes are small
- When studies are small
- When fields are “hot” (many furtively competitively teams work on them)
- When there is strong interest in the results
- When databases are large
- When analyses are more flexible

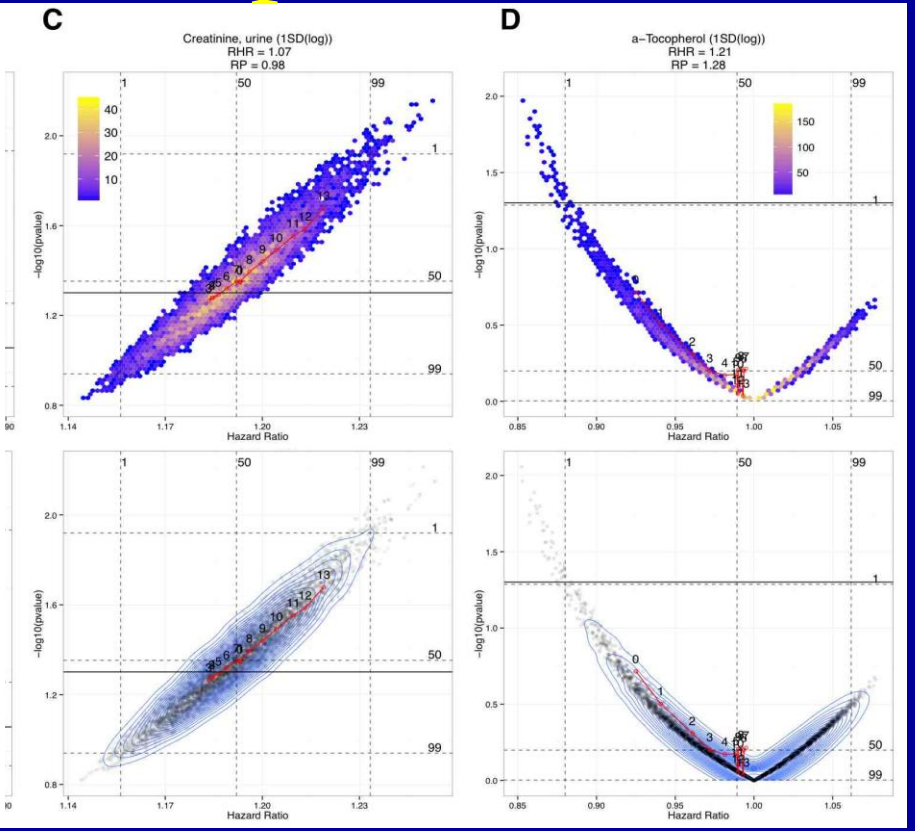
Ioannidis JP. PLoS Medicine 2005

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

Katherine S. Button[1,2], John P. A. Ioannidis[3], Claire Mokrysz[1], Brian A. Nosek[4], Jonathan Flint[5], Emma S. J. Robinson[6] and Marcus R. Munafò[1]

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.

Analytical flexibility: Vibration of effects and the Janus phenomenon



Patel, Burford, Ioannidis. JCE 2015

Conflicts of interest

- 185 MEDLINE-listed meta-analyses evaluating antidepressants for depression published in 1/2007-3/2014.
- Only 58 of the 185 meta-analyses on antidepressants for depression (31%) had any negative statements in the concluding statement of the abstract.
- Meta-analyses including an author who were employees of the manufacturer of the assessed drug were 22-times less likely to have negative statements about the drug than other meta-analyses (1/54 [2%] vs. 57/131 [44%], $p<0.001$).

Shanil Ebrahim, Sheena Bance, Abha Athale, Cindy Malachowski, John P.A. Ioannidis

Registration

- Level 0: no registration
- Level 1: registration of dataset
- Level 2: registration of protocol
- Level 3: registration of analysis plan
- Level 4: registration of analysis plan and raw data
- Level 5: open live streaming

Public Availability of Published Research Data in High-Impact Journals

Alawi A. Alsheikh-Ali[1,2§], Waqas Qureshi[3§], Mouaz H. Al-Mallah[4,5], John P. A. Ioannidis[2,6,7,8,9*]

Journal	Impact Factor	Policy of Required Public Deposition for Types of Data: Microarray	Nucleic Acid	Protein	Macromolecular	Policy of Provision of Materials and Methods: Materials upon request	Protocols upon request	Condition of publication	Full data deposited: Percentage of papers
New England Journal of Medicine	52.589								0
Cell	29.887								1
Nature	28.751								0
Lancet	28.638								0
Nature Medicine	26.382								0
Science	26.372								1
Nature Immunology	26.218								9
Nature Genetics	25.556								0
JAMA	25.547								1
Nature Biotechonology	22.848								5
Nature Materials	19.782								0
Immunity	19.266								0
Nature Cell Biology	17.623								0
Journal of Clinical Investigation	16.915								0
Archives of General Psychiatry	15.976								0
Journal of the National Cancer Institute	15.678								0
Nature Neuroscience	15.664								1
Journal of Experimental Medicine	15.612								0
Annals of Internal Medicine	15.516								0
Journal of Clinical Oncology	15.484								0
Nature Methods	15.478								6
Genes and Development	14.795								3
Nature Physics	14.677								0
PLoS Biology	13.501								2
Neuron	13.41								0
Molecular Cell	13.156								0
Circulation	12.755								0
PLoS Medicine	12.601								0
Developmental Cell	12.436								0
Gastroenterology	11.673								0
Genome Research	11.224								6
American Journal of Human Genetics	11.092								3
Nature Structural and Molecular Biology	11.085								0
Journal of the American College of Cardiology	11.054								0
Blood	10.896								0
Hepatology	10.734								0
Current Biology	10.539								0
Gut	10.015								0
British Medical Journal	9.723								0
Circulation Research	9.721								1
Plant Cell	9.653								0
Nano Letters	9.627								0
Journal of Cell Biology	9.598								0
PNAS	9.598								1
Molecular and Cellular Proteomics	9.425								7
PLoS Pathogens	9.336								0
American Journal of Psychiatry	9.127								0
American Journal of Respiratory and Critical Care Medicine	9.074								0
Annals of Neurology	8.813								0
PLoS Genetics	8.721								0

A research finding cannot reach credibility over 50% unless

$u < R$

i.e. bias must be less than the pre-study odds

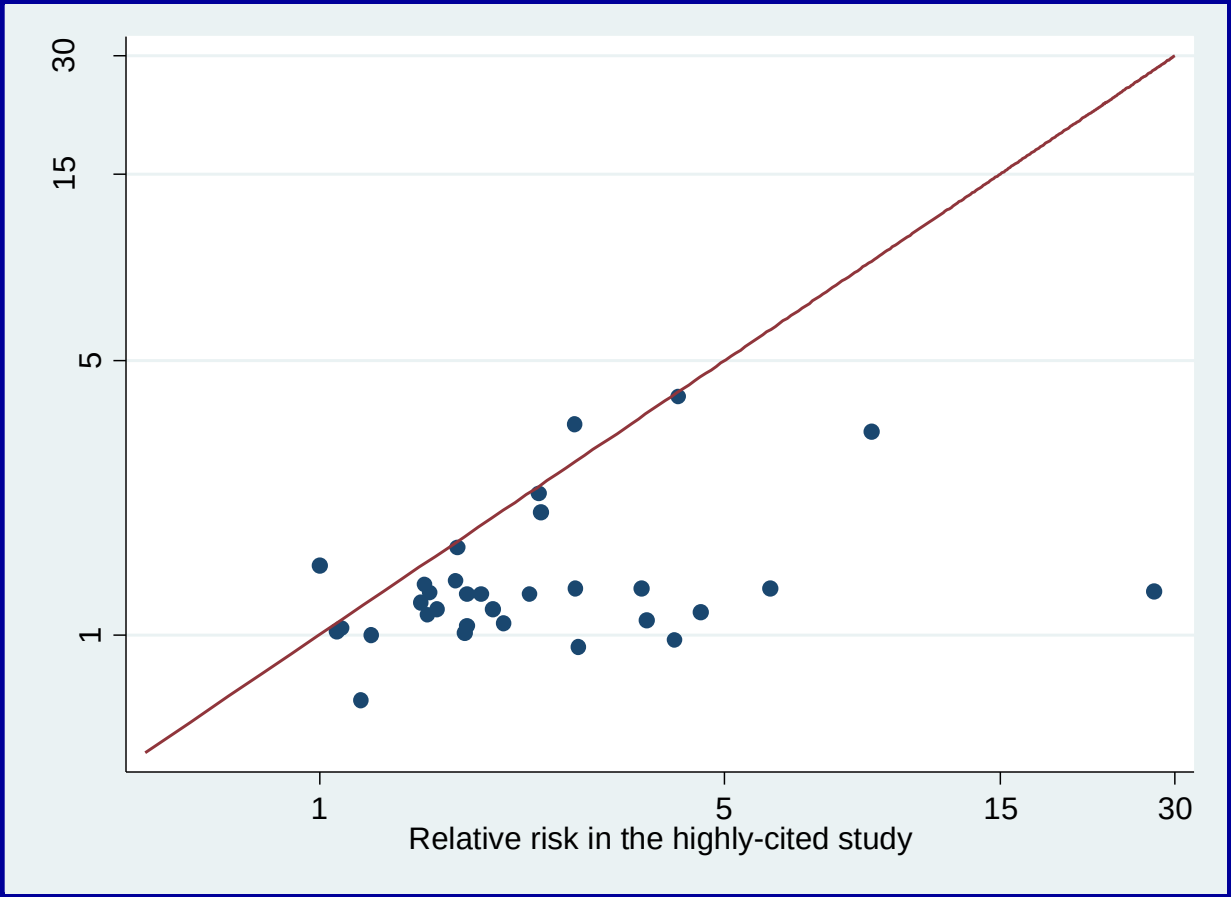
Early discovered true associations are inflated

TABLE 1. Selected Evaluations Suggesting That Early Discovered Effects Are Inflated

Research Field	Theoretical Work or Empirical Evidence and References
Highly cited clinical research	A quarter of most-cited clinical trials and 5/6 most-cited epidemiological studies were either fully contradicted or found to have exaggerated results[2]
Early stopped clinical trials	Early stopping results in inflated effects in theory[3,4] and shown also in practice[5]
Clinical trials of mental health interventions	More likely for effect sizes of pharmacotherapies to diminish than to increase over time[6]
Clinical trials on heart failure interventions	"Regression to the truth" in phase III trials for interventions with early promising results[7]
Clinical trials on diverse interventions	Effectiveness shown to fade over time[8]
Multiple meta-analyses on effectiveness	Eleven independent meta-analyses on acetylcysteine show decreasing effects over time[9]
Epidemiologic associations	Expected to be inflated in multiple testing with significance threshold; empirical demonstration for occupational carcinogens[10]
Pharmacoepidemiology	"Phantom ship" associations that don't stand upon further evaluation[11]
Gene-disease associations	Several empirical evaluations showing dissipation of effect sizes over time[12–15]
Linkage studies in humans	Theory anticipates large upward bias ("winner's curse") in effects of discovered loci[16–18]
Genetic traits in experimental crosses	As above (actually literature on the "Beavis effect" precedes literature on humans)[19–22]
Genome-wide associations	Large winner's curse anticipated for discovered effects in underpowered conditions[23,24]
Ecology and evolution	Empirical demonstration that relationships fade over time[25,26]
Psychology	Replication studies in psychology failing to confirm true effects because the new studies were underpowered due to reliance on the estimate of effect from the original positive study[27]
Early repeated data peaking in general	Simulations to model inflation of effects with repeated data peaking[28]
Prognostic models	Overestimated prognostic performance with step-wise selection of variables based on significance thresholds[29–32]
Regression models in general	Exaggerated effects (coefficients) with stepwise selection based on significance thresholds and small datasets[32–34]; may correct substantially if a very lenient alpha = 0.20 is used for selection[34] [thus having enough power]

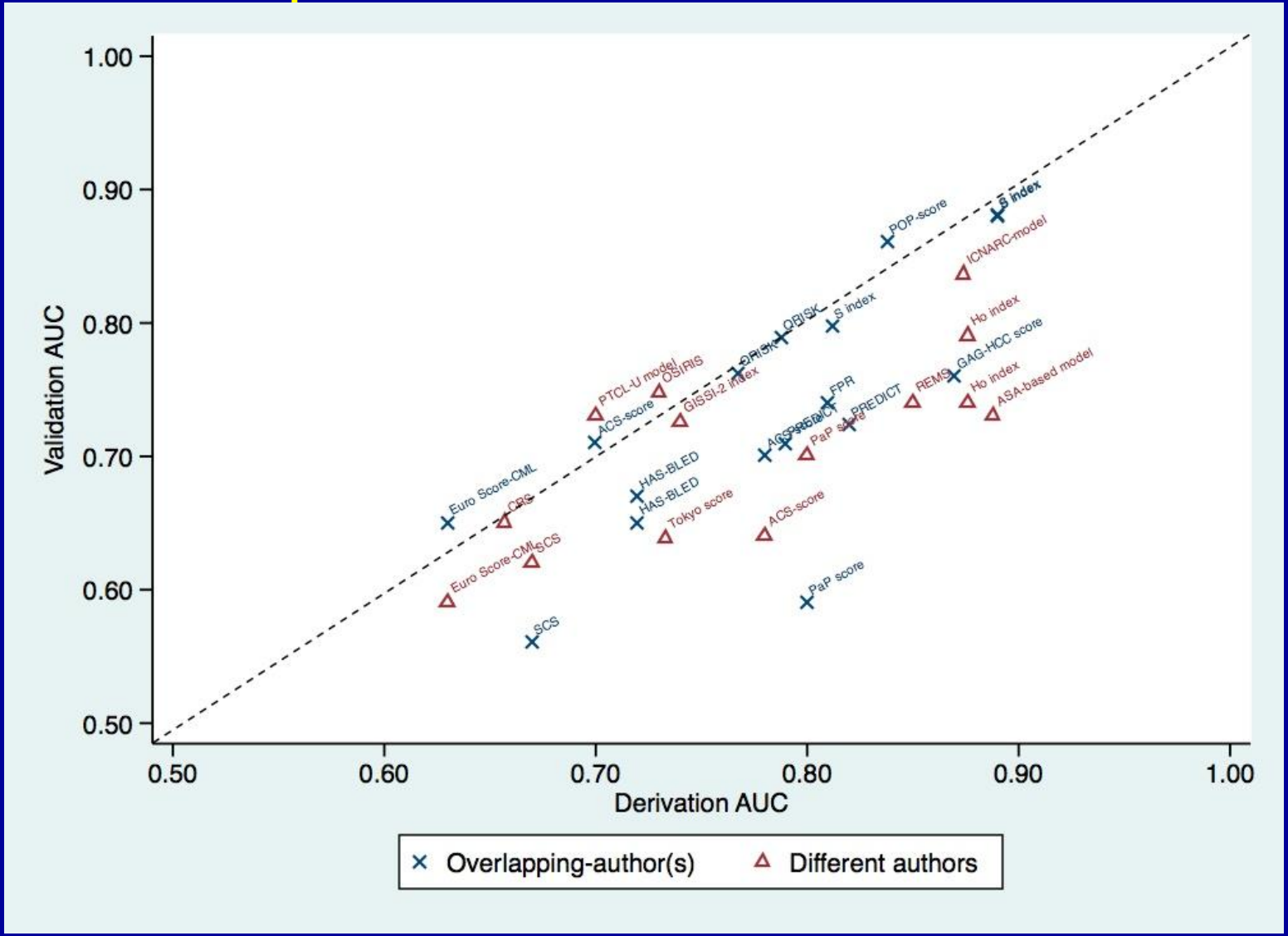
Ioannidis, Why most true discovered associations are inflated. Epidemiology 2008

Effect sizes for the top-cited biomarkers in the biomedical literature



Ioannidis and Panagiotou, JAMA 2011

Decrease in AUC of predictive models upon external validation



Siontis, Castaldi, Tzoulaki, Ioannidis, JCE 2015

Adjusting effects downwards

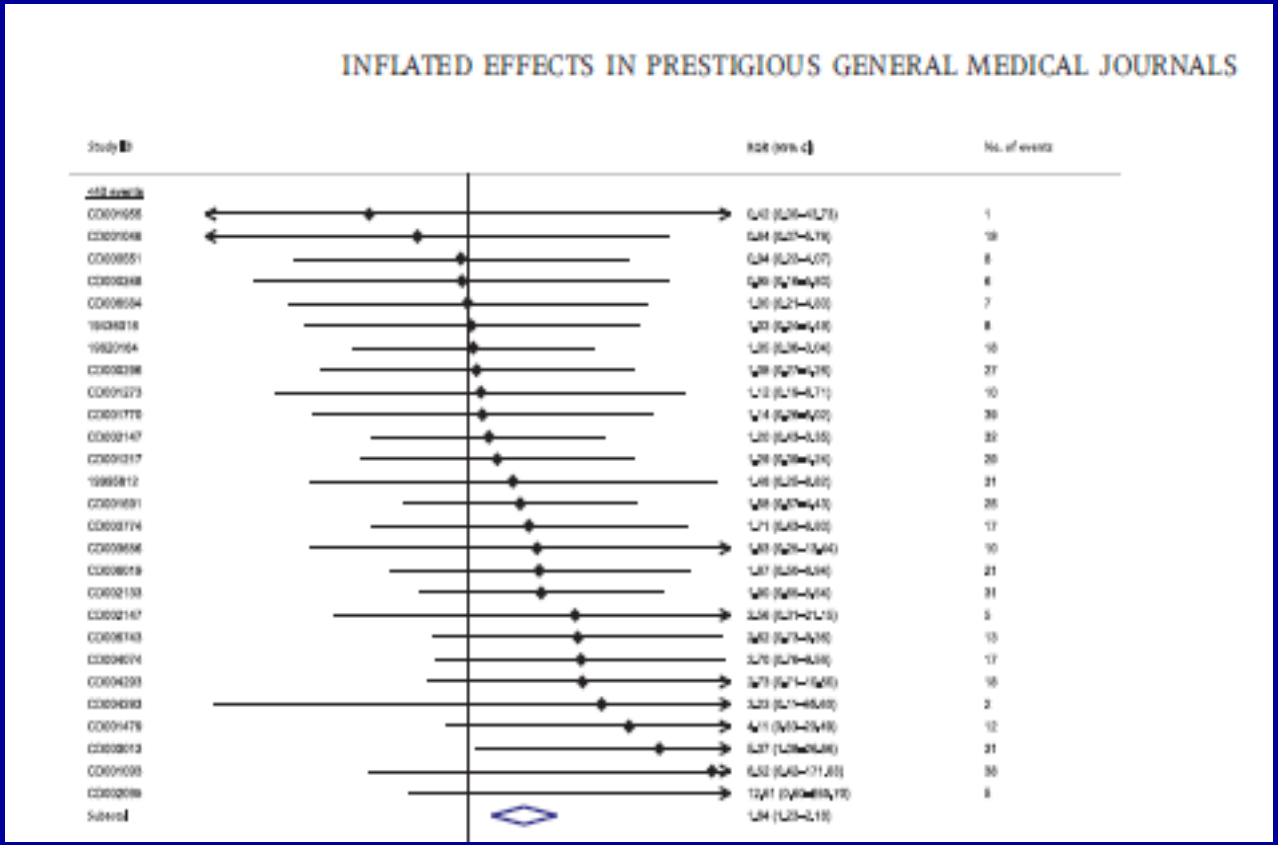
Published by Oxford University Press on behalf of the International Epidemiological Association
© The Author 2011; all rights reserved. Advance Access publication 8 September 2011

International Journal of Epidemiology 2011;**40**:1280–1291
doi:10.1093/ije/dyr095

METHODOLOGY

Magnitude of effects in clinical trials published in high-impact general medical journals

Konstantinos CM Siontis,[1] Evangelos Evangelou[1] and John PA Ioannidis[1,2,3,4]*



Repeatability

nature genetics

ANALYSIS

Repeatability of published microarray gene expression analyses

John P A Ioannidis[1–3], David B Allison[4], Catherine A Ball[5], Issa Coulibaly[4], Xiangqin Cui[4], Aedín C Culhane[6,7], Mario Falchi[8,9], Cesare Furlanello[10], Laurence Game[11], Giuseppe Jurman[10], Jon Mangion[11], Tapan Mehta[4], Michael Nitzberg[5], Grier P Page[4,12], Enrico Petretto[11,13] & Vera van Noort[14]

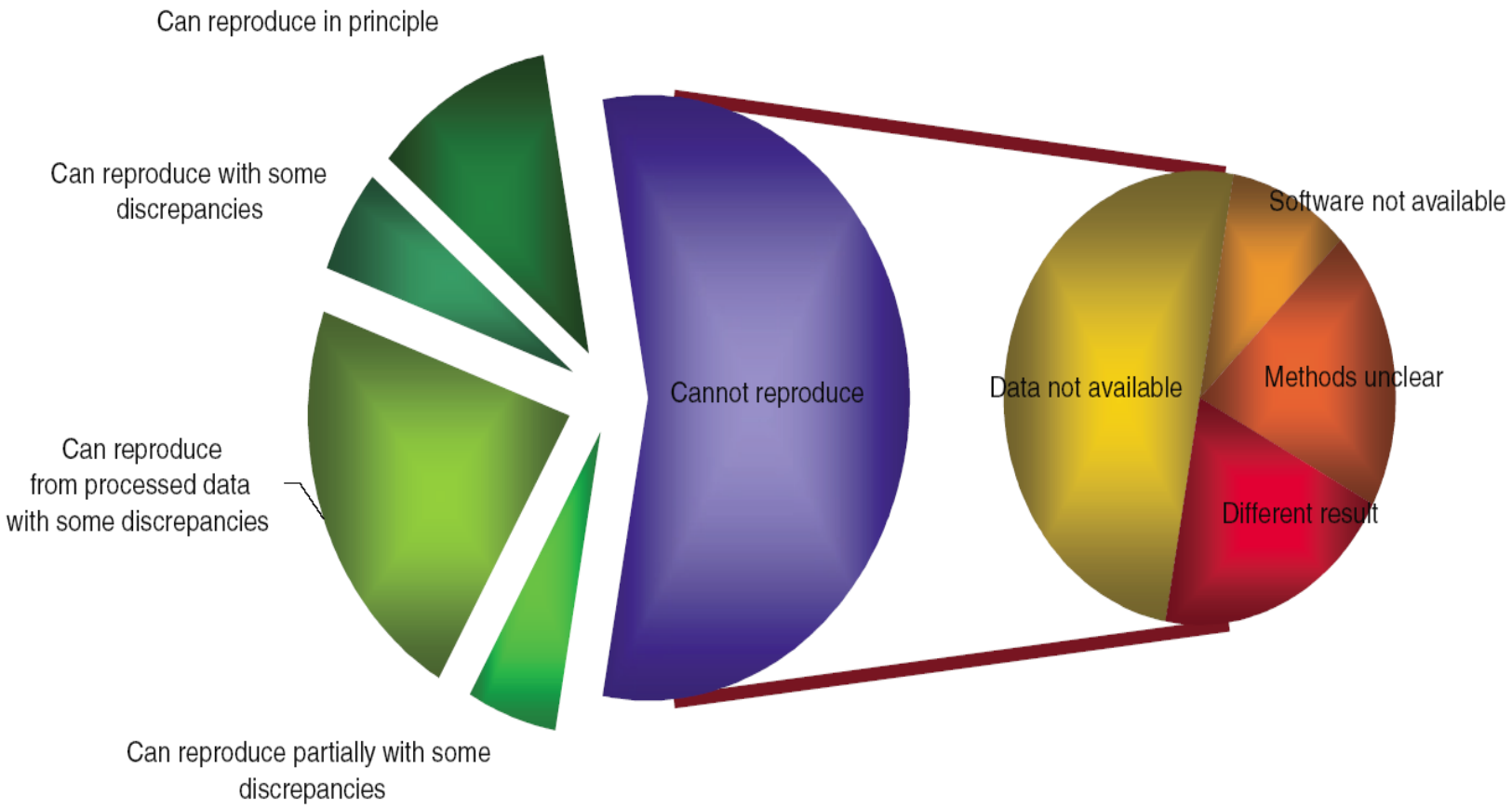


Figure 1 Summary of the efforts to replicate the published analyses.

REPRODUCIBILITY

Enhancing Reproducibility for Computational Methods

Data, code and workflows should be available and cited.

By Victoria Stodden, Marcia McNutt, David H. Bailey, Ewa Deelman, Yolanda Gil, Brooks Hanson, Michael A. Heroux, John P.A. Ioannidis, Michela Taufer

Science, December 2, 2016

Box 1. Some Research Practices that May Help Increase the Proportion of True Research Findings

- Large-scale collaborative research
- Adoption of replication culture
- Registration (of studies, protocols, analysis codes, datasets, raw data, and results)
- Sharing (of data, protocols, materials, software, and other tools)
- Reproducibility practices
- Containment of conflicted sponsors and authors
- More appropriate statistical methods
- Standardization of definitions and analyses
- More stringent thresholds for claiming discoveries or "successes"
- Improvement of study design standards
- Improvements in peer review, reporting, and dissemination of research
- Better training of scientific workforce in methods and statistical literacy

Ioannidis, PLoS Medicine 2014

Guidelines as a marketing tool and as a potential threat to patients

CLINICAL GUIDELINES

Ensuring the integrity of clinical practice guidelines: a tool for protecting patients

Jeanne Lenzer, **Jerome Hoffman**, **Curt Furberg**, and **John Ioannidis** pull together a large expert working group to offer a manifesto for clinical guidelines

Box 1: Red flags that should raise substantial skepticism among guideline readers (and medical journals)

- Sponsor(s) is a professional society that receives substantial industry funding;
- Sponsor is a proprietary company, or is undeclared or hidden
- Committee chair(s) have any financial conflict*
- Multiple panel members have any financial conflict*
- Any suggestion of committee stacking that would pre-ordain a recommendation regarding a controversial topic
- No or limited involvement of an expert in methodology in the evaluation of evidence
- No external review
- No inclusion of non-physician experts/patient representative/community stakeholders

*Includes a panelist with either or both a financial relationship with a proprietary healthcare company and/or whose clinical practice/specialty depends on tests or interventions covered by the guideline

Why Most Clinical Research Is Not Useful

John P. A. Ioannidis[1,2]*

1 Stanford Prevention Research Center, Department of Medicine and Department of Health Research and Policy, Stanford University School of Medicine, Palo Alto, California, United States of America, **2** Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Palo Alto, California, United States of America

* jioannid@stanford.edu

Summary Points

- Blue-sky research cannot be easily judged on the basis of practical impact, but clinical research is different and should be useful. It should make a difference for health and disease outcomes or should be undertaken with that as a realistic prospect.
- Many of the features that make clinical research useful can be identified, including those relating to problem base, context placement, information gain, pragmatism, patient centeredness, value for money, feasibility, and transparency.
- Many studies, even in the major general medical journals, do not satisfy these features, and very few studies satisfy most or all of them. Most clinical research therefore fails to be useful not because of its findings but because of its design.
- The forces driving the production and dissemination of nonuseful clinical research are largely identifiable and modifiable.
- Reform is needed. Altering our approach could easily produce more clinical research that is useful, at the same or even at a massively reduced cost.

S

A (2016) Why Most Clinical
l. PLoS Med 13(6): e1002049.
ned.1002049

2016

Table 1. Features to consider in appraising whether clinical research is useful.

Feature	Questions to Ask
Problem base	Is there a health problem that is big/important enough to fix?
Context placement	Has prior evidence been systematically assessed to inform (the need for) new studies?
Information gain	Is the proposed study large and long enough to be sufficiently informative?
Pragmatism	Does the research reflect real life? If it deviates, does this matter?
Patient centeredness	Does the research reflect top patient priorities?
Value for money	Is the research worth the money?
Feasibility	Can this research be done?
Transparency	Are methods, data, and analyses verifiable and unbiased?

doi:10.1371/journal.pmed.1002049.t001

Concluding comments

- Type of design, sample size (power), presence of conflicts of interest (financial or other), analytical flexibility, and potential for (hidden) multiplicity may influence the credibility of different types of research
- Each paper/study may have a multitude of other features that may help understand how credible it is
- Significant does not mean credible
- Credible does not mean useful

Special thanks

- Daniele Fanelli, Stanford University
- Steve Goodman, Stanford University
- Shanil Ebrahim, Stanford University
- Despina Contopoulos-Ioannidis, Stanford University
- Robert Tibshirani, Stanford University
- Chirag Patel, Stanford University and Harvard University
- David Chavalarias, ISC, Paris
- Lamberto Manzoli, University of Chieti
- Maria Elen Flacco, University of Chieti
- Paolo Villari, University of Rome La Sapienza
- Fainia Kavvoura, Oxford University
- Kostas Siontis, Mayo Clinic
- George Siontis, University of Bern
- Vangelis Evangelou, University of Ioannina
- Muin Khoury, CDC and NCI
- Orestis Panagiotou, National Cancer Institute
- Kevin Boyack, Map of Science, SciTech
- Glenn Begley, Amgen, TetraLogic
- Joseph Lau, Brown University
- Tiago Pereira, U Sao Paolo
- Malcolm MacLeod, University of Edinburgh
- Kostas Tsilidis, University of Ioannina
- David Allison, University of Alabama at Birmingham
- Brian Nosek, Center for Open Science
- Belinda Burford, University of Melbourne
- David Goldstein, Duke University
- Jeanne Lenzer, BMJ