

# What Does Research Reproducibility Mean?

Health Effects Institute Conference

April 30, 2018

Chicago, IL

**Steven Goodman, MD, MHS, PhD**

Assoc. Dean, Clinical and Translational Research

Professor of Medicine and Epidemiology

Chief, Division of Epidemiology

Co-director, Meta-research Innovation Center at Stanford (METRICS)

**[steve.goodman@stanford.edu](mailto:steve.goodman@stanford.edu)**



---

## Commentary

---

# Reproducible Epidemiologic Research

**Roger D. Peng, Francesca Dominici, and Scott L. Zeger**

From the Biostatistics Department, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD.

*Received for publication November 4, 2005; accepted for publication January 10, 2006.*

---

The replication of important findings by multiple independent investigators is fundamental to the accumulation of scientific evidence. Researchers in the biologic and physical sciences expect results to be replicated by independent data, analytical methods, laboratories, and instruments. Epidemiologic studies are commonly used to quantify small health effects of important, but subtle, risk factors, and replication is of critical importance where results can inform substantial policy decisions. However, because of the time, expense, and opportunism of many current epidemiologic studies, it is often impossible to fully replicate their findings. An attainable minimum standard is “reproducibility,” which calls for data sets and software to be made available for verifying published findings and conducting alternative analyses. The authors outline a standard for reproducibility and evaluate the reproducibility of current epidemiologic research. They also propose methods for reproducible research and implement them by use of a case study in air pollution and health.

air pollution; information dissemination; models, statistical

---



## Commentary

### Reproducible Epic

**Roger D. Peng, Franc**

From the Biostatistics Dep

*Received for publication N*

**TABLE 1. Criteria for reproducible epidemiologic research**

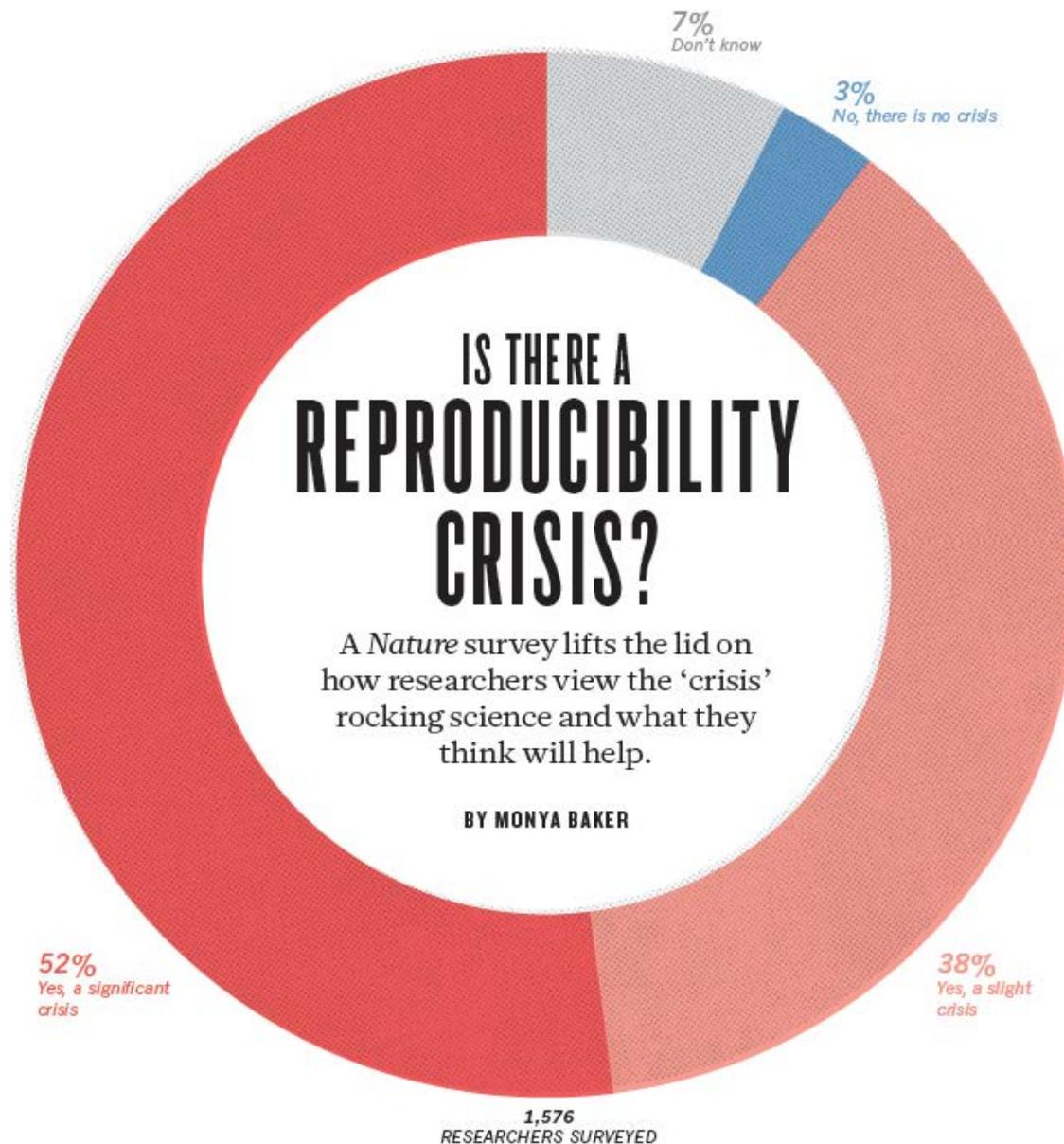
Research component	Requirement
Data	Analytical data set is available.
Methods	Computer code underlying figures, tables, and other principal results is made available in a human-readable form. In addition, the software environment necessary to execute that code is available.
Documentation	Adequate documentation of the computer code, software environment, and analytical data set is available to enable others to repeat the analyses and to conduct other similar ones.
Distribution	Standard methods of distribution are used for others to access the software, data, and documentation.

accumulation of  
 ted by indepen-  
 used to quantify  
 here results can  
 of many current  
 um standard is  
 ed findings and

The replication o  
 scientific evidence  
 dent data, analytica  
 small health effects  
 inform substantial  
 epidemiologic stud  
 “reproducibility,” w

conducting alternative analyses. The authors outline a standard for reproducibility and evaluate the reproducibility of current epidemiologic research. They also propose methods for reproducible research and implement them by use of a case study in air pollution and health.

air pollution; information dissemination; models, statistical







# Policy: NIH plans to enhance reproducibility

**Francis S. Collins & Lawrence A. Tabak**

27 January 2014

**Francis S. Collins and Lawrence A. Tabak discuss initiatives that the US National Institutes of Health is exploring to restore the self-correcting nature of preclinical research.**

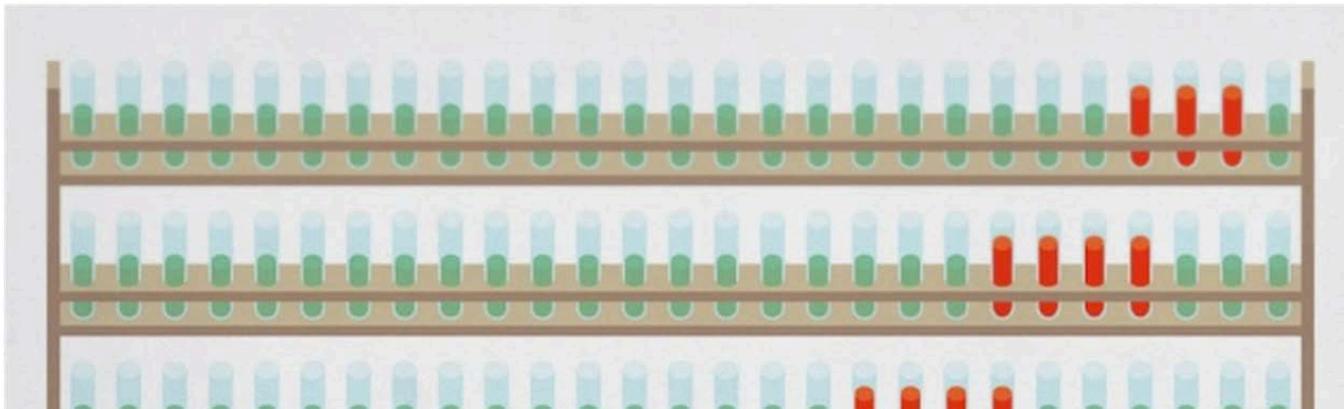


PDF



Rights & Permissions

**Subject terms:** [Biological techniques](#) · [Lab life](#) · [Peer review](#) · [Research management](#)



# Collins/Tabak on Reproducibility

...a complex array of other factors seems to have contributed to the lack of reproducibility. Factors include poor training of researchers in experimental design; increased emphasis on making provocative statements rather than presenting technical details; and publications that do not report basic elements of experimental design.

Some irreproducible reports are probably the result of coincidental findings that happen to reach statistical significance, coupled with publication bias.

Another pitfall is overinterpretation of creative ‘hypothesis-generating’ experiments, which are designed to uncover new avenues of inquiry rather than to provide definitive proof for any single question. Still, there remains a troubling frequency of published reports that claim a significant result, but fail to be reproducible.



# Collins/Tabak on Reproducibility

...a complex array of other factors seems to have contributed to the lack of reproducibility. Factors include **poor training of researchers in experimental design**; increased emphasis on **making provocative statements** rather than presenting technical details; and publications that **do not report basic elements of experimental design**.

Some irreproducible reports are probably the result of **coincidental findings that happen to reach statistical significance**, coupled with **publication bias**.

Another pitfall is **overinterpretation** of creative ‘hypothesis-generating’ experiments, which are designed to uncover new avenues of inquiry rather than to provide definitive proof for any single question. Still, there remains a troubling frequency of published reports **that claim a significant result, but fail to be reproducible**.

## RESEARCH ARTICLE

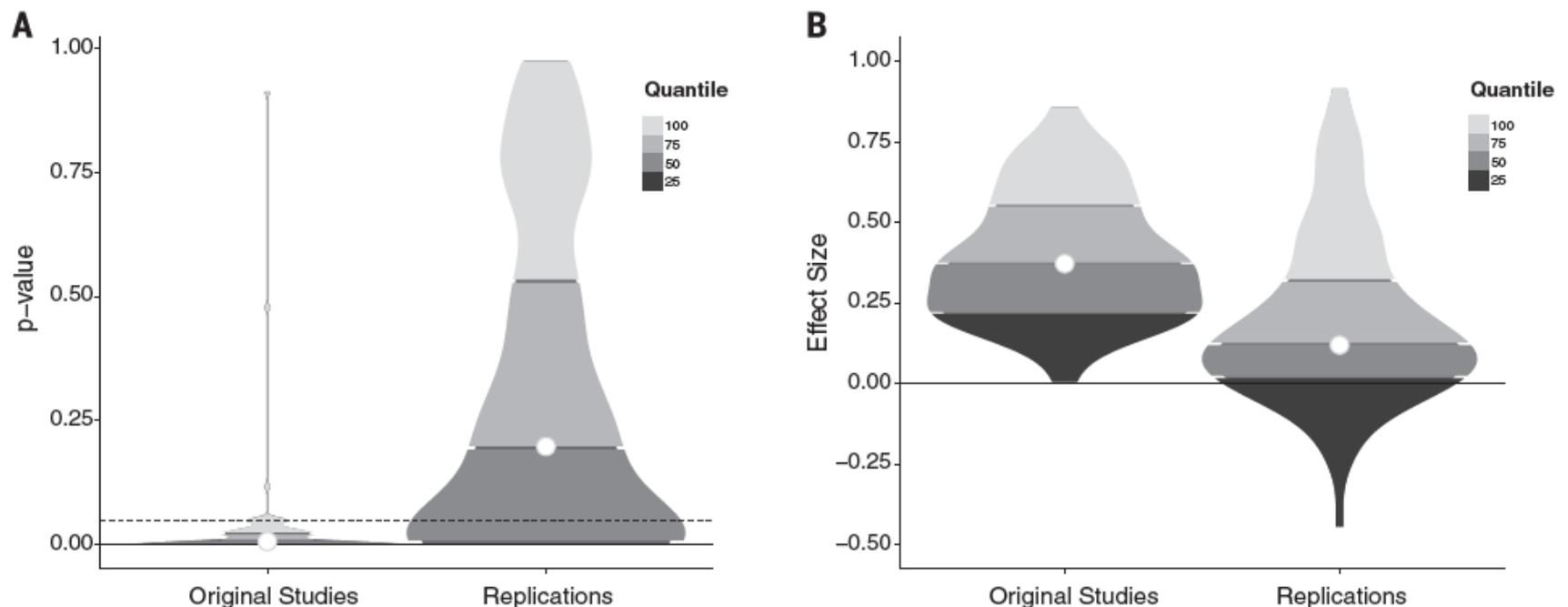
## PSYCHOLOGY

# Estimating the reproducibility of psychological science

Open Science Collaboration\*†

facilitated each step of the process and maintained the protocol and project resources. Replication materials and data were required to be archived publicly in order to maximize transparency, accountability, and reproducibility of the project (<https://osf.io/ezcu>).

In total, 100 replications were completed by 270 contributing authors. There were many different research designs and analysis strategies in the original research. Through consultation with original authors, obtaining original materials, and internal review, replications maintained high fidelity to the original designs. Analyses con-



**Fig. 1. Density plots of original and replication  $P$  values and effect sizes. (A)  $P$  values. (B) Effect sizes (correlation coefficients). Lowest quantiles for  $P$  values are not visible because they are clustered near zero.**

# nature

International weekly journal of science

[Home](#)

[News & Comment](#)

[Research](#)

[Careers & Jobs](#)

[Current Issue](#)

[Archive](#)

[Audio & Video](#)

[For Authors](#)

[News & Comment](#)

[News](#)

[2016](#)

[March](#)

[Article](#)

NATURE | NEWS



## Over half of psychology studies fail reproducibility test

Largest replication study to date casts doubt on many published positive results.

**Monya Baker**

27 August 2015



# Many Psychology Findings Not as Strong as Claimed, Study Says

By BENEDICT CAREY AUG. 27, 2015

**Par 2:** “Their conclusions, have confirmed the worst fears of scientists...”

**Par 5:** “More than 60 of the studies did not hold up.”

**Par 9:** “The new analysis...found no evidence ...that any original study was definitively false. Rather, it concluded that the evidence for most published findings was not nearly as strong as originally claimed.

**Par 11:** The report appears at a time when the number of retractions of published papers is rising sharply in a wide variety of disciplines.

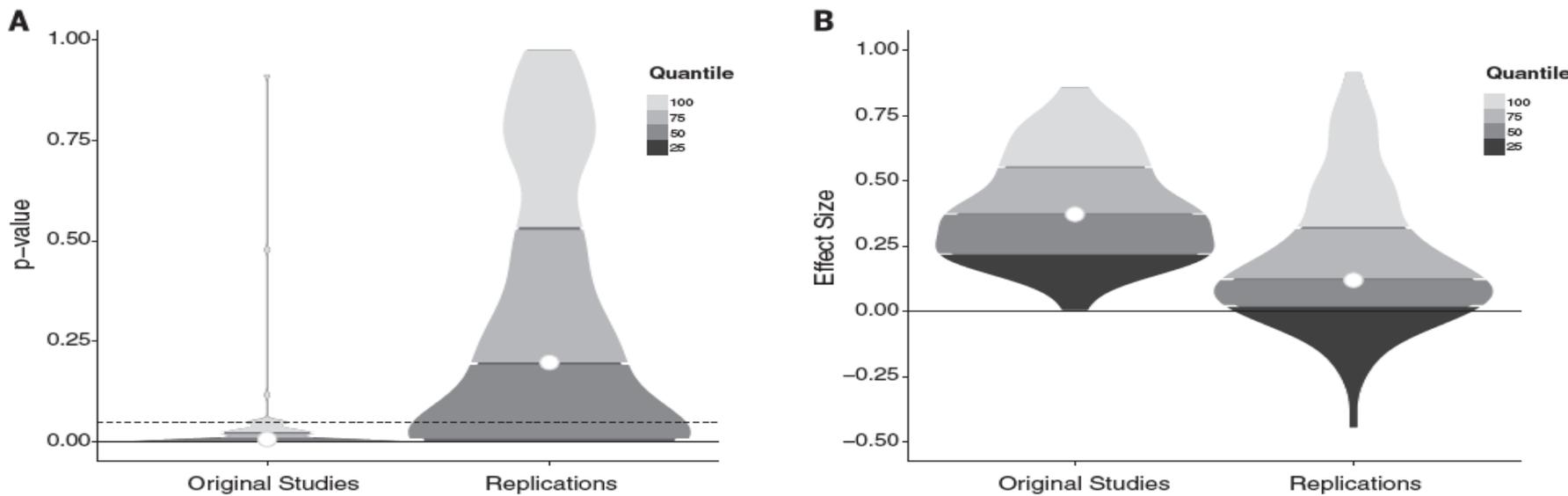
**Pars 19-20:** Yet very few of the redone studies contradicted the original ones; their results were simply weaker. “We think of these findings as two data points, not in terms of true or false,” Dr. Nosek said.

# OSC Definitions of Reproducibility

1. Significance levels (36%)
2. Whether >50% of replication effect sizes exceeded the original. (11%)
3. Whether effect size was within the confidence interval of replication study. (47%)
4. Whether the combined estimate of the original and replication studies was statistically significant. (68%)
5. “Subjective impression” (39%)

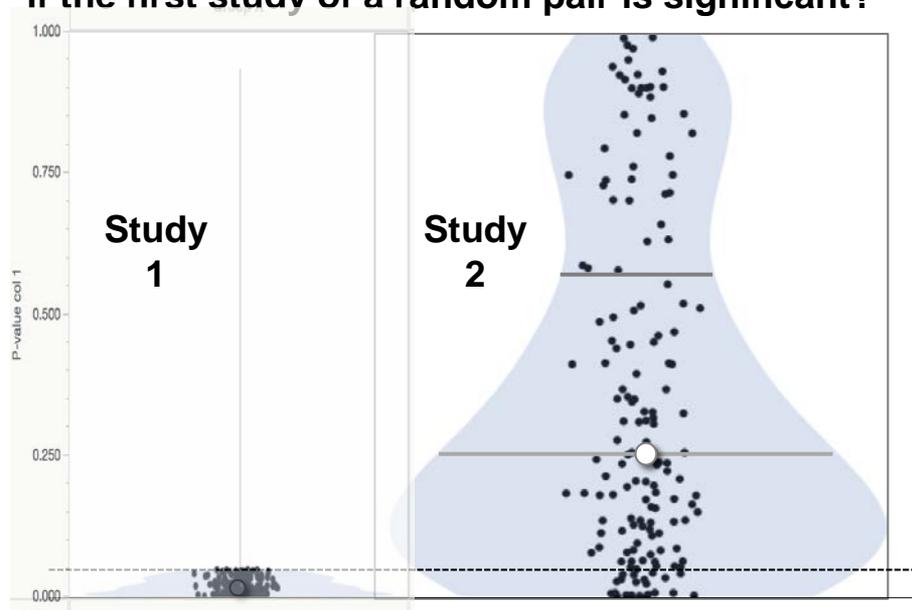
OSC = Open Science Collaboration





**Fig. 1. Density plots of original and replication  $P$  values and effect sizes. (A)  $P$  values. (B) Effect sizes (correlation coefficients). Lowest quantile  $P$  values are not visible because they are clustered near zero.**

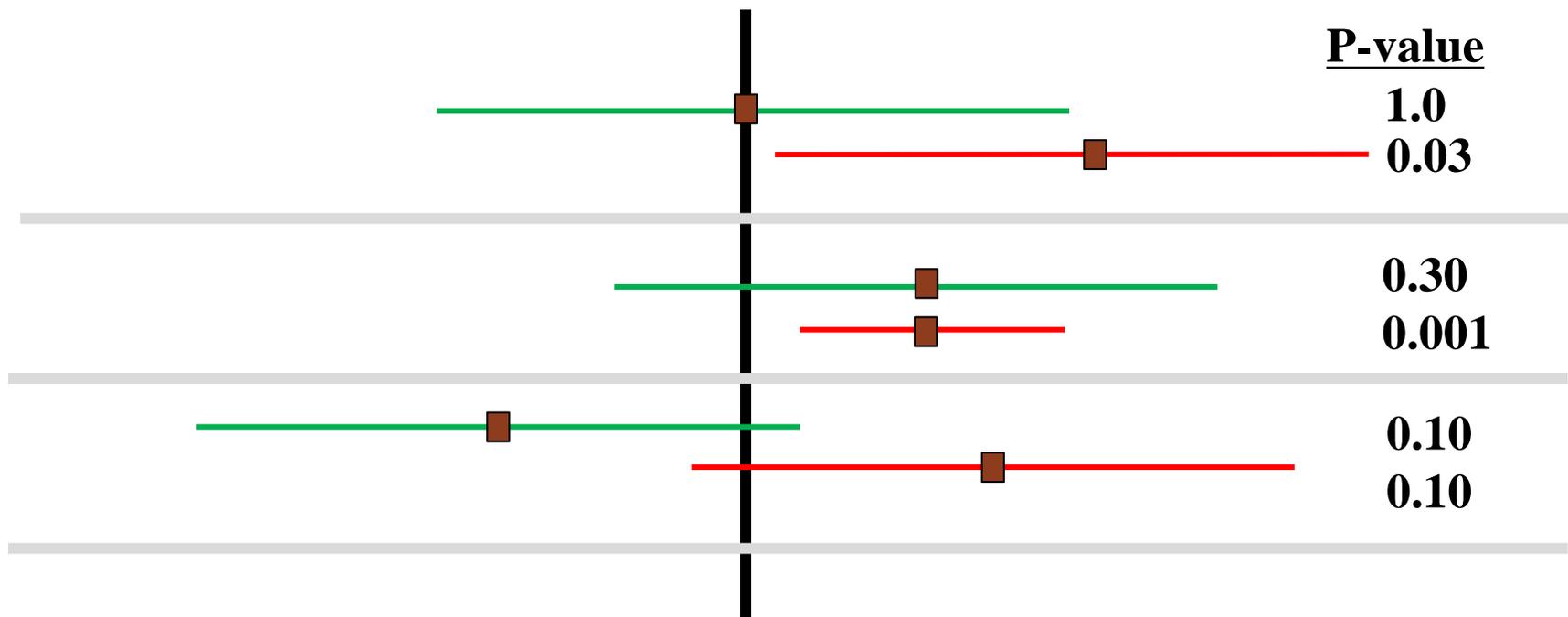
**What if we assume a 1 SE effect, but only “publish” if the first study of a random pair is significant?**



**The pattern is recreated by:**  
**1.) Publication bias**  
**2.) Regression to the mean**

**All of these estimates are from the same “truth”!**

# Which results didn't "reproduce"?



# The New England Journal of Medicine

©Copyright, 1988, by the Massachusetts Medical Society

Volume 319

DECEMBER 29, 1988

Number 26

## EFFECTS OF ADJUVANT TAMOXIFEN AND OF CYTOTOXIC THERAPY ON MORTALITY IN EARLY BREAST CANCER

### An Overview of 61 Randomized Trials among 28,896 Women

EARLY BREAST CANCER TRIALISTS' COLLABORATIVE GROUP

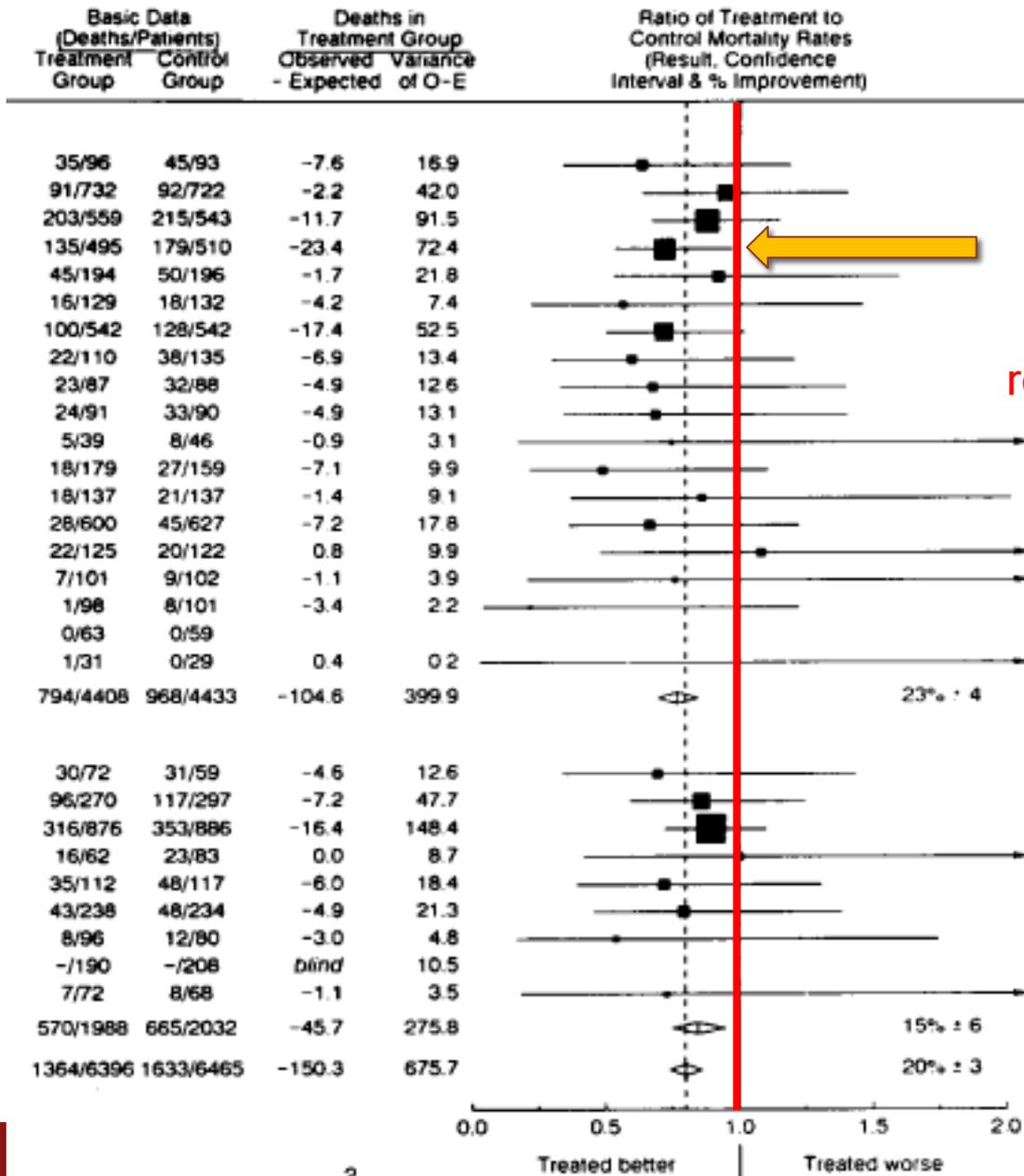
**Abstract** We sought information worldwide on mortality according to assigned treatment in all randomized trials that began before 1985 of adjuvant tamoxifen or cytotoxic therapy for early breast cancer (with or without regional lymph-node involvement). Coverage was reasonably complete for most countries. In 28 trials of tamoxifen nearly 4000 of 16,513 women had died, and in 40 chemotherapy trials slightly more than 4000 of 13,442 women had died. The 8106 deaths were approximately evenly distributed over years 1, 2, 3, 4, and 5+ of follow-up, with little useful information beyond year 5.

Systematic overviews of the results of these trials demonstrated reductions in mortality due to treatment that were significant when tamoxifen was compared with no tamoxifen ( $P < 0.0001$ ), any chemotherapy with no chemotherapy ( $P = 0.003$ ), and polychemotherapy with single-agent chemotherapy ( $P = 0.001$ ). In tamoxifen trials,

there was a clear reduction in mortality only among women 50 or older, for whom assignment to tamoxifen reduced the annual odds of death during the first five years by about one fifth. In chemotherapy trials there was a clear reduction only among women under 50, for whom assignment to polychemotherapy reduced the annual odds of death during the first five years by about one quarter. Direct comparisons showed that combination chemotherapy was significantly more effective than single-agent therapy, but suggested that administration of chemotherapy for 8 to 24 months may offer no survival advantage over administration of the same chemotherapy for 4 to 6 months.

Because it involved several thousand women, this overview was able to demonstrate particularly clearly that both tamoxifen and cytotoxic therapy can reduce five-year mortality. (N Engl J Med 1988; 319:1681-92.)

**(B) Women aged ≥ 50 years at entry**



22 of 23 trials  
have  
P > 0.05

Was a null-effect  
reproduced 22 times?



Test for heterogeneity:  $X^2_{26} = 19.7$ : NS

# Making sense of replications

**Abstract** The first results from the Reproducibility Project: Cancer Biology suggest that there is scope for improving reproducibility in pre-clinical cancer research.

DOI: [10.7554/eLife.23383.001](https://doi.org/10.7554/eLife.23383.001)

**BRIAN A NOSEK AND TIMOTHY M ERRINGTON\***

**eLIFE** Feature article

Repro

**There is no straightforward answer to the question "what counts as a successful replication of an original result?"**

# Making sense of replications

**Abstract** The first results from the Reproducibility Project: Cancer Biology suggest that there is scope for improving reproducibility in pre-clinical cancer research.

DOI: [10.7554/eLife.23383.001](https://doi.org/10.7554/eLife.23383.001)

**BRIAN A NOSEK AND TIMOTHY M ERRINGTON\***

**“Scientific claims gain credibility by accumulating evidence from multiple experiments, and a single study cannot provide conclusive evidence for or against a claim. Equally, a single replication cannot make a definitive statement about the original finding. However, the new evidence provided by a replication can increase or decrease confidence in the reproducibility of the original finding. When a replication “fails” it can spur productive theorizing about the source of that irreproducibility.”**

# What does research reproducibility mean?

Steven N. Goodman,\* Daniele Fanelli, John P. A. Ioannidis

The language and conceptual framework of “research reproducibility” are nonstandard and unsettled across the sciences. In this Perspective, we review an array of explicit and implicit definitions of reproducibility and related terminology, and discuss how to avoid potential misunderstandings when these terms are used as a surrogate for “truth.”

[www.ScienceTranslationalMedicine.org](http://www.ScienceTranslationalMedicine.org) 1 June 2016 Vol 8 Issue 341 341ps12

# Meanings of reproducibility

## Methods reproducibility

- With same data, can the analytic findings be reproduced?
- Includes computational reproducibility
- Related to *processes* of science, addresses issues of trust
- *Transparency, methods reporting, data and code sharing*

## Results reproducibility

- Related to *results of science*
- *New evidence, confirmation?*

## Inferential reproducibility

- Related to *interpretation of results*
- Strength of claims
- Truth?

# What are “data?”

- Raw data
- Abstracted data
- Coded data
- Computerized data
- Edited, a.k.a. “cleaned” data
- Derived, transformed data
- Analyzable data
- Analyzed data, data summaries



# Raw Data

Death certificates

Sensor readings

CT Scan

Pulmonary Function Test

Air sample

Medical record

Clinical exam video



# Raw Data

STATE OF TEXAS **031-02-2** **03100** CERTIFICATE OF DEATH

STATE FILE NO. [REDACTED]

4/10x

0413-

Texas Department of Health — BUREAU OF VITAL STATISTICS

1. NAME OF DECEASED [Type or print]			[a] First	[b] Middle	[c] Last	2. SEX Male	3. DATE OF DEATH [REDACTED], 19[REDACTED]			
4. RACE White	5a. WAS THE DECEDENT OF SPANISH ORIGIN? NO	5b. IF YES, SPECIFY MEXICAN, CUBAN, PUERTO RICAN, ETC.			6. DATE OF BIRTH	7. AGE [In years last birthday]	IF UNDER 1 YEAR Months	Days	IF UNDER 24 HRS. Hours	Minutes
8a. PLACE OF DEATH — COUNTY Cameron			8b. CITY OR TOWN [If outside city limits, give precinct no.] Harlingen			8c. NAME OF [If not in hospital, give street address] HOSPITAL OR INSTITUTION Valley Bapt. Med. Cen.			8d. INSIDE CITY LIMITS? Yes	
9. MARRIED, NEVER MARRIED, WIDOWED, DIVORCED [Specify] Married	10. BIRTHPLACE [State or foreign country] Wisconsin	11. CITIZEN OF WHAT COUNTRY? USA			12. WAS DECEDENT EVER IN U.S. ARMED FORCES? Yes	13. SURVIVING SPOUSE [If wife, give maiden name] 1 2				
14. SOCIAL SECURITY NO.		15a. USUAL OCCUPATION [Give kind of work done during most of working life, even if retired] Salesman-Retired				15b. KIND OF BUSINESS OR INDUSTRY Hardware				
16a. RESIDENCE — STATE Texas	16b. COUNTY Cameron		16c. CITY OR TOWN [If outside city limits, show rural] Harlingen Rural			16d. STREET ADDRESS [If rural, give location]			16e. INSIDE CITY LIMITS? No	
17. FATHER'S NAME [REDACTED]			18. MOTHER'S MAIDEN NAME [REDACTED]			19. SIGNATURE OF INFORMANT [REDACTED]				
20. PART I Conditions, if any, which gave rise to immediate cause stating the underlying cause last	20. IMMEDIATE CAUSE [Enter only one cause per line for (a), (b), (c)]								Interval between onset and death	
	(a) NATURAL CAUSES DUE TO, OR AS A CONSEQUENCE OF:								Interval between onset and death	
	(b) MOST PROBABLE HEART ATTACK DUE TO, OR AS A CONSEQUENCE OF:								Interval between onset and death	
20. PART II	OTHER SIGNIFICANT CONDITIONS — CONDITIONS CONTRIBUTING TO DEATH BUT NOT RELATED TO CAUSE GIVEN IN PART I (a)								21. AUTOPSY? No	
22a. ACC., SUICIDE, HOM., UNDET., OR PENDING INVEST. [Specify]	22b. DATE OF INJURY [Mo., Day, Yr.]	22c. HOUR OF INJURY	22d. DESCRIBE HOW INJURY OCCURRED							
22e. INJURY AT WORK [Specify yes or no]	22f. PLACE OF INJURY — At home, farm, street, factory, office building, etc. [Specify]			22g. LOCATION	STREET OR R.F.D. NO.	CITY OR TOWN	STATE			
23a. To the best of my knowledge, death occurred at the time, date, and place and					23b. On the basis of examination and					

# Abstracted data

- Questionnaire
- Length of stay
- CT scan reading
- PFTs: FEV1, Tidal Volume
- Cause(s) of death

PFTs = Pulmonary Function Tests



# What are “data?”

- Raw data
- Abstracted data
- Coded data
- Computerized data
- Edited, a.k.a. “cleaned” data
- Derived, transformed data
- Analyzable data
- Analyzed data, data summaries



# Reproducibility, et al.

Reproducibility

Replicability

Repeatability

Reliability

Robustness

Generalizability



# Reproducibility, et al.

Reproducibility

Replicability

Repeatability

Reliability

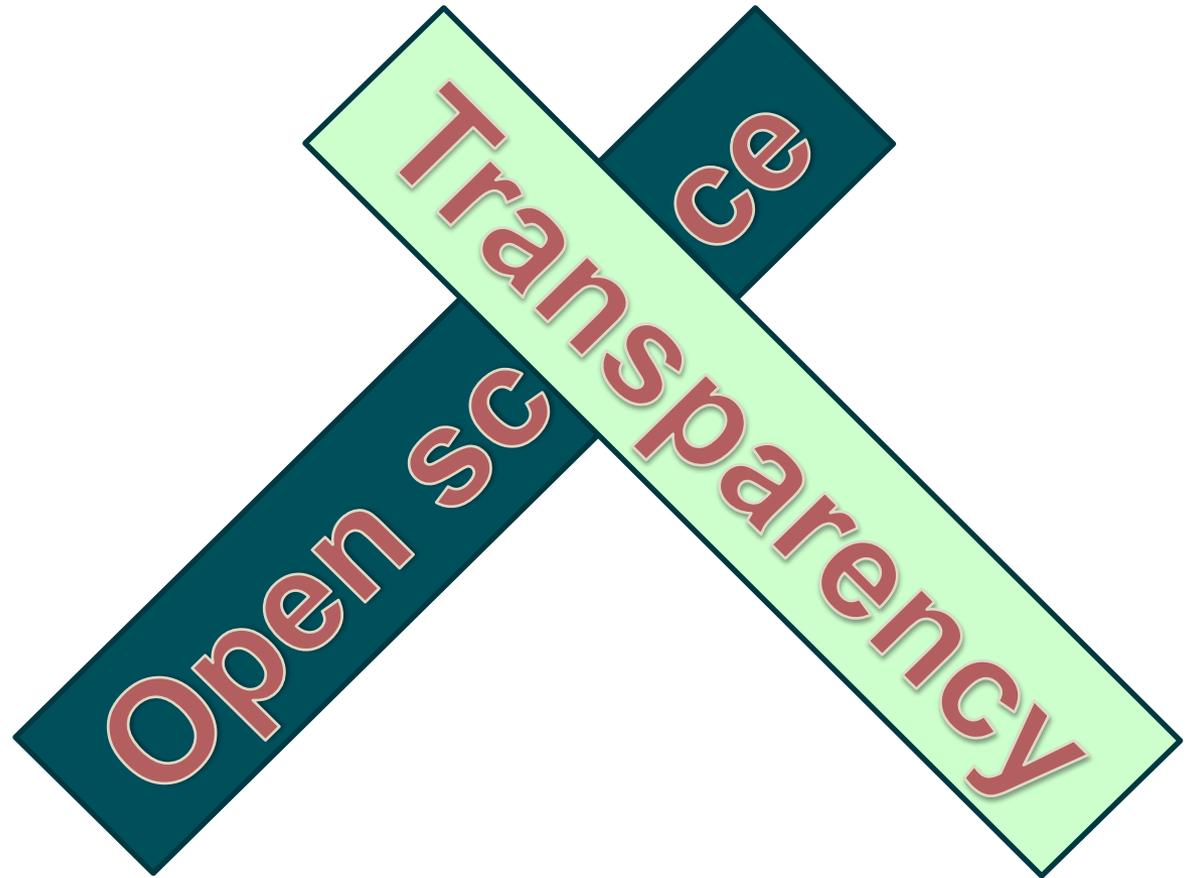
Robustness

Generalizability



# Reproducibility, et al.

- Reproducibility
- Replicability
- Repeatability
- Reliability
- Robustness
- Generalizability



# Reproducibility, et al.

**Truth!**

Video removed: “you can't handle the truth” scene from “A Few Good Men”



# Why can't we handle the truth?

- **Traditional statistical methods have no language or measure for truth.**
- **Many judgments are made in the design and analysis whose effects on proximity to truth cannot be quantified.**
- **What we have is a set of operational procedures and social conventions for when knowledge claims are permitted.**

# What is “statistics”?

- Guide to reasoning under uncertainty.
- #1 goal is to *get the uncertainty right*.
  
- *Uncertainty about what?*
  - ❖ *The truth!*

# Two forms of uncertainty

## ➤ Stochastic – Chance

- ❖ Deductive.

- ❖ Stochastic uncertainty measures (e.g. SEs, CIs) represent the *minimum* uncertainty.

## ➤ Epistemic - Degree of belief (bias, causality, plausibility of effect or effect size, relevance of external information, some design effects, model uncertainty, data quality(?))

- ❖ Inductive

- ❖ Due in part to uncertainty in assumptions.

- ❖ *The degree of uncertainty in any conclusion, qualitative or quantitative, is epistemic.*

# Three sources of truth deviations

➤ **Random error**

➤ **Bias**

➤ **Generalizability / transportability**

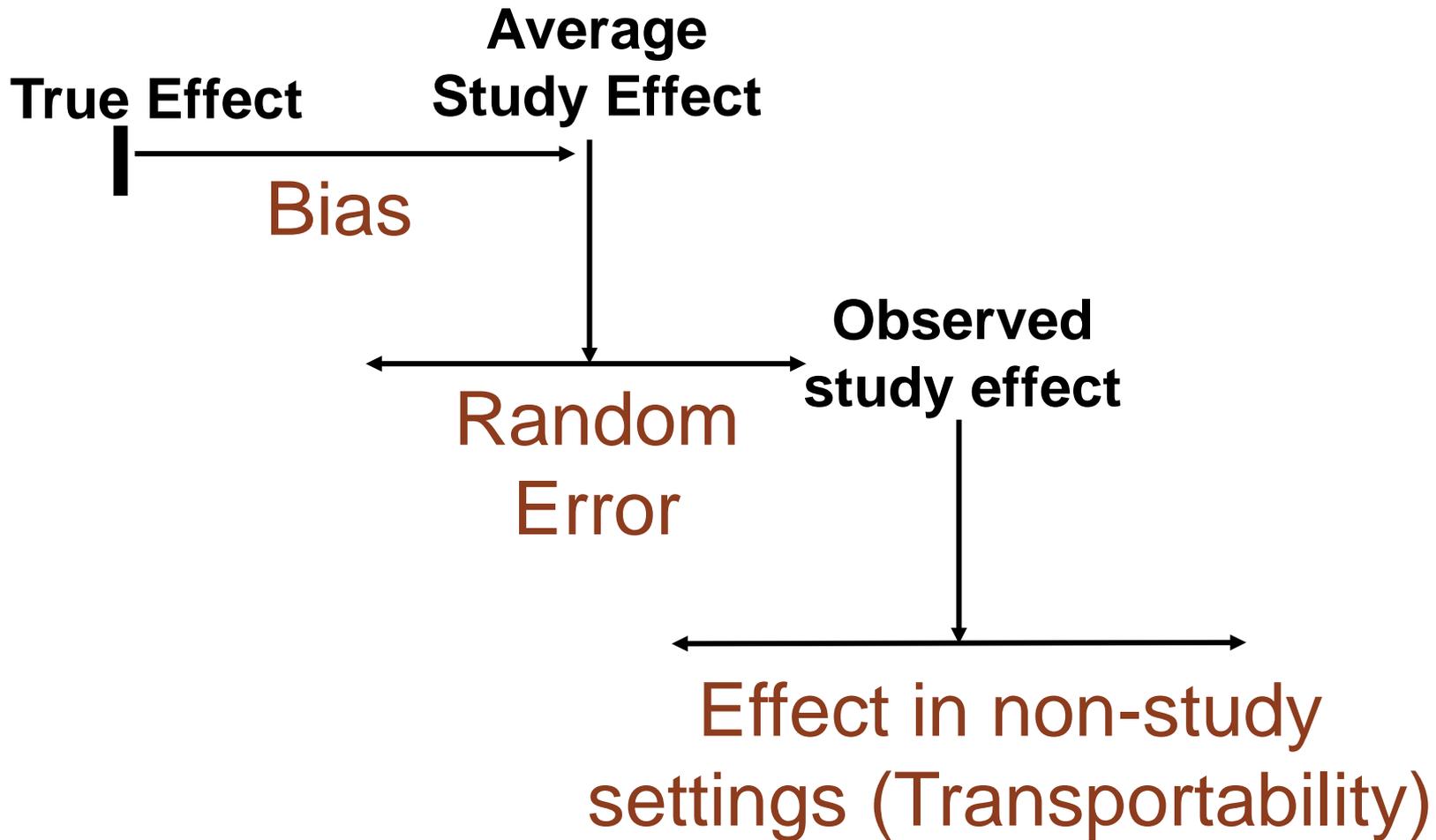
❖ Other people

❖ Other places

❖ Other times

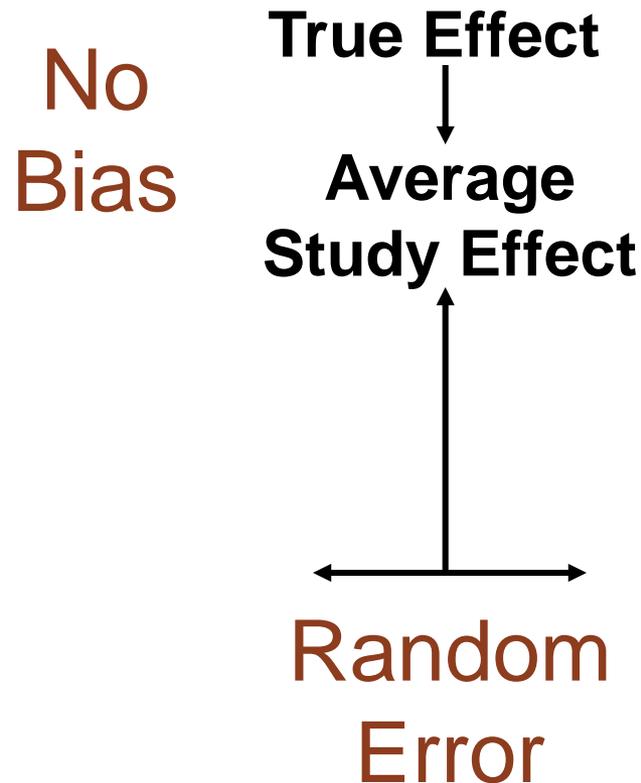
❖ Other exposures

# Effect of Random and Systematic Error



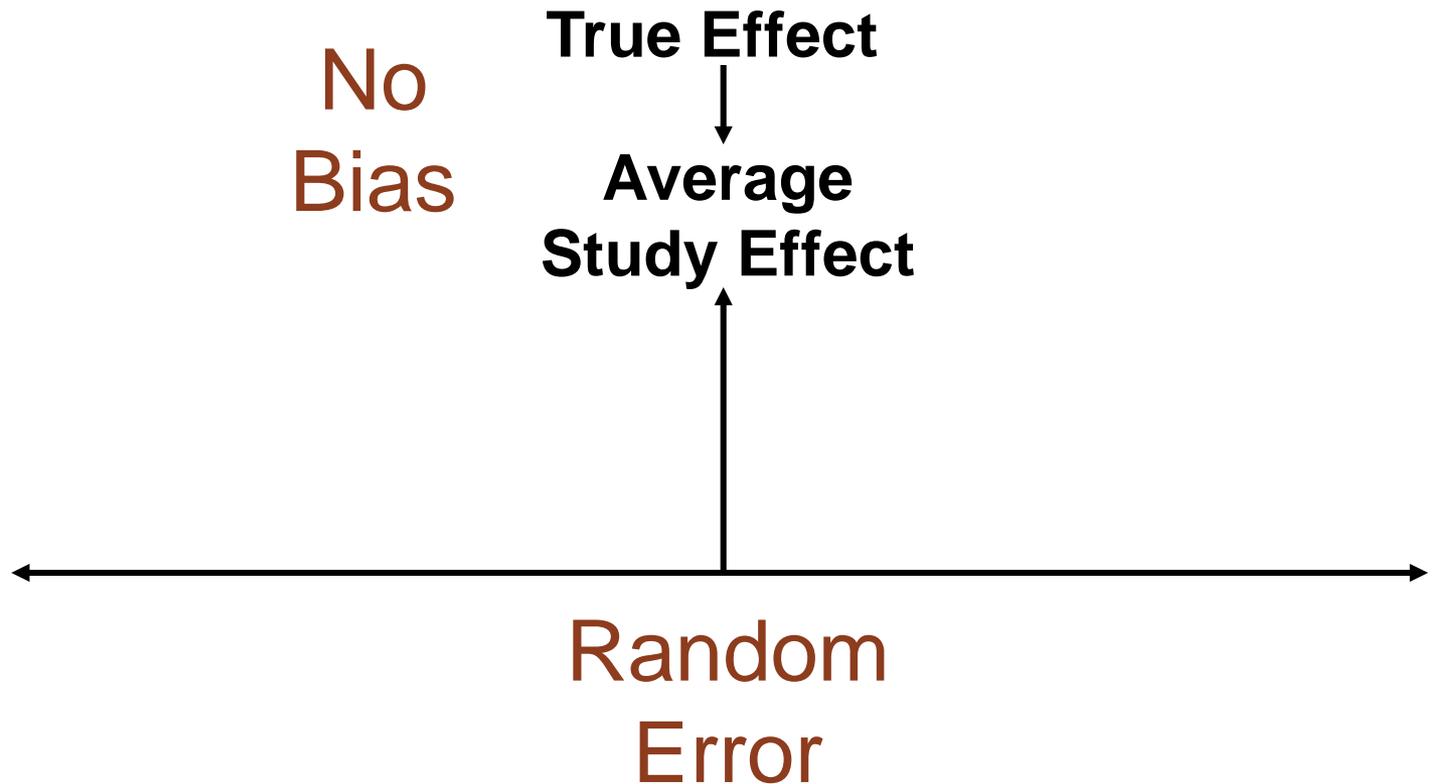
# Large Study - Good Design

e.g. Large, multicenter randomized controlled trial (RCT)



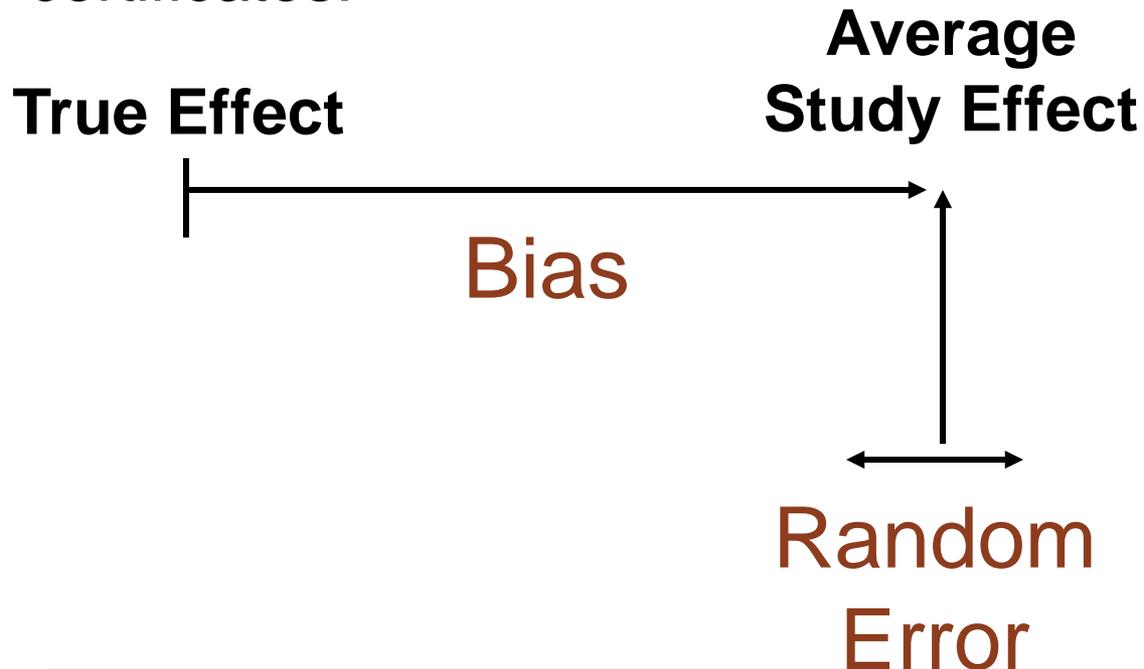
# Small Study - Good Design

e.g. Small, single site RCT



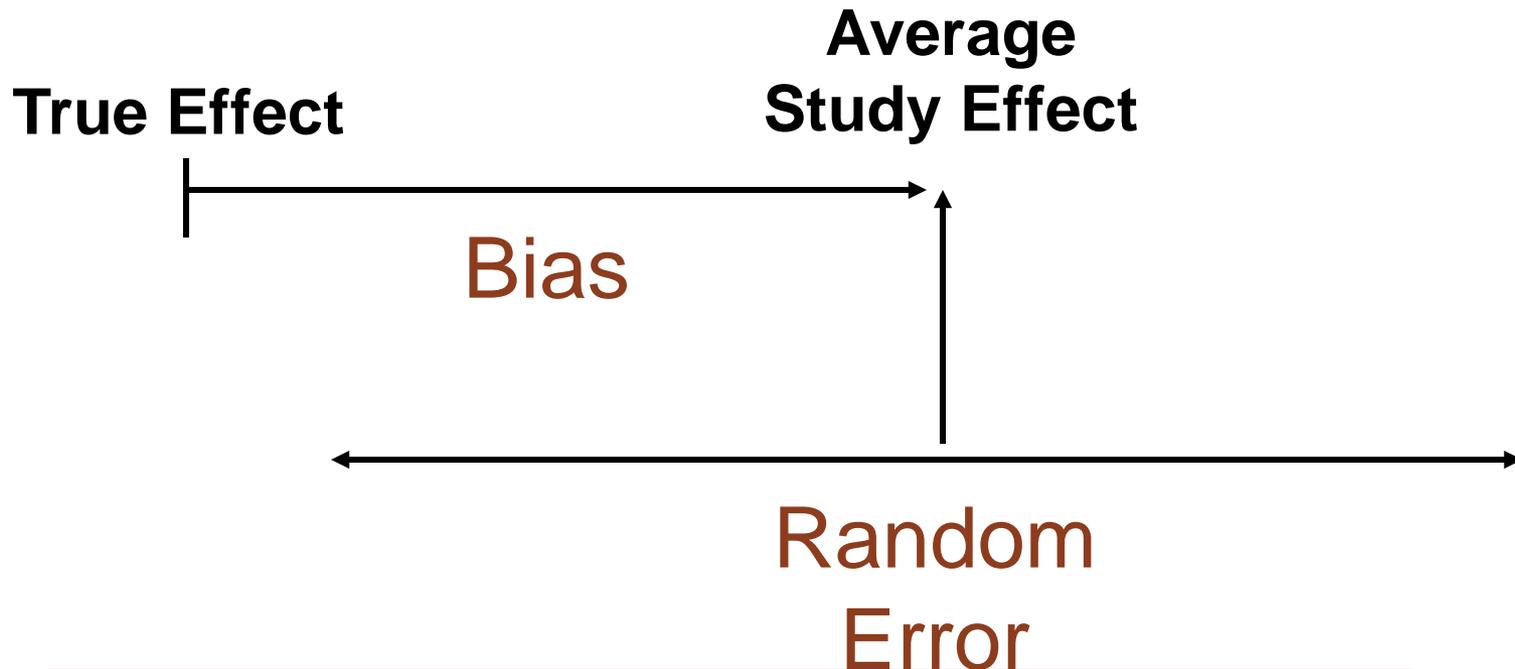
# Large Study - Poor Design

e.g. Correlation of prescriptions from health plan reimbursement files and cause of death from death certificates.

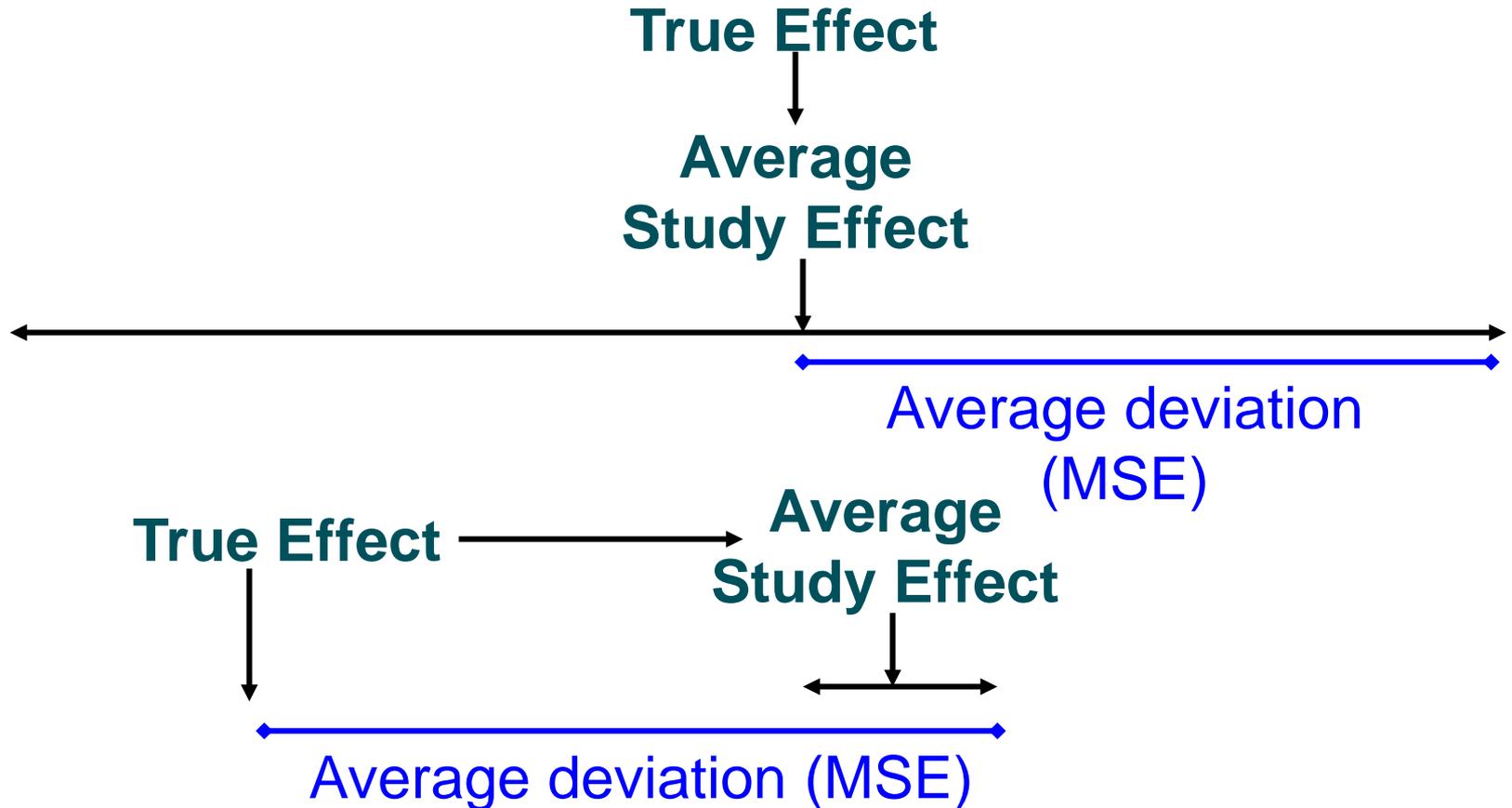


# Small Study - Poor Design

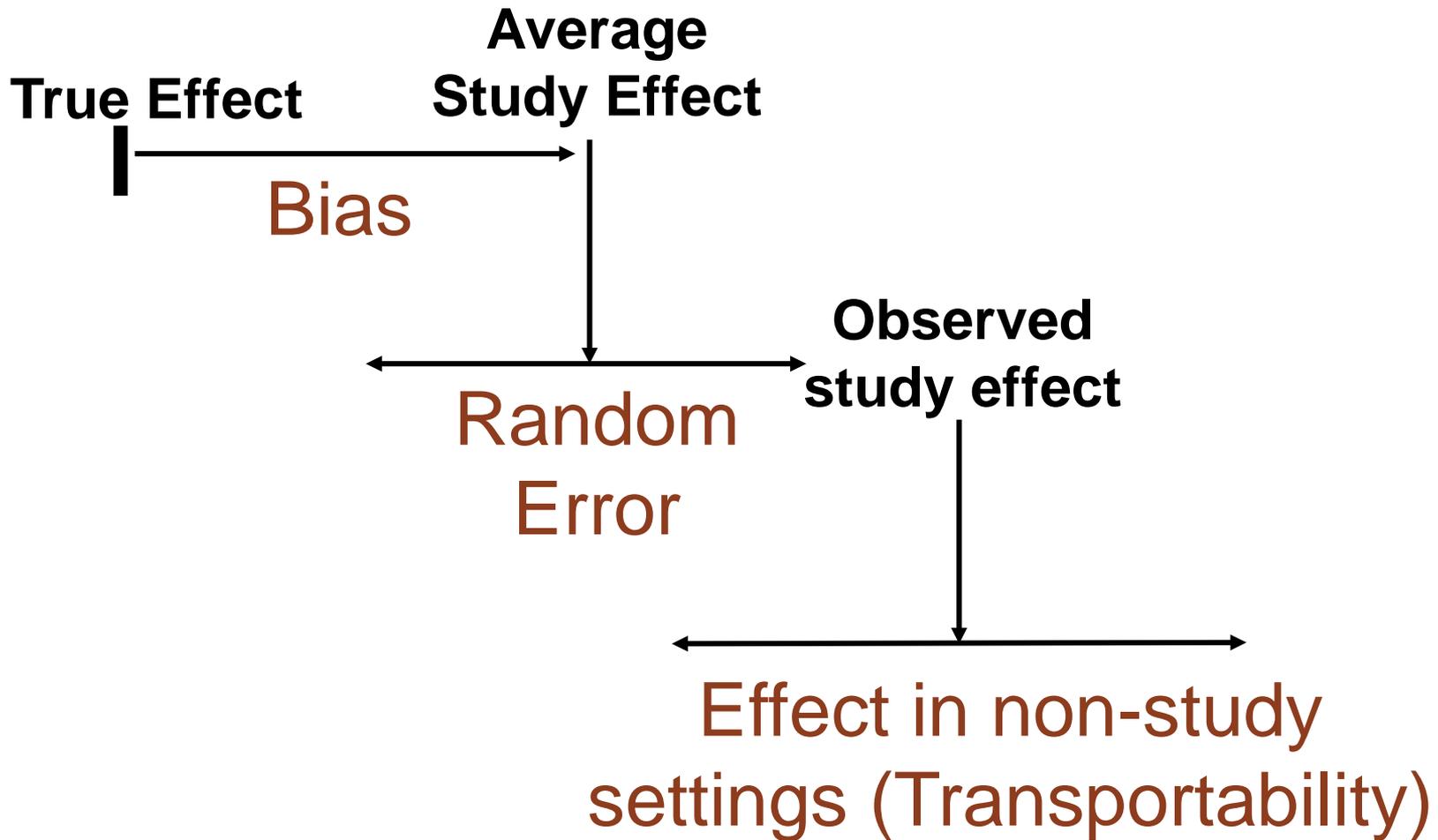
e.g. Single city, one year study, with historical controls



# Which is better?



# Effect of Random and Systematic Error



# On P-values and Truth



# ASA Statement, 2016

“Researchers should bring many contextual factors into play to derive scientific inferences, including the design of a study, the quality of the measurements, the external evidence for the phenomenon under study, and the validity of assumptions that underlie the data analysis. ... ***The widespread use of “statistical significance” (generally interpreted as “ $p \leq 0.05$ ”) as a license for making a claim of a scientific finding (or implied truth) leads to considerable distortion of the scientific process.***”

# Scientific Conclusions Are...

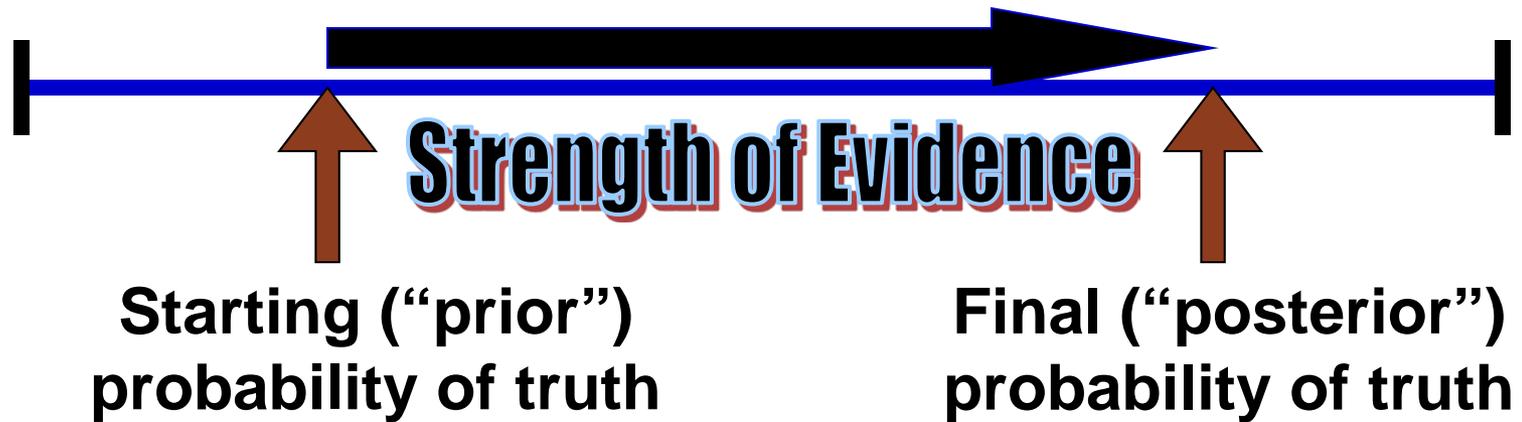
# NOT



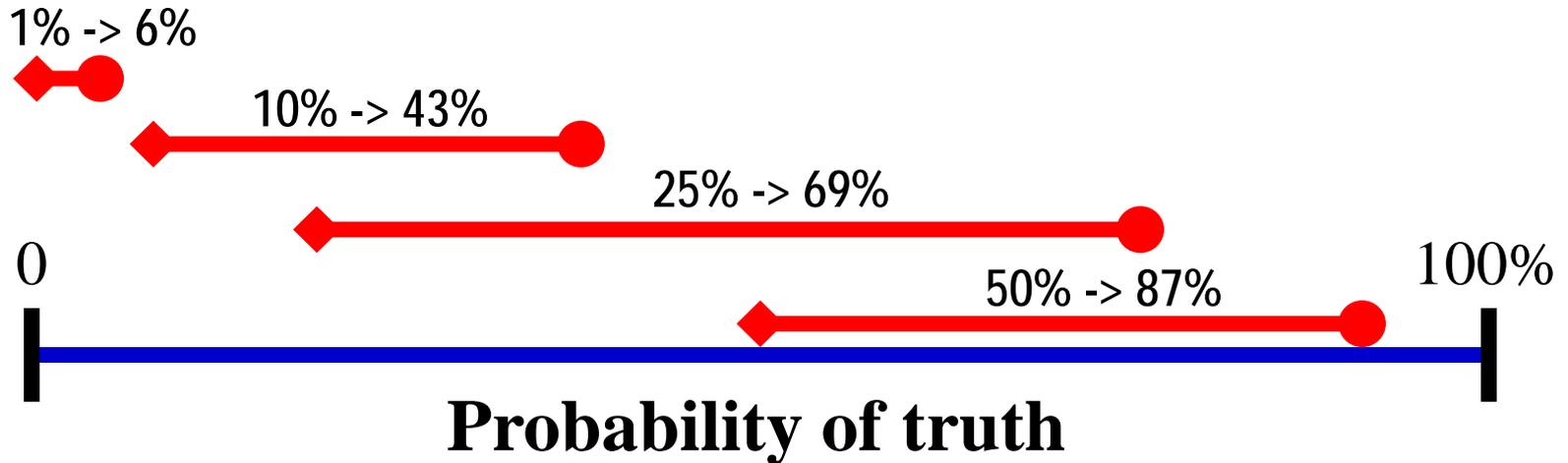
# Rather...



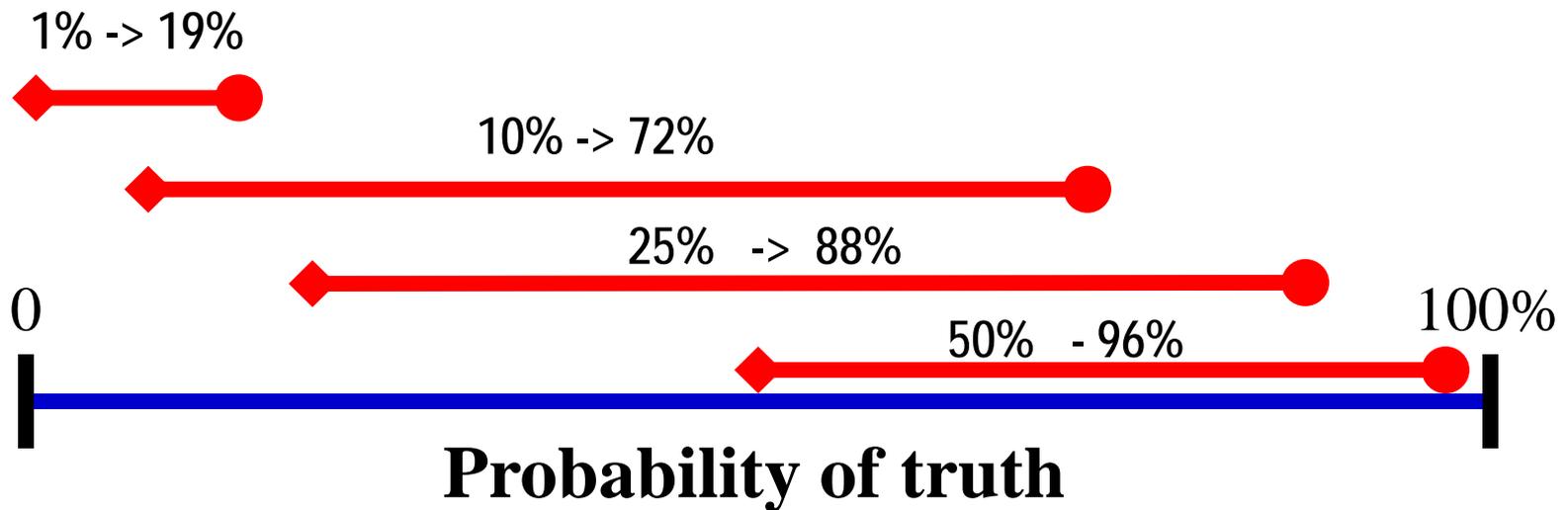
# Bayes Theorem



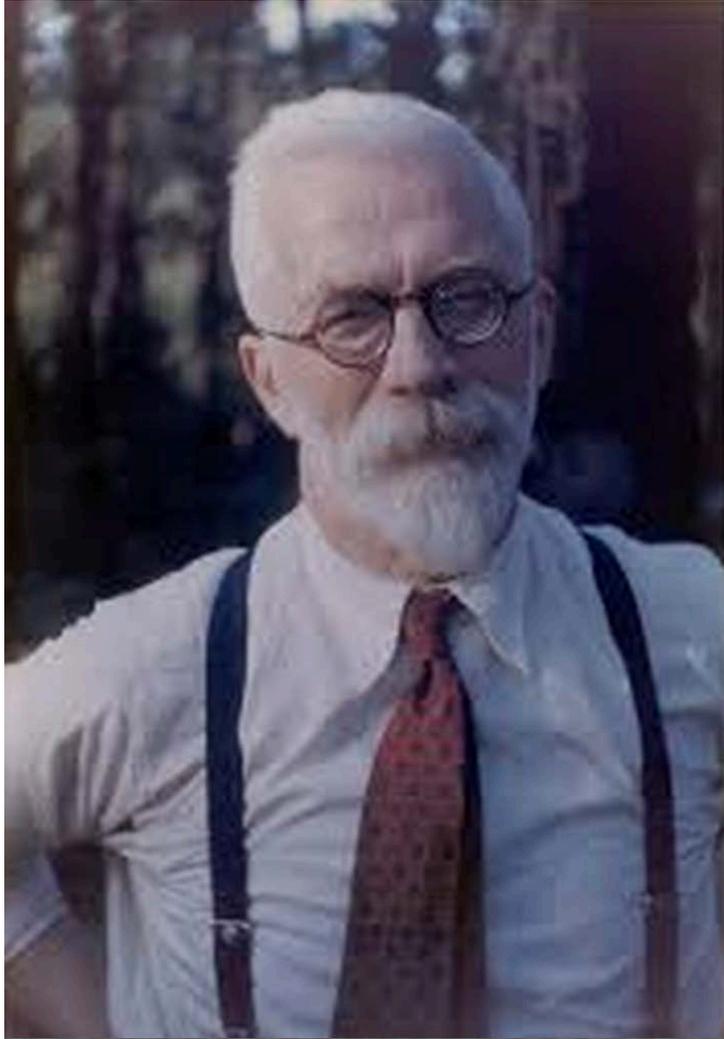
# What does $P=0.05$ do? (at most)



# What does $P=0.01$ do? (at most)



# Cumulative evidence



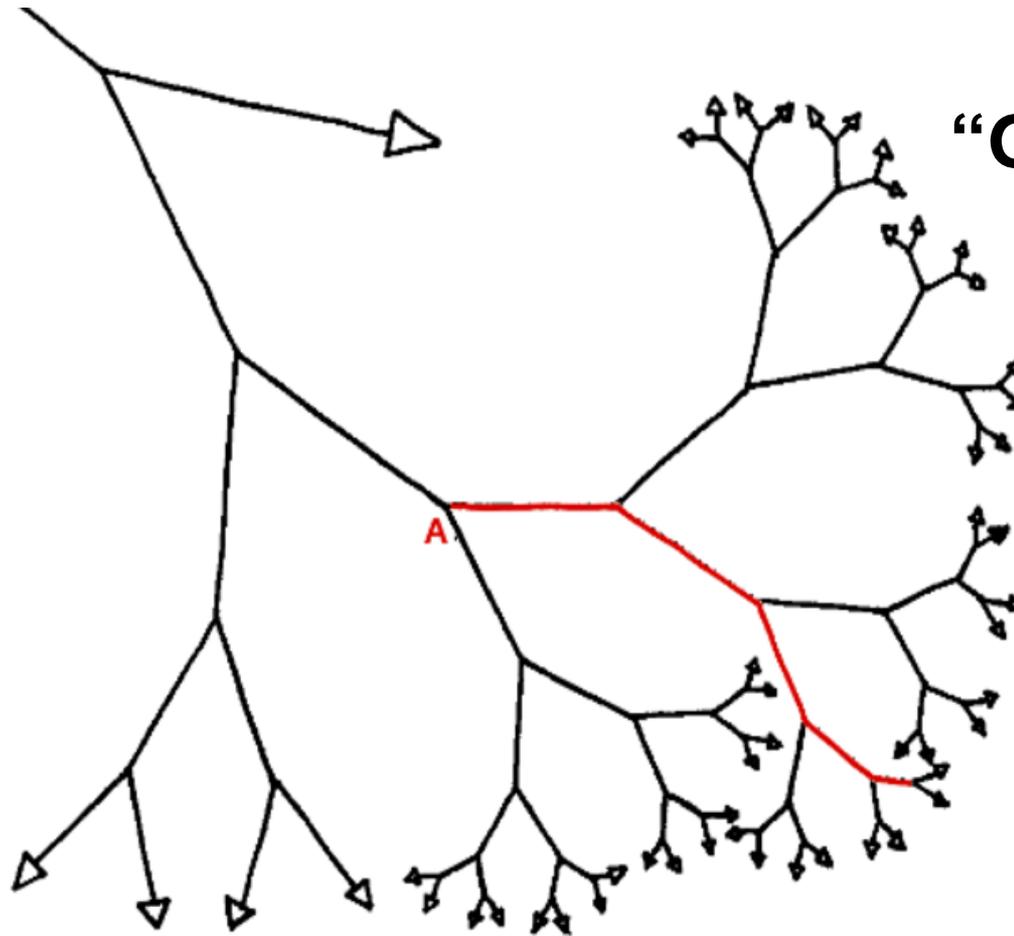
RA Fisher (1890 – 1966)

“Personally, the writer prefers to set a low standard of significance at the 5 percent point . . . ”

“A scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give this level of significance.”

RA Fisher, The arrangement of field experiments. *Journal of the Ministry of Agriculture*, 1926, 33, 503-513.

# Researcher degrees of freedom



“Garden of Forking Paths”

Gelman & Loken (2014)

# Consilience



William Whewell (1794-1866)

English philosopher and polymath

Created the word “scientist,” as well as physicist, ion, anode, cathode and dielectric.

Profoundly influenced Darwin, Faraday, Babbage and JS Mill

*Consilience* – proving a theory or demonstrating a property by measuring it in multiple different ways that did not share sources of error



Photo source: <https://24.p3k.hu/app/uploads/sites/11/2017/11/sallysellsinprogressblog.jpg>

# Summary

- “Methods reproducibility” is a form of confirmation that partly addresses issues of trust. If there is zero trust, there will always be more questions about the underlying data and what was done.
- Methods reproducibility does not directly address optimal analytic methods, but can allow different analytic approaches.
- “Results reproducibility” almost defies formal definition. Most complex science is not a series of “proofs” and “disproofs”.
- We develop networks of cumulative evidence that strengthen or weaken causal claims. (“Consilience”) Single studies will always involve data, design, or analysis issues subject to questioning.
- **Virtually all theories are underdetermined by the underlying evidence. Absolute certainty is rarely possible, but degrees of certainty sufficient for action is.**



ALESZU BAJAK



## Lectures Aren't Just Boring, They're Ineffective, Too, Study Finds

12 May 2014 3:00 pm | [122 Comments](#)



Wikimedia

**Blah?** Traditional lecture classes have higher undergraduate failure rates than those using active learning techniques, new research finds.

Are your lectures droning on? Change it up every 10 minutes with more active teaching techniques and more students will succeed, researchers say. [A new study](#) finds that

# New study says studies are wrong

AFP RELAXNEWS / Friday, August 28, 2015, 8:37 AM

AAA



SHARE THIS URL

nydn.us/1Jq6sXi

COPY



HALFPOINT/SHUTTERSTOCK.COM

**Some studies aren't worth stressing over.**

Scientific studies about how people act or think can rarely be replicated by outside experts, said a study Thursday that raised new questions about the seriousness of psychology research.

A team of 270 scientists tried reproducing 100 psychology and social science studies that had been published in three top peer-reviewed U.S. journals in

*Thank you!*

