METRICS
META-RESEARCH INNOVATION CENTER AT STANFORD

NCP Forum

Improving Cancer Diagnosis and Care: Clinical Application of Computational Methods in Precision Oncology

Conceptual Strategies for Reproducibility of Precision Oncology Data and Methods

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

STANFORD UNIVERSITY

Disclosures

- **Advisory roles for:**
  - Annals of Internal Medicine
  - PCORI (Patient-centered Outcome Research Institute)
  - Technology Assessment program, national Blue Cross-Blue Shield
- **NASEM is supporting travel for this meeting, and the lunch sandwiches were pretty good.**
- **Many medical schools rejected me because I said I was interested in the application of math to medicine.**

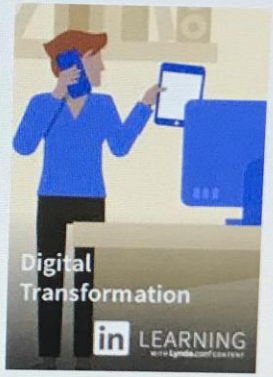
We are awash in messages (a.k.a. hype) about the enormous value of genomics for health…

3

watch: featured content

Business Etiquette: Mee···

Data and Medicine



Digital Transformation

Dining on a Dime: S4 E1

Dining on a Dime: S4 E2

watch | talk | learn | eat | shop | kids play | quicknav | info / help | media player

TV | MAP

watch: details

Data and Medicine

Data and Medicine

Price: Free
Rating: N/R
Duration: 00:06:07
Language: English
Director: N/A
Cast: N/A

Computer science is saving lives through genetic sequencing and harnessing the power of data to fight disease.


HiSeq™ 1500


DANIELA
PROFESSOR OF BIOSTATISTICS
TV
MAP


BILL
CREATED MICROSOFT

Traditional guarantors of research quality are less effective…

10

2014

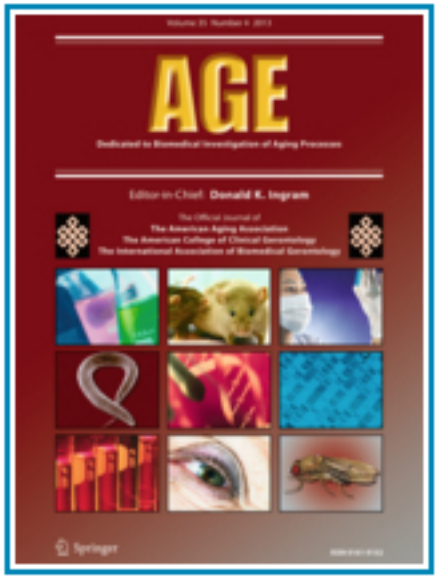
Retraction Watch

Tracking retractions as

NIH/Harvard team loses aging study to manipulated data

with 14 comments

Age has retracted a 2012 article by a group of scientists from the National Institutes of Health and Brigham and Women's Hospital in Boston after an NIH inquiry turned up evidence of data manipulation in the work.



The article, "Aging decreases rate of docosahexaenoic acid synthesis-secretion from circulating unesterified α-linolenic acid by rat liver," came from the lab of Stanley Rapoport, chief of the brain physiology and metabolism section of the National Institute on Aging.

As the abstract explained:

> Docosahexaenoic acid (DHA, 22:6n-3), an n-3 polyunsaturated fatty acid (PUFA) found at high concentrations in brain and retina and critical to their

2015

Retraction Watch

Tracking retractions as

NIH neuroscientist loses second paper, again the result of first author misconduct

with 5 comments

Stanley Rapoport, a neuroscientist in the National Institute on Aging, isn't having a lot of luck with his first authors. One committed misconduct and cost him a paper in the journal *Age* last year, and now he's lost another paper with a different first author, but for the exact same reason.

The latest paper, in *Neurochemical Research*, examined whether chronic doses of aspirin reduce brain inflammation. It has been cited 14 times, according to Thomson Scientific's Web of Knowledge.

Here's more from the note:

> This article has been retracted on request of the Editor-in-Chief. The National Institutes of Health (NIH) has found that Dr. Mireille Basselin engaged in research misconduct by fabricating and/or falsifying data in

Stanley Rapoport. Source: NIH

2016

Retraction Watch

Tracking retractions as

More co-author misconduct raises NIH neuroscientist's retraction count to 8

with 8 comments

Not again.

That's the sound of learning that a third scientist you worked with committed misconduct.

In the last two years, we reported on two retractions for neuroscientist Stanley Rapoport, the result of misconduct by two different first authors. We've since discovered more retractions resulting from those cases — and a new retraction stemming from the actions of yet another co-author.

Stanley Rapoport. Source: NIH

Although the latest retraction notice doesn't reveal the reason for retraction, both the journal editor and Rapoport — based at the National Institute on Aging (NIA), part of the National Institutes of Health (NIH) — confirmed to us that it is the result of misconduct by the last author, Jagadeesh Rao. According to Rapoport, a "number of retractions [for] Rao are still in the works."

2017

Retraction Watch

Tracking retractions as

Prominent NIH researcher up to a dozen retractions

with 10 comments

Neuroscientist Stanley Rapoport hasn't had much luck with his co-authors.

Recently, we've reported on multiple retractions of papers co-authored by Rapoport after three different first authors were found to have committed misconduct. Now, the fallout from one of those cases had led to four more retractions, bringing Rapoport's total to 12.

The latest batch of retractions stem from the actions of Jagadeesh Rao.

Here's the first notice, issued by *Psychopharmacology:*

Stanley Rapoport. Source: NIH

> The National Institutes of Health has found that Dr. Jagadeesh S. Rao engaged in research misconduct by falsifying data in this article. Data in in Figures 4A–B were falsified. Therefore, Dr. Stanley I. Rapoport has requested a full retraction.

of misconduct by the last author, Jagadeesh Rao. According to Rapoport, a "number of retractions [for] Rao are still in the works."

June 1. 2017

NIH neuroscientist up to 16 retractions

Neuroscientist Stanley Rapoport just can't catch a break.

Stanley Rapoport. Source: NIH

Rapoport, who's based at National Institute on Aging, is continuing to experience fallout from his research collaborations, after multiple co-authors have been found to have committed misconduct.

Most recently, Rapoport has had four papers retracted in three journals, citing falsified data in a range of figures. Although the notices do not specify how the data falsification occurred, Jagadeesh Rao, who was recently found guilty of research misconduct, is corresponding author on all four papers.

Back in December, Rapoport told us that a "number of retractions [for] Rao are still in the works:"

July 24, 2017

NIH section chief with 19 retractions

A former section chief at the National Institutes of Health who has had 19 papers retracted is no longer running a lab, Retraction Watch has learned.

Stanley Rapoport. Source: NIH

In the last three years, Stanley Rapoport, who is based at the U.S. National Institute on Aging (NIA), has lost 19 papers due to the misconduct of three different co-authors—Jagadeesh Rao, Fei Gao and Mireille Basselin.

And now, Rapoport, who was chief of the brain physiology and metabolism section of the NIA, no longer runs a lab.

Bad Luck?

"The misconduct, as I now understand it, **was very technical and outside of my areas of expertise**. In retrospect, I don't think I could have spotted the misconduct earlier. Data were presented at internal meetings, when the misconduct was not identified. Basselin and Gao and Rao had PhDs and strong letters of recommendation."

"**In these days of complex interdisciplinary research, one depends on the trustworthiness of colleagues who use methodologies with which one has no personal experience.** I regret missing the falsifications by Dr. Rao…"

9/21/2018

Special Issue Metaresearch

FEATURE

Research on research

BY MARTIN ENSERINK
SCIENCE | 21 SEP 2018 : 1178-1179 |

A growing number of scientists is studying science itself.

Summary Full Text PDF

Journals under the microscope

BY JENNIFER COUZIN-FRANKEL
SCIENCE | 21 SEP 2018 : 1180-1183 |

"Journalologists" use scientific methods to study publishing. Is their work improving science?

Summary Full Text PDF

The metawars

BY JOP DE VRIEZE
SCIENCE | 21 SEP 2018 : 1184-1188 |

Meta-analyses were supposed to end scientific debates. Often, they only cause more controversy.

Summary Full Text PDF

The truth squad

BY ERIK STOKSTAD
SCIENCE | 21 SEP 2018 : 1189-1191 |

In its drive to expose weaknesses in science, an up-and-coming research group doesn't mind stepping on some toes.

Summary Full Text PDF

A recipe for rigor

BY KAI KUPFERSCHMIDT
SCIENCE | 21 SEP 2018 : 1192-1193 |

A simple strategy to avoid bias—declaring in advance what you will study, and how—is rapidly catching on.

Summary Full Text PDF

POLICY FORUM

Toward a more scientific science

BY PIERRE AZOULAY, JOSHUA GRAFF-ZIVIN, BRIAN UZZI, DASHUN WANG, HEIDI WILLIAMS, JAMES A. EVANS, GINGER ZHE JIN, SUSAN FENG LU, BENJAMIN F. JONES, KATY BÖRNER, KARIM R. LAKHANI, KEVIN J. BOUDREAU, EVA C. GUINAN
SCIENCE | 21 SEP 2018 : 1194-1197 |

Summary Full Text PDF

nature International weekly journal of science

NATURE | COMMENT

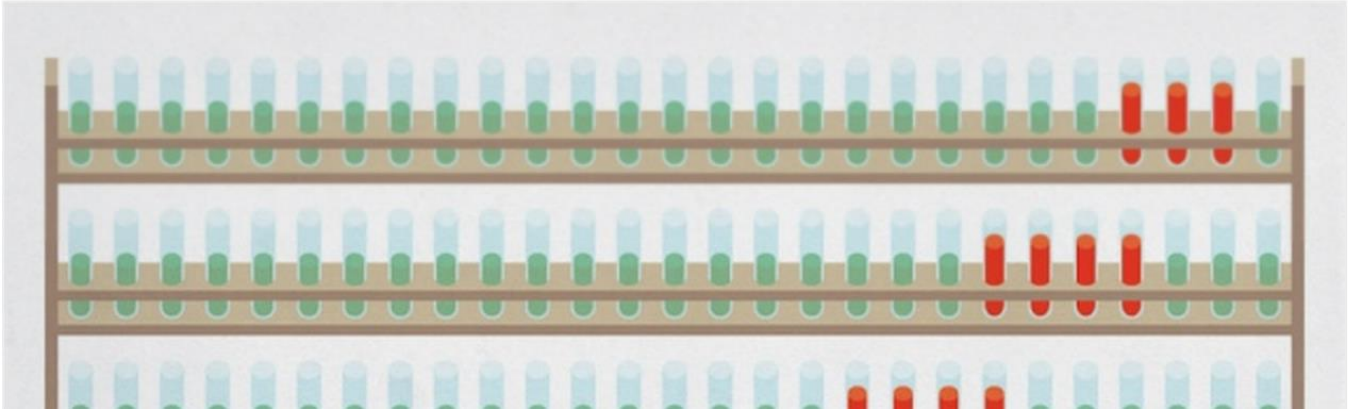
Policy: NIH plans to enhance reproducibility

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.

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.

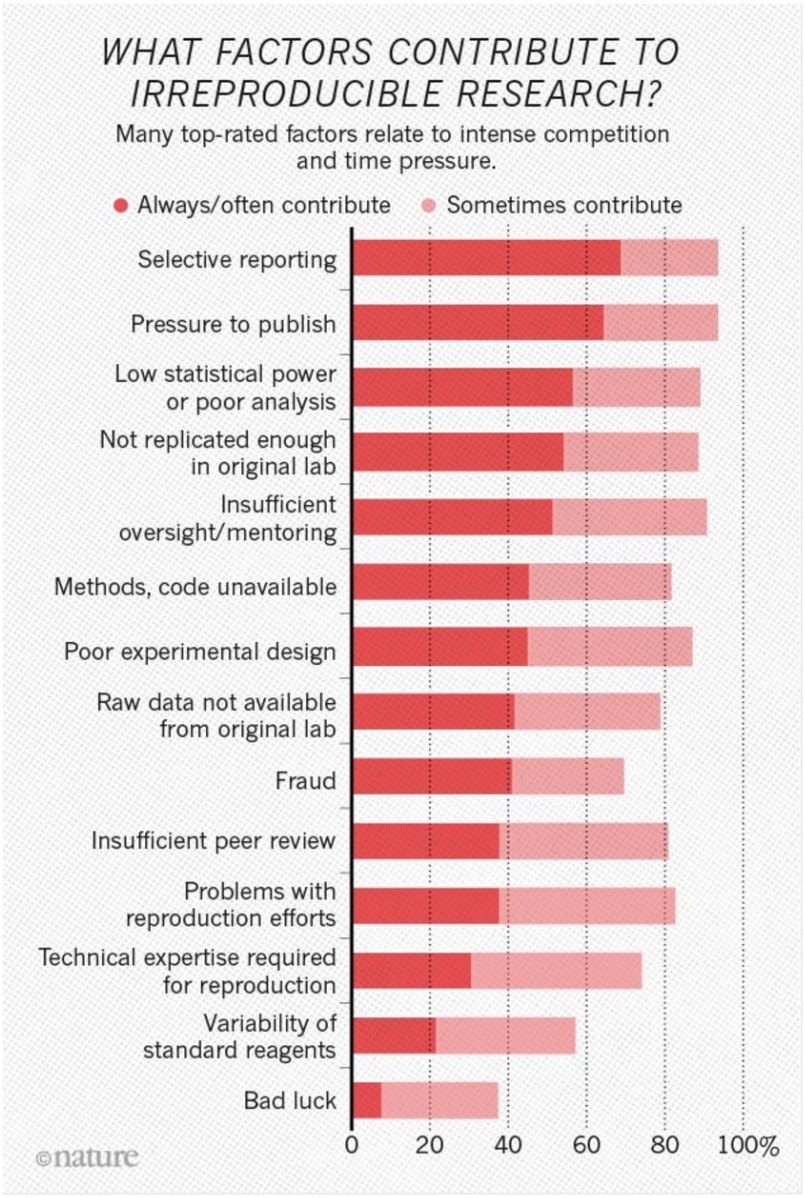
Another pitfall is overinterpretation of creative 'hypothesis-generating' experiments, which are designed to uncover new avenues of inquiry rather than to provide definitive proof for any single question. Still, there remains a troubling frequency of published reports that claim a significant result,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

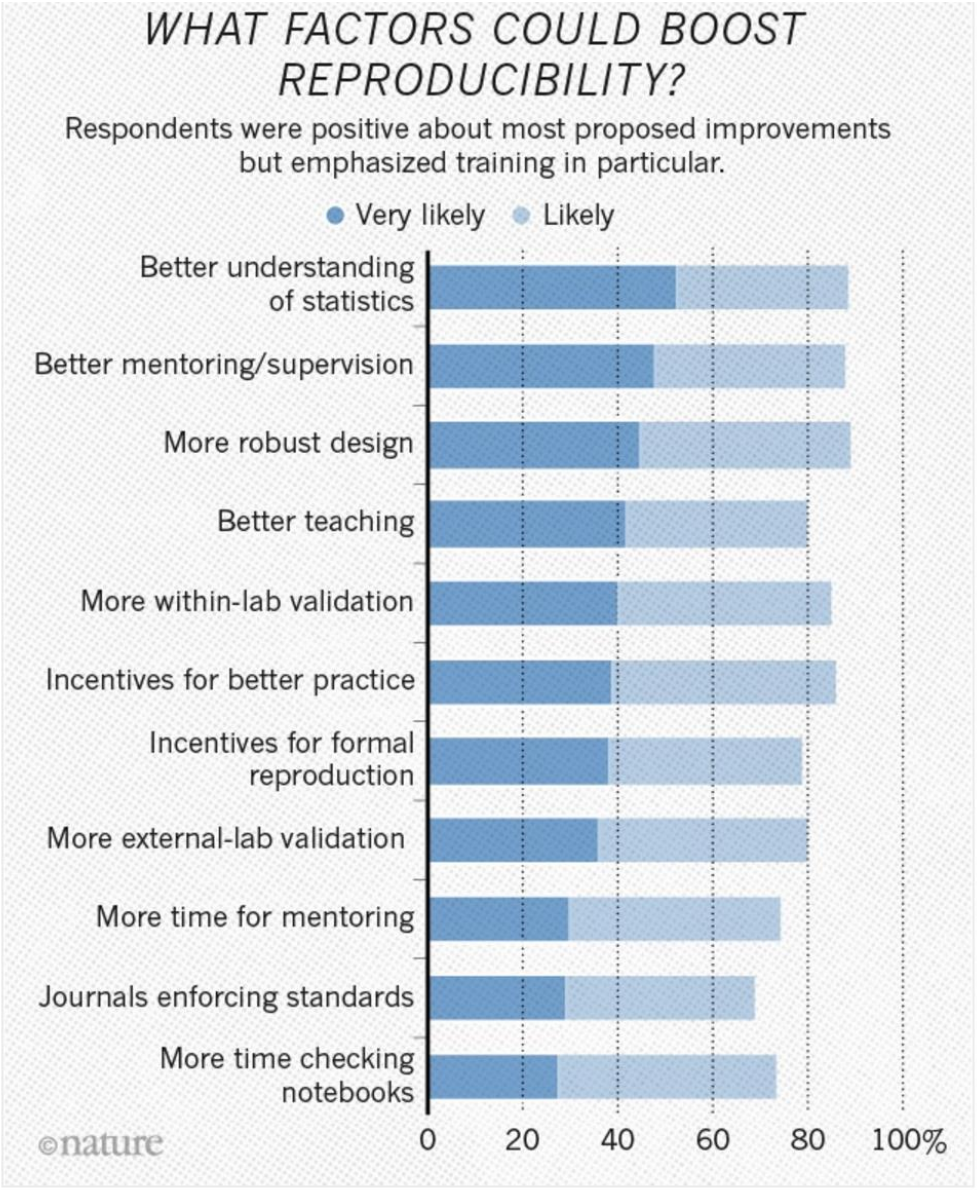
Collins/Tabak on Reproducibility

...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.**


WHAT FACTORS CONTRIBUTE TO IRREPRODUCIBLE RESEARCH?
Many top-rated factors relate to intense competition and time pressure.
Always/often contribute
Sometimes contribute
Selective reporting
Pressure to publish
Low statistical power or poor analysis
Not replicated enough in original lab
Insufficient oversight/mentoring
Methods, code unavailable
Poor experimental design
Raw data not available from original lab
Fraud
Insufficient peer review
Problems with reproduction efforts
Technical expertise required for reproduction
Variability of standard reagents
Bad luck
0
20
40
60
80
100%
©nature


WHAT FACTORS COULD BOOST REPRODUCIBILITY?
Respondents were positive about most proposed improvements but emphasized training in particular.
Very likely
Likely
Better understanding of statistics
Better mentoring/supervision
More robust design
Better teaching
More within-lab validation
Incentives for better practice
Incentives for formal reproduction
More external-lab validation
More time for mentoring
Journals enforcing standards
More time checking notebooks
0
20
40
60
80
100%
©nature

22

PERSPECTIVE

SCIENTIFIC INTEGRITY

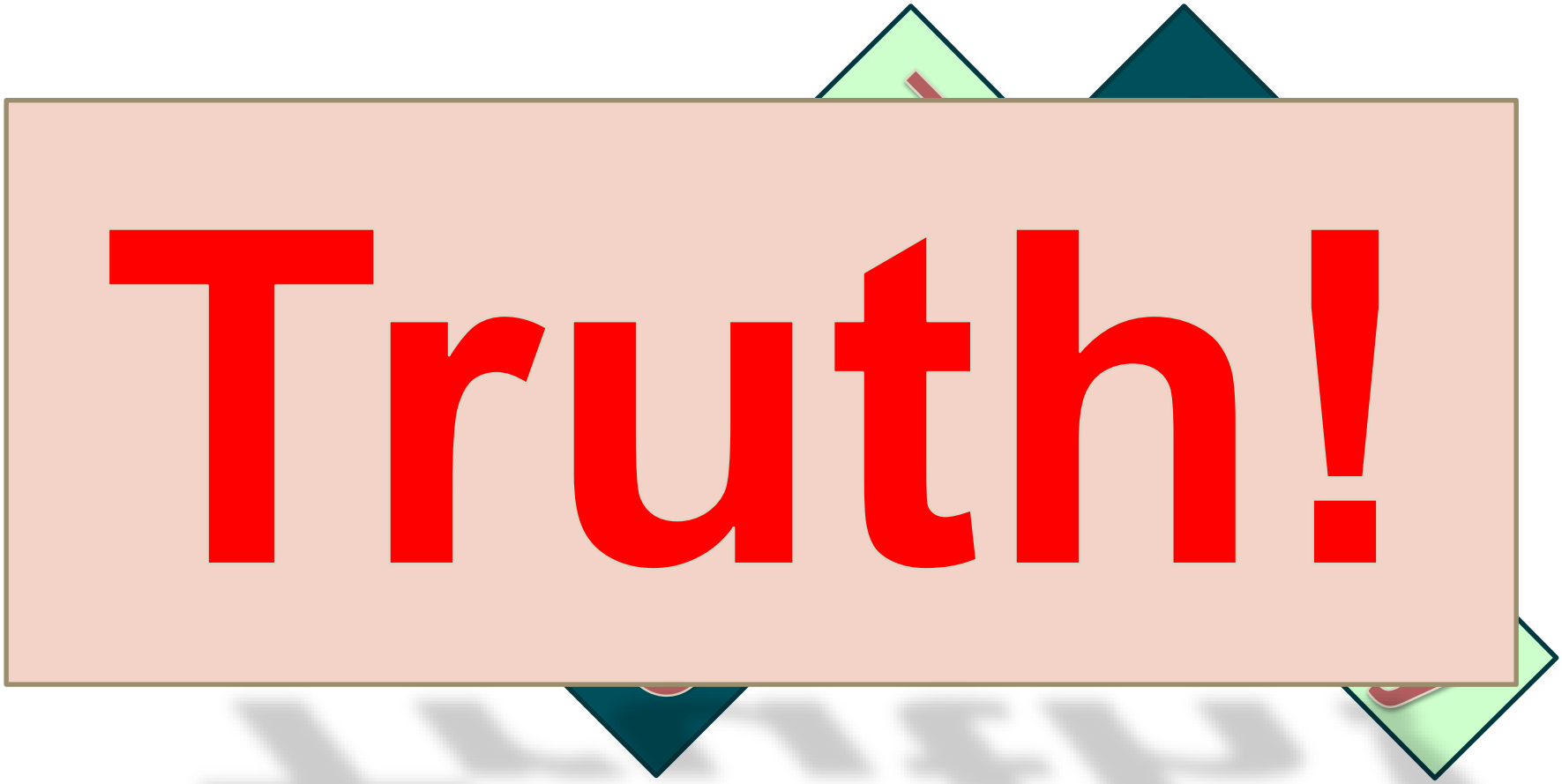
What does research reproducibility mean?

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 1 June 2016 Vol 8 Issue 341 341ps12

Reproducibility, et al.



Which truth?

- **What is the actual claim?**
- **There are "degrees of truth" or confidence in a claim. We have poorly developed tools or language for this; closest is evidential quality.**
- **True enough to take the next step in development.**
  - Prospective validation?
  - Independent validation?
  - Develop targeted therapy?
  - Clinical trial?
  - Clinical practice?
- **Ultimate truth is whether or not patients will be better off.**

Meanings of reproducibility

Methods reproducibility

- Whether we can tell what was done, enough to assess validity, the sources of variation, nature of confirmation/validation.
- Related to scientific *processes.*
- Includes computational reproducibility
- *Transparency, methods reporting, data and code sharing,*

Results reproducibility

- Degree of support that subsequent studies provide for the original claim
- Related to *results of* science.
- *Additional evidence, validation, confirmation.*

Inferential reproducibility

- Are the results *interpreted* the same way by different people?
- Strength of claims, degree of proof/validation/generalizability.
- Truth? For whom?

Methods Reproducibility

American Journal of Epidemiology
Vol. 163, No. 9
DOI: 10.1093/aje/kwj093
Advance Access publication March 1, 2006

Commentary

Reproducible Epic

Roger D. Peng, Franc

From the Biostatistics Dep

Received for publication M

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.

The replication c accumulation of scientific evidence ted by indepen- dent data, analytica used to quantify small health effects here results can inform substantial of many current epidemiologic stuc um standard is "reproducibility," w ed 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

Use of proteomic patterns in serum to identify ovarian cancer

Emanuel F Petricoin III, Ali M Ardekani, Ben A Hitt, Peter J Levine, Vincent A Fusaro, Seth M Steinberg, Gordon B Mills, Charles Simone, David A Fishman, Elise C Kohn, Lance A Liotta

Summary

Background New technologies for the detection of early-stage ovarian cancer are urgently needed. Pathological changes within an organ might be reflected in proteomic patterns in serum. We developed a bioinformatics tool and used it to identify proteomic patterns in serum that distinguish neoplastic from non-neoplastic disease within the ovary.

Methods Proteomic spectra were generated by mass spectroscopy (surface-enhanced laser desorption and ionisation). A preliminary "training" set of spectra derived from analysis of serum from 50 unaffected women and 50 patients with ovarian cancer were analysed by an iterative searching algorithm that identified a proteomic pattern that completely discriminated cancer from non-cancer. The discovered pattern was then used to classify an independent set of 116 masked serum samples: 50 from women with ovarian cancer, and 66 from unaffected women or those with non-malignant disorders.

Findings The algorithm identified a cluster pattern that, in the training set, completely segregated cancer from non-cancer. The discriminatory pattern correctly identified all 50 ovarian cancer cases in the masked set, including all 18 stage I cases. Of the 66 cases of non-malignant disease, 63 were recognised as not cancer. This result yielded a sensitivity of 100% (95% CI 93–100), specificity of 95% (87–99), and positive predictive value of 94% (84–99).

Introduction

Application of new technologies for detection of ovarian cancer could have an important effect on public health,[1] but to achieve this goal, specific and sensitive molecular markers are essential.[1–5] This need is especially urgent in women who have a high risk of ovarian cancer due to family or personal history of cancer, and for women with a genetic predisposition to cancer due to abnormalities in predisposition genes such as *BRCA1* and *BRCA2*. There are no effective screening options for this population.

Ovarian cancer presents at a late clinical stage in more than 80% of patients,[1] and is associated with a 5-year survival of 35% in this population. By contrast, the 5-year survival for patients with stage I ovarian cancer exceeds 90%, and most patients are cured of their disease by surgery alone.[1–6] Therefore, increasing the number of women diagnosed with stage I disease should have a direct effect on the mortality and economics of this cancer without the need to change surgical or chemotherapeutic approaches.

Cancer antigen 125 (CA125) is the most widely used biomarker for ovarian cancer.[1–6] Although concentrations of CA125 are abnormal in about 80% of patients with advanced-stage disease, they are increased in only 50–60% of patients with stage I ovarian cancer.[1–6] CA125 has a positive predictive value of less than 10% as a single marker, but the addition of ultrasound screening to CA125 measurement has improved the

Short timeline of Ovacheck

- The researchers, from NIH & FDA, won widespread acclaim.
- Congressional resolution urged further funding for their research.
- The magazine “Health” named the test one of the top ten medical advances of the year.
- Commercial rights to develop the test were licensed from the US government to Correlogic Systems.
- Correlogic granted licenses to Quest Diagnostics and the Laboratory Corporation of America, hoping to market the test under the brand name OvaCheck.

31

BIOINFORMATICS

Vol. 20 no. 5 2004, pages 777–785
DOI: 10.1093/bioinformatics/btg484

Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments

Keith A. Baggerly, Jeffrey S. Morris and Kevin R. Coombes*

Department of Biostatistics, U.T. M.D. Anderson Cancer Center, 1515 Holcombe Blvd, Box 447, Houston, TX 77030-4009, USA

Received on July 14, 2003; revised on October 14, 2003; accepted on October 16, 2003
Advance Access publication January 29, 2004

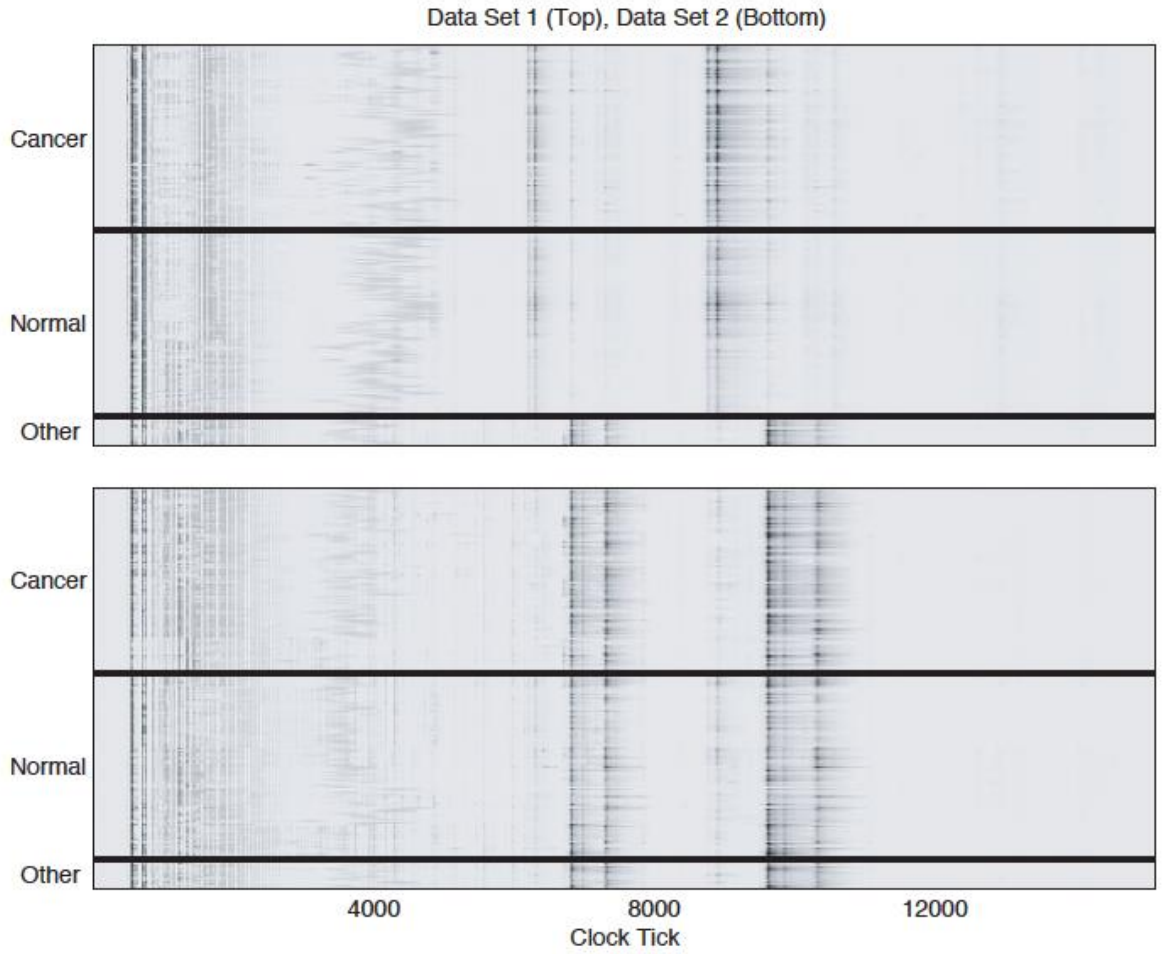


Fig. 2. Heat map of all 216 samples from dataset 1 (top), which were run on the H4 chip, and of all 216 samples from dataset 2 (bottom), which are the same biological samples as dataset 1, just run on the WCX2 chip. The gross break at the 'benign disease' juncture in dataset 1, and the similarity of the profiles to those in dataset 2, suggests that a change in protocol occurred in the middle of the first experiment.

33

“We can achieve perfect classification with noise”

Baggerly, 2004

To achieve the level of reproducibility required .., careful attention must be paid to **measuring and controlling sources of variation in the procedure**. A (very) incomplete list of such sources includes time (since results from a single instrument can drift), temperature, humidity, the instrument used and the laboratory in which the experiment is conducted. A more complete list must be established, and experiments must be performed to estimate the magnitude of each effect.

Whenever possible**, standard protocols should be drawn up to minimize the effect of irrelevant sources of variation**. Sources that cannot be controlled must be repeatedly measured to account for them. Samples where these conditions have been altered should be included in the training set so that these changes do not drive the classification. The goal, of course, is to prevent major technological differences from overwhelming the biology associated with the outcome of interest.

34

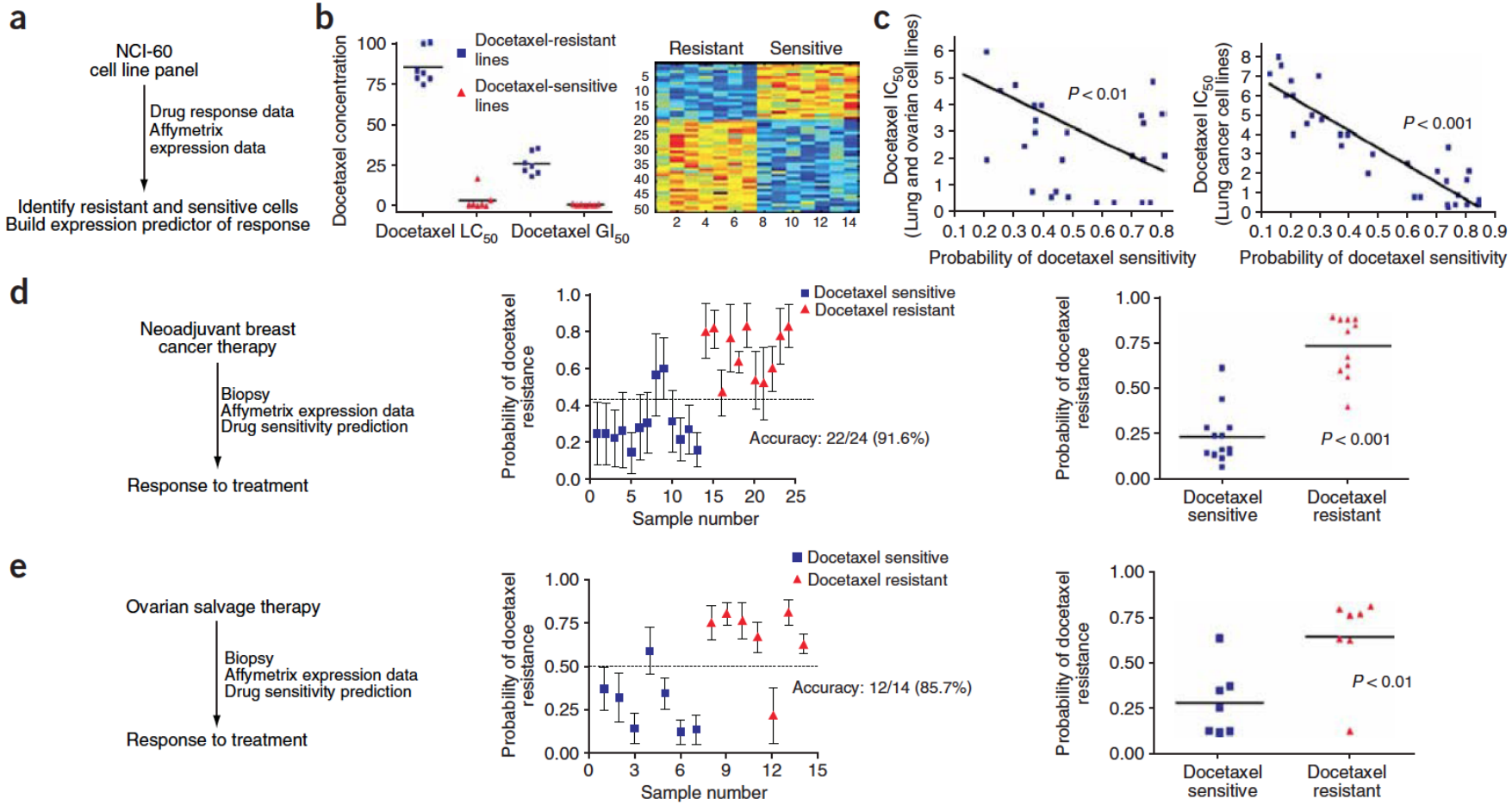


Figure 1 A gene expression signature that predicts sensitivity to docetaxel. (**a**) Strategy for generating the chemotherapeutic response predictor. (**b**) Left, cell lines from the NCI-60 panel used to develop the *in vitro* signature of docetaxel sensitivity. There is a statistically significant difference (Mann-Whitney *U*-test) in the IC_{50} and LC_{50} of the cell lines chosen to represent the sensitive and resistant subsets. Right, expression plots for genes selected for discriminating the docetaxel-resistant and docetaxel-sensitive NCI-60 cell lines, with blue representing the lowest expression and red the highest. Each column in the figure represents an individual sample. Each row represents an individual gene, ordered from top to bottom according to regression coefficient. (**c**) Left, validation of the docetaxel response prediction model in an independent set of lung and ovarian cancer cell line samples. A collection of lung and ovarian cell lines were used in a cell proliferation assay to determine the IC_{50} of docetaxel in the individual cell lines. Right, validation of the docetaxel response prediction model in another independent set of 29 lung cancer cell line samples. (**d**) Left, a strategy for assessment of the docetaxel response predictor as a function of clinical response in the breast neoadjuvant setting. Middle, predicted probability of docetaxel sensitivity in a collection of samples from a breast cancer single-agent neoadjuvant study. Right, a single-variable scatter plot of a significance test of the predicted probabilities of sensitivity to docetaxel in the sensitive and resistant tumors ($P < 0.001$, Mann-Whitney *U*-test). (**e**) Left, a strategy for assessment of the docetaxel response predictor as a function of clinical response in advanced ovarian cancer. Middle, predicted probability of docetaxel sensitivity in a collection of samples from a prospective single agent salvage therapy study. Right, a single-variable scatter plot showing statistical significance ($P < 0.01$, Mann-Whitney *U*-test).

ORIGINAL ARTICLE

A Genomic Strategy to Refine Prognosis in Early-Stage Non–Small-Cell Lung Cancer

Anil Potti, M.D., Sayan Mukherjee, Ph.D., Rebecca Petersen, M.D., Holly K. Dressman, Ph.D., Andrea Bild, Ph.D., Jason Koontz, M.D., Robert Kratzke, M.D., Mark A. Watson, M.D., Ph.D., Michael Kelley, M.D., Geoffrey S. Ginsburg, M.D., Ph.D., Mike West, Ph.D., David H. Harpole, Jr., M.D., and Joseph R. Nevins, Ph.D.

From the Institute for Genome Sciences and Policy (A.P., S.M., H.K.D., A.B., J.K., G.S.G., M.W., J.R.N.) and the Institute of Statistics and Decision Sciences (S.M., M.W.), Duke University; and the Departments of Medicine (A.P., J.K., M.K., G.S.G.), Surgery (R.P., D.H.H.), and Molecular Genetics and Microbiology (H.K.D., A.B., J.R.N.), Duke University Medical Center

ABSTRACT

BACKGROUND

Clinical trials have indicated a benefit of adjuvant chemotherapy for patients with stage IB, II, or IIIA — but not stage IA — non–small-cell lung cancer (NSCLC). This classification scheme is probably an imprecise predictor of the prognosis of an individual patient. Indeed, approximately 25 percent of patients with stage IA disease have a recurrence after surgery, suggesting the need to identify patients in this subgroup for more effective therapy.

Forensic Bioinformatician Aims To Solve Mysteries of Biomarker Studies

By Liz Savage

Keith Baggerly, Ph.D., is not a detective in the traditional sense—he doesn't

that we've had a good deal of trouble figuring out in some instances," Baggerly said.

validated. "A lot of these findings that he's reported on … are things that people are

What is the number-one problem that he encounters? Bookkeeping. "It's not sexy, it's not higher mathematics. It's bookkeeping …keeping track of the labels and keeping track of what goes where…

"I'm not really worried about particularly esoteric mathematics. Most of the stuff that I'm worried about is very clearly understanding what was done at each of the steps involved."

Baggerly comes in.

To correctly apply the new genomic findings to the patients at hand, the analyses that produced these findings must be understood in detail. Thus, for the last few years Baggerly and his colleagues at the University of Texas M. D. Anderson Cancer Center in Houston have tackled the difficult job of reproducing some complex analyses. Using the raw data and

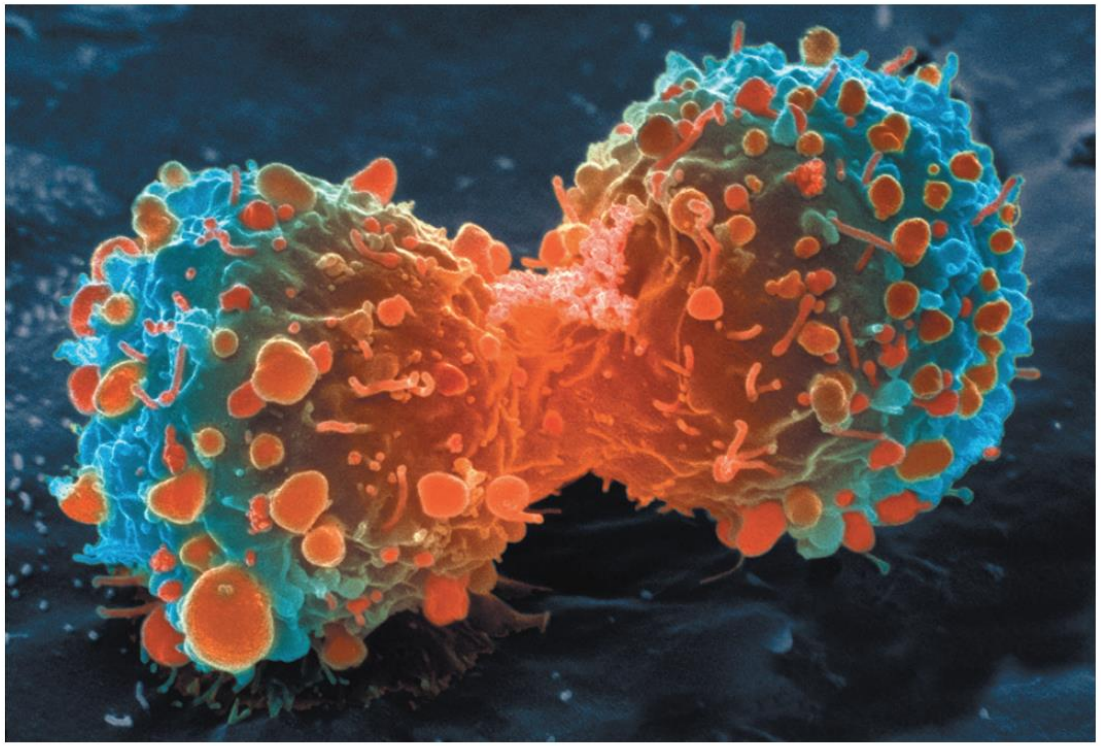
molecular profiling results [for example] can be as dangerous as bad treatments. It's so critical that someone really take a close look at these things."

detect ovarian cancer in its early stages. Nonetheless, there was concern over the reproducibility of the results, and the case encouraged debate about the validation of proteomic studies more generally.

M. D. Anderson investigators, like many others, wanted to try out the new test for themselves, so they enlisted Baggerly and his colleagues to show them how to reproduce

Keith Baggerly, Ph.D.

Although Baggerly's colleagues are quick to praise him, they also point out that, in an


S. GSCHMEISSNER/SPL

Many landmark findings in preclinical oncology research are not reproducible, in part because of inadequate cell lines and animal models.

Raise standards for preclinical cancer research

C. Glenn Begley and **Lee M. Ellis** propose how methods, publications and incentives must change if patients are to benefit.

Efforts over the past decade to characterize the genetic alterations in human cancers have led to a better understanding of molecular drivers of this complex set of diseases. Although we in the cancer field hoped that this would lead to more effective drugs, historically, our ability to translate cancer research to clinical success has been remarkably low[1]. Sadly, clinical trials in oncology have the highest failure rate compared with other therapeutic areas. Given the high unmet need in oncology, it is understandable that barriers to clinical development may be lower than for other disease areas, and a larger number of drugs with suboptimal preclinical validation will enter oncology trials. However, this low success rate is not sustainable or acceptable, and investigators must reassess their approach to translating discovery research into greater clinical success and impact.

Many factors are responsible for the high failure rate, notwithstanding the inherently difficult nature of this disease. Certainly, the limitations of preclinical tools such as inadequate cancer-cell-line and mouse models[2] make it difficult for even ▶

The findings of only 6 of 53 (11%) “landmark” preclinical experiments could be replicated with repeated experimentation.

STANFORD UNIVERSITY

Begley on reproducibility

"...when findings could not be reproduced, an attempt was made to contact the original authors, discuss the discrepant findings, exchange reagents and repeat experiments under the authors' direction, occasionally even in the laboratory of the original investigator. ..."

Reuters article:

Begley met for breakfast at a cancer conference with the lead scientist of one of the problematic studies...."We went through the paper line by line, figure by figure," said Begley. "I explained that we re-did their experiment 50 times and never got their result. He said they'd done it six times and got this result once, but put it in the paper because it made the best story."

Training set

Validation set!

Validation set

- Group independent of the training set on whom you apply the training model, and report numbers
- Ideally, a completely new population, in another place, gathered for different reasons, with clear eligibility criteria, defining the test's target population.
- All real-world variation - in the testing procedure (e.g. sample prep, transport, test failures) - should be reflected.
- If the model or test procedure is tweaked based on the validation set, *it becomes a new training set.* You must validate on a new population.
- Watch out for factors that might not be in eligibility criteria that make populations different (e.g. race, ethnicity, age, disease factors)

Hayes *BMC Medicine* 2013, 11:221
http://www.biomedcentral.com/1741-7015/11/221

BMC Medicine

COMMENTARY Open Access

OMICS-based personalized oncology: if it is worth doing, it is worth doing well!

Daniel F Hayes

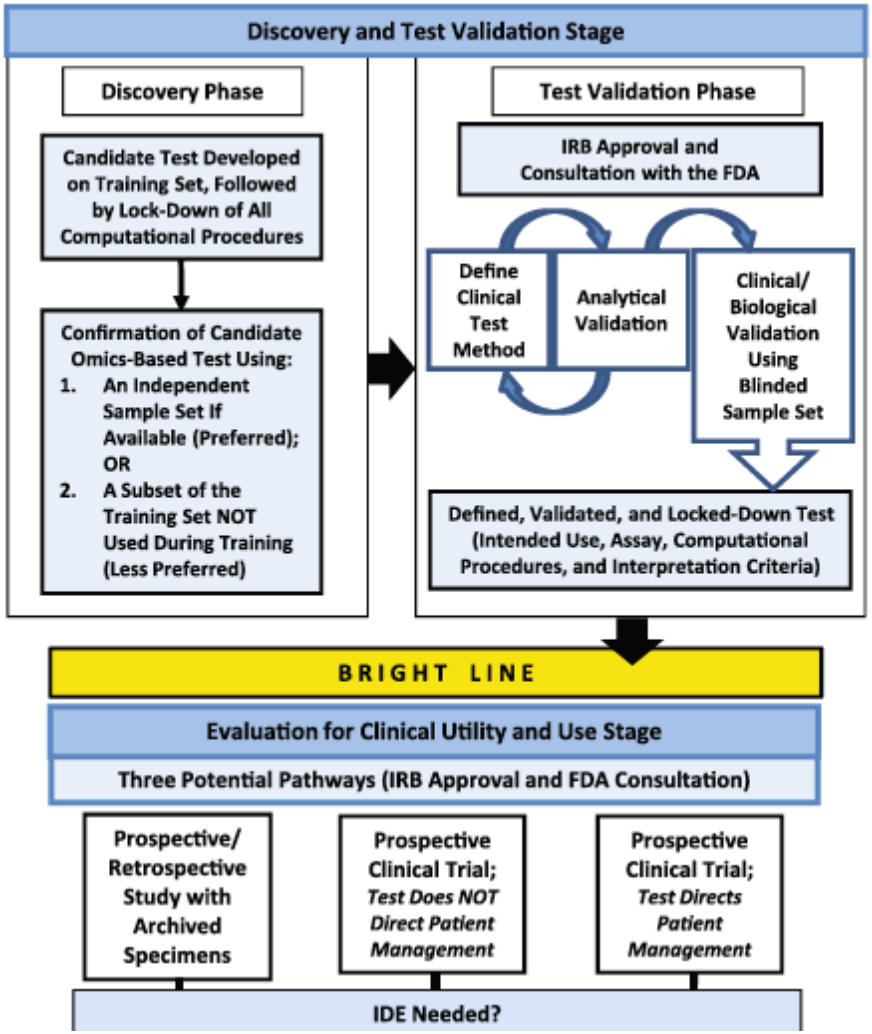


Table 1. Elements of tumor marker studies that constitute Levels of Evidence determination*

Category	A	B	C	D
Element	**Prospective**	**Prospective using archived samples**	**Prospective/ observational**	**Retrospective/ observational**
Clinical trial	PCT designed to address tumor marker	Prospective trial not designed to address tumor marker, but design accommodates tumor marker utility Accommodation of predictive marker requires PRCT	Prospective observational registry, treatment and follow-up not dictated	No prospective aspect to study
Patients and patient data	Prospectively enrolled, treated, and followed in PCT	Prospectively enrolled, treated, and followed in clinical trial and, especially if a predictive utility is considered, a PRCT addressing the treatment of interest	Prospectively enrolled in registry, but treatment and follow-up standard of care	No prospective stipulation of treatment or follow-up; patient data collected by retrospective chart review
Specimen collection, processing, and archival	Specimens collected, processed, and assayed for specific marker in real time	Specimens collected, processed, and archived prospectively using generic SOPs. Assayed after trial completion	Specimens collected, processed, and archived prospectively using generic SOPs. Assayed after trial completion	Specimens collected, processed and archived with no prospective SOPs
Statistical design and analysis	Study powered to address tumor marker question	Study powered to address therapeutic question and underpowered to address tumor marker question	Study not prospectively powered at all. Retrospective study design confounded by selection of specimens for study	Study not prospectively powered at all. Retrospective study design confounded by selection of specimens for study
		Focused analysis plan for marker question developed before doing assays	Focused analysis plan for marker question developed before doing assays	No focused analysis plan for marker question developed before doing assays
Validation	Result unlikely to be play of chance	Result more likely to be play of chance that A but less likely than C	Result very likely to be play of chance	Result very likely to be play of chance
	Although preferred, validation not required	Requires one or more validation studies	Requires subsequent validation studies	Requires subsequent validation

* PCT = prospective controlled trial; PRCT = prospective randomized controlled trial; SOPs = standard operating practices.

Conclusions/recs of Dx studies by development phase

Phase	Question/design	Primary measures
0 Early developmental	Technical issues; reliability, validity, procedures, transportability, training of tester or reader	Ready to take to clinical application.
1a Early developmental **aka "Clinical validity" (1a-2b)**	Test values different between cases and normals?	If yes, how to define cutoffs for "test"?
1b Early developmental	Test positivity different between cases and normals?	Test ready to be tested in subjects needing test (i.e. not "normals")
2a Middle developmental	Test positivity in cases and those *suspected of disease*	Population needs expansion to real actual target population
2b Middle developmental	+ comorbidities, expanded spectrum of disease	Must assess whether it is valuable clinically in the context of the total information at the time of decision.
3a Clinical Utility	Value of test with <u>all information available at time of diagnosis.</u> Do testing procedures mimic real use? Is decision specified?	Need to assess all health outcomes that follow testing in real world conditions.
3b Clinical utility	Actual testing conditions with decision specified. Does test lead to better net health outcomes?	**Ready for clinical use. What are optimal testing/decision pathways? Are payors likely to cover? Ready for clinical guidelines?**
4 (Societal utility)	Better net health outcomes at acceptable	Cost-benefit, cost-effectiveness

A few observations on methods reproducibility

- There are standard developmental phases of therapeutic agents that we recognize and rarely skip.
- These include manufacturing, quality control of formulation, dosing, PK/PD, safety, preliminary biologic activity/efficacy, and finally efficacy and benefit vs. safety.
- With computationally based diagnostic/prognostic/predictive models, we often don't recognize developmental phases, with recommendations for clinical application based on the equivalent of pre-clinical development.
- The reliability/reproducibility of claims re performance, in what patients and for what therapeutic decisions need to proceed in a systematic fashion, with patient clinical outcomes as the ultimate determinant of value.

Results Reproducibility

The New England
Journal of Medicine

Volume 319 DECEMBER 29, 1988 Number 26

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

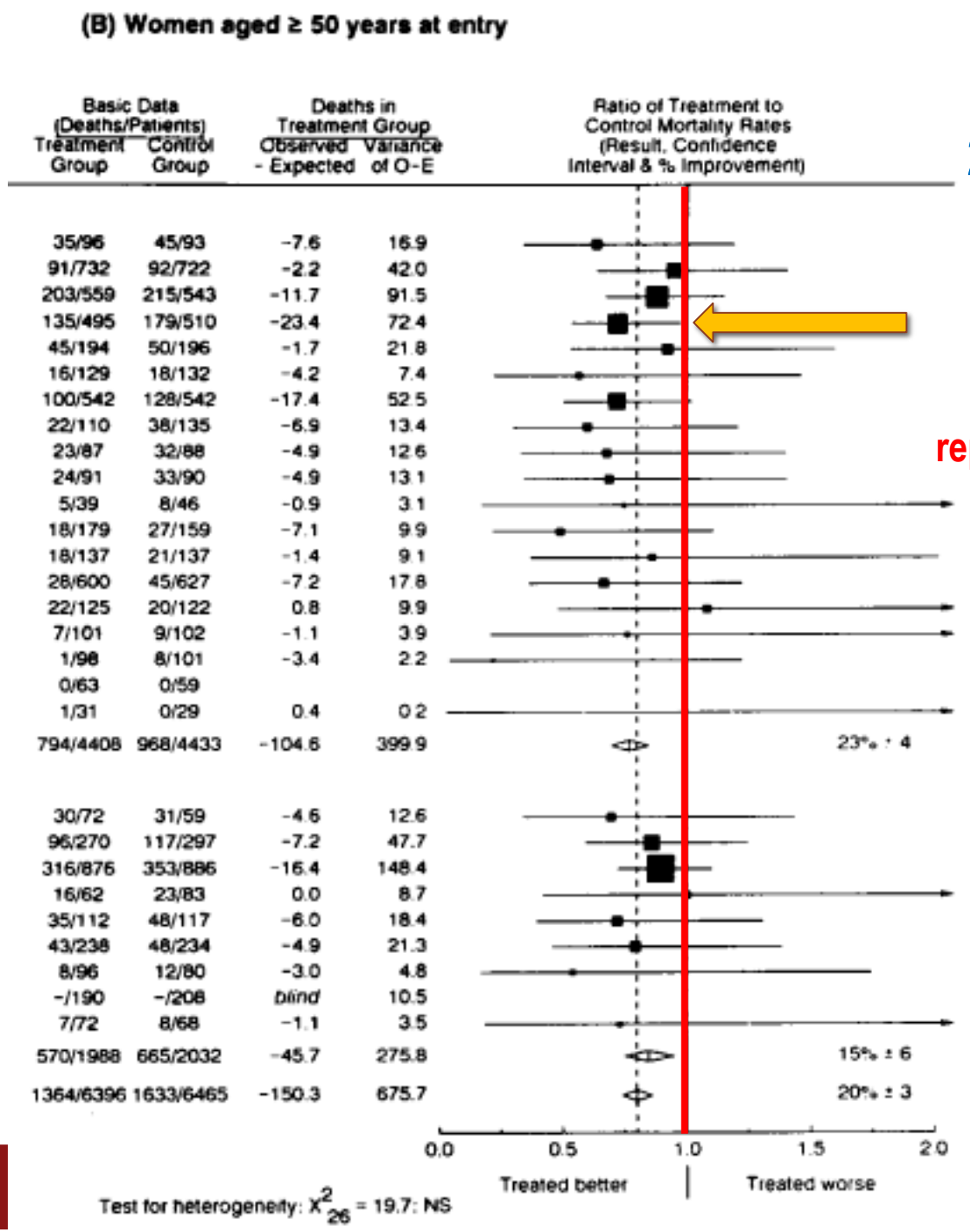
An Overview of 61 Randomized Trials among 28,896 Women

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.)



22 of 23 trials have P > 0.05

Was a null-effect reproduced 22 times?

STANFORD UNIVERSITY

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/ezcuj).

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-

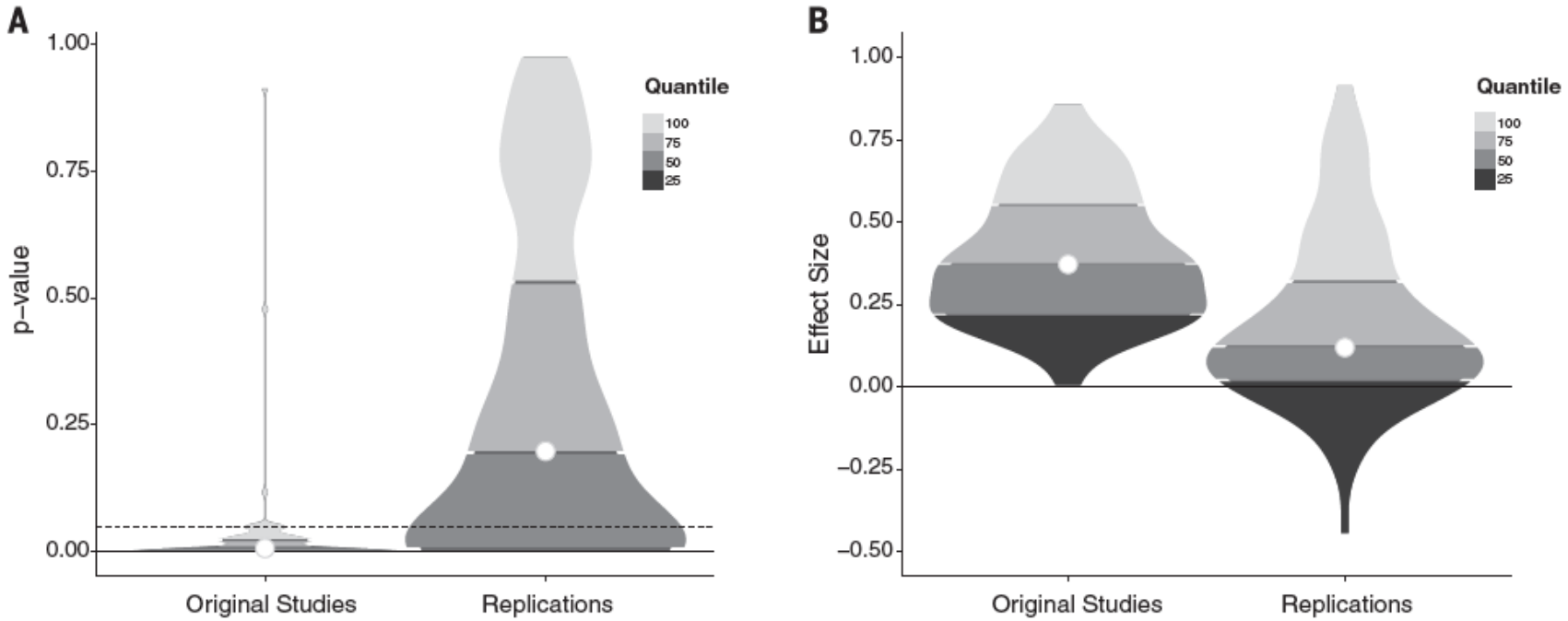


Fig. 1. Density plots of original and replication *P* values and effect sizes. (**A**) *P* values. (**B**) Effect sizes (correlation coefficients). Lowest quantiles for *P* values are not visible because they are clustered near zero.

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%)

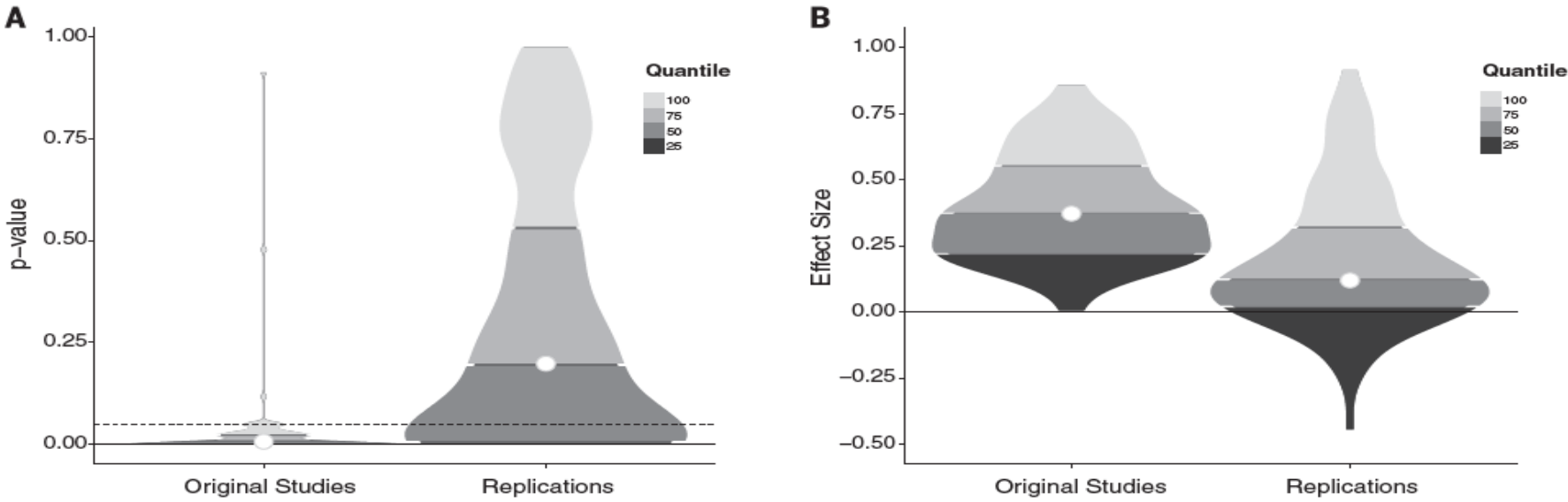
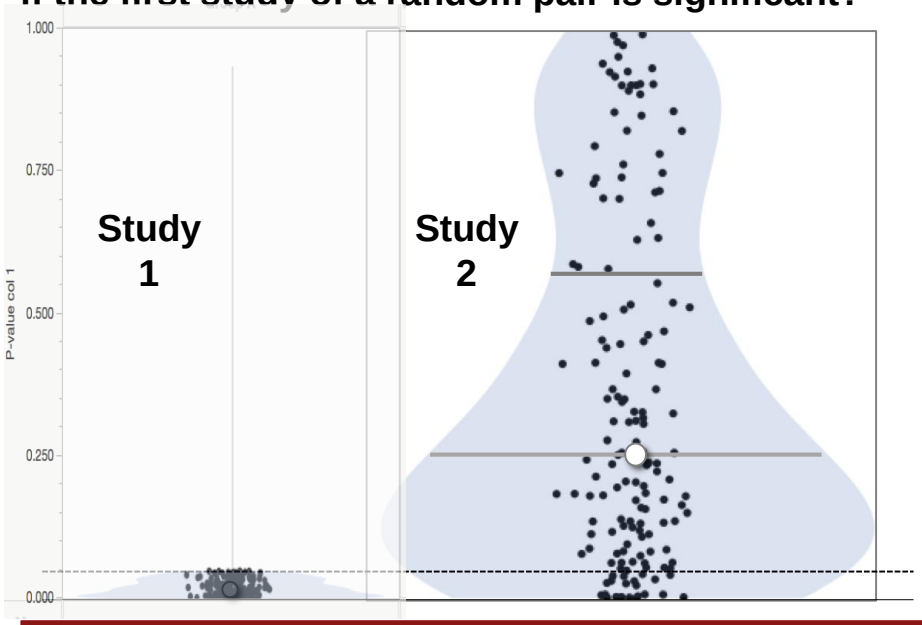


Fig. 1. Density plots of original and replication *P* values and effect sizes. (**A**) *P* values. (**B**) Effect sizes (correlation coefficients). Lowest quantile *P* values are not visible because they are clustered near zero.

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"!

Cancer Reproducibility Research: A clash of cultures?

57

Search results

Items: 1 to 20 of 57

1. **Registered report**: Melanoma exosomes educate bone marrow progenitor cells toward a pro-metastatic phenotype through MET.

   Lesnik J, Antes T, Kim J, Griner E, Pedro L; **Reproducibility Project**: Cancer Biology; **Reproducibility Project** Cancer Biology.

   Elife. 2016 Jan 29;5:e07383. doi: 10.7554/eLife.07383.

   PMID: 26826285 **Free PMC Article**

   Similar articles

2. **Registered report**: IDH mutation impairs histone demethylation and results in a block to cell differentiation.

   Richarson AD, Scott DA, Zagnitko O, Aza-Blanc P, Chang CC, Russler-Germain DA; **Reproducibility Project**: Cancer Biology.

   Elife. 2016 Mar 11;5:e10860. doi: 10.7554/eLife.10860.

   PMID: 26971564 **Free PMC Article**

   Similar articles

3. **Registered report**: Kinase-dead BRAF and oncogenic RAS cooperate to drive tumor progression through CRAF.

   Bhargava A, Anant M, Mack H; **Reproducibility Project**: Cancer Biology; **Reproducibility Project** Cancer Biology.

   Elife. 2016 Feb 17;5. pii: e11999. doi: 10.7554/eLife.11999.

   PMID: 26885666 **Free PMC Article**

   Similar articles

4. **Registered report**: Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma.

   Repass J, Maherali N, Owen K; **Reproducibility Project**: Cancer Biology; **Reproducibility Project** Cancer Biology.

   Elife. 2016 Feb 11;5. pii: e10012. doi: 10.7554/eLife.10012.

   PMID: 26882501 **Free PMC Article**

   Similar articles

5. **Registered report**: A chromatin-mediated reversible drug-tolerant state in cancer cell subpopulations.

   Haven B, Heilig E, Donham C, Settles M, Vasilevsky N, Owen K; **Reproducibility Project**: Cancer Biology; **Reproducibility Project** Cancer Biology.

   Elife. 2016 Feb 23;5. pii: e09462. doi: 10.7554/eLife.09462.

   PMID: 26905833 **Free PMC Article**

eLIFE elifesciences.org

REPLICATION STUDY

Replication Study: The CD47-signal regulatory protein alpha (SIRPa) interaction is a therapeutic target for human solid tumors

Stephen K Horrigan, Reproducibility Project: Cancer Biology*

Noble Life Sciences, Gaithersburg, United States

REPRODUCIBILITY PROJECT CANCER BIOLOGY

Abstract In 2015, as part of the Reproducibility Project: Cancer Biology, we published a Registered Report (Chroscinski et al., 2015) that described how we intended to replicate selected experiments from the paper "The CD47-signal regulatory protein alpha (SIRPa) interaction is a therapeutic target for human solid tumors "(Willingham et al., 2012). Here we report the results of

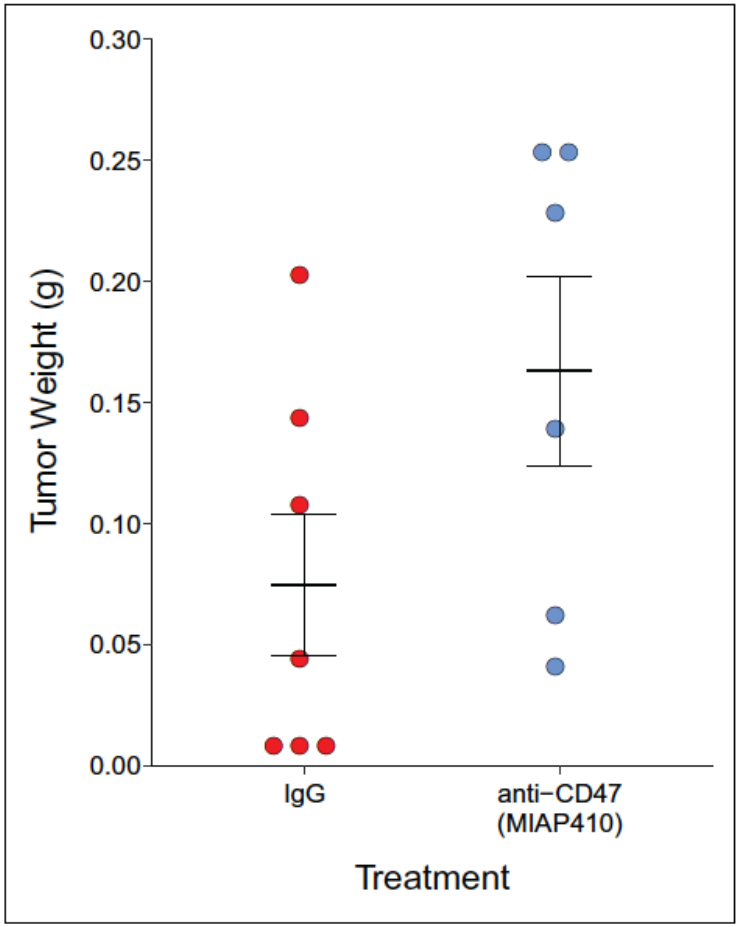


Figure 2. Final tumor weights of immune competent hosts treated with control or CD47 targeted antibodies. At the end of the predefined study period (Day 31), tumors from mice bearing orthotopic MT1A2 breast tumors treated every other day with IgG isotype control (IgG) (n = 7) or anti-mouse CD47 (anti-CD47) (n = 6) antibodies were excised and weighed. Dot plot with means reported as crossbars and error bars represent s.e.m. Two-tailed Welch's t-test between IgG and anti-CD47 treated tumors; t(9.66) = 1.796, p=0.104. Additional details for this experiment can be found at https://osf.io/g57ch/.

DOI: 10.7554/eLife.18173.004

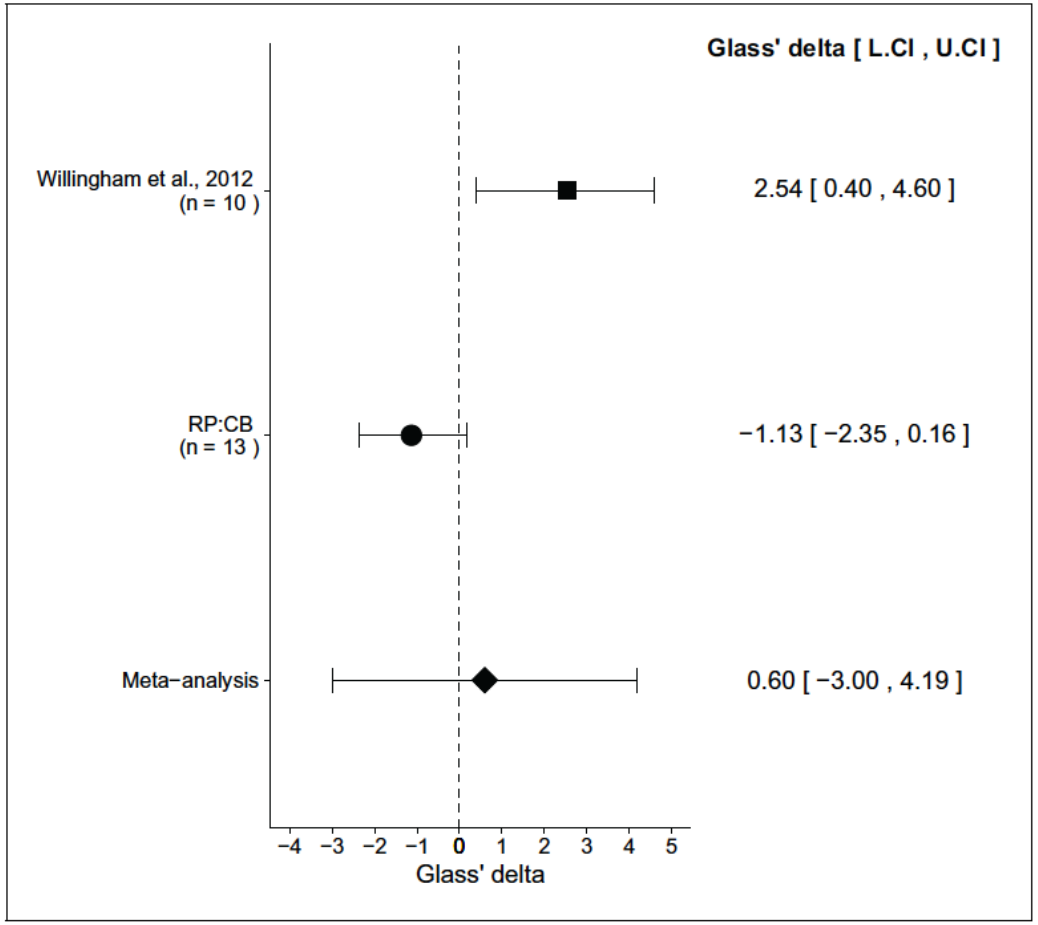


Figure 3. Meta-analysis of effect. Effect size (Glass' Δ) and 95% confidence interval are presented for ***Willingham et al. (2012)***, this replication attempt (RP:CB), and a meta-analysis to combine the two effects of tumor weight comparisons. Sample sizes used in ***Willingham et al. (2012)*** and this replication attempt are reported under the study name. Random effects meta-analysis of tumors treated with IgG compared to anti-CD47 (meta-analysis p=0.745). Additional details for this meta-analysis can be found at https://osf.io/ha2bx/.

DOI: 10.7554/eLife.18173.007

60

Editor's letter

- *….we note that the whole purpose of the replication study is to provide conclusive evidence for or against the originally reported claims. …*
- *For publication in* eLife *the results must be conclusive. Editorially we are willing to consider one last round of revision of your manuscript. However, we want to be clear that eLife will consider a revised manuscript only if it contains enough independent experiments, replicates and internal controls, to arrive at conclusive results. You would need to test therapeutic effects with control tumors that are allowed to grow much larger (e.g. at least 1cm in diameter) than before. You would also need to take into account that responses to immunotherapies in mouse models (and in the clinic) can be delayed.*
- *In sum, you need to provide sufficient experimental evidence to render the replication study conclusive. eLife would consider a revised manuscript only if and when this is achieved.*

Response

"This replication attempt, like all of the replication attempts in this project, are designed to perform independent replications with a calculated sample size to detect the originally reported effect size with at least 80% power. Further, this project will report the cumulative evidence across multiple independent replications among multiple studies. Thus, no single replication from this project, just like no original experiment or study, can provide conclusive evidence for or against an effect; rather, it's the cumulative evidence that forms the foundation of scientific knowledge."

Response, cont.

"However, we understand the desire to perform the experiment independently again, but with modifications to the design outlined and peer reviewed in the Registered Report – before the results were known. While it's not within the scope of this project, or as part of this publishing model, to also conduct these studies, the results of this replication bring variables not previously thought to influence the experiment into question (size of the control tumors at the end of the study, length of treatment, etc). Importantly though, it is only because of the results that these and other aspects now become targets for hypothesizing and investigation."

REPRODUCIBILITY IN CANCER BIOLOGY

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

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?"

REPRODUCIBILITY IN CANCER BIOLOGY

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

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."

Moving away from "reproducibility"

- Most empirical science is not a series of "proofs" and "disproofs".
- Our (flawed) intuition about "reproducibility" derives from experiments (or mathematics) where the signal/noise ratio is quite high, e.g. cold fusion, cloning.
- **The better question is how *efficient* is the scientific enterprise in generating reliable knowledge, what affects that efficiency, and how we can improve it. (Includes % of research that is *uninformative*.)**

Inferential Reproducibility

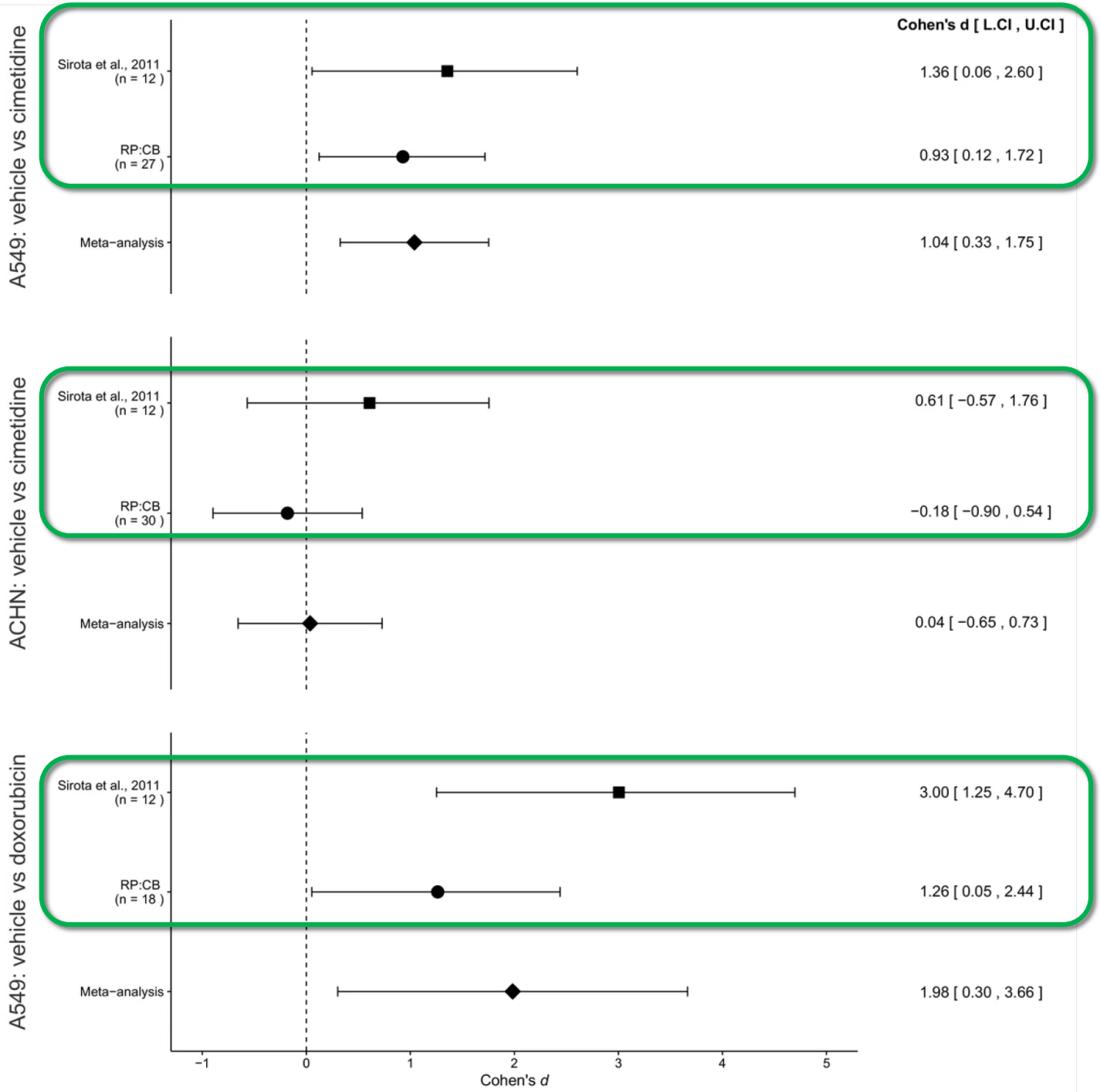
Replication Study: Discovery and preclinical validation of drug indications using compendia of public gene expression data

Irawati Kandela, Fraser Aird, Reproducibility Project: Cancer Biology*

Developmental Therapeutics Core, Northwestern University, Evanston, United States

REPRODUCIBILITY PROJECT CANCER BIOLOGY

Abstract In 2015, as part of the Reproducibility Project: Cancer Biology, we published a Registered Report (Kandela et al., 2015) that described how we intended to replicate selected experiments from the paper "Discovery and Preclinical Validation of Drug Indications Using Compendia of Public Gene Expression Data" (Sirota et al., 2011). Here we report the results of those experiments. We found that cimetidine treatment in a xenograft model using A549 lung adenocarcinoma cells resulted in decreased tumor volume compared to vehicle control; however, while the effect was in the same direction as the original study (Figure 4C; Sirota et al., 2011), it was not statistically significant. Cimetidine treatment in a xenograft model using ACHN renal cell carcinoma cells did not differ from vehicle control treatment, similar to the original study (Supplemental Figure 1; Sirota et al., 2011). Doxorubicin treatment in a xenograft model using A549 lung adenocarcinoma cells did not result in a statistically significant difference compared to vehicle control despite tumor volume being reduced to levels similar to those reported in the original study (Figure 4C; Sirota et al., 2011). Finally, we report a random effects meta-analysis for each result. These meta-analyses show that the inhibition of A549 derived tumors by cimetidine resulted in a statistically significant effect, as did the inhibition of A549 derived tumors by doxorubicin. The effect of cimetidine on ACHN derived tumors was not statistically significant, as predicted.


A549: vehicle vs cimetidine
ACHN: vehicle vs cimetidine
A549: vehicle vs doxorubicin
Cohen's d [L.CI , U.CI]
Sirota et al., 2011 (n = 12)
RP:CB (n = 27)
Meta−analysis
1.36 [0.06 , 2.60]
0.93 [0.12 , 1.72]
1.04 [0.33 , 1.75]
Sirota et al., 2011 (n = 12)
RP:CB (n = 30)
Meta−analysis
0.61 [−0.57 , 1.76]
−0.18 [−0.90 , 0.54]
0.04 [−0.65 , 0.73]
Sirota et al., 2011 (n = 12)
RP:CB (n = 18)
Meta−analysis
3.00 [1.25 , 4.70]
1.26 [0.05 , 2.44]
1.98 [0.30 , 3.66]
−1
0
1
2
3
4
5
Cohen's d

We found that ….**while the effect was in the same direction as the original study (Figure 4C; Sirota et al., 2011), it was not statistically significant.** …

Doxorubicin treatment in a xenograft model using A549 lung adenocarcinoma cells did not result in a statistically significant difference compared to vehicle control despite tumor volume being reduced to levels similar to those reported in the original study (Figure 4C; Sirota et al., 2011).

Finally, we report a random effects meta-analysis for each result. These meta-analyses show that the inhibition of A549 derived tumors by cimetidine resulted in a statistically significant effect, as did the inhibition of A549 derived tumors by doxorubicin.

70

Review comment

2) The original study observed a statistically significant reduction in A549 tumor volume while the current study did not, although the direction and magnitude of changes are similar. …. The current study conducted pre-planned contrasts on log transformed data within the framework of ANOVA, and the p-values are Bonferroni corrected, while the previous study performed t-tests on untransformed data without Bonferroni correction.

Why is Bonferroni correction used in the new study as opposed to testing directly the single original hypothesis, i.e. that tumor growth of cimetidine treated mice at the highest concentration is reduced compared to PBS/vehicle treatment? We don't see the rationale here for introducing a Bonferroni correction. In addition, it is well known that Bonferroni is an ultra-conservative way to account for multiple hypothesis testing. Without Bonferroni correction, the p-value is significant (p = 0.035) despite the larger error size. Please explain and discuss.

Authors' response

We performed Bonferroni correction because of the multiple comparisons (3 in total) that we performed on this experimental design. While this is a more conservative approach to adjust for multiple testing, we accounted for this in our power calculations to ensure the sample size was sufficient. ….. However, the difference between significant and not-significant is not necessarily statistically significant (see: Gelman and Stern, 2006; Nieuwenhuis et al., 2011), which is why the replication study designed and performed the statistical tests this way.

To allow readers to interpret the test results, we report the corrected and uncorrected p values. To further clarify why we performed the ANOVA and multiple comparison corrections we revised the manuscript to explain the analysis and how we accounted for the Bonferroni approach in our sample size calculations.

Review comment

- **3)** **Please note that there are more effective models in which the curves can be easily fit by regression analysis (e.g. MANOVA or Regression with RE/AR errors). Such models could be used here. Using the last time point is especially sensitive to measurement errors. …Please explain these issues and provide the data plotted as individual xenografts rather than averages in the supplemental data so that readers can examine the extent of mouse to mouse variations.**

- **Thank you for this suggestion. We agree that different models to analyze the data could be used and we encourage others to explore the data this way. …However, for this specific project we have restricted our analysis to what we specified in the Registered Report and how the sample size was determined. The final caliper measurement was also the approach taken in the original paper. Also, during peer review of the Registered Report we proposed performing an exploratory analysis of all the data, but removed this as suggested by a reviewer since in vivo caliper measurements are noisy, especially with the earlier time-points when the tumors are smaller in size. We have revised the manuscript to communicate how multiple approaches could be taken, but we have limited our analysis to what we proposed.**

eLIFE elifesciences.org

INSIGHT

REPRODUCIBILITY IN CANCER BIOLOGY

Mixed outcomes for computational predictions

REPRODUCIBILITY PROJECT CANCER BIOLOGY

Experimental efforts to validate the output of a computational model that predicts new uses for existing drugs highlights the inherently complex nature of cancer biology.

CHI VAN DANG

Dang. eLife 2017;6:e22661. DOI: 10.7554/eLife.22661

PLOS | BIOLOGY

Accuracy was related to expertise: experts with higher h-indices were more accurate, whereas experts with more topic-specific expertise were less accurate. Our findings suggest that experts, especially those with specialized knowledge, were overconfident about the RP:CB replicating individual experiments within published reports; researcher optimism likely reflects a combination of overestimating the validity of original studies and underestimating the difficulties of repeating their methodologies.

OPEN ACCESS

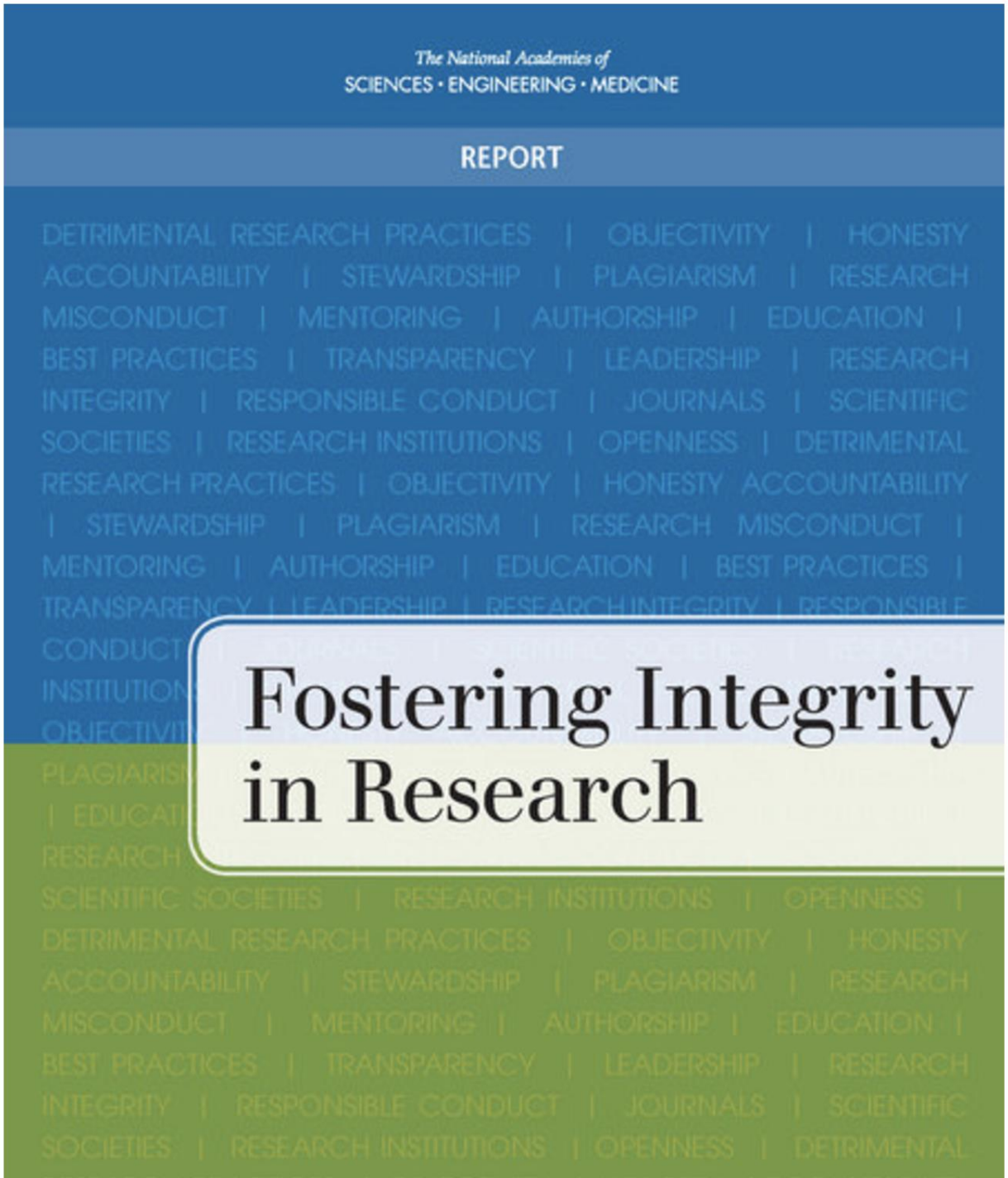
Citation: Benjamin D, Mandel DR, Kimmelman J (2017) Can cancer researchers accurately judge whether preclinical reports will reproduce? PLoS Biol 15(6): e2002212. https://doi.org/10.1371/journal.pbio.2002212

Academic Editor: Lisa Bero, University of Sydney, Australia

Received: February 12, 2017

Accepted: May 18, 2017

...er biology. ...original findings ...mining the pace at which science self-corrects. We collected forecasts fr... ...nical cancer researchers on the first 6 replication studies conducted by the Re... ...roject: Cancer Biology (RP:CB) to assess the accuracy of expert judgments on ... ...lication outcomes. On average, researchers forecasted a 75% probability of replica... ...statistical significance and a 50% probability of replicating the effect size, yet none of th... ...studies successfully replicated on either criterion (for the 5 studies with results reported). Accuracy was related to expertise: experts with higher h-indices were more accurate, whereas experts with more topic-specific expertise were less accurate. Our findings suggest that experts, especially those with specialized knowledge, were overconfident about the RP:CB replicating individual experiments within published reports; researcher optimism likely reflects a combination of overestimating the validity of original studies and underestimating the difficulties of repeating their methodologies.



Included DRPs in their definition of research integrity, and made recommendations for institutional changes.

STANFORD UNIVERSITY

“Misused statistics” is standard methodology within many disciplines….

- Reliance on single studies to establish major claims
- Routine underpowering of experiments.
- Selective reporting of experiments that ”work.”
- Failure to account adequately for massive multiplicity and high “research degrees of freedom”.
- Poor reporting and handling of missing data
- Strong publication bias in high tier journals.
- Failure to conduct sensitivity (aka robustness) analyses.
- Inadequate internal or external validation.
- Use of rigid significance verdicts. Minimal reporting of precision.

Reproducibility and precision oncology

- Because these predictors almost defy assessment of ”face validity,” attention to rigorous development, complete and transparent reporting and proper validation is more important than for medical interventions whose mechanism is at least partly understood.
- Transparency about developmental methods is less important than those around validation.
- Multiple validation studies should be done.
- RCTs maybe be the only way to assess the whole decisional-therapeutic pathway.
- Continued performance monitoring, e.g. Phase IV, should be done with these markers just as with therapeutics.
- Patient outcomes are the bottom line.

ScienceInsider
Breaking news and analysis from the world of science policy

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

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!