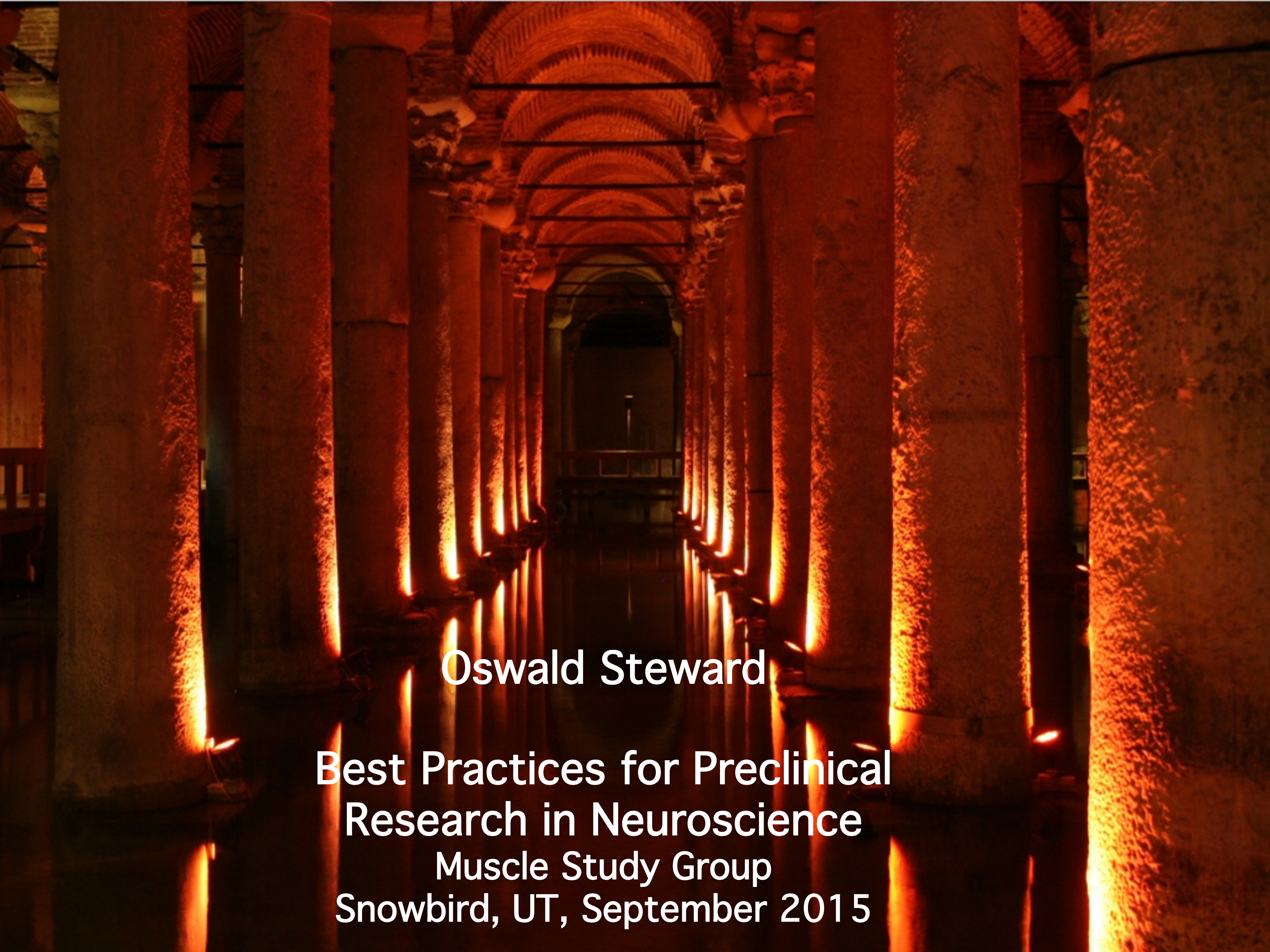
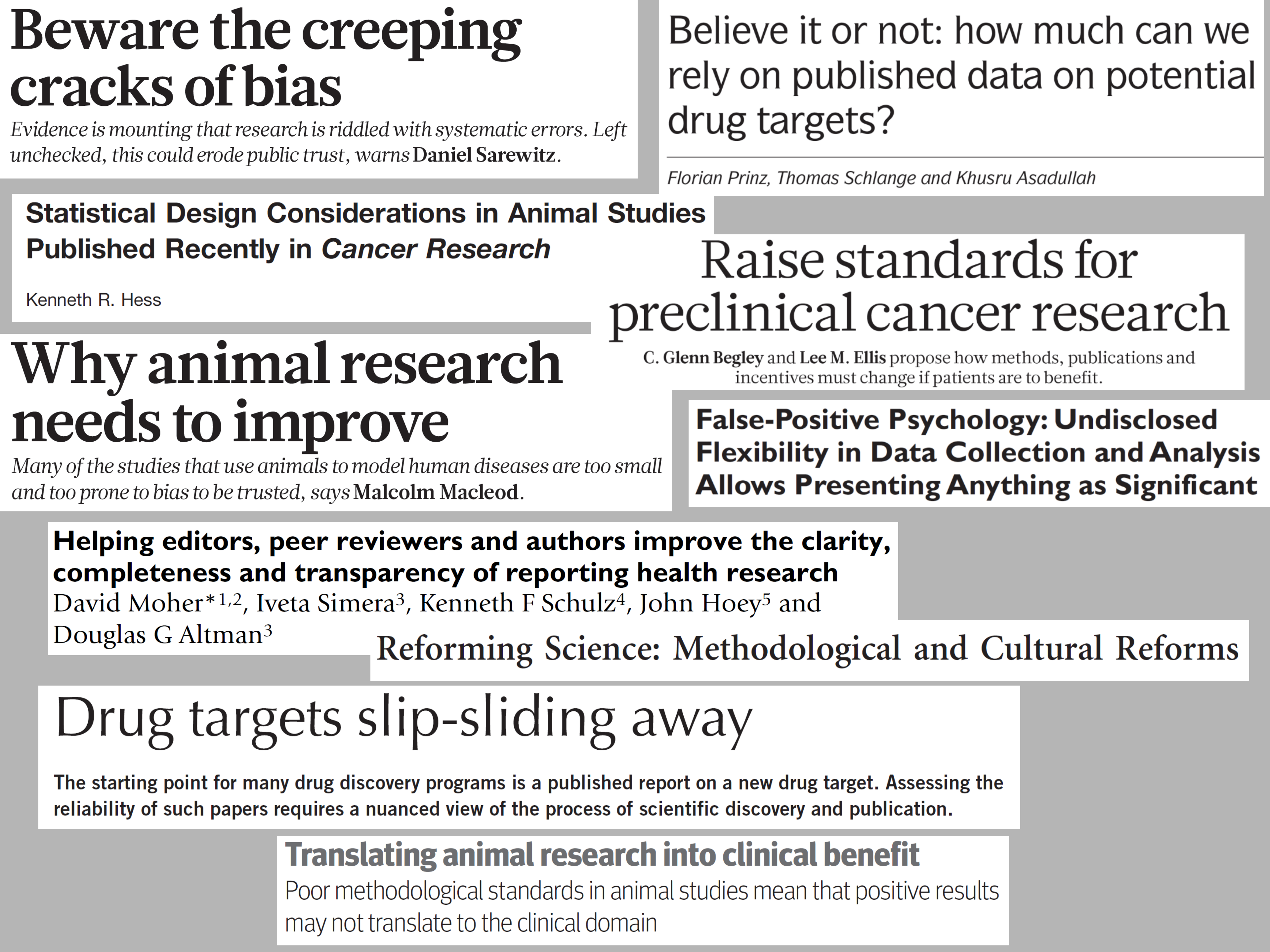


Oswald Steward

Best Practices for Preclinical Research in Neuroscience

Muscle Study Group

Snowbird, UT, September 2015



Beware the creeping cracks of bias

Evidence is mounting that research is riddled with systematic errors. Left unchecked, this could erode public trust, warns **Daniel Sarewitz**.

Believe it or not: how much can we rely on published data on potential drug targets?

Florian Prinz, Thomas Schlange and Khusru Asadullah

Statistical Design Considerations in Animal Studies Published Recently in *Cancer Research*

Kenneth R. Hess

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.

Why animal research needs to improve

Many of the studies that use animals to model human diseases are too small and too prone to bias to be trusted, says **Malcolm Macleod**.

False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant

Helping editors, peer reviewers and authors improve the clarity, completeness and transparency of reporting health research

David Moher*[1,2], Iveta Simera[3], Kenneth F Schulz[4], John Hoey[5] and Douglas G Altman[3]

Reforming Science: Methodological and Cultural Reforms

Drug targets slip-sliding away

The starting point for many drug discovery programs is a published report on a new drug target. Assessing the reliability of such papers requires a nuanced view of the process of scientific discovery and publication.

Translating animal research into clinical benefit

Poor methodological standards in animal studies mean that positive results may not translate to the clinical domain

- The Problem: An increasing number of recent reports of lack of reproducibility of published findings.

- Several high profile replication attempts have been unable to reproduce most studies that were examined.

- The issue is especially critical for preclinical research, which can be the basis for clinical trials that are doomed to fail.

- Congress is taking notice.

- NIH and journals are revising review criteria.

Concerns identified in preclinical cancer studies:

1) Prinz et al. (2011) Believe it or not: How much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery.* Inconsistencies in 2/3 studies

2) Begley and Ellis (2012) Raise standards for preclinical cancer research. *Nature.* 6/53 “landmark” papers replicated

3) Begley (2013) Six red flags for suspect work. *Nature.*
-Were experiments performed blinded?
-Were basic experiments repeated?
-Were all the results presented?
-Were there positive and negative controls?
-Were reagents validated?
-Were statistical tests appropriate?

An example of the problem in Neuroscience: Our experience with the FORE-SCI Replication Contract.

In 2003, NINDS issued an RFP for contracts to replicate SCI models and treatment. Fore-SCI contracts were funded for a total of 10 years and 20 promising, high profile studies were repeated, 18 of which have been published.

Only about 10% of the published findings were replicated. A major problem was lack of transparency of experimental details in published papers.

Our interim report:

Steward et al (2012) Replication and reproducibility in spinal cord research. *Exp. Neurol. Special Issue, 233, 597-605.*

Findings and conclusions from the FORE-SCI project

- Surprising preponderance of failures to replicate (16/18)
  - What does a failure to replicate actually mean?

- Methods sections are often misleading
  - Randomization is rarely explained and often is NOT DONE.
  - Communication with original authors often reveals that the experiment was NOT done as the Methods imply.

- Publishing negative results is doable and generally well-received by the field.

  - Over the past 10 years, I've published 13 papers reporting failure to replicate including one in Cell and 2 in Neuron.

Important methodological issues we discovered

- Papers describe work carried out over prolonged time periods, sometimes several years. Experimental groups are sometimes NOT run simultaneously, but this is not described in the Methods. This is not unique to SCI research or to preclinical studies.

- Batching of animals/non-simultaneity of group assessment is almost never explained.

- In some cases, there is no practical alternative (for example with complicated protocols in which only a few animals can be done at any time).

In preclinical research the problems seem to be more about experimental design and executioin than post-hoc quantitative or statistical analysis. Problems include:

1-Pooling data from experiments done over time and then compiling groups at the end. Often the subjects in individual experiments in the compilation were not randomized, and sometimes different groups may be done on different days. This is especially problematic for interventions that take time to produce (like a spinal cord injury).

2-Testing to a foregone conclusion: This involves doing interim statistical analyses and increasing "n" until a significant effect is seen. This related to the first because of the lack of clear stopping rules in studies done over prolonged time periods.

Continued:

3-Searching for the positive result (multiple comparisons until you find a measure on which groups differ).

4-Publication bias for positive results, and failure to report the entire collection of analyses in a particular study.

5- Lack of self-replication prior to publication.

6-Failure to report methods completely and transparently, especially in terms of pooling data from different experiments, randomization, and group compilation.

The most common criticism of reports of failure to replicate is that the replication wasn’t done in exactly the same way. This is invariably true.

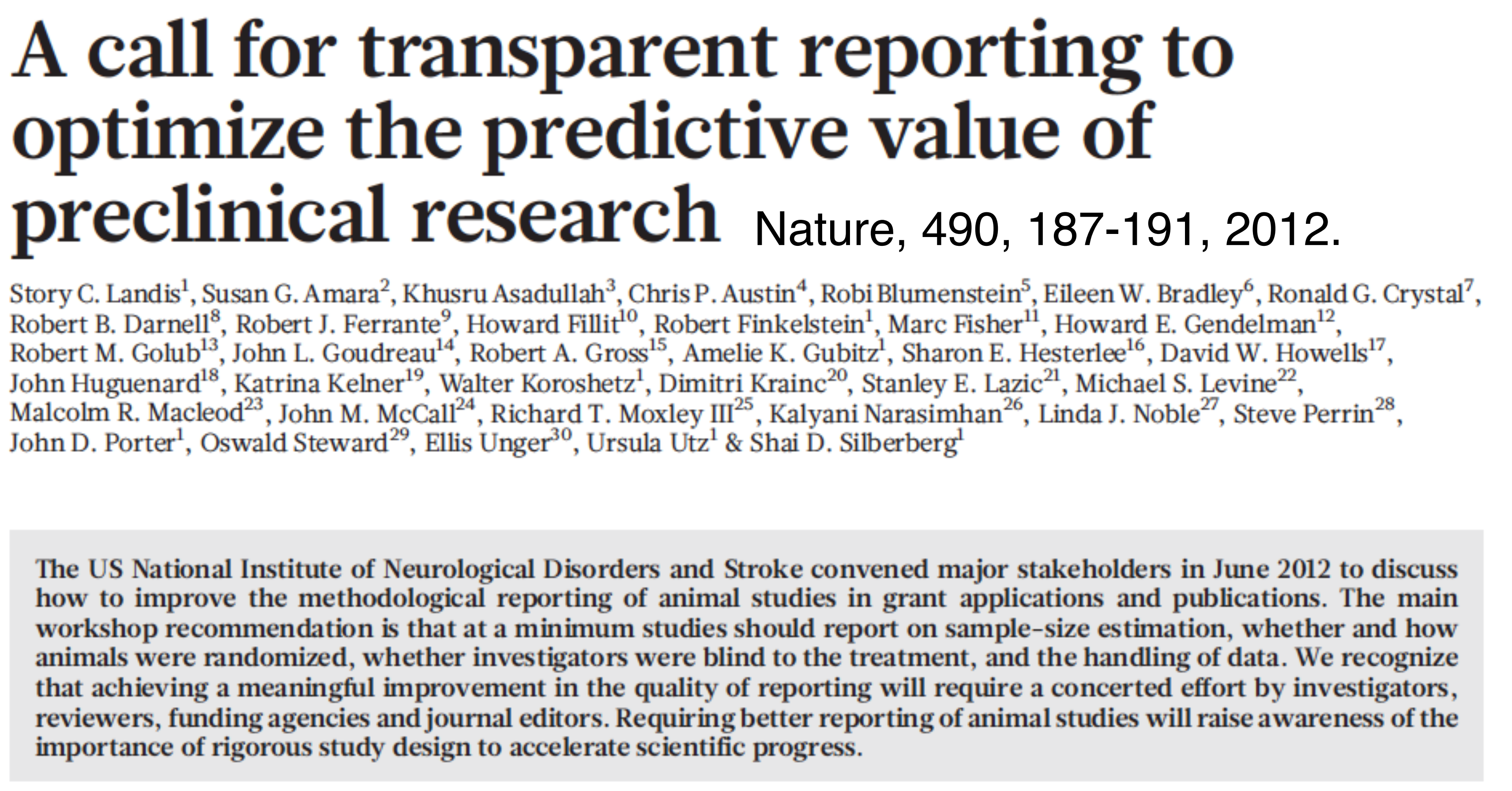
BUT, whatever happened to the “caveats” section of Discussions? If there is reason to believe that findings only apply in a highly constrained set of circumstances, it’s important to say that. At the very least, until proven otherwise, it’s important to say that the findings MIGHT only apply in a highly constrained circumstances

AND for preclinical studies that are presented as pointing the way to therapies, if things only work in highly constrained settings, the approach is NOT going to be translatable.

The problem is that the culture of science and the reward structure of academics emphasizes “high profile” journals. Noting caveats doesn’t get you there.

Due in part to the results from the FORE-SCI Contracts and other reports, NINDS convened a workshop in June 2012.

Minimal requirements from NINDS workshop: sample size estimation, whether and how animals were randomized, blinding, appropriate data handling (data inclusion, exclusion) and thorough and transparent reporting.



A call for transparent reporting to optimize the predictive value of preclinical research

Nature, 490, 187-191, 2012.

Story C. Landis[1], Susan G. Amara[2], Khusru Asadullah[3], Chris P. Austin[4], Robi Blumenstein[5], Eileen W. Bradley[6], Ronald G. Crystal[7], Robert B. Darnell[8], Robert J. Ferrante[9], Howard Fillit[10], Robert Finkelstein[1], Marc Fisher[11], Howard E. Gendelman[12], Robert M. Golub[13], John L. Goudreau[14], Robert A. Gross[15], Amelie K. Gubitz[1], Sharon E. Hesterlee[16], David W. Howells[17], John Huguenard[18], Katrina Kelner[19], Walter Koroshetz[1], Dimitri Krainc[20], Stanley E. Lazic[21], Michael S. Levine[22], Malcolm R. Macleod[23], John M. McCall[24], Richard T. Moxley III[25], Kalyani Narasimhan[26], Linda J. Noble[27], Steve Perrin[28], John D. Porter[1], Oswald Steward[29], Ellis Unger[30], Ursula Utz[1] & Shai D. Silberberg[1]

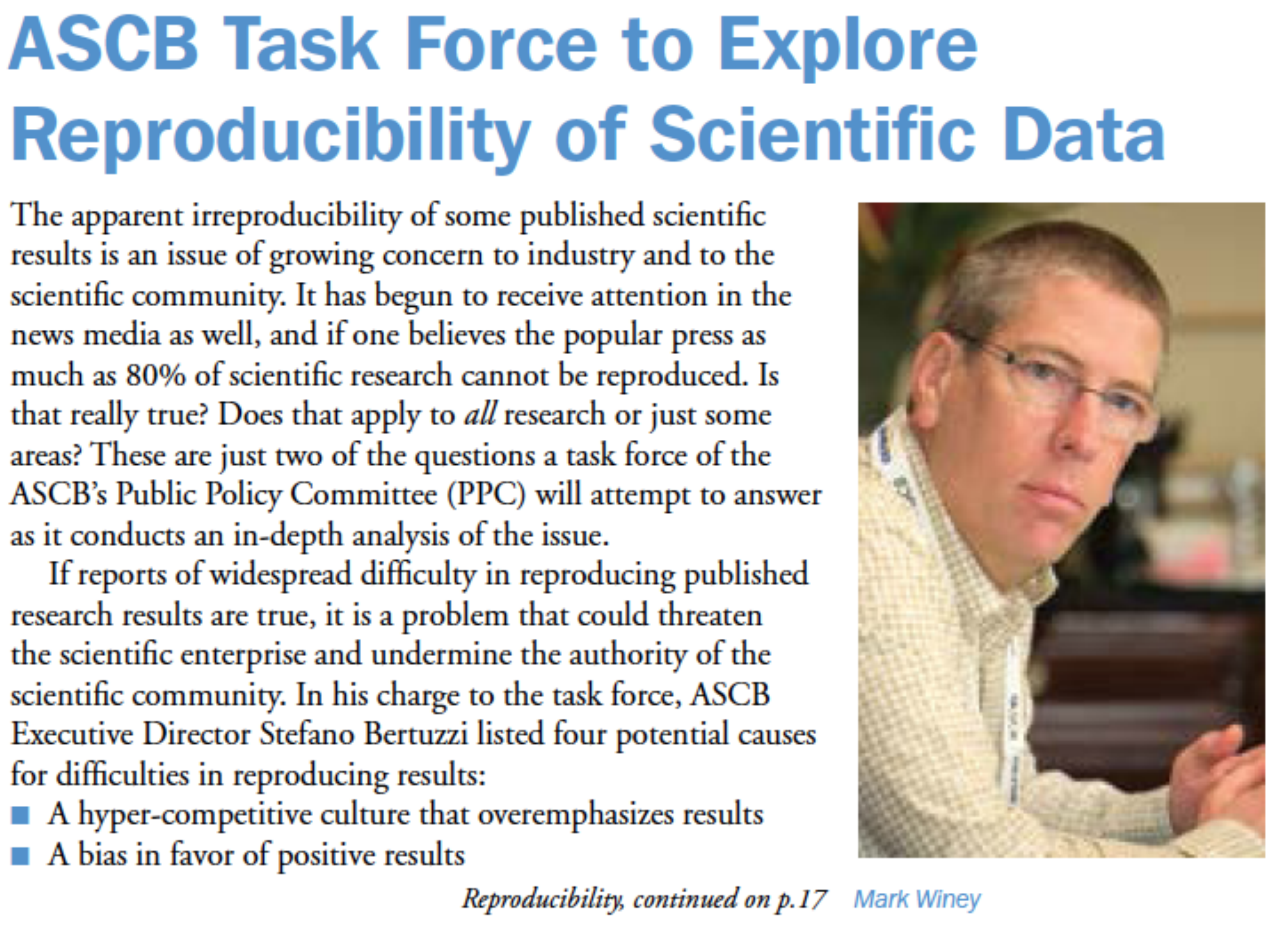
The US National Institute of Neurological Disorders and Stroke convened major stakeholders in June 2012 to discuss how to improve the methodological reporting of animal studies in grant applications and publications. The main workshop recommendation is that at a minimum studies should report on sample-size estimation, whether and how animals were randomized, whether investigators were blind to the treatment, and the handling of data. We recognize that achieving a meaningful improvement in the quality of reporting will require a concerted effort by investigators, reviewers, funding agencies and journal editors. Requiring better reporting of animal studies will raise awareness of the importance of rigorous study design to accelerate scientific progress.

In 2013, The Society for Neuroscience established a Scientific Rigor Working Group (O. Steward and E. Dicco-Bloom, Co-Chairs).

Through the efforts of the Working Group, there were two symposia related to scientific rigor at the SFN meeting in 2014 including: *Reliability of research findings: Emerging best practices to improve rigor:* Participants included Story Landis, Tom Insel, Francis Collins, Huda Zoghbi, and John Morrison.

SFN has received a grant to produce training modules in best practices to enhance scientific rigor.

There are ongoing efforts by other scientific societies to improve reproducibility



ASCB Task Force to Explore Reproducibility of Scientific Data

The apparent irreproducibility of some published scientific results is an issue of growing concern to industry and to the scientific community. It has begun to receive attention in the news media as well, and if one believes the popular press as much as 80% of scientific research cannot be reproduced. Is that really true? Does that apply to *all* research or just some areas? These are just two of the questions a task force of the ASCB's Public Policy Committee (PPC) will attempt to answer as it conducts an in-depth analysis of the issue.

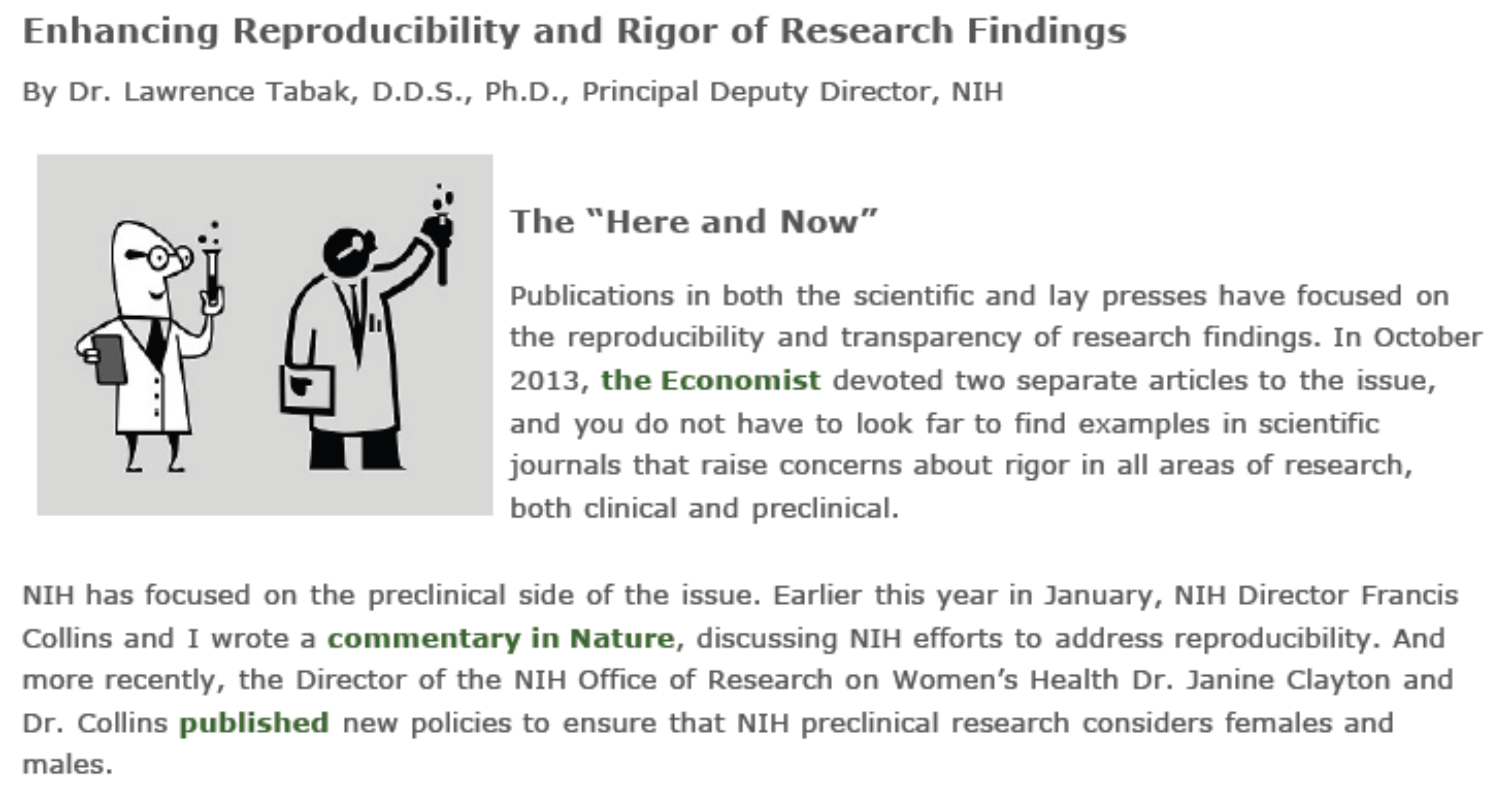
If reports of widespread difficulty in reproducing published research results are true, it is a problem that could threaten the scientific enterprise and undermine the authority of the scientific community. In his charge to the task force, ASCB Executive Director Stefano Bertuzzi listed four potential causes for difficulties in reproducing results:

- A hyper-competitive culture that overemphasizes results
- A bias in favor of positive results

Reproducibility, continued on p.17

Mark Winey

Actions by NIH



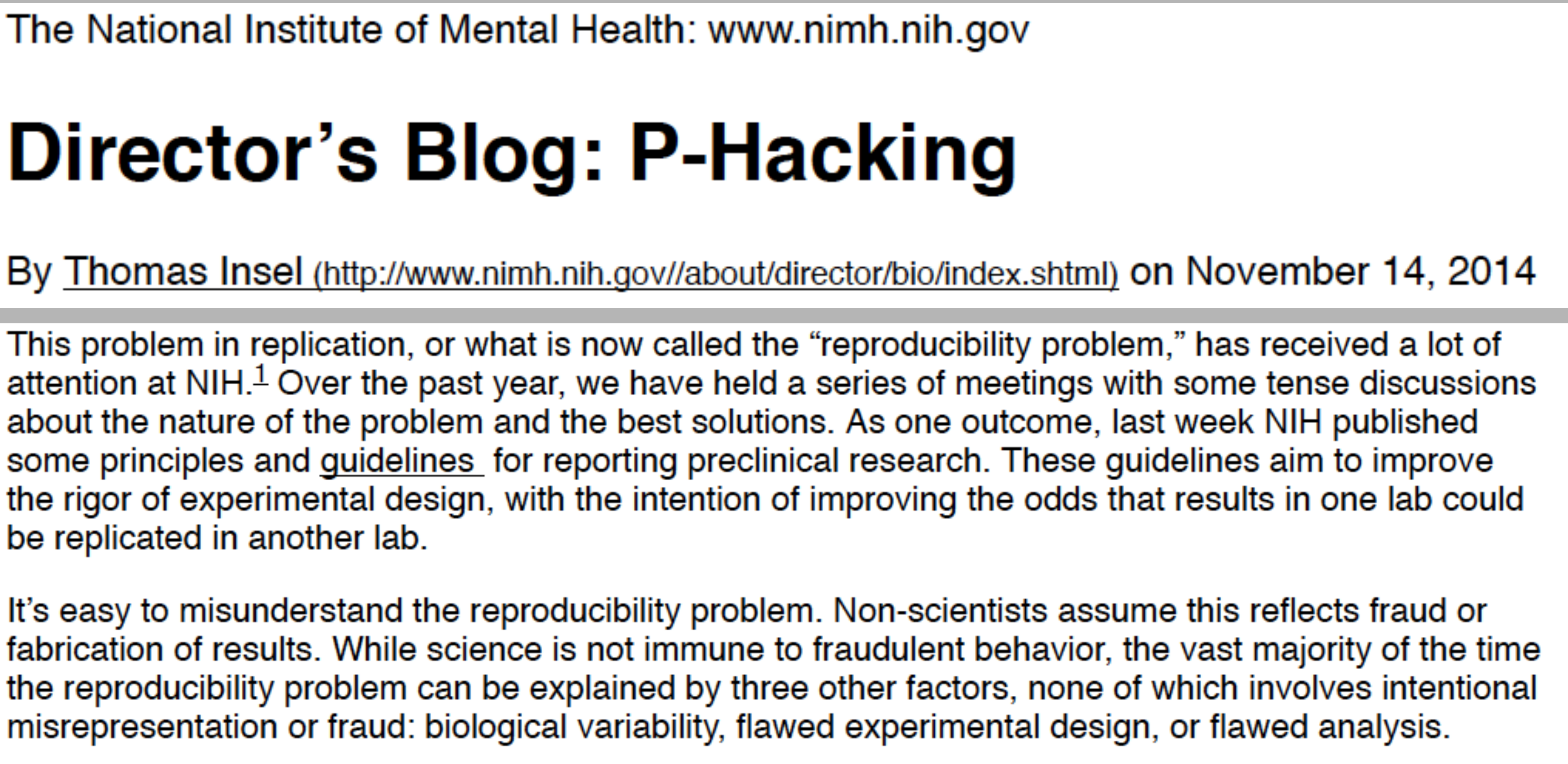
Enhancing Reproducibility and Rigor of Research Findings

By Dr. Lawrence Tabak, D.D.S., Ph.D., Principal Deputy Director, NIH

The "Here and Now"

Publications in both the scientific and lay presses have focused on the reproducibility and transparency of research findings. In October 2013, **the Economist** devoted two separate articles to the issue, and you do not have to look far to find examples in scientific journals that raise concerns about rigor in all areas of research, both clinical and preclinical.

NIH has focused on the preclinical side of the issue. Earlier this year in January, NIH Director Francis Collins and I wrote a **commentary in Nature**, discussing NIH efforts to address reproducibility. And more recently, the Director of the NIH Office of Research on Women's Health Dr. Janine Clayton and Dr. Collins **published** new policies to ensure that NIH preclinical research considers females and males.



The National Institute of Mental Health: www.nimh.nih.gov

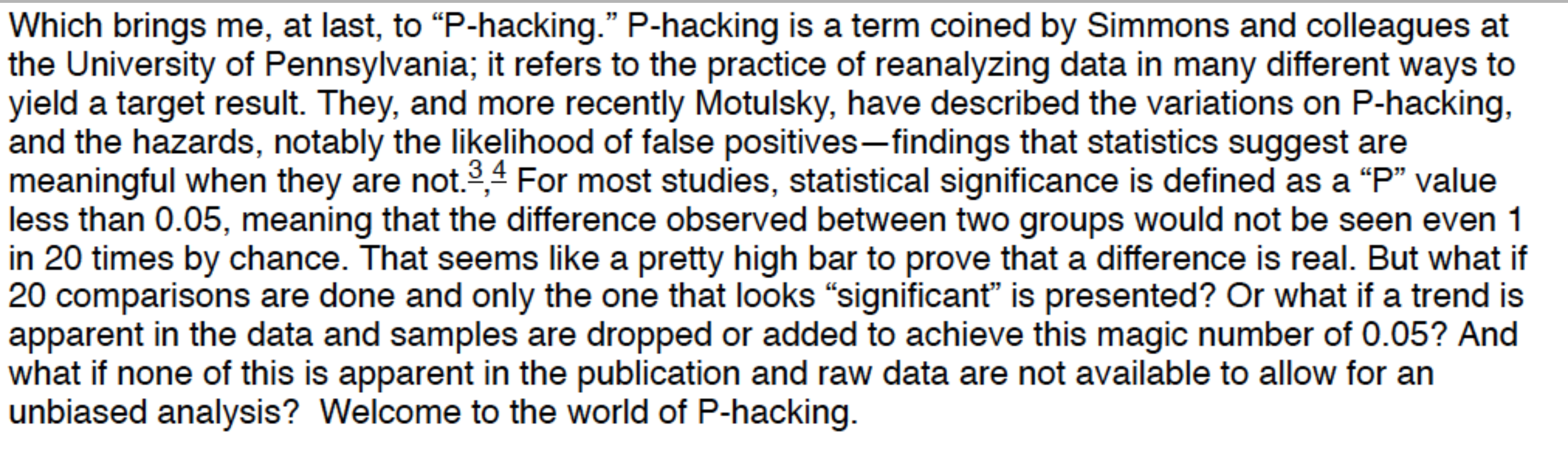
Director’s Blog: P-Hacking

By Thomas Insel (http://www.nimh.nih.gov//about/director/bio/index.shtml) on November 14, 2014

This problem in replication, or what is now called the “reproducibility problem,” has received a lot of attention at NIH.[1] Over the past year, we have held a series of meetings with some tense discussions about the nature of the problem and the best solutions. As one outcome, last week NIH published some principles and guidelines for reporting preclinical research. These guidelines aim to improve the rigor of experimental design, with the intention of improving the odds that results in one lab could be replicated in another lab.

It’s easy to misunderstand the reproducibility problem. Non-scientists assume this reflects fraud or fabrication of results. While science is not immune to fraudulent behavior, the vast majority of the time the reproducibility problem can be explained by three other factors, none of which involves intentional misrepresentation or fraud: biological variability, flawed experimental design, or flawed analysis.

P-hacking

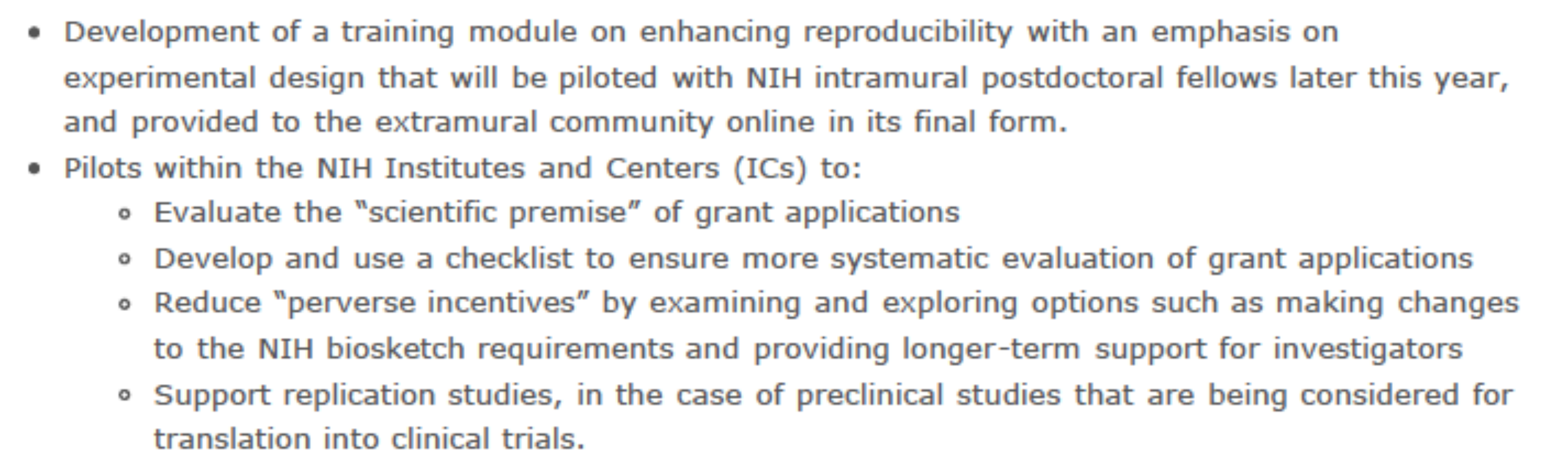


Which brings me, at last, to “P-hacking.” P-hacking is a term coined by Simmons and colleagues at the University of Pennsylvania; it refers to the practice of reanalyzing data in many different ways to yield a target result. They, and more recently Motulsky, have described the variations on P-hacking, and the hazards, notably the likelihood of false positives—findings that statistics suggest are meaningful when they are not.[3],[4] For most studies, statistical significance is defined as a “P” value less than 0.05, meaning that the difference observed between two groups would not be seen even 1 in 20 times by chance. That seems like a pretty high bar to prove that a difference is real. But what if 20 comparisons are done and only the one that looks “significant” is presented? Or what if a trend is apparent in the data and samples are dropped or added to achieve this magic number of 0.05? And what if none of this is apparent in the publication and raw data are not available to allow for an unbiased analysis? Welcome to the world of P-hacking.

A core set of reporting standards for rigorous study design

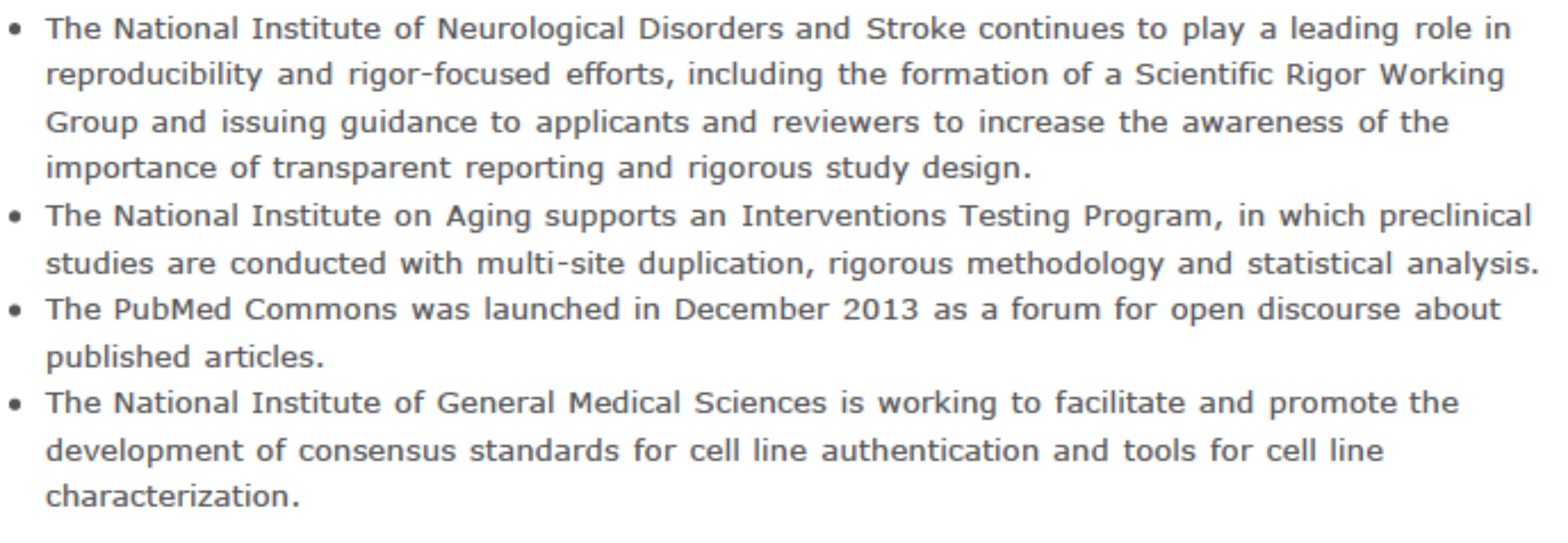
- Randomization.
- Blinding
  - Allocation concealment
  - Blinded testing
  - Blinded outcome assessment
- Sample size determination (pre experiment power calculations.
- Data handling
  - Stopping rules
  - Prospective inclusion/exclusion criteria
  - Handling of outliers
  - Endpoint selection (avoiding testing to a foregone conclusion)
  - Defining what constitutes an “experiment” for purposes of analysis.

NIH Initiatives include:



- Development of a training module on enhancing reproducibility with an emphasis on experimental design that will be piloted with NIH intramural postdoctoral fellows later this year, and provided to the extramural community online in its final form.
- Pilots within the NIH Institutes and Centers (ICs) to:
  - Evaluate the "scientific premise" of grant applications
  - Develop and use a checklist to ensure more systematic evaluation of grant applications
  - Reduce "perverse incentives" by examining and exploring options such as making changes to the NIH biosketch requirements and providing longer-term support for investigators
  - Support replication studies, in the case of preclinical studies that are being considered for translation into clinical trials.

NIH Initiatives include:



- The National Institute of Neurological Disorders and Stroke continues to play a leading role in reproducibility and rigor-focused efforts, including the formation of a Scientific Rigor Working Group and issuing guidance to applicants and reviewers to increase the awareness of the importance of transparent reporting and rigorous study design.
- The National Institute on Aging supports an Interventions Testing Program, in which preclinical studies are conducted with multi-site duplication, rigorous methodology and statistical analysis.
- The PubMed Commons was launched in December 2013 as a forum for open discourse about published articles.
- The National Institute of General Medical Sciences is working to facilitate and promote the development of consensus standards for cell line authentication and tools for cell line characterization.

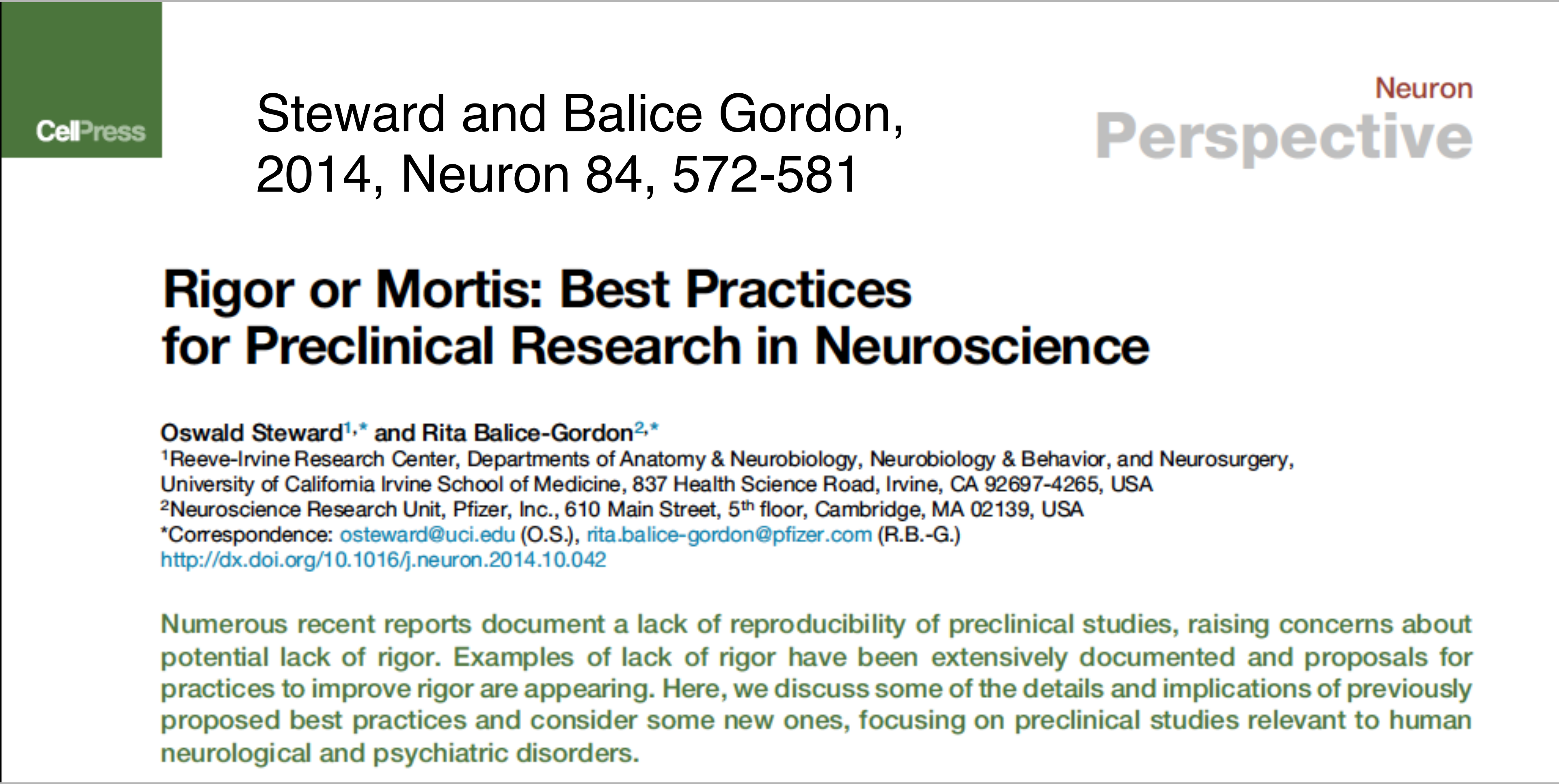
EDITORIAL

Journals unite for reproducibility

> "...scientific journals are standing together in their conviction that reproducibility and transparency are important..."

Marcia McNutt
Editor-in-Chief
Science *Journals*

Some recommendations for best practices for preclinical research in neuroscience



Steward and Balice Gordon, 2014, Neuron 84, 572-581

CellPress

Neuron
Perspective

Rigor or Mortis: Best Practices for Preclinical Research in Neuroscience

Oswald Steward[1,*] **and Rita Balice-Gordon**[2,*]
[1]Reeve-Irvine Research Center, Departments of Anatomy & Neurobiology, Neurobiology & Behavior, and Neurosurgery, University of California Irvine School of Medicine, 837 Health Science Road, Irvine, CA 92697-4265, USA
[2]Neuroscience Research Unit, Pfizer, Inc., 610 Main Street, 5th floor, Cambridge, MA 02139, USA
*Correspondence: osteward@uci.edu (O.S.), rita.balice-gordon@pfizer.com (R.B.-G.)

http://dx.doi.org/10.1016/j.neuron.2014.10.042

Numerous recent reports document a lack of reproducibility of preclinical studies, raising concerns about potential lack of rigor. Examples of lack of rigor have been extensively documented and proposals for practices to improve rigor are appearing. Here, we discuss some of the details and implications of previously proposed best practices and consider some new ones, focusing on preclinical studies relevant to human neurological and psychiatric disorders.

Latin scholars will note this should be "Rigor or Mort".

Reliability of research findings: Emerging best practices to improve rigor

Neuron
Rigor or Mortis: Best Practices for Preclinical Research in Neuroscience

NIH Director Frances Collins talking about our paper.

Frances Collins

Co-moderators at the table were Story Landis, recently retired Director of the National Institutes of Neurological Disorders and Stroke and Tom Insel, Director, National Institute of Mental Health.

The biopharma definition of “preclinical research”:
Everything done prior to human biology validation studies, i.e., everything done in cells and animals.

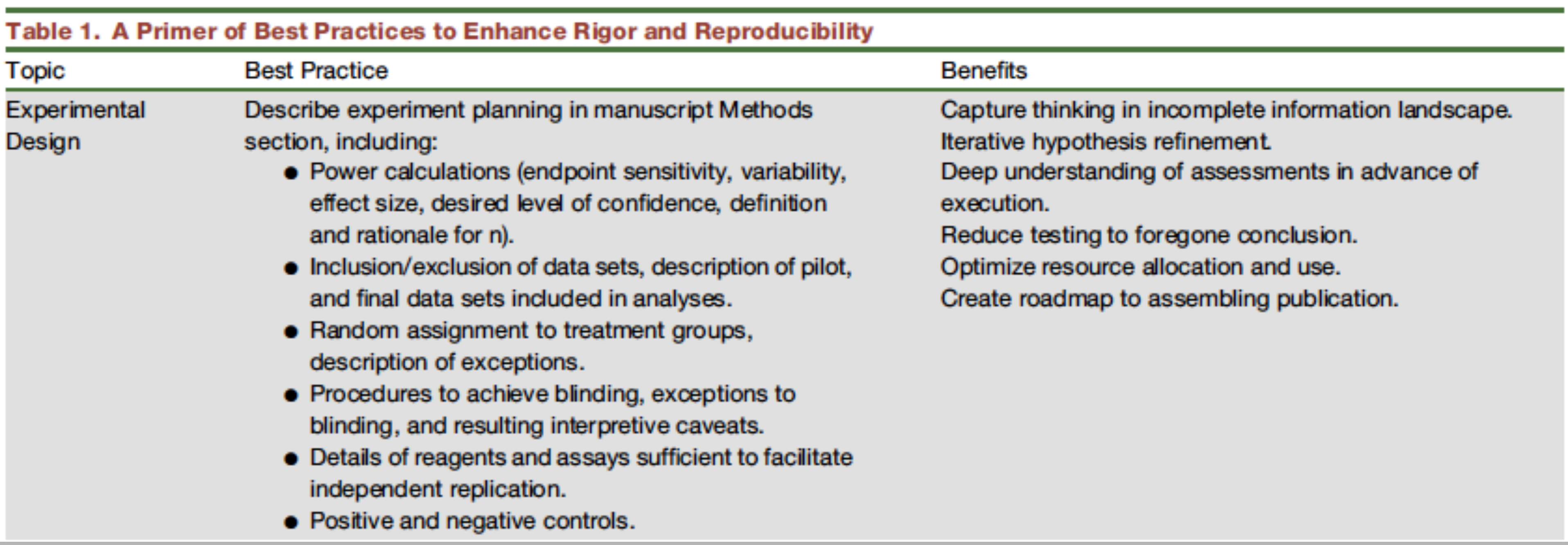


Table 1. A Primer of Best Practices to Enhance Rigor and Reproducibility

Topic	Best Practice	Benefits
Experimental Design	Describe experiment planning in manuscript Methods section, including: ● Power calculations (endpoint sensitivity, variability, effect size, desired level of confidence, definition and rationale for n). ● Inclusion/exclusion of data sets, description of pilot, and final data sets included in analyses. ● Random assignment to treatment groups, description of exceptions. ● Procedures to achieve blinding, exceptions to blinding, and resulting interpretive caveats. ● Details of reagents and assays sufficient to facilitate independent replication. ● Positive and negative controls.	Capture thinking in incomplete information landscape. Iterative hypothesis refinement. Deep understanding of assessments in advance of execution. Reduce testing to foregone conclusion. Optimize resource allocation and use. Create roadmap to assembling publication.

Analysis and Statistics	Describe statistical analysis plan in manuscript Methods section, including: ● Methods to test for significance. ● Interim analyses, futility assessments. ● Data inclusion/exclusion, attrition. ● Statistical treatment of technical and biological replicates. ● Test-retest approaches. ● Statement of central tendency, variance, statistical test, and p value for significant and nonsignificant differences. ● Descriptive statistics for groups as well as pooled values.	Enhance awareness of and reduce sources of potential unconscious bias. Minimize type 1 error.

Data Management	Develop lab standards for indexing and maintaining information, including: ● Recording of key experimental design and execution parameters. ● Archiving raw data and at least one backup with appropriate frequency. ● Curation of process from raw data to summary figure to conclusion.	Ensure all information supporting a conclusion can be located during and after study completion.

Resource Sharing	Include lists of resources in manuscripts that will be made available and point of contact for requests. Indicate time limit for resource availability, if any. Include budget line item to support resource sharing in funding applications. Deposit animal lines at commercial vendor within 3 months of publication. Provide raw data upon request.	Simplify sharing of reagents, protocols, raw data to facilitate replication, interpretation of data. Help distinguish lack of conceptual validation versus lack of replication. Enable meta-analyses and data basing.
Publication and Reporting	Provide comprehensive review checklist for methodology, reagents, and resource sharing. Two-stage review: if manuscript meets general journal criteria (novelty, impact, general interest), initiate second stage of review for technical merit including details relating to rigor.	Raise awareness of key metrics for determining rigor. Facilitate replication of key findings.

One approach to training at the institution
Building upon the required RCR course

John H. Morrison, PhD

Professor of Neuroscience

Dean, Basic Sciences and the Graduate School of Biomedical Sciences

Icahn School of Medicine at Mount Sinai

Responsible Conduct of Research: RCR

- NIH requires that all trainees, …receiving support through any NIH grant… must receive instruction in responsible conduct of research.”

- Format: Substantial face-to-face discussions…are highly encouraged. Online instruction is not considered adequate.

- Instruction must be undertaken at least once during each career stage, and at a frequency of no less than once every four years.

RCR: NIH Conduct Issues of Concern

RCR at Mount Sinai includes eight modules, 7 on how to be a good scientist

- Research Misconduct
- Lab notebooks
- Conflict of Interest
- Human Subjects
- Animal Welfare
- Publication Practices and Responsible Authorship
- Mentor / Trainee Responsibilities
- Peer Review
- Collaborative Science

RCR: NIH Conduct Issues of Concern
Build in Best Practices (Charles Mobbs, PhD)

- Research Misconduct: Clarify what it is and is not
- Lab notebooks: Data acquisition, management, ownership, sharing; Detailed protocols
- Conflict of Interest: Unintentional bias
- Human Subjects: Randomization, blind, stats
- Animal Welfare: Quality and choice of animal model
- Publication Practices and Responsible Authorship: Transparency, detailed methods, full reporting
- Mentor / Trainee Responsibilities: Appropriate incentives; testing vs proving hypothesis
- Peer Review: Balancing rigor and novelty
- Collaborative Science: Lab visits and independent corroboration

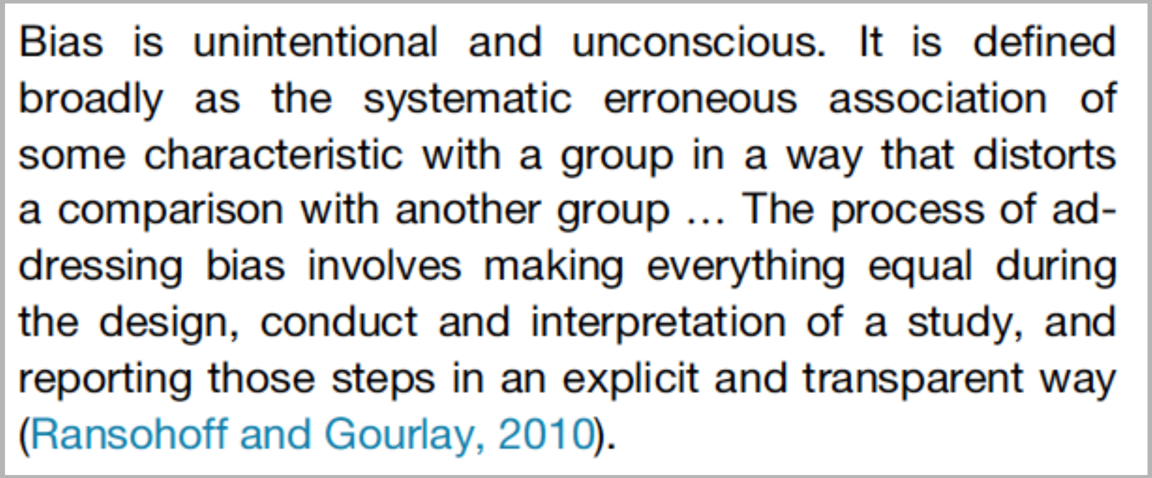
Steward and Balice-Gordon is the Syllabus at Mount Sinai

The culture of the lab: Are best practices discussed and put at a high priority?

1) Is the experimental design from start to finish laid out prospectively?

2) Is every detail sufficiently noted to allow for replication in and outside the lab?

3) Is bias minimized through blinding, recoding, and systematic random sampling?



Bias is unintentional and unconscious. It is defined broadly as the systematic erroneous association of some characteristic with a group in a way that distorts a comparison with another group … The process of addressing bias involves making everything equal during the design, conduct and interpretation of a study, and reporting those steps in an explicit and transparent way (Ransohoff and Gourlay, 2010).

4) Is the guiding philosophy to “test” an hypothesis or “prove” an hypothesis?

NIH will announce new guidelines

NOT-OD-15-103

- The National Institutes of Health (NIH) Office of Extramural Research (OER) plans to clarify and revise application instructions and review criteria to enhance reproducibility of research findings through increased scientific rigor and transparency. These updates, pending approval by the White House Office of Management and Budget (OMB), will take effect for applications submitted for the January 25, 2016, due date and beyond.
-
- file:///replication%20and%20rigor/NOT-OD-15-103_%20Enhancing%20Reproducibility%20through%20Rigor%20and%20Transparency.html#sthash.dBfVN2VM.dpuf