Research: increasing value, reducing waste 2

Increasing value and reducing waste in research design, conduct, and analysis

John P A Ioannidis, Sander Greenland, Mark A Hlatky, Muin J Khoury, Malcolm R Macleod, David Moher, Kenneth F Schulz, Robert Tibshirani

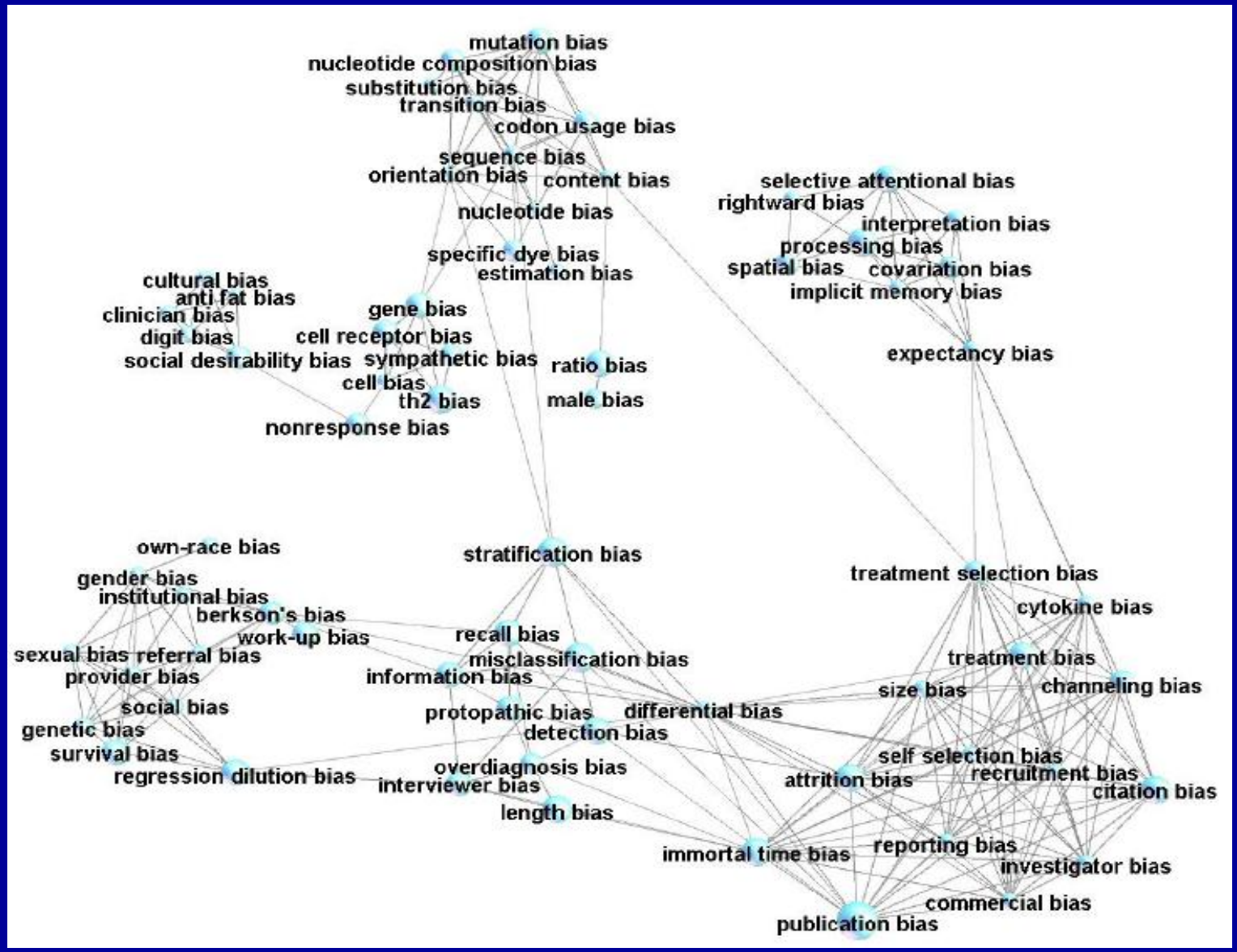
Main issues

- Effect-size ratio
- Development of protocols and improvement of designs
- Research workforce and stakeholders
- Reproducibility practices and reward systems

Effect-size ratio

- Many effects of interest are relatively small.
- Small effects are difficult to distinguish from biases.
- There are just too many biases (see next slide on mapping 235 biomedical biases).
- Design choices can affect both the signal and the noise.
- Design features can impact on the magnitude of effect estimates.
- In randomized trials, allocation concealment, blinding, and mode of randomization may influence effect estimates, especially for subjective outcomes.
- In case-control designs, the spectrum of disease may influence estimates of diagnostic accuracy; and choice of population (derived from randomized or observational datasets) can influence estimates of predictive discrimination.
- Design features are often very suboptimal, in both human and animal studies (see slide on animal studies).

Mapping 235 biases in 17 million Pub Med papers



Chavalarias and Ioannidis, JCE 2010

Very large effects are extremely uncommon

ORIGINAL CONTRIBUTION

Empirical Evaluation of Very Large Treatment Effects of Medical Interventions

Tiago V. Pereira, PhD
Ralph I. Horwitz, MD
John P. A. Ioannidis, MD, DSc

Context Most medical interventions have modest effects, but occasionally some clinical trials may find very large effects for benefits or harms.

Objective To evaluate the frequency and features of very large effects in medicine.

Data Sources Cochrane Database of Systematic Reviews (CDSR, 2010, issue 7).

Study Selection We separated all binary-outcome CDSR forest plots with comparisons of interventions according to whether the first published trial, a subsequent trial (not the first), or no trial had a nominally statistically significant ($P<.05$) very large effect (odds ratio [OR], ≥5). We also sampled randomly 250 topics from each group for further in-depth evaluation.

Data Extraction We assessed the types of treatments and outcomes in trials with very large effects, examined how often large-effect trials were followed up by other trials on the same topic, and how these effects compared against the effects of the respective meta-analyses.

Results Among 85 002 forest plots (from 3082 reviews), 8239 (9.7%) had a significant very large effect in the first published trial, 5158 (6.1%) only after the first published trial, and 71 605 (84.2%) had no trials with significant very large effects. Nominally significant very large effects typically appeared in small trials with median number of events: 18 in first trials and 15 in subsequent trials. Topics with very large effects were less likely than other topics to address mortality (3.6% in first trials, 3.2% in subsequent trials, and 11.6% in no trials with significant very large effects) and were more likely to address laboratory-defined efficacy (10% in first trials,10.8% in subsequent, and 3.2% in no trials with significant very large effects). First trials with very large effects were as likely as trials with no very large effects to have subsequent published trials. Ninety percent and 98% of the very large effects observed in first and subsequently published trials, respectively, became smaller in meta-analyses that included other trials; the median odds ratio decreased from 11.88 to 4.20 for first trials, and from 10.02 to 2.60 for subsequent trials. For 46 of the 500 selected topics (9.2%; first and subsequent trials) with a very large-effect trial, the meta-analysis maintained very large effects with $P<.001$ when additional trials were included, but none pertained to mortality-related outcomes. Across the whole CDSR, there was only 1 intervention with large beneficial effects on mortality, $P<.001$, and no major concerns about the quality of the evidence (for a trial on extracorporeal oxygenation for severe respiratory failure in newborns).

Conclusions Most large treatment effects emerge from small studies, and when additional trials are performed, the effect sizes become typically much smaller. Well-validated large effects are uncommon and pertain to nonfatal outcomes.

JAMA. 2012;308(16):1676-1684 www.jama.com

MOST EFFECTIVE INTERVENTIONS in health care confer modest, incremental benefits.[1,2] Randomized trials, the gold standard to evaluate medical interventions, are ideally conducted under the principle of equipoise[3]: the compared groups are not perceived to have a clear advantage; thus, very large treatment effects are usually not anticipated. However, very large treatment effects are observed occasionally in some trials. These effects may include both anticipated and unexpected treatment benefits, or they may involve harms.

Large effects are important to document reliably because in a relative scale they represent potentially the cases in which interventions can have the most impressive effect on health outcomes and because they are more likely to be adopted rapidly and with less evidence. Consequently, it is important to know whether, when observed, very large effects are reliable and in what sort of experimental outcomes they are commonly observed. The importance of very large effects has drawn attention mostly in observational studies[4,5] but has not been well studied in randomized evidence. It is unknown how often very large effects are replicated in

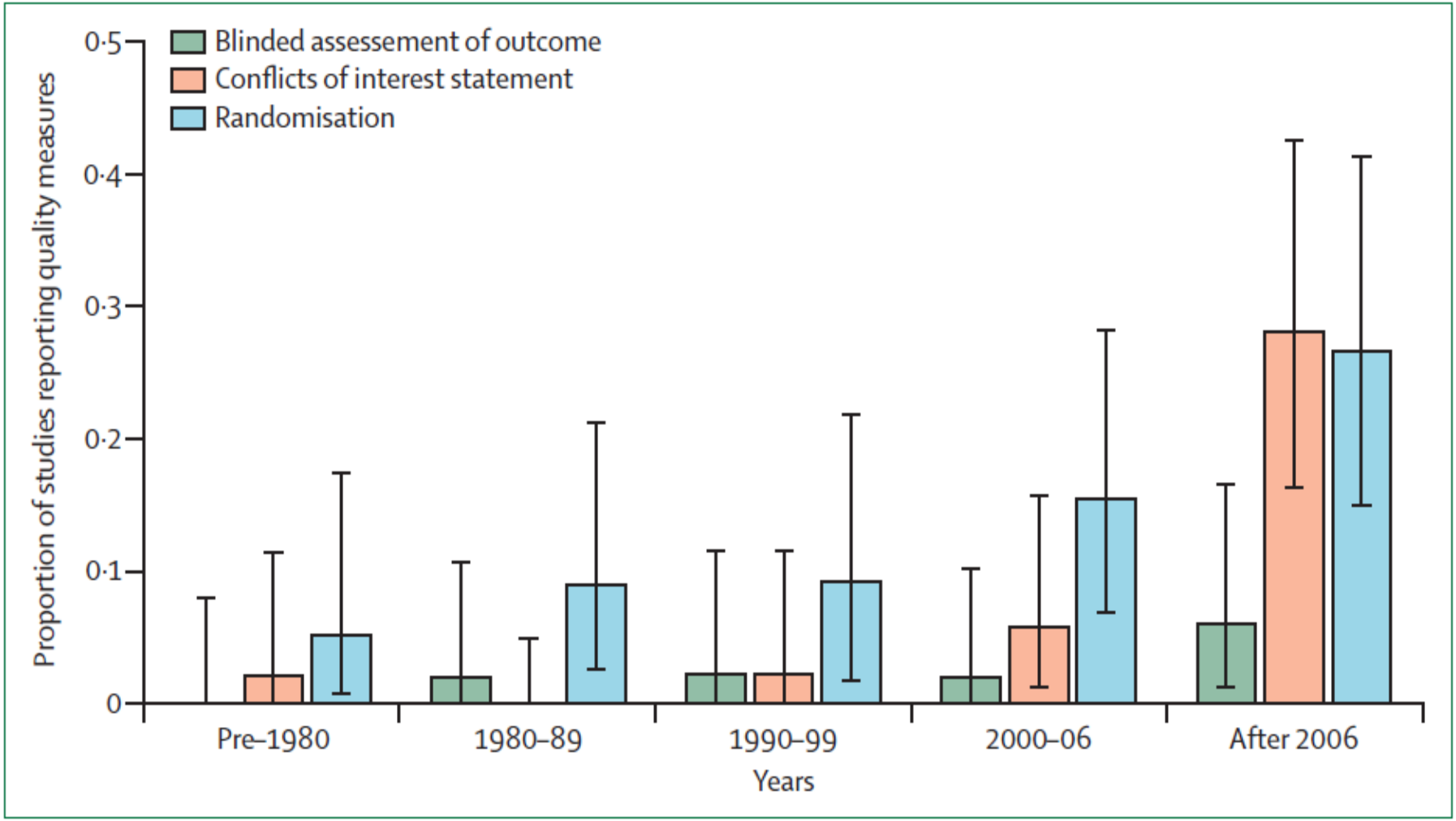


Figure: **Trends in three methodological quality indicators for reports of in-vivo studies**
We randomly sampled 2000 records from PubMed (published 1960–2012) on the basis of their PubMed ID (see appendix for details and the study dataset). 254 reports described in-vivo, ex-vivo, or in-vitro experiments involving non-human animals. Two investigators independently judged whether blinded assessment of outcome, randomisation, or a conflicts of interest statement were included. The proportion reports including this information is

Effect-size ratio – options for improvement

- Design research to either involve larger effects and/or diminish biases.
- In the former case, the effect may not be generalizable.
- Anticipating the magnitude of the effect-to-bias ratio is needed to decide whether the proposed research is justified.
- The minimum acceptable effect-to-bias ratio may vary in different types of designs and research fields.
- Criteria may rank the credibility of the effects by considering what biases might exist and how they may have been handled (e.g GRADE).
- Improving the conduct of studies, not just reporting, to maximize the effect-to-bias ratio. Journals may consider setting minimal design prerequisites for accepting papers.
- Funding agencies can also set minimal standards to reduce the effect-to-bias threshold to acceptable levels.

Developing protocols and improving designs

- Poor protocols and documentation
- Poor utility of information
- Statistical power and outcome misconceptions
- Lack of consideration of other evidence
- Subjective, non-standardized definitions and ‘vibration of effects’

Options for improvement

- Public availability/registration of protocols or complete documentation of exploratory process
- A priori examination of the utility of information: power, precision, value of information, plans for future use, heterogeneity considerations
- Avoid statistical power and outcome misconceptions
- Consideration of both prior and ongoing evidence
- Standardization of measurements, definitions and analyses, whenever feasible

Research workforce and stakeholders

- Statisticians and methodologists: only sporadically involved in design, poor statistics in much of research
- Clinical researchers: often have poor training in research design and analysis
- Laboratory scientists: perhaps even less well equipped in methodological skills.
- Conflicted stakeholders (academic clinicians or laboratory scientists, or corporate scientists with declared or undeclared financial or other conflicts of interest, ghost authorship by industry)

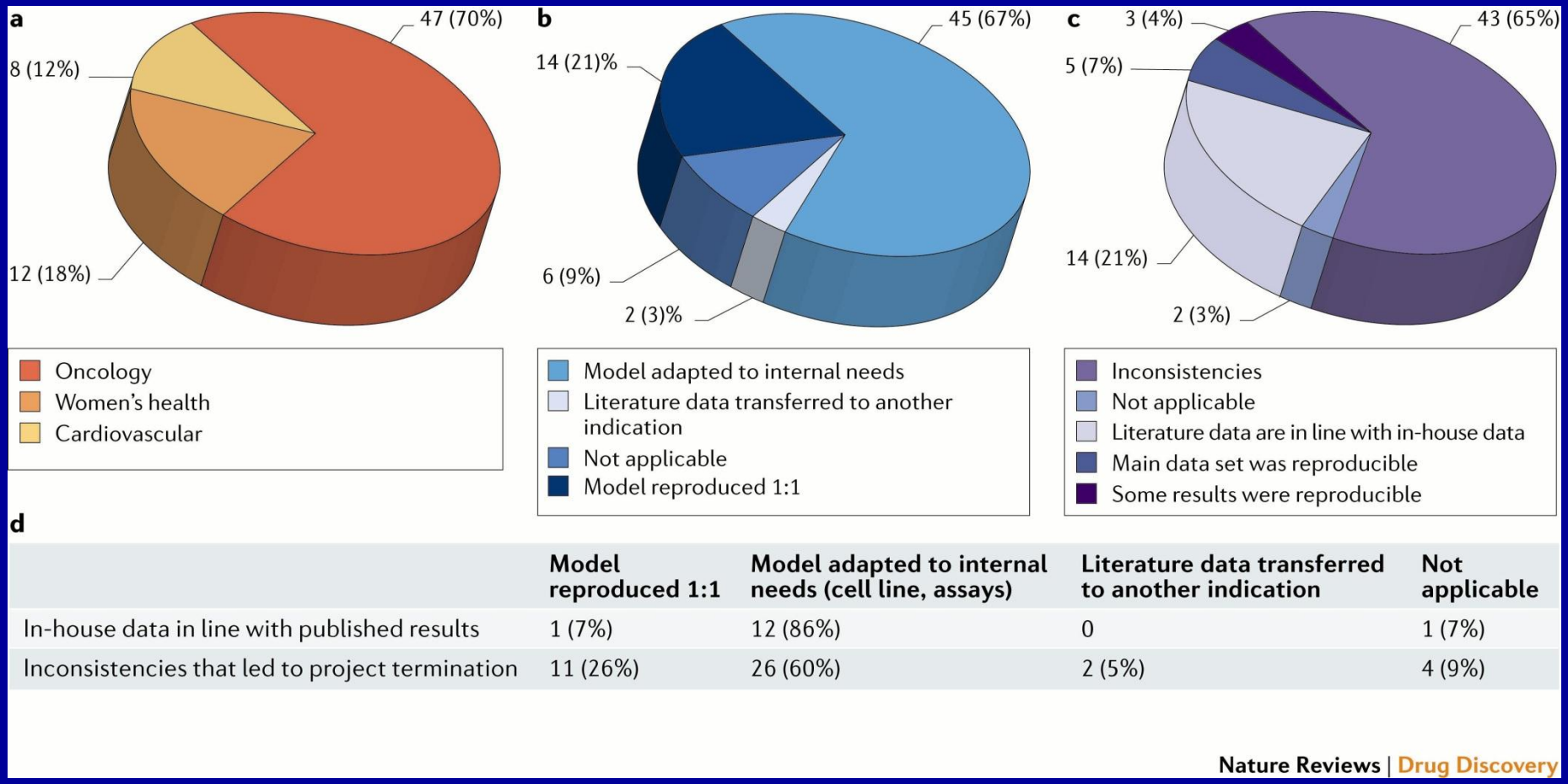
Options for improvement

- Research workforce: more methodologists should be involved in all stages of research; enhance communication of investigators with methodologists.
- Enhance training of clinicians and scientists in quantitative research methods and biases; opportunities may exist in medical school curricula, and licensing examinations
- Reconsider expectations for continuing professional development, reflective practice and validation of investigative skills; continuing methodological education.
- Conflicts: involve stakeholders without financial conflicts in choosing design options; consider patient involvement

Reproducibility practices and reward systems

- Usually credit is given to the person who first claims a new discovery, rather than replicators who assess its scientific validity.
- Empirically, it is often impossible to repeat published results by independent scientists (see next 2 slides).
- Original data are difficult or impossible to obtain or analyze.
- Reward mechanisms focus on the statistical significance and newsworthiness of results rather than study quality and reproducibility.
- Promotion committees misplace emphasis on quantity over quality.
- With thousands of biomedical journals in the world, virtually any manuscript can get published.
- Researchers are tempted to promise and publish exaggerated results to continue getting funded for “innovative” work.
- Researchers face few negative consequences result from publishing flawed or incorrect results or for making exaggerated claims.

A pleasant surprise: the industry championing replication



d

	Model reproduced 1:1	Model adapted to internal needs (cell line, assays)	Literature data transferred to another indication	Not applicable
In-house data in line with published results	1 (7%)	12 (86%)	0	1 (7%)
Inconsistencies that led to project termination	11 (26%)	26 (60%)	2 (5%)	4 (9%)

Nature Reviews | Drug Discovery

Prinz et al., Nature Reviews Drug Discovery 2011

Repeatability

nature genetics

ANALYSIS

Repeatability of published microarray gene expression analyses

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

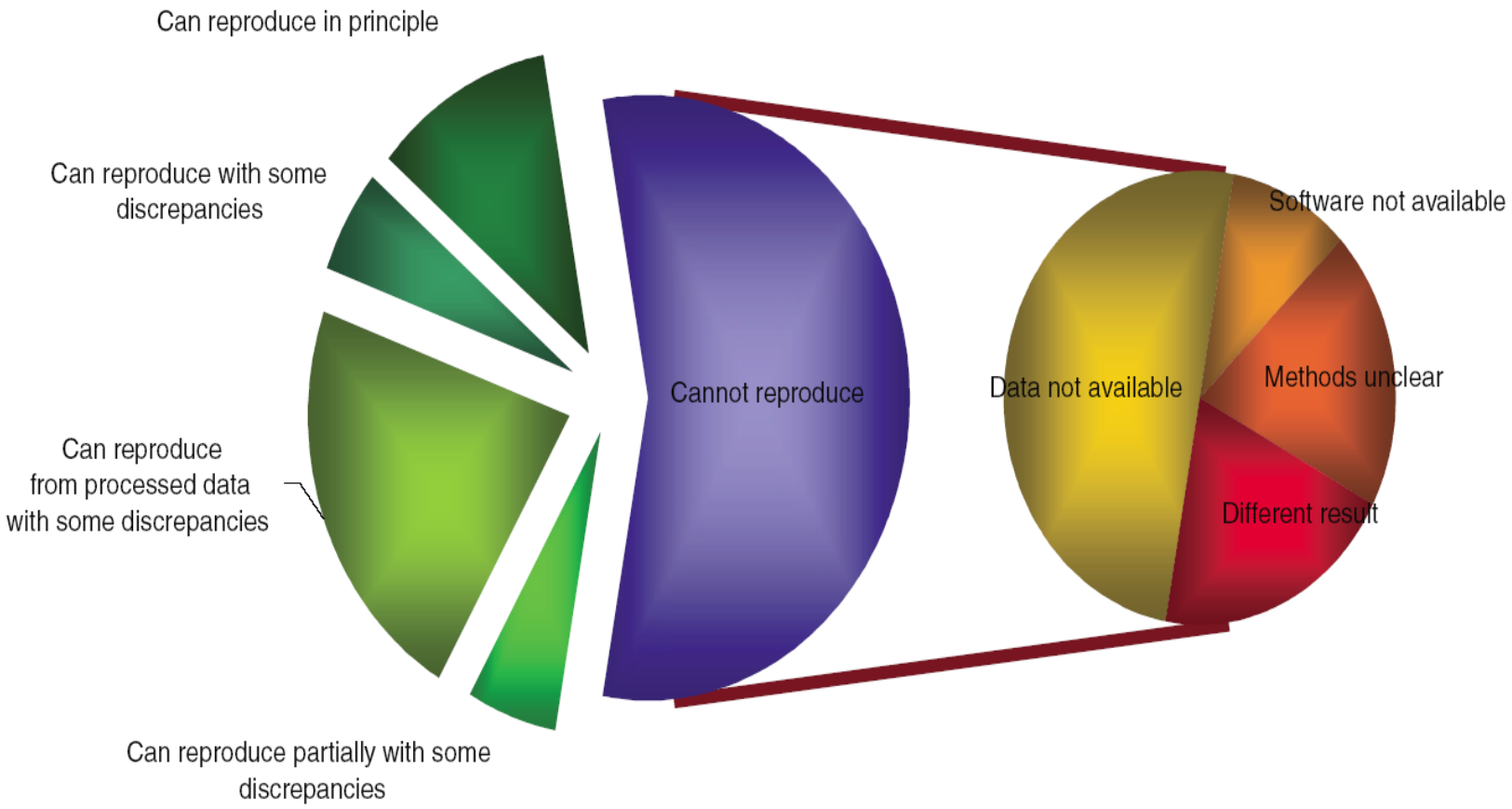


Figure 1 Summary of the efforts to replicate the published analyses.

Options for improvement

- Support and reward (at funding and/or publication level) quality, transparency, data sharing, reproducibility
- Encouragement and publication of reproducibility checks
- Adoption of software systems that encourage accuracy and reproducibility of scripts.
- Public availability of raw data
- Improved scientometric indices; reproducibility indices.
- Post-publication peer-review, ratings and comments

PERSPECTIVE

Improving Validation Practices in "Omics" Research

John P. A. Ioannidis[1] and Muin J. Khoury[2]*

"Omics" research poses acute challenges regarding how to enhance validation practices and eventually the utility of this rich information. Several strategies may be useful, including routine replication, public data and protocol availability, funding incentives, reproducibility rewards or penalties, and targeted repeatability checks.

The exponential growth of the "omics" fields (genomics, transcriptomics, proteomics, metabolomics, and others) fuels expectations for a new era of personalized medicine. ation of the predictive value in real-practice populations, whereas clinical utility requires evaluation of the balance of benefits and harms associated with the adoption of these technologies issues (*8*, *9*), because information on available classifiers constantly changes and new classifiers are proposed. There is at least one recent unfortunate example, where gene signatures were moved into clinical trial experimentation with insufficient previous validation. Three trials of gene signatures to predict outcomes of chemotherapy in treating non–small-cell lung cancer and breast cancer were suspended in 2011 after the realization that their supporting published evidence was nonreproducible (*10*).

Many scientists now demand reproducible omics research (*11*). This requires access to the full data, protocols, and analysis codes for published studies so that other scientists can repeat analyses and verify results. Fortunately, several public data repositories exist, such as the Gene Expression Omnibus, ArrayExpress, and the Stanford Microarray Database. There have also

Science, December 2, 2011

Levels of registration

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

Recommendations and monitoring

1 Make publicly available the full protocols, analysis plans or sequence of analytical choices, and raw data for all designed and undertaken biomedical research

- Monitoring—proportion of reported studies with publicly available (ideally preregistered) protocol and analysis plans, and proportion with raw data and analytical algorithms publicly available within 6 months after publication of a study report

2 Maximise the effect-to-bias ratio in research through defensible design and conduct standards, a well trained methodological research workforce, continuing professional development, and involvement of non-conflicted stakeholders

- Monitoring—proportion of publications without conflicts of interest, as attested by declaration statements and then checked by reviewers; the proportion of publications with involvement of scientists who are methodologically well qualified is also important, but difficult to document

3 Reward (with funding, and academic or other recognition) reproducibility practices and reproducible research, and enable an efficient culture for replication of research

- Monitoring—proportion of research studies undergoing rigorous independent replication and reproducibility checks, and proportion replicated and reproduced

Tailored recommendations per field: e.g. animal research

***Panel 2*: Ten options to improve the quality of animal research**

Protocols and optimum design

1 Creation of a publicly accessible date-stamped protocol preceding data collection and analysis, or clear documentation that research was entirely exploratory
2 Use of realistic sample size calculations
3 Focus on relevance, not only statistical efficiency

Effect-to-bias ratio

4 Random assignment of groups
5 Incorporation of blind observers
6 Incorporation of heterogeneity into the design, whenever appropriate, to enhance generalisability
7 Increase in multicentre studies
8 Publishers should adopt and implement the ARRIVE (Animal Research: Reporting In Vivo Experiments) guidelines

Workforce and stakeholders

9 Programmes for continuing professional development for researchers

Reproducibility and reward systems

10 Funders should increase attention towards quality and enforce public availability of raw data and analyses