Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

Lancet. Author manuscript; available in PMC 2016 January 01.

Published in final edited form as:

Lancet. 2014 January 11; 383(9912): 166–175. doi:10.1016/S0140-6736(13)62227-8.

Research: increasing value, reducing waste 2:

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

Prof. John P A Ioannidis, MD, **Prof. Sander Greenland, DrPH**, **Prof. Mark A Hlatky, MD**, **Muin J Khoury, MD**, **Prof. Malcolm R Macleod, PhD**, **Prof. David Moher, PhD**, **Prof. Kenneth F Schulz, PhD**, and **Prof. Robert Tibshirani, PhD**

Stanford Prevention Research Center (Prof J P A Ioannidis MD), and Division of Cardiovascular Medicine (Prof M A Hlatky MD), Department of Medicine, School of Medicine, Stanford University, Stanford, CA, USA; Division of Epidemiology, School of Medicine, Stanford University, Stanford, CA, USA (Prof J P A Ioannidis); Division of Health Services Research (Prof M A Hlatky) and Department of Health Research and Policy (Prof R Tibshirani PhD), Stanford University, Stanford, CA, USA; Department of Statistics, School of Humanities and Sciences, Stanford University, Stanford, CA, USA (Prof J P A Ioannidis, Prof R Tibshirani); Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, CA, USA (Prof J P A Ioannidis); Department of Epidemiology and Department of Statistics, UCLA School of Public Health, Los Angeles, CA, USA (Prof S Greenland DrPH); Office of Public Health Genomics, Centers for Disease Control and Prevention, Atlanta, GA, USA (M J Khoury MD); Epidemiology and Genomics Research Program, National Cancer Institute, Rockville, MD, USA (M J Khoury); Department of Clinical Neurosciences, University of Edinburgh School of Medicine, Edinburgh, UK (Prof M R Macleod PhD); Clinical Epidemiology Program, Ottawa Hospital Research Institute (Prof D Moher PhD), and Department of Epidemiology and Community Medicine, Faculty of Medicine (Prof D Moher), University of Ottawa, Ottawa, ON, Canada; FHI 360, Durham, NC, USA (Prof K F Schulz PhD); and Department of Obstetrics and Gynecology, University of North Carolina School of Medicine, Chapel Hill, NC, USA (Prof K F Schulz)

Abstract

Correctable weaknesses in the design, conduct, and analysis of biomedical and public health research studies can produce misleading results and waste valuable resources. Small effects can be difficult to distinguish from bias introduced by study design and analyses. An absence of detailed written protocols and poor documentation of research is common. Information obtained might not be useful or important, and statistical precision or power is often too low or used in a misleading way. Insufficient consideration might be given to both previous and continuing studies. Arbitrary choice of analyses and an overemphasis on random extremes might affect the reported findings.

Correspondence to: Prof John P A Ioannidis, Stanford Prevention Research Center, Stanford, CA 94350, USA, jioannid@stanford.edu.

See Online for appendix

Contributors

All authors participated in the development of the report, including conception, provision of data and references, writing of the manuscript, revision of the draft, and approval of the final version. JPAI wrote the first complete draft, which was improved and revised by all other authors.

Conflicts of interest

We declare that we have no conflicts of interest.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Several problems relate to the research workforce, including failure to involve experienced statisticians and methodologists, failure to train clinical researchers and laboratory scientists in research methods and design, and the involvement of stakeholders with conflicts of interest. Inadequate emphasis is placed on recording of research decisions and on reproducibility of research. Finally, reward systems incentivise quantity more than quality, and novelty more than reliability. We propose potential solutions for these problems, including improvements in protocols and documentation, consideration of evidence from studies in progress, standardisation of research efforts, optimisation and training of an experienced and non-conflicted scientific workforce, and reconsideration of scientific reward systems.

Introduction

Design, conduct, and analysis of biomedical and public health research form an interdependent continuum. Some specialties have more efficient mechanisms than others to optimise the design, conduct, and analysis of studies, providing the opportunity for different specialties to learn from successful approaches and avoid common pitfalls. The rapid introduction of new biological measurement methods involving genomes, gene products, biomarkers, and their interactions has promoted novel and complex analysis methods that are incompletely understood by many researchers and might have their own weaknesses. Additionally, biomedical and public health research increasingly interacts with many disciplines, using methods and collaborating with scientists from other sciences, such as economics, operational research, behavioural sciences, and informatics,[1] heightening the need for careful study design, conduct, and analysis.

Recommendations

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

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

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

These issues are often related to misuse of statistical methods, which is accentuated by inadequate training in methods. For example, a study[2] of reports published in 2001 showed that p values did not correspond to the given test statistics in 38% of articles published in *Nature* and 25% in the *British Medical Journal*. Prevalent conflicts of interest can also affect the design, analysis, and interpretation of results. Problems in study design go beyond statistical analysis, and are shown by the poor reproducibility of research. Researchers at Bayer[3] could not replicate 43 of 67 oncological and cardiovascular findings reported in academic publications. Researchers at Amgen could not reproduce 47 of 53 landmark oncological findings for potential drug targets.[4] The scientific reward system places insufficient emphasis on investigators doing rigorous studies and obtaining reproducible results.

Problems related to research methodology are intricately linked to the training and composition of the scientific workforce, to the scientific environment, and to the reward system. We discuss the problems and suggest potential solutions from all these perspectives. We provide examples from randomised trials, traditional epidemiology studies, systematic reviews, genetic and molecular epidemiology studies, so-called omics, and animal studies. Further reading for each section is provided in the appendix.

Effect-to-bias ratio

The problem

In research, many effects of interest are fairly small, including those seen in clinical trials and meta-analyses,[5] biomarker studies,[6] traditional[7–10] and genome[11] epidemiology studies, and omics.[12] Small effects are difficult to distinguish from biases (information, selection, confounding, etc).[8,13] When effects and biases are potentially of similar magnitude, the validity of any signal is questionable. Design choices can increase the signal, decrease the noise, or both. For example, investigators might enhance the signal in a clinical trial by inclusion of only high-risk groups,[14] but this design choice could reduce the generalisability of the study results. Sometimes, issues might differ for signals of effectiveness versus signals of adverse events.[14] Many biases might inflate estimates of effectiveness, but underestimate adverse effects, especially when financial conflicts of interest exist.

Several meta-epidemiological studies have shown that design features can affect the magnitude of effect estimates. For randomised trials, allocation concealment, blinding, and method of randomisation might modify effect estimates, especially for subjective outcomes.[15] For case-control study designs, the range of disease can affect estimates of diagnostic accuracy,[16,17] and choice of population (derived from randomised or observational datasets) can affect estimates of predictive discrimination for biomarkers.[18] Modelling is open to a wide range of subtle biases in model specification.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Options for improvement

For exposures or interventions for which the existence of an effect is unknown or controversial, the effect-to-bias ratio might be improved by research involving large effects and by reduction of biases. For research involving large effects, investigators should acknowledge that the effect is being documented in favourable circumstances. If an effect is documented in design conditions that have been chosen to inflate the effect size, then generalisation to other settings or unselected populations should take account of this selection. Anticipation, acknowledgment, and estimation of the magnitude of the effect-to-bias ratio are all needed to decide whether undertaking of the proposed research is even justifiable. The minimum acceptable effect-to-bias ratio can vary in different types of designs and research specialties. Research efforts in domains in which the effect-to-bias ratio is low might be futile and need to await reductions in biases. For example, results from tens of thousands of candidate gene studies yielded little reliable information at a time when genotyping errors, population stratification, selective reporting, and other biases were extremely large compared with genetic effects.[19]

In several specialties, criteria are being developed that try to rank the credibility of effects on the basis of what biases might exist and how they might have been handled. Examples are GRADE (Grading of Recommendations Assessment, Development and Evaluation) for clinical evidence;[20] the American Heart Association criteria for novel biomarkers;[21] and the Venice criteria for genetic associations,[22] and their extrapolation to gene–environment interactions.[23]

There is a pressing need to improve the quality of research to minimise biases. Some of the benefits might occur indirectly, through pressure to improve reporting (see Chan and colleagues[24] in this Series). However, additional efforts should focus directly on improvements in conduct of studies to maximise the effect-to-bias ratio. Journals should consider setting some design prerequisites for particular types of studies before they accept reports for publication. This requirement goes beyond simply asking for transparency from investigators in reporting of what was done. Examples include the MIAME (Minimum Information About a Microarray Experiment) guidelines for microarray experiments[25] and similar guidelines for other types of experiments.[26,27] Some of the reporting guidelines for animal studies have suggestions for improved study design and conduct.[28,29]

Finally, funders could help to promote high-quality research. Many expert panels note the large number of excellent applications that they receive, but this perception is not consistent with the quality of reported research (figure). Members of funding panels, often drawn from the research community itself, might be reluctant to impose a high quality threshold that could disadvantage many investigators. The scientific and administrative leadership of funding agencies could clarify the great importance that they attach to study quality and the minimum standards that they require to reduce the effect-to-bias threshold to an acceptable level.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Development of protocols and improvement of designs

Problem 1: poor protocols and designs

The extent to which research is done on the basis of a rudimentary protocol or no protocol at all is unknown, because even when protocols are written, they are often not publicly available. Consequently, researchers might improvise during the conduct of their studies, and place undue emphasis on chance findings. Although some improvisation is unavoidable because of unanticipated events during a study (eg, an unexpectedly high dropout rate, or unpredicted adverse events), changes in the research plan are often poorly documented[30] and not present in formal data analyses (eg, non-response and refusal data might be neither reported nor used to adjust formal uncertainty measures).

Problem 2: poor utility of information

Studies are often designed without proper consideration of the value or usefulness of the information that they will produce. Although replication of previous research is a core principle of science, at some point, duplicative investigations contribute little additional value. Conversely, one study in a previously understudied domain might supply too little information to be useful, and small, uninformative studies remain common in several specialties.[31–34] In principle, analysis of the expected information content of a study (value of information) might guide judgments about whether it is reasonable to initiate or fund a particular study, but there has been very little experience with this technique.[35]

Problem 3: statistical power and outcome misconceptions

Calculations of power needed to reject the null hypothesis are conventional, but they can mislead because they assume that no problem will occur during the study, no other evidence will be available to inform decision makers, and that the arbitrary α 0·05 strikes the proper balance between false acceptance or rejection of the null hypothesis. These conditions hardly ever exist. Moreover, a study with high power to reject the null hypothesis that fails to reject it at the conventional (5%) α-error-level might still support the alternative hypothesis better than it does the null.[36]

The quest for adequate statistical power might lead researchers to choose outcome measures that are clinically trivial or scientifically irrelevant.[37] For example, trials of cholinesterase inhibitors in Alzheimer's disease have used cognitive function scales that allow detection of small, yet clinically meaningless, changes.[38] Researchers often use composite outcomes in an attempt to boost statistical power, but the components of the composite might not show the same underlying disease process, or might be subject to clinically subjective decisions—eg, the composite of death, myocardial infarction, or repeat revascularisation.[39,40]

In animal studies investigators commonly use less clinically relevant outcome measures than in human trials, which could lead to statistically robust effect sizes, or they use experimental injury models primed to have large all-or-nothing treatment effects. Such statistical optimisation comes at the cost of generalisability, because extrapolation might be required not only across species, but also across doses, sometimes by many orders of magnitude.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Problem 4: insufficient consideration of other evidence

Typically, every study is designed, done, and discussed in isolation[41] (see Chalmers and colleagues[42] in this Series). Moreover, most research designs do not take account of similar studies being done at the same time.[43] The total sample size of clinical trials that are in progress might exceed the total sample size of all completed trials.[44] The launch of yet another trial might be unnecessary. The need for replication of biomedical research should be balanced with the avoidance of mere repetition.

Problem 5: subjective, non-standardised definitions and vibration of effects

Many definitions and most analyses involve subjective judgments that leave much room for the so-called vibration of effects during statistical analysis.[43] Vibration of effects means that results can differ (they vibrate over a wide possible range), dependent on how the analysis is done. This effect occurs when many variations can be used in analyses—eg, many variables to include in, or exclude from, statistical adjustments; different statistical models to be used; different definitions of outcomes and predictors; and use of different inclusion and exclusion criteria for the study population. Many combinations of these analysis options can be used, and the results can vary, depending on which are chosen. This variance can lead to bias if only a few chosen analyses are reported, especially if the investigators have a preference for a particular result or are influenced by optimism bias.[45]

Systematic reviews of previous data are an interesting challenge in this regard, because they are done retrospectively, and investigators might have some knowledge of the data, even as they design the review. Conflicted investigators (eg, those with professional conflicts or industry support) could design the study protocol in a way that would favour the outcomes that they want to obtain.[46] Industry-supported systematic reviews obtain favourable results more often than do other systematic reviews, although the difference lies more in the interpretation of the results rather than in the actual numbers.[47,48]

Options for improvement: protocols and documentation

Clinical trials and other studies that are not exploratory research should be done in accordance with detailed written plans, with advance public deposition of the protocol.[49–51] Use of a strict preconceived protocol might not be feasible for some exploratory research, but nonetheless investigators should rigorously document the sequence of decisions and findings made in the course of the study, and reasons for those decisions. Even in randomised trials and other extensively designed studies, some post-hoc decisions might need to be made. Reports of such studies should distinguish clearly between prespecified analyses and post-hoc explorations of the data. Systematic reviews with written protocols detailing prespecified steps can now be registered prospectively.[52,53] Protocol registration will not avoid the need for unanticipated deviations from the protocol, but would make deviations more visible and open to public judgment (panel 1).

Registration of clinical trials became widespread only when it became a prerequisite for publication in most major journals. Similarly, protocol or dataset registration and deposition is likely to become widely adopted only with similar incentives—eg, if a prerequisite for funding and publication of research reports. For some types of studies, especially those

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

involving microarray and macromolecular data, public deposition of protocols, data, and analyses have already become stated prerequisites for most major journals, but these practices are inadequately en forced.[54] Another option is to encourage or require full external peer review and publication of protocols in journals. Funding agencies or institutional review boards peer review some research protocols, but many are not reviewed. Public review might enhance the relevance and quality of these investigations, although empirical evidence is needed. Periodic comparisons of study protocols with published results[30,55] could provide useful feedback to investigators, journals, and funding agencies.

For important epidemiological research that must be done with use of highly exploratory analyses (eg, routine database screening for identification of adverse events), documentation of how studies were done, including decisions made along the way, is essential to provide transparency. Information about key study events, such as refusal rates and dropouts, should be incorporated into data analyses and reporting. Methods for analysis of missing data show promise for incorporation of these and other uncertainty sources,[56,57] although prevention of missing data is, by far, the most preferable solution. No methods to address missing data can definitively eliminate or adjust for potential selection bias in a randomised trial with substantial losses after randomisation when those losses differ substantially between the randomised groups.

Panel 1: Protocols for systematic reviews and their registration

Protocols for systematic reviews, like any other research endeavour, are important. They provide the researchers with an explicit research plan and allow others to find possible discrepancies between the final review publication and its protocol. In a study of 213 systematic reviews indexed in PubMed in November, 2004, examining therapeutic effectiveness, investigators of almost all the Cochrane reviews reported use of a protocol (122 (98%) of 125), whereas only some investigators of non-Cochrane reviews did so (ten (11%) of 88). Similar findings have been reported elsewhere. Although some of the researchers who did not report use of a protocol might have used one, it is unlikely that all of them did so.

To help to overcome reporting biases and other problems, such as unnecessary duplication, and improve transparency, an international register for systematic reviews was launched in February, 2011, called PROSPERO.[53] PROSPERO is an international database of prospectively registered systematic reviews. Key features from the review protocol are recorded and maintained as a permanent record. The register is internet-based, free to search, and open for free registration to anyone doing a systematic review for which there is a health-related outcome. The aim of the register is to provide transparency in the review process, help to reduce unplanned duplication of reviews, and enable comparison of reported review findings with what was planned in the protocol. This register might serve to discourage bias in the conduct and reporting of systematic reviews.

As of May, 2013, investigators had registered more than 1000 systematic review protocols from 27 countries. The National Institutes for Health Research, UK, have mandated the registration of systematic reviews that they fund. The Canadian Institutes

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

of Health Research are working on a similar policy initiative. *Systematic Reviews*, an open-access Medline-indexed journal, publishes systematic review protocols. Since launch in February, 2012, the journal has published 89 protocols (as of November, 2013). These initiatives will also enable researchers to periodically assess the association between review protocols and final publications.

For preclinical laboratory and animal studies, pre-specified protocols that are publicly deposited might also be desirable. Researchers who do this research have little experience of using protocols, and feasibility needs to be probed. Date-stamped study protocols—including a statement of purpose or the hypotheses to be tested, power calculations, methods for data collection, and a statistical analysis plan—could be made available to journal reviewers on request.

Options for improvement: use of information, and precision, power, and outcomes

Whenever appropriate, proposals for study funding should include realistic calculations of power or expected precision based on clinically relevant outcomes. Investigators designing clinical trials should consider pragmatic designs[58,59] and patient-centred outcomes,[60,61] which would be important to the end-users of the research (see also Chalmers and colleagues[42] in this Series). When feasible, investigators might consider the value of information anticipated from the study in view of its anticipated cost.

Biobanks and clinical data registries are constructed to have several uses; some uses might be predicted and others will arise with emergence of new technologies and new questions of interest. Nevertheless, some rational design calculations (power or sample size, or precision) should be done, at least those based on uses foreseeable at the time the study was designed.[62] Protocols should be written prospectively for studies on the basis of such data repositories. Translational research for new technologies (eg, the development of diagnostic tests from genomics) might benefit from value of information analysis because of rapidly evolving information and the inability to do randomised clinical trials in many cases. Formal methods of modelling have been suggested.[63]

For animal studies, effects of realistic treatment doses might be small, and therefore appropriately powered studies will have to be large. To increase the generalisability of findings, investigators should plan for heterogeneity in the circumstances of testing.[64,65] For these large studies to be feasible, consideration should be given to the development of multicentre animal studies.[66,67]

Options for improvement: consideration of evidence

Researchers should anticipate evidence from continuing research when they design new studies. For example, investigators designing new randomised trials should consider previous trials and trials that are in progress to identify the most important remaining questions and comparisons.[42,68] This awareness of all evidence[44] might improve efficiency in use of resources and provide more informative results. Several specialties have been transformed by large-scale collaborative consortia of investigators working on the same subject, particularly human genome epidemiology.[69,70] Large and inclusive consortia can

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

have a more comprehensive view of what is already available or underway in the specialty through the enhancement of communication between investigators. New and interesting ideas can be proposed by individual investigators and then tested efficiently at the consortium level. Consortia have been particularly successful in situations in which there is consensus that maximisation of sample size is paramount, such as in genome-wide association studies.

Options for improvement: standardisation of efforts

Full standardisation of definitions and analytical procedures could be feasible for new research efforts. For existing datasets and studies, harmonisation attempts to achieve some, but not necessarily perfect, homogeneity of definitions might need substantial effort and coordination.[71] Large consortia and collaborations can allow the use of a common language among investigators for clinical definitions, laboratory measurements, and statistical analyses. Some specialties have improved and standardised their operating outcome definitions through international collaboration, both for clinical research (eg, OMERACT [Outcome Measures in Rheumatology] in rheumatic diseases) and for animal studies (eg, the European Mouse Phenotyping Resource of Standardised Screens). Different specialties face different challenges in the standardisation of analysis practices. For example, randomised clinical trials have a very long standing history and some methods are widely accepted as the standard—eg, the intention-to-treat principle, or use of Kaplan-Meier plots with the log-rank test. Deviations from these standards usually have to be explained. Conversely, in other specialties, many alternative methods exist, none with strong enough rationale to be preferred. Irrespective of which method is used, all results should be presented, not only the most interesting results (see Chan and colleagues[24] in this Series).

Research workforce and stakeholders

Problems

Statistical methods can be complex, and continue to evolve in many specialties, particularly novel ones such as omics. However, statisticians and methodologists are only sporadically involved, often leading to flawed designs and analyses.[72] Much flawed and irreproducible work has been published, even when only simple statistical tests are involved. Investigators of one study[73] examined the use of Fisher's exact test in 71 articles from six major medical journals. When a statistician was a member of the team, the test was used more appropriately than when one was not. Data that are multidimensional (ie, contain many features) are particularly at risk of false positives and overfitting, particularly when analysed by inexperienced or untrained analysts. Problems with statistical analysis might not be identified in peer review, especially when the report is not assessed by a statistician or methodologist.

Many biomedical researchers have poor training in research design and analysis. Although physicians must pass rigorous examinations to practise medicine, they can practise medical research with nearly no training. In many countries, physician investigators have a short introduction to biostatistics early in medical school, and receive no formal training in clinical research thereafter. The little training that they receive often focuses on data

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

analysis, and rarely includes study design, which is arguably the most crucial element in research methods.[74]

Little evidence exists about the research training of laboratory scientists. The way that many laboratory studies are reported suggests that scientists are unaware that their methodological approach is without rigour (figure). Many laboratory scientists have insufficient training in statistical methods and study design. This issue might be a more important deficiency than is poor training in clinical researchers, especially for laboratory investigation done by one scientist in an isolated laboratory—by contrast, many people would examine a clinical study protocol and report.

Research is often done by stakeholders with conflicts of interest that favour specific results. These stakeholders could be academic clinicians, laboratory scientists, or corporate scientists, with declared or undeclared financial or other conflicts of interest. Much clinical research is designed and done under the supervision of the industry, with little or no input from independent researchers. Clinicians might participate in this process simply through the recruitment of study participants, without making any meaningful contribution to the design, analysis, or even writing of the research reports, which might be done by company ghost writers.

Options for improvement

Statisticians and methodologists should be involved in all stages of research. This recommendation has been repeatedly discussed, mostly for clinical trials, but it applies to all types of studies. Enhancement of communication between methodologists and other health scientists is also important. Medical and public health schools and graduate programmes should enhance training of clinicians and scientists in quantitative research methods, sensitise them to biases and ways to avoid or minimise bias in study design and conduct, and provide them with methods to account or adjust for biases in obtained data. Medical school students are exposed to a large amount of information and focus on what they will be tested on to practise medicine. Even those students who will not become researchers need to understand research design and analysis sufficiently to critically assess research relevant to their clinical practice. Medical licensing examinations could include a reasonable amount of material to test clinical research methods. For young investigators, formal training in clinical research methods, including obtaining of a masters degree, is already an important credential and a key component of some career development awards.

Expectations for continued professional development, reflective practice, and validation of investigative skills should be reconsidered. The medical profession recognises the need for continued medical education, and even revalidation, to ensure the highest quality of clinical practice. Clinical and laboratory researchers might also benefit from an opportunity to update their skills in view of newer methodological developments, perhaps through short courses and novel approaches to continued methodological education.

Suggestions to minimise influence of potential conflicts

The influence of conflicted stakeholders in the design, conduct, and analysis of studies should be diminished through the involvement of stakeholders without financial conflicts in

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

decisions about design options—eg, which treatments should be compared in clinical trials, or which outcomes are relevant. Relevant stakeholders can include patients,[75] who are key end-users of biomedical research, and their co-involvement in prioritisation of research needs better analysis.[42] Even apparently shared research values such as objectivity might mean different things to different people[76,77]—eg, funders and policy makers might need higher thresholds than do researchers, commercial developers of tests and interventions, or even patients.[76]

Reproducibility practices and reward systems

Replication and repeatability

In most research specialties, great credit is given to the person who first claims a new discovery, with few accolades given to those who endeavour to replicate findings to assess their scientific validity. Cross-validation of a single dataset might yield inflated results because of biases.[78] Replication of findings in new samples is often done by the same researcher who made the original claim; this type of replication might be subject to optimism and allegiance biases, might perpetrate the same errors as the original work, and might have low generalisability. External validation by independent teams is essential, yet infrequent in many specialties.

When systematic efforts are made by independent scientists, empirical studies indicate that it is often impossible to repeat published results.[3,4,79,80] Original data might be difficult or impossible to obtain or analyse. Only two of 18 microarray-related research articles published in *Nature Genetics* had their analyses repeated by independent analysts, even though availability of raw data, protocols, and analysis codes was a prerequisite for publication of these studies.[80] Some research groups even undermine the goal of independent replication and data-sharing mandates by forcing replicators to use their proprietary analysis system and use coauthors from the original group to minimise any conflict with previous reports.

Reward mechanisms

Reward mechanisms (eg, prestigious publications, funding, and promotion) often focus on the statistical significance and newsworthiness of results rather than the quality of the design, conduct, analysis, documentation, and reproducibility of a study. Similarly, statistically significant results, prestigious authors or journals, and well connected research groups attract more citations than do studies without these factors, creating citation bias.[81,82]

Many appointment and promotion committees function under a misplaced emphasis on number of publications. Although publication of research is essential, use of number of publications as an indicator of scholarly accomplishment stresses quantity rather than quality. With thousands of biomedical journals, nearly any manuscript can get published. Almost 20 years ago, Altman[83] noted that there was a need for less research, better research, and research done for the right reasons. In some environments, accomplishment is judged on the basis of funding rather than publications, but funding is a means, not an end product.[84] Researchers are tempted to promise and publish exaggerated results to continue their funding for what they think of as innovative work. At present, researchers face few negative

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

consequences from publication of flawed or incorrect results or for inclusion of exaggerated claims. Even when errors are detected in published articles, detection is often a long time after publication, and refuted results might be cited for many years after they have been discredited.[85]

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

Suggestions to improve reproducibility and reward systems—Designs could be supported and rewarded (at funding or publication level, or both) that foster careful documentation and allow testing of repeatability and external validation efforts, including datasets being made available to research groups that are independent of the original group.[86–90] It is important to reward scientists on the basis of good quality of research and documentation, and reproducibility of results, rather than statistical significance.[91] With use of online publication space, journals could promote repeatability and replication—eg, *PLOS One* has pledged to publish reproducibility checks done by contracted independent laboratories as part of the reproducibility initiative.[92] So-called statistical shops could adopt software systems that encourage accuracy and reproducibility of their software scripts, such as Sweave. Public availability of raw data and complete scripts of statistical analyses could be required by journals and funding agencies sponsoring new research—eg, as the Institute of Medicine recommended in a report on omics.[93]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Scientific productivity cannot be judged simply by number of publications. Publication of many low-quality articles is worse than is production of none. Scientometrics has led to the development of several rigorous quantitative indices, some of which might allow correction for self-citations, relative effect compared with other scientists in the same specialty, and large amounts of coauthorship. Although sophisticated citation indices are an improvement compared with publication counts, they do not account necessarily for the reproducibility of the work. Post-publication peer review might provide further insights about study quality and reproducibility, but few data exist for the effectiveness of this approach.

The development of electronic publishing could allow for post-publication ratings and comments on scientific work. One author (RT) has helped to create such a system at PubMed, which is called PubMed Commons. It is a new feature built into PubMed, researchers can add a comment to any publication, and read the comments of others. PubMed Commons is a forum for open and constructive criticism and discussion of scientific issues. At present, comments are not anonymous to maintain the quality of the interchange. There is a need to understand better how to quantify scientific reputation.[94] There should be sufficient transparency and scientific reasoning in the process to avoid systematic manipulations of reputation,[95] or other gaming of publication and citation systems, in which authors can artificially inflate their productivity metrics. Some metrics are easier to game (eg, number of publications), whereas others are more difficult (eg, number of publications with >300 citations).

Conclusions and recommendations

We have outlined several problems and solutions to reduce waste in the design, conduct, and analysis of research. Not all these solutions are equally relevant or practicable for all research disciplines, and each specialty might need to prioritise which changes are most crucial. For example, panel 2 lists the ten most important priorities for animal research.

To maximise motivation for change, reductions of waste in research will need behavioural changes, not only from researchers, but also from publishers and regulators. These changes will need external pressure from stakeholders such as funding agencies. Funders are eager to ensure that they get a good return on their investments; inadequate research diminishes the fiscal investment that they have made. Patients and the public also have an important voice.[96] Science is a global, multidisciplinary, loosely organised, and heterogeneous endeavour. Hopefully, funders that insist on high-quality study design, institutions that have clear expectations for studies occurring in their name, and publishers that insist on rigorous and transparent presentation of research studies will, in time, fund and publish research of the highest quality and, thereby, obtain a competitive advantage. The more systematic stakeholders can be in these efforts, the better the quality of the science—in which all of us have a stake—will be.

On the first page of this report, we list some recommendations and metrics that can be used to measure improvements in the future. These improvements are likely to be inter-related, so that advances in one aspect can be expected also to improve other aspects of biomedical research. Additionally, the proposed metrics of success might have some interdependence. The current picture is probably not very favourable for these metrics, leaving much room for

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

improvement. Preregistration of studies and their protocols varies widely for different designs. For clinical trials, registration has been successful overall, but in several specialties, only a few trials might be registered. For example, although 80% in a sample of trials of new oncology drugs reported in 2007–08 had been registered,[97] the respective figure was only 34% randomised trials of physical therapy reported in 2009.[98] Randomised trials constitute less than 5% of all biomedical articles, and most other studies are not registered at present. For example, ClinicalTrials.gov included 26 449 records of registered observational studies as of April 5, 2013 (10 636 were still open). Thus, less than 5% of published studies overall are likely to be pre-registered at present.

Detailed analysis plans and raw data become publicly available even less frequently than do protocols for trials, but registration varies between different specialties. For example, there is a long and successful tradition for registration of measurement protocols and data from microarray experiments—eg, Gene Expression Omnibus, which, as of April 5, 2013, contained information about 11 376 measurement platforms, 906 634 samples, 37 339 series, and 3200 datasets. Conversely, this registration is not yet the case for most traditional epidemiology studies.

Many studies involve authors with some financial conflicts of interests, especially those in clinical research. For example, among reports published in the *Journal of Clinical Oncology* in 2006–07, authors of 69% of clinical trial articles and 51% of editorials declared financial conflicts of interest.[99] Other empirical assessments have shown more prevalent conflicts of interests in treatment studies.[100] The reported prevalence of conflicts of interest is probably underestimated because of non-disclosures, although journal disclosure policies have been strengthened.[101]

Similarly, replication practices and standards vary substantially between specialties, so the relevant metric would be for more specialties to routinely adopt replication practices. Reproducibility checks are nearly non-existent, with few exceptions, and so any improvement would be positive. Accurate field surveys and reanalysis of these metrics might offer some evidence for whether design, conduct, and analysis of biomedical research are improving with time.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

1. Khoury MJ, Clauser SB, Freedman AN, et al. Population sciences, translational research, and the opportunities and challenges for genomics to reduce the burden of cancer in the 21st century. Cancer Epidemiol Biomarkers Prev. 2011; 20:2105–14. [PubMed: 21795499]
2. García-Berthou E, Alcaraz C. Incongruence between test statistics and P values in medical papers. BMC Med Res Methodol. 2004; 4:13. [PubMed: 15169550]
3. Prinz F, Schlange T, Asadullah K. Believe it or not: how much can we rely on published data on potential drug targets? Nat Rev Drug Discov. 2011; 10:712–13. [PubMed: 21892149]
4. Begley CG, Ellis LM. Drug development: raise standards for preclinical cancer research. Nature. 2012; 483:531–33. [PubMed: 22460880]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

5. Pereira TV, Ioannidis JP. Statistically significant meta-analyses of clinical trials have modest credibility and inflated effects. J Clin Epidemiol. 2011; 64:1060–69. [PubMed: 21454050]
6. Ioannidis JP, Panagiotou OA. Comparison of effect sizes associated with biomarkers reported in highly cited individual articles and in subsequent meta-analyses. JAMA. 2011; 305:2200–10. [PubMed: 21632484]
7. Bracken MB. Why are so many epidemiology associations inflated or wrong? Does poorly conducted animal research suggest implausible hypotheses? Ann Epidemiol. 2009; 19:220–24. [PubMed: 19217006]
8. Greenland, S.; Lash, TL. Bias analysis. In: Rothman, KJ.; Greenland, S.; Lash, TL., editors. Modern epidemiology. 3. Philadelphia: Lippincott–Williams–Wilkins; 2008. p. 345-80.Chapter 19
9. Greenland S. Multiple comparisons and association selection in general epidemiology. Int J Epidemiol. 2008; 37:430–34. [PubMed: 18453632]
10. Ioannidis JP. Why most discovered true associations are inflated. Epidemiology. 2008; 19:640–48. [PubMed: 18633328]
11. Hindorff LA, Sethupathy P, Junkins HA, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci USA. 2009; 106:9362–67. [PubMed: 19474294]
12. Ioannidis JP. Expectations, validity, and reality in omics. J Clin Epidemiol. 2010; 63:945–49. [PubMed: 20573481]
13. Chavalarias D, Ioannidis JP. Science mapping analysis characterizes 235 biases in biomedical research. J Clin Epidemiol. 2010; 63:1205–15. [PubMed: 20400265]
14. Senn, S. Statistical issues in drug development. 2. New York: Wiley; 2007.
15. Savović J, Jones HE, Altman DG, et al. Influence of reported study design characteristics on intervention effect estimates from randomized, controlled trials. Ann Intern Med. 2012; 157:429–38. [PubMed: 22945832]
16. Lijmer JG, Mol BW, Heisterkamp S, et al. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA. 1999; 282:1061–66. [PubMed: 10493205]
17. Rutjes AW, Reitsma JB, Di Nisio M, Smidt N, van Rijn JC, Bossuyt PM. Evidence of bias and variation in diagnostic accuracy studies. CMAJ. 2006; 174:469–76. [PubMed: 16477057]
18. Tzoulaki I, Siontis KC, Ioannidis JP. Prognostic effect size of cardiovascular biomarkers in datasets from observational studies versus randomised trials: meta-epidemiology study. BMJ. 2011; 343:d6829. [PubMed: 22065657]
19. Ioannidis JP, Tarone R, McLaughlin JK. The false-positive to false-negative ratio in epidemiologic studies. Epidemiology. 2011; 22:450–56. [PubMed: 21490505]
20. Guyatt GH, Oxman AD, Vist GE, et al. for the GRADE Working Group. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ. 2008; 336:924–26. [PubMed: 18436948]
21. Hlatky MA, Greenland P, Arnett DK, et al. for the American Heart Association Expert Panel on Subclinical Atherosclerotic Diseases and Emerging Risk Factors and the Stroke Council. Criteria for evaluation of novel markers of cardiovascular risk: a scientific statement from the American Heart Association. Circulation. 2009; 119:2408–16. [PubMed: 19364974]
22. Ioannidis JP, Boffetta P, Little J, et al. Assessment of cumulative evidence on genetic associations: interim guidelines. Int J Epidemiol. 2008; 37:120–32. [PubMed: 17898028]
23. Boffetta P, Winn DM, Ioannidis JP, et al. Recommendations and proposed guidelines for assessing the cumulative evidence on joint effects of genes and environments on cancer occurrence in humans. Int J Epidemiol. 2012; 41:686–704. [PubMed: 22596931]
24. Chan, A-W.; Song, F.; Vickers, A., et al. Increasing value and reducing waste: addressing inaccessible research. Lancet. 2014. published online Jan 8. http://dx.doi.org/10.1016/S0140-6736(13)62296-5
25. Brazma A, Hingamp P, Quackenbush J, et al. Minimum information about a microarray experiment (MIAME)–toward standards for microarray data. Nat Genet. 2001; 29:365–71. [PubMed: 11726920]
26. Simera I, Moher D, Hoey J, Schulz KF, Altman DG. A catalogue of reporting guidelines for health research. Eur J Clin Invest. 2010; 40:35–53. [PubMed: 20055895]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

27. Moher D, Weeks L, Ocampo M, et al. Describing reporting guidelines for health research: a systematic review. J Clin Epidemiol. 2011; 64:718–42. [PubMed: 21216130]
28. Hooijmans CR, Leenaars M, Ritskes-Hoitinga M. A gold standard publication checklist to improve the quality of animal studies, to fully integrate the three Rs, and to make systematic reviews more feasible. Altern Lab Anim. 2010; 38:167–82. [PubMed: 20507187]
29. Kilkenny C, Browne WJ, Cuthill IC, Emerson M, Altman DG. Improving bioscience research reporting: the ARRIVE guidelines for reporting animal research. PLoS Biol. 2010; 8:e1000412. [PubMed: 20613859]
30. Dwan K, Altman DG, Cresswell L, Blundell M, Gamble CL, Williamson PR. Comparison of protocols and registry entries to published reports for randomised controlled trials. Cochrane Database Syst Rev. 2011; 1:MR000031. [PubMed: 21249714]
31. Halpern SD, Karlawish JH, Berlin JA. The continuing unethical conduct of underpowered clinical trials. JAMA. 2002; 288:358–62. [PubMed: 12117401]
32. Keen HI, Pile K, Hill CL. The prevalence of underpowered randomized clinical trials in rheumatology. J Rheumatol. 2005; 32:2083–88. [PubMed: 16265683]
33. Maxwell SE. The persistence of underpowered studies in psychological research: causes, consequences, and remedies. Psychol Methods. 2004; 9:147–63. [PubMed: 15137886]
34. Moher D, Dulberg CS, Wells GA. Statistical power, sample size, and their reporting in randomized controlled trials. JAMA. 1994; 272:122–24. [PubMed: 8015121]
35. Soares MO, Welton NJ, Harrison DA, et al. An evaluation of the feasibility, cost and value of information of a multicentre randomised controlled trial of intravenous immunoglobulin for sepsis (severe sepsis and septic shock): incorporating a systematic review, meta-analysis and value of information analysis. Health Technol Assess. 2012; 16:1–186.
36. Greenland S. Nonsignificance plus high power does not imply support for the null over the alternative. Ann Epidemiol. 2012; 22:364–68. [PubMed: 22391267]
37. Ocana A, Tannock IF. When are “positive” clinical trials in oncology truly positive? J Natl Cancer Inst. 2011; 103:16–20. [PubMed: 21131576]
38. Qaseem A, Snow V, Cross JT Jr, et al. for the American College of Physicians and American Academy of Family Physicians Panel on Dementia. Current pharmacologic treatment of dementia: a clinical practice guideline from the American College of Physicians and the American Academy of Family Physicians. Ann Intern Med. 2008; 148:370–78. [PubMed: 18316755]
39. Kazi DS, Hlatky MA. Repeat revascularization is a faulty end point for clinical trials. Circ Cardiovasc Qual Outcomes. 2012; 5:249–50. [PubMed: 22592752]
40. Ferreira-González I, Busse JW, Heels-Ansdell D, et al. Problems with use of composite end points in cardiovascular trials: systematic review of randomised controlled trials. BMJ. 2007; 334:786. [PubMed: 17403713]
41. Robinson KA, Goodman SN. A systematic examination of the citation of prior research in reports of randomized, controlled trials. Ann Intern Med. 2011; 154:50–55. [PubMed: 21200038]
42. Chalmers, I.; Bracken, MB.; Djulbegovic, B., et al. How to increase value and reduce waste when research priorities are set. Lancet. 2014. published online Jan 8. http://dx.doi.org/10.1016/S0140-6736(13)62229-1
43. Ioannidis JP. Perfect study, poor evidence: interpretation of biases preceding study design. Semin Hematol. 2008; 45:160–66. [PubMed: 18582622]
44. Ioannidis JP, Karassa FB. The need to consider the wider agenda in systematic reviews and meta-analyses: breadth, timing, and depth of the evidence. BMJ. 2010; 341:c4875. [PubMed: 20837576]
45. Chalmers I, Matthews R. What are the implications of optimism bias in clinical research? Lancet. 2006; 367:449–50. [PubMed: 16473106]
46. Mann, H.; Djulbegovic, B. [accessed Nov 29, 2013] Comparator bias: why comparisons must address genuine uncertainties. JLL Bulletin: commentaries on the history of treatment evaluation. http://www.jameslindlibrary.org/illustrating/articles/comparator-bias-why-comparisons-must-address-genuine-uncertaint
47. Jørgensen AW, Hilden J, Gøtzsche PC. Cochrane reviews compared with industry supported meta-analyses and other meta-analyses of the same drugs: systematic review. BMJ. 2006; 333:782. [PubMed: 17028106]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

48. Yank V, Rennie D, Bero LA. Financial ties and concordance between results and conclusions in meta-analyses: retrospective cohort study. BMJ. 2007; 335:1202–05. [PubMed: 18024482]
49. Greenland S. Commentary: on 'Quality in epidemiological research: hould we be submitting papers before we have the results and submitting more hypothesis generating research?'. Int J Epidemiol. 2007; 36:944–45. [PubMed: 17954715]
50. McNamee D. Review of clinical protocols at The Lancet. Lancet. 2001; 357:1819–20. [PubMed: 11410187]
51. Palepu A, Kendall C, Moher D. Open Medicine endorses PROSPERO. Open Med. 2011; 5:e65–66. [PubMed: 22046223]
52. Stewart L, Moher D, Shekelle P. Why prospective registration of systematic reviews makes sense. Syst Rev. 2012; 1:7. [PubMed: 22588008]
53. Booth A, Clarke M, Dooley G, et al. The nuts and bolts of PROSPERO: an international prospective register of systematic reviews. Syst Rev. 2012; 1:2. [PubMed: 22587842]
54. Alsheikh-Ali AA, Qureshi W, Al-Mallah MH, Ioannidis JP. Public availability of published research data in high-impact journals. PLoS One. 2011; 6:e24357. [PubMed: 21915316]
55. Al-Marzouki S, Roberts I, Evans S, Marshall T. Selective reporting in clinical trials: analysis of trial protocols accepted by The Lancet. Lancet. 2008; 372:201. [PubMed: 18640445]
56. Scharfstein DO, Rotnitsky A, Robins JM. Adjusting for nonignorable drop-out using semiparametric nonresponse models. J Am Stat Assoc. 1999; 94:1096–120.
57. Little, RJA.; Rubin, DB. Statistical analysis with missing data. 2. New York: Wiley; 2002.
58. Roland M, Torgerson DJ. What are pragmatic trials? BMJ. 1998; 316:285. [PubMed: 9472515]
59. Ware JH, Hamel MB. Pragmatic trials—guides to better patient care? N Engl J Med. 2011; 364:1685–87. [PubMed: 21542739]
60. Gandhi GY, Murad MH, Fujiyoshi A, et al. Patient-important outcomes in registered diabetes trials. JAMA. 2008; 299:2543–49. [PubMed: 18523223]
61. Montori VM, Gandhi GY, Guyatt GH. Patient-important outcomes in diabetes—time for consensus. Lancet. 2007; 370:1104–06. [PubMed: 17905147]
62. Burton PR, Hansell AL, Fortier I, et al. Size matters: just how big is BIG? Quantifying realistic sample size requirements for human genome epidemiology. Int J Epidemiol. 2009; 38:263–73. [PubMed: 18676414]
63. Veenstra DL, Roth JA, Garrison LP Jr, Ramsey SD, Burke W. A formal risk-benefit framework for genomic tests: facilitating the appropriate translation of genomics into clinical practice. Genet Med. 2010; 12:686–93. [PubMed: 20808229]
64. Richter SH, Garner JP, Zipser B, et al. Effect of population heterogenization on the reproducibility of mouse behavior: a multi-laboratory study. PLoS One. 2011; 6:e16461. [PubMed: 21305027]
65. Richter SH, Garner JP, Auer C, Kunert J, Würbel H. Systematic variation improves reproducibility of animal experiments. Nat Methods. 2010; 7:167–68. [PubMed: 20195246]
66. Bath PM, Macleod MR, Green AR. Emulating multicentre clinical stroke trials: a new paradigm for studying novel interventions in experimental models of stroke. Int J Stroke. 2009; 4:471–79. [PubMed: 19930059]
67. Dirnagl U, Fisher M. International, multicenter randomized preclinical trials in translational stroke research: it's time to act. J Cereb Blood Flow Metab. 2012; 32:933–35. [PubMed: 22510602]
68. Salanti G, Kavvoura FK, Ioannidis JP. Exploring the geometry of treatment networks. Ann Intern Med. 2008; 148:544–53. [PubMed: 18378949]
69. Khoury MJ, Gwinn M, Clyne M, Yu W. Genetic epidemiology with a capital E, ten years after. Genet Epidemiol. 2011; 35:845–52. [PubMed: 22125223]
70. Seminara D, Khoury MJ, O'Brien TR, et al. the Human Genome Epidemiology Network, and the Network of Investigator Networks. The emergence of networks in human genome epidemiology: challenges and opportunities. Epidemiology. 2007; 18:1–8. [PubMed: 17179752]
71. Fortier I, Burton PR, Robson PJ, et al. Quality, quantity and harmony: the DataSHaPER approach to integrating data across bioclinical studies. Int J Epidemiol. 2010; 39:1383–93. [PubMed: 20813861]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

72. Hu J, Coombes KR, Morris JS, Baggerly KA. The importance of experimental design in proteomic mass spectrometry experiments: some cautionary takes. Briefings Functional Genomics. 2005; 3:322–31.
73. McKinney WP, Young MJ, Hartz A, Lee MB. The inexact use of Fisher's Exact Test in six major medical journals. JAMA. 1989; 261:3430–33. [PubMed: 2724487]
74. Altman DG. Poor-quality medical research: what can journals do? JAMA. 2002; 287:2765–67. [PubMed: 12038906]
75. Methodology Committee of the Patient-Centered Outcomes Research Institute (PCORI). Methodological standards and patient-centeredness in comparative effectiveness research: the PCORI perspective. JAMA. 2012; 307:1636–40. [PubMed: 22511692]
76. Deverka PA, Schully SD, Ishibe N, et al. Stakeholder assessment of the evidence for cancer genomic tests: insights from three case studies. Genet Med. 2012; 14:656–62. [PubMed: 22481130]
77. Greenland S. Transparency and disclosure, neutrality and balance: shared values or just shared words? J Epidemiol Community Health. 2012; 66:967–70. [PubMed: 22268131]
78. Castaldi PJ, Dahabreh IJ, Ioannidis JP. An empirical assessment of validation practices for molecular classifiers. Brief Bioinform. 2011; 12:189–202. [PubMed: 21300697]
79. Hothorn T, Leisch F. Case studies in reproducibility. Brief Bioinform. 2011; 12:288–300. [PubMed: 21278369]
80. Ioannidis JP, Allison DB, Ball CA, et al. Repeatability of published microarray gene expression analyses. Nat Genet. 2009; 41:149–55. [PubMed: 19174838]
81. Gøtzsche PC. Reference bias in reports of drug trials. Br Med J (Clin Res Ed). 1987; 295:654–56.
82. Greenberg SA. How citation distortions create unfounded authority: analysis of a citation network. BMJ. 2009; 339:b2680. [PubMed: 19622839]
83. Altman DG. The scandal of poor medical research. BMJ. 1994; 308:283–84. [PubMed: 8124111]
84. Ioannidis JP. Research needs grants, funding and money—missing something? Eur J Clin Invest. 2012; 42:349–51. [PubMed: 22050119]
85. Tatsioni A, Bonitsis NG, Ioannidis JP. Persistence of contradicted claims in the literature. JAMA. 2007; 298:2517–26. [PubMed: 18056905]
86. Diggle PJ, Zeger SL. Embracing the concept of reproducible research. Biostatistics. 2010; 11:375. [PubMed: 20538869]
87. Keiding N. Reproducible research and the substantive context. Biostatistics. 2010; 11:376–78. [PubMed: 20498225]
88. Laine C, Goodman SN, Griswold ME, Sox HC. Reproducible research: moving toward research the public can really trust. Ann Intern Med. 2007; 146:450–53. [PubMed: 17339612]
89. Peng RD. Reproducible research and biostatistics. Biostatistics. 2009; 10:405–08. [PubMed: 19535325]
90. Peng RD. Reproducible research in computational science. Science. 2011; 334:1226–27. [PubMed: 22144613]
91. Ioannidis JP, Khoury MJ. Improving validation practices in "omics" research. Science. 2011; 334:1230–32. [PubMed: 22144616]
92. Ioannidis JP, Nosek B, Iorns E. Reproducibility concerns. Nat Med. 2012; 18:1736–37. [PubMed: 23223056]
93. Ommen, G.; DeAngelis, CD.; DeMets, DL., et al. Evolution of translational omics: lessons learned and the path forward. Washington, DC: Institute of Medicine National Academies of Sciences Press; 2012.
94. Witten DM, Tibshirani R. Scientific research in the age of omics: the good, the bad, and the sloppy. J Am Med Inform Assoc. 2013; 20:125–27. [PubMed: 23037799]
95. Solove, DJ. The future of reputation. New Haven: Yale University Press; 2007.
96. Evans, I.; Thornton, H.; Chalmers, I. [accessed Nov 29, 2013] Testing treatments: better research for better healthcare. 22010. http://www.jameslindlibrary.org/pdf/testing-treatments.pdf
97. Rasmussen N, Lee K, Bero L. Association of trial registration with the results and conclusions of published trials of new oncology drugs. Trials. 2009; 10:116. [PubMed: 20015404]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

98. Pinto RZ, Elkins MR, Moseley AM, et al. Many randomized trials of physical therapy interventions are not adequately registered: a survey of 200 published trials. Phys Ther. 2013; 93:299–309. [PubMed: 23125281]

99. Riechelmann RP, Wang L, O'Carroll A, Krzyzanowska MK. Disclosure of conflicts of interest by authors of clinical trials and editorials in oncology. J Clin Oncol. 2007; 25:4642–47. [PubMed: 17925561]

100. Jagsi R, Sheets N, Jankovic A, Motomura AR, Amarnath S, Ubel PA. Frequency, nature, effects, and correlates of conflicts of interest in published clinical cancer research. Cancer. 2009; 115:2783–91. [PubMed: 19434666]

101. Bosch X, Pericas JM, Hernández C, Doti P. Financial, nonfinancial and editors' conflicts of interest in high-impact biomedical journals. Eur J Clin Invest. 2013; 43:660–67. [PubMed: 23550719]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

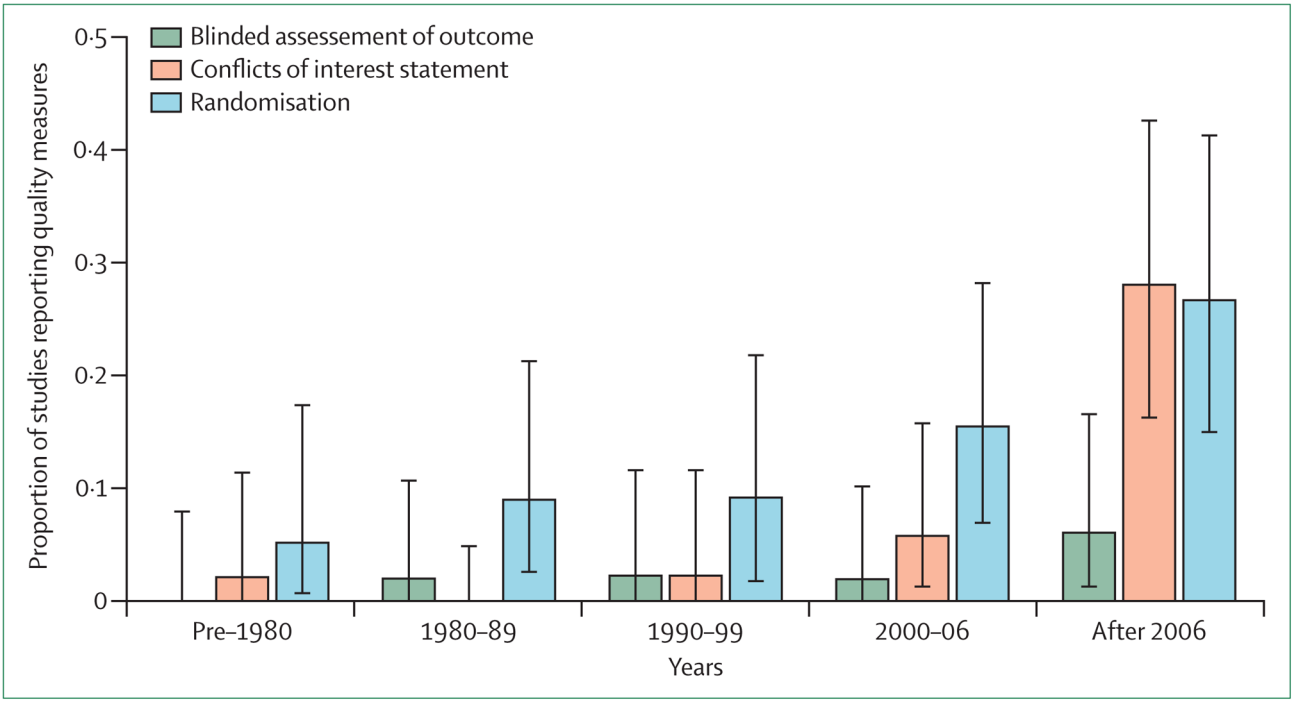

Figure. Trends in three methodological quality indicators for reports of in-vivo studies

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 described in quintiles of publication year, along with their 95% CI. Sample size calculation, concealment of allocation sequence, or blinded conduct of the experiment were not reported for any study, so are not shown. The appendix contains detailed protocol, data extraction process, flow diagram, and raw data.