A manifesto for reproducible science

Article *in* Nature Human Behaviour · January 2017

DOI: 10.1038/s41562-016-0021

CITATIONS
922

READS
2,968

10 authors, including:

Brian Nosek
University of Virginia
267 PUBLICATIONS **34,636** CITATIONS
SEE PROFILE

Katherine S Button
University of Bath
77 PUBLICATIONS **4,913** CITATIONS
SEE PROFILE

Nathalie Percie du Sert
National Centre for the Replacement, Refinement and Reduction of Animals in Re…
37 PUBLICATIONS **1,373** CITATIONS
SEE PROFILE

nature human behaviour

PERSPECTIVE

PUBLISHED: 10 JANUARY 2017 | VOLUME: 1 | ARTICLE NUMBER: 0021

OPEN

A manifesto for reproducible science

Marcus R. Munafò[1,2]*, Brian A. Nosek[3,4], Dorothy V. M. Bishop[5], Katherine S. Button[6], Christopher D. Chambers[7], Nathalie Percie du Sert[8], Uri Simonsohn[9], Eric-Jan Wagenmakers[10], Jennifer J. Ware[11] and John P. A. Ioannidis[12,13,14]

Improving the reliability and efficiency of scientific research will increase the credibility of the published scientific literature and accelerate discovery. Here we argue for the adoption of measures to optimize key elements of the scientific process: methods, reporting and dissemination, reproducibility, evaluation and incentives. There is some evidence from both simulations and empirical studies supporting the likely effectiveness of these measures, but their broad adoption by researchers, institutions, funders and journals will require iterative evaluation and improvement. We discuss the goals of these measures, and how they can be implemented, in the hope that this will facilitate action toward improving the transparency, reproducibility and efficiency of scientific research.

What proportion of published research is likely to be false? Low sample size, small effect sizes, data dredging (also known as *P*-hacking), conflicts of interest, large numbers of scientists working competitively in silos without combining their efforts, and so on, may conspire to dramatically increase the probability that a published finding is incorrect[1]. The field of metascience — the scientific study of science itself — is flourishing and has generated substantial empirical evidence for the existence and prevalence of threats to efficiency in knowledge accumulation (refs 2–7; Fig. 1).

Data from many fields suggests reproducibility is lower than is desirable[8–14]; one analysis estimates that 85% of biomedical research efforts are wasted[14], while 90% of respondents to a recent survey in *Nature* agreed that there is a 'reproducibility crisis'[15]. Whether 'crisis' is the appropriate term to describe the current state or trajectory of science is debatable, but accumulated evidence indicates that there is substantial room for improvement with regard to research practices to maximize the efficiency of the research community's use of the public's financial investment in research.

Here we propose a series of measures that we believe will improve research efficiency and robustness of scientific findings by directly targeting specific threats to reproducible science. We argue for the adoption, evaluation and ongoing improvement of these measures to optimize the pace and efficiency of knowledge accumulation. The measures are organized into the following categories[16]: methods, reporting and dissemination, reproducibility, evaluation and incentives. They are not intended to be exhaustive, but provide a broad, practical and evidence-based set of actions that can be implemented by researchers, institutions, journals and funders. The measures and their current implementation are summarized in Table 1.

The problem

A hallmark of scientific creativity is the ability to see novel and unexpected patterns in data. John Snow's identification of links between cholera and water supply[17], Paul Broca's work on language lateralization[18] and Jocelyn Bell Burnell's discovery of pulsars[19] are examples of breakthroughs achieved by interpreting observations in a new way. However, a major challenge for scientists is to be open to new and important insights while simultaneously avoiding being misled by our tendency to see structure in randomness. The combination of apophenia (the tendency to see patterns in random data), confirmation bias (the tendency to focus on evidence that is in line with our expectations or favoured explanation) and hindsight bias (the tendency to see an event as having been predictable only after it has occurred) can easily lead us to false conclusions[20]. Thomas Levenson documents the example of astronomers who became convinced they had seen the fictitious planet Vulcan because their contemporary theories predicted its existence[21]. Experimenter effects are an example of this kind of bias[22].

Over-interpretation of noise is facilitated by the extent to which data analysis is rapid, flexible and automated[23]. In a high-dimensional dataset, there may be hundreds or thousands of reasonable alternative approaches to analysing the same data[24,25]. For example, in a systematic review of functional magnetic resonance imaging (fMRI) studies, Carp showed that there were almost as many unique analytical pipelines as there were studies[26]. If several thousand potential analytical pipelines can be applied to high-dimensional data, the generation of false-positive findings is highly likely. For example, applying almost 7,000 analytical pipelines to a single fMRI dataset resulted in over 90% of brain voxels showing significant activation in at least one analysis[27].

[1]MRC Integrative Epidemiology Unit, University of Bristol, Bristol BS8 2BN, UK. [2]UK Centre for Tobacco and Alcohol Studies, School of Experimental Psychology, University of Bristol, 12a Priory Road, Bristol BS8 1TU, UK. [3]Department of Psychology, University of Virginia, Charlottesville, Virginia 22904, USA. [4]Center for Open Science, Charlottesville, Virginia 22903, USA. [5]Department of Experimental Psychology, University of Oxford, 9 South Parks Road, Oxford OX1 3UD, UK. [6]Department of Psychology, University of Bath, Bath BS2 7AY, UK. [7]Cardiff University Brain Research Imaging Centre, School of Psychology, Cardiff University, Cardiff CF24 4HQ, UK. [8]National Centre for the Replacement, Refinement and Reduction of Animals in Research (NC3Rs), London NW1 2BE, UK. [9]The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. [10]Department of Psychology, University of Amsterdam, Amsterdam 1018 WT, Netherlands. [11]CHDI Management/CHDI Foundation, New York, New York 10001, USA. [12]Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford 94304, California, USA. [13]Stanford Prevention Research Center, Department of Medicine and Department of Health Research and Policy, Stanford University School of Medicine, Stanford 94305, California, USA. [14]Department of Statistics, Stanford University School of Humanities and Sciences, Stanford 94305, California, USA.
*e-mail: marcus.munafo@bristol.ac.uk

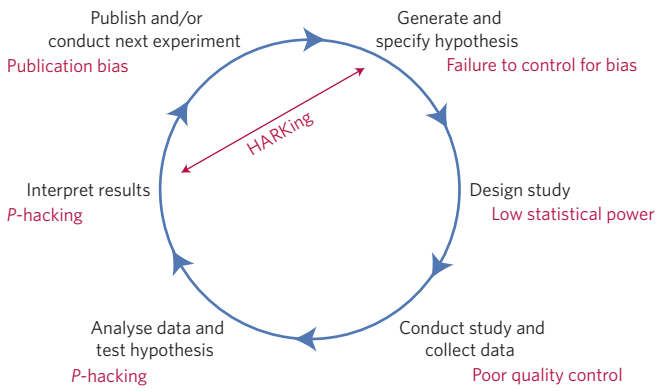

Figure 1 | Threats to reproducible science. An idealized version of the hypothetico-deductive model of the scientific method is shown. Various potential threats to this model exist (indicated in red), including lack of replication[5], hypothesizing after the results are known (HARKing)[7], poor study design, low statistical power[2], analytical flexibility[51], *P*-hacking[4], publication bias[3] and lack of data sharing[6]. Together these will serve to undermine the robustness of published research, and may also impact on the ability of science to self-correct.

During data analysis it can be difficult for researchers to recognize *P*-hacking[28] or data dredging because confirmation and hindsight biases can encourage the acceptance of outcomes that fit expectations or desires as appropriate, and the rejection of outcomes that do not as the result of suboptimal designs or analyses. Hypotheses may emerge that fit the data and are then reported without indication or recognition of their *post hoc* origin[7]. This, unfortunately, is not scientific discovery, but self-deception[29]. Uncontrolled, it can dramatically increase the false discovery rate. We need measures to counter the natural tendency of enthusiastic scientists who are motivated by discovery to see patterns in noise.

Methods

In this section we describe measures that can be implemented when performing research (including, for example, study design, methods, statistics, and collaboration).

Protecting against cognitive biases. There is a substantial literature on the difficulty of avoiding cognitive biases. An effective solution to mitigate self-deception and unwanted biases is blinding. In some research contexts, participants and data collectors can be blinded to the experimental condition that participants are assigned to, and to the research hypotheses, while the data analyst can be blinded to key parts of the data. For example, during data preparation and cleaning, the identity of experimental conditions or the variable labels can be masked so that the output is not interpretable in terms of the research hypothesis. In some physical sciences this approach has been extended to include deliberate perturbations in or masking of data to allow data preparation (for example, identification of outliers) to proceed without the analyst being able to see the corresponding results[30]. Pre-registration of the study design, primary outcome(s) and analysis plan (see 'Promoting study pre-registration' section, below) is a highly effective form of blinding because the data do not exist and the outcomes are not yet known.

Improving methodological training. Research design and statistical analysis are mutually dependent. Common misperceptions, such as the interpretation of *P* values[31], limitations of null-hypothesis significance testing[32], the meaning and importance of statistical power[2], the accuracy of reported effect sizes[33], and the likelihood that a sample size that generated a statistically significant finding will also be adequate to replicate a true finding[34], could all be addressed through improved statistical training. Similarly, basic design principles are important, such as blinding to reduce experimenter bias, randomization or counterbalancing to control for confounding, and the use of within-subjects designs, where possible, to maximize power. However, integrative training in research practices that can protect oneself against cognitive biases and the effects of distorted incentives is arguably more important. Moreover, statistical and methodological best practices are under constant revision and improvement, so that senior as well as junior researchers need continuing methodological education, not least because much training of early-career researchers is informal and flows from their supervisor or mentor. A failure to adopt advances in methodology — such as the very slow progress in increasing statistical power[35,36] — may be partly a function of failing to inculcate a continuing professional education and development ethic.

Without formal requirements for continuing education, the most effective solutions may be to develop educational resources that are accessible, easy-to-digest and immediately and effectively applicable to research (for example, brief, web-based modules for specific topics, and combinations of modules that are customized for particular research applications). A modular approach simplifies the process of iterative updating of those materials. Demonstration software and hands-on examples may also make the lessons and implications particularly tangible to researchers at any career stage: the Experimental Design Assistant (https://eda.nc3rs.org.uk) supports research design for whole animal experiments, while *P*-hacker (http://shinyapps.org/apps/p-hacker/) shows just how easy it is to generate apparently statistically significant findings by exploiting analytic flexibility.

Implementing independent methodological support. The need for independent methodological support is well-established in some areas — many clinical trials, for example, have multidisciplinary trial steering committees to provide advice and oversee the design and conduct of the trial. The need for these committees grew out of the well-understood financial conflicts of interest that exist in many clinical trials. The sponsor of a trial may be the company manufacturing the product, and any intentional or unintentional influence can distort the study design, analysis and interpretation of results for the ultimate financial benefit of the manufacturer at the cost of the accuracy of the science and the health benefit to the consumers[37,38]. Non-financial conflicts of interest also exist, such as the beliefs and preconceptions of individual scientists and the stakes that researchers have in obtaining publishable results in order to progress their career[39,40]. Including independent researchers (particularly methodologists with no personal investment in a research topic) in the design, monitoring, analysis or interpretation of research outcomes may mitigate some of those influences, and can be done either at the level of the individual research project or through a process facilitated by a funding agency (see Box 1).

Encouraging collaboration and team science. Studies of statistical power persistently find it to be below (sometimes well below) 50%, across both time and the different disciplines studied[2,35,36]. Low statistical power increases the likelihood of obtaining both false-positive and false-negative results[2], meaning that it offers no advantage if the purpose is to accumulate knowledge. Despite this, low-powered research persists because of dysfunctional incentives, poor understanding of the consequences of low power, and lack of resources to improve power. Team science is a solution to the latter problem — instead of relying on the limited resources of single investigators, distributed collaboration across many study sites facilitates high-powered designs and greater potential for testing generalizability across the settings and populations sampled. This also brings greater scope for multiple theoretical and disciplinary perspectives, and a diverse range of research cultures and experiences, to be incorporated into a research project.

Table 1 | A manifesto for reproducible science.

Theme	Proposal	Examples of initiatives/potential solutions (extent of current adoption)	Stakeholder(s)
Methods	Protecting against cognitive biases	All of the initiatives listed below (* to ****) Blinding (**)	J, F
	Improving methodological training	Rigorous training in statistics and research methods for future researchers (*) Rigorous continuing education in statistics and methods for researchers (*)	I, F
	Independent methodological support	Involvement of methodologists in research (**) Independent oversight (*)	F
	Collaboration and team science	Multi-site studies/distributed data collection (*) Team-science consortia (*)	I, F
Reporting and dissemination	Promoting study pre-registration	Registered Reports (*) Open Science Framework (*)	J, F
	Improving the quality of reporting	Use of reporting checklists (**) Protocol checklists (*)	J
	Protecting against conflicts of interest	Disclosure of conflicts of interest (***) Exclusion/containment of financial and non-financial conflicts of interest (*)	J
Reproducibility	Encouraging transparency and open science	Open data, materials, software and so on (* to **) Pre-registration (**** for clinical trials, * for other studies)	J, F, R
Evaluation	Diversifying peer review	Preprints (* in biomedical/behavioural sciences, **** in physical sciences) Pre- and post-publication peer review, for example, Publons, PubMed Commons (*)	J
Incentives	Rewarding open and reproducible practices	Badges (*) Registered Reports (*) Transparency and Openness Promotion guidelines (*) Funding replication studies (*) Open science practices in hiring and promotion (*)	J, I, F

Estimated extent of current adoption: *, <5%; **, 5–30%; ***, 30–60%; ****, >60%. Abbreviations for key stakeholders: J, journals/publishers; F, funders; I, institutions; R, regulators.

Multi-centre and collaborative efforts have a long and successful tradition in fields such as randomized controlled trials in some areas of clinical medicine, and in genetic association analyses, and have improved the robustness of the resulting research literatures. Multi-site collaborative projects have also been advocated for other types of research, such as animal studies[41–43], in an effort to maximize their power, enhance standardization, and optimize transparency and protection from biases. The Many Labs projects illustrate this potential in the social and behavioural sciences, with dozens of laboratories implementing the same research protocol to obtain highly precise estimates of effect sizes, and evaluate variability across samples and settings[44,45]. It is also possible, and desirable, to incorporate a team science ethos into student training (see Box 2).

Reporting and dissemination

In this section we describe measures that can be implemented when communicating research (including, for example, reporting standards, study pre-registration, and disclosing conflicts of interest).

Promoting study pre-registration. Pre-registration of study protocols for randomized controlled trials in clinical medicine has become standard practice[46]. In its simplest form it may simply comprise the registration of the basic study design, but it can also include a detailed pre-specification of the study procedures, outcomes and statistical analysis plan. It was introduced to address two problems: publication bias and analytical flexibility (in particular outcome switching in the case of clinical medicine). Publication bias[47], also known as the file drawer problem[48], refers to the fact that many more studies are conducted than published. Studies that obtain positive and novel results are more likely to be published than studies that obtain negative results or report replications of prior results[47,49,50]. The consequence is that the published literature indicates stronger evidence for findings than exists in reality. Outcome switching refers to the possibility of changing the outcomes of interest in the study depending on the observed results. A researcher may include ten variables that could be considered outcomes of the research, and — once the results are known — intentionally or unintentionally select the subset of outcomes that show statistically significant results as the outcomes of interest. The consequence is an increase in the likelihood that reported results are spurious by leveraging chance, while negative evidence gets ignored. This is one of several related research practices that can inflate spurious findings when analysis decisions are made with knowledge of the observed data, such as selection of models, exclusion rules and covariates. Such data-contingent analysis decisions constitute what has become known as *P*-hacking[51], and pre-registration can protect against all of these.

The strongest form of pre-registration involves both registering the study (with a commitment to make the results public) and closely pre-specifying the study design, primary outcome and analysis plan in advance of conducting the study or knowing the outcomes of the research. In principle, this addresses publication bias by making all research discoverable, whether or not it is ultimately published, allowing all of the evidence about a finding to be obtained and evaluated. It also addresses outcome switching, and *P*-hacking more generally, by requiring the researcher to articulate analytical decisions prior to observing the data, so that these decisions remain data-independent. Critically, it also makes clear the distinction

Box 1 | Independent oversight: the case of CHDI Foundation.

CHDI Foundation — a privately-funded non-profit drug-development organization targeting Huntington's disease — convened a working group in 2013 to identify practical and viable steps that could be taken to help ensure the rigor of their research[76]. One concrete product of this meeting was the establishment of the Independent Statistical Standing Committee (ISSC; http://chdifoundation.org/independent-statistical-standing-committee/) designed to provide independent, unbiased and objective evaluation and expert advice regarding all aspects of experimental design and statistics. CHDI has made this resource available to the wider Huntington's disease research community on a priority basis. The ISSC is comprised of individuals with specific expertise in research design and statistics. Critically, committee members are not themselves engaged in Huntington's disease research, and have no investment in study results, or other conflicts of interest. The committee provides a number of services, including (but not limited to) provision of expert assistance in developing protocols and statistical analysis plans, and evaluation of prepared study protocols. Their oversight and input, particularly at the study design stage, may mitigate low statistical power, inadequate study design, and flexibility in data analysis and reporting[6,71]. As recently highlighted, "asking questions at the design stage can save headaches at the analysis stage: careful data collection can greatly simplify analysis and make it more rigorous"[77].

Box 2 | Distributed student projects.

Student assessment requirements, and limited access to populations of interest, may hinder extensive collaboration within a single institution, but it could be achieved across multiple institutions in the form of a distributed student project. Under this model, academics and students from several institutions would form a consortium, collaboratively develop a research question, protocol and analysis plan, and publicly pre-register it prior to data collection. The protocol would be implemented by each student at each participating centre, and the resulting data pooled for analysis. Consortium meetings before and after data collection could be used to integrate training in research design, while offering opportunities for creative input from the students. Conclusions based on results would be mutually agreed in preparation for wider dissemination, using inclusive authorship conventions such as those adopted by genetic consortia. Students would learn rigorous research methods through active participation in research that is sufficiently well designed and conducted to be genuinely meaningful. Critically, collaborative team science would be instilled at an early stage of training.

The Collaborative Replications and Education Project (CREP; https://osf.io/wfc6u/) is an example of this concept in psychology, albeit in a more centralized form. A coordinating team identifies recently published research that could be replicated in the context of a semester-long undergraduate course on research methods. A central commons provides the materials and guidance to incorporate the replications into projects or classes, and the data collected across sites are aggregated into manuscripts for publication. The Pipeline[78] and Many Labs[44,45] projects also offer opportunities to contribute to large-scale replication efforts with coordinated data collection across many locations simultaneously.

between data-independent confirmatory research that is important for testing hypotheses, and data-contingent exploratory research that is important for generating hypotheses.

While pre-registration is now common in some areas of clinical medicine (due to requirements by journals and regulatory bodies, such as the Food and Drug Administration in the United States and the European Medicines Agency in the European Union), it is rare in the social and behavioural sciences. However, support for study pre-registration is increasing; websites such as the Open Science Framework (http://osf.io/) and AsPredicted (http://AsPredicted.org/) offer services to pre-register studies, the Preregistration Challenge offers education and incentives to conduct pre-registered research (http://cos.io/prereg), and journals are adopting the Registered Reports publishing format[52,53] to encourage pre-registration and add results-blind peer review (see Box 3).

Improving the quality of reporting. Pre-registration will improve discoverability of research, but discoverability does not guarantee usability. Poor usability reflects difficulty in evaluating what was done, in reusing the methodology to assess reproducibility, and in incorporating the evidence into systematic reviews and meta-analyses. Improving the quality and transparency in the reporting of research is necessary to address this. The Transparency and Openness Promotion (TOP) guidelines offer standards as a basis for journals and funders to incentivize or require greater transparency in planning and reporting of research[54]. TOP provides principles for how transparency and usability can be increased, while other guidelines provide concrete steps for how to maximize the quality of reporting in particular areas. For example, the Consolidated Standards of Reporting Trials (CONSORT) statement provides guidance for clear, complete and accurate reporting of randomized controlled trials[55–57]. Over 300 reporting guidelines now exist for observational studies, prognostic studies, predictive models, diagnostic tests, systematic reviews and meta-analyses in humans, a large variety of studies using different laboratory methods, and animal studies. The Equator Network (http://www.equator-network.org/) aggregates these guidelines to improve discoverability[58]. There are also guidelines for improving the reporting of research planning; for example, the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) statement for reporting of systematic reviews and meta-analyses[59], and PRISMA-P for protocols of systematic reviews[60]. The Preregistration Challenge workflow and the pre-registration recipe for social-behavioural research[61] also illustrate guidelines for reporting research plans.

The success of reporting guidelines depends on their adoption and effective use. The social and behavioural sciences are behind the biomedical sciences in their adoption of reporting guidelines for research, although with rapid adoption of the TOP guidelines and related developments by journals and funders that gap may be closing. However, improved reporting may be insufficient on its own to maximize research quality. Reporting guidelines are easily perceived by researchers as bureaucratic exercises rather than means of improving research and reporting. Even with pre-registration of clinical trials, one study observed that just 13% of trials published outcomes completely consistent with the pre-registered commitments. Most publications of the trials did not report pre-registered outcomes and added new outcomes that were not part of the registered design (see www.COMPare-trials.org). Franco and colleagues observed similar findings in psychology[62]; using protocol pre-registrations and public data from the Time-sharing Experiments for the Social Sciences project (http://www.tessexperiments.org/), they found that 40% of published reports failed to mention one or more of the experimental conditions of the experiments, and approximately 70% of published reports failed to mention one or more of the outcome measures included in the study. Moreover, outcome measures that were not included were much more likely to be negative results and associated with smaller effect sizes than outcome measures that were included.

Box 3 | Registered Reports.

The Registered Reports (RR) initiative seeks to eliminate various forms of bias in hypothesis-driven research[52,53], and in particular, the evaluation of a study based on the results. Unlike conventional journal articles, RRs split the peer review process into two stages, before and after results are known. At the first stage, reviewers and editors assess a detailed protocol that includes the study rationale, procedure and a detailed analysis plan. Following favourable reviews (and probably revision to meet strict methodological standards), the journal offers in-principle acceptance: publication of study outcomes is guaranteed provided the authors adhere to the approved protocol, the study meets pre-specified quality checks, and conclusions are appropriately evidence-bound. Once the study is completed, the authors resubmit a complete manuscript that includes the results and discussion. The article is published at the end of this two-stage process. By accepting articles before results are known, RRs prevent publication bias. By reviewing the hypotheses and analysis plans in advance, RRs should also help neutralize *P*-hacking and HARKing (hypothesizing after the results are known) by authors, and CARKing (critiquing after the results are known) by reviewers with their own investments in the research outcomes, although empirical evidence will be required to confirm that this is the case.

Perhaps the most commonly voiced objection to RRs is that the format somehow limits exploration or creativity by requiring authors to adhere to a pre-specified methodology. However, RRs place no restrictions on creative analysis practices or serendipity. Authors are free to report the outcomes of any unregistered exploratory analyses, provided such tests are clearly labelled as *post hoc*. Thus, the sole requirement is that exploratory outcomes are identified transparently as exploratory (for a list of frequently asked questions see https://cos.io/rr/#faq). Of course, RRs are not intended for research that is solely exploratory.

As of November 2016, RRs have been adopted by over 40 journals, including *Nature Human Behaviour*, covering a wide range of life, social and physical sciences (for a curated list see https://cos.io/rr/#journals). The concept also opens the door to alternative forms of research funding that place a premium on transparency and reproducibility. For example, authors could submit a detailed proposal before they have funding for their research. Following simultaneous review by both the funder and the journal, the strongest proposals would be offered financial support by the funder and in-principle acceptance for publication by the journal (https://cos.io/rr/#funders).

Box 4 | Evidence for the effectiveness of reporting guidelines.

In medicine there is strong evidence for the effectiveness of CONSORT guidelines — journals that do not endorse the CONSORT statement show poorer reporting quality compared with endorsing journals[79]. For the ARRIVE (Animal Research: Reporting of *In Vivo* Experiments) guidelines[80], studies comparing the reporting of ARRIVE items in specific fields of research before and after the guidelines were published report mixed results[81–83]. A randomized controlled trial is in progress to assess the impact of mandating a completed ARRIVE checklist with manuscript submissions on the quality of reporting in published articles (https://ecrf1.clinicaltrials.ed.ac.uk/iicarus). The success of these efforts will require journals and funders to adopt guidelines and support the community's iterative evaluation and improvement cycle.

The positive outcome of reporting guidelines is that they make it possible to detect and study these behaviours and their impact. Otherwise, these behaviours are simply unknowable in any systematic way. The negative outcome is the empirical evidence that reporting guidelines may be necessary, but will not alone be sufficient, to address reporting biases. The impact of guidelines and how best to optimize their use and impact will be best assessed by randomized trials (see Box 4).

Reproducibility

In this section we describe measures that can be implemented to support verification of research (including, for example, sharing data and methods).

Promoting transparency and open science. Science is a social enterprise: independent and collaborative groups work to accumulate knowledge as a public good. The credibility of scientific claims is rooted in the evidence supporting them, which includes the methodology applied, the data acquired, and the process of methodology implementation, data analysis and outcome interpretation. Claims become credible by the community reviewing, critiquing, extending and reproducing the supporting evidence. However, without transparency, claims only achieve credibility based on trust in the confidence or authority of the originator. Transparency is superior to trust.

Open science refers to the process of making the content and process of producing evidence and claims transparent and accessible to others. Transparency is a scientific ideal, and adding 'open' should therefore be redundant. In reality, science often lacks openness: many published articles are not available to people without a personal or institutional subscription, and most data, materials and code supporting research outcomes are not made accessible, for example, in a public repository (refs 63,64; Box 5).

Very little of the research process (for example, study protocols, analysis workflows, peer review) is accessible because, historically, there have been few opportunities to make it accessible even if one wanted to do so. This has motivated calls for open access, open data and open workflows (including analysis pipelines), but there are substantial barriers to meeting these ideals, including vested financial interests (particularly in scholarly publishing) and few incentives for researchers to pursue open practices. For example, current incentive structures promote the publication of 'clean' narratives, which may require the incomplete reporting of study procedures or results. Nevertheless, change is occurring. The TOP guidelines[54,65] promote open practices, while an increasing number of journals and funders require open practices (for example, open data), with some offering their researchers free, immediate open-access publication with transparent post-publication peer review (for example, the Wellcome Trust, with the launch of Wellcome Open Research). Policies to promote open science can include reporting guidelines or specific disclosure statements (see Box 6). At the same time, commercial and non-profit organizations are building new infrastructure such as the Open Science Framework to make transparency easy and desirable for researchers.

Evaluation

In this section we describe measures that can be implemented when evaluating research (including, for example, peer review).

Diversifying peer review. For most of the history of scientific publishing, two functions have been confounded — evaluation and dissemination. Journals have provided dissemination via sorting and delivering content to the research community, and gatekeeping via peer review to determine what is worth disseminating. However, with the advent of the internet, individual researchers are no longer

Box 5 | Data sharing.

Sharing data in public repositories offers field-wide advantages in terms of accountability, data longevity, efficiency and quality (for example, reanalyses may detect crucial mistakes or even data fabrication)[84]. Unfortunately, many scientific disciplines, including most of those devoted to the study of human behaviour, do not have a culture that values open data[6]. In the past, data sharing has rarely been enforced or facilitated. Recent initiatives, however, aim to change the normative culture. Hopefully, these initiatives will change the culture on data sharing. Once accepted as the norm, we doubt that data sharing will ever go out of fashion.

Transparency and Openness Promotion (TOP)

In 2015, Nosek and colleagues[54] proposed author guidelines to help journals and funders adopt transparency and reproducibility policies. As of November 2016 there were 757 journal and 64 organization signatories to the TOP guidelines. For example, the journal *Science* decided to "publish papers only if the data used in the analysis are available to any researcher for purposes of reproducing or extending the analysis"[65] and the conglomerate of Springer Nature journals adopted similar data-sharing policies.

Badges to acknowledge open-science practices

The Center for Open Science has suggested that journals assign a badge to articles with open data (as well as to other open practices such as pre-registration and open materials). The main purpose of the badges is to signal that the journal values these practices. The journal *Psychological Science* has adopted these badges, and there is evidence that the open data badge has had a positive effect, increasing data sharing by more than tenfold (Fig. 2).

The Peer Reviewers' Openness Initiative

Researchers who sign this initiative (https://opennessinitiative.org) pledge that as reviewers they will not offer comprehensive review for any manuscript that does not make data publicly available without a clear reason[85].

Requirements from funding agencies

In recent years, prominent funding agencies such as Research Councils UK in the United Kingdom and the National Institutes of Health (NIH) and National Science Foundation (NSF) in the United States have increased pressure on researchers to share data. For instance, the 2015 NIH Public Access Plan (https://grants.nih.gov/grants/NIH-Public-Access-Plan.pdf) states: "NIH intends to make public access to digital scientific data the standard for all NIH-funded research". Since 2010, NSF requires submission of a data-management plan that stipulates how data will be stored and shared.

Box 6 | Disclosure.

Full disclosure refers to the process of describing in full the study design and data collected that underlie the results reported, rather than a curated version of the design, and/or a subset of the data collected. The need for disclosure is clear: in order to adequately evaluate results we need to know how they were obtained. For example, the informational value of a dependent variable exhibiting an effect of interest is different if only one variable was collected or if fifteen were. The probability of a single variable achieving $P < 0.05$ just by chance is 5%, but the probability of one of fifteen variables achieving $P < 0.05$ is 54%[1]. It is obvious that cherry-picking one from fifteen variables invalidates the results unless it is clear that this has happened. If readers know, then they can adjust their interpretation accordingly. From this simple fact it follows that if authors do not tell us whether they collected one or fifteen variables readers cannot evaluate their research[51].

The simplest form of disclosure is for authors to assure readers via an explicit statement in their article that they are disclosing the data fully. This can be seen as a simple item of reporting guidance where extra emphasis is placed on some aspects that are considered most essential to disclose. For example, including the following 21-word statement: "We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study"[86]. Alternatively, a more complex, but also more enforceable and accountable process is for journals to require explicit and specific disclosure statements. The journal *Psychological Science*, for example, now requires authors to "Confirm that (a) the total number of excluded observations and (b) the reasons for making these exclusions have been reported in the Method section(s)"[87].

dependent on publishers to bind, print and mail their research to subscribers. Dissemination is now easy and can be controlled by researchers themselves. For example, preprint services (arXiv for some physical sciences, bioRxiv and PeerJ for the life sciences, engrXiv for engineering, PsyArXiv for psychology, and SocArXiv and the Social Science Research Network (SSRN) for the social sciences) facilitate easy sharing, sorting and discovery of research prior to publication. This dramatically accelerates the dissemination of information to the research community.

With increasing ease of dissemination, the role of publishers as a gatekeeper is declining. Nevertheless, the other role of publishing — evaluation — remains a vital part of the research enterprise. Conventionally, a journal editor will select a limited number of reviewers to assess the suitability of a submission for a particular journal. However, more diverse evaluation processes are now emerging, allowing the collective wisdom of the scientific community to be harnessed[66]. For example, some preprint services support public comments on manuscripts, a form of pre-publication review that can be used to improve the manuscript. Other services, such as PubMed Commons and PubPeer, offer public platforms to comment on published works facilitating post-publication peer review. At the same time, some journals are trialling 'results-free' review, where editorial decisions to accept are based solely on review of the rationale and study methods alone (that is, results-blind)[67].

Both pre- and post-publication peer review mechanisms dramatically accelerate and expand the evaluation process[68]. By sharing preprints, researchers can obtain rapid feedback on their work from a diverse community, rather than waiting several months for a few reviews in the conventional, closed peer review process. Using post-publication services, reviewers can make positive and critical commentary on articles instantly, rather than relying on the laborious, uncertain and lengthy process of authoring a commentary and submitting it to the publishing journal for possible publication, eventually.

As public forms of pre- and post-publication review, these new services introduce the potential for new forms of credit and reputation enhancement[69]. In the conventional model, peer review is done privately, anonymously and purely as a service. With public commenting systems, a reviewer that chooses to be identifiable may gain (or lose) reputation based on the quality of review. There are a number of possible and perceived risks of non-anonymous reviewing that reviewers must consider and research must evaluate, but there is evidence that open peer review improves the quality of reviews received[70]. The opportunity for accelerated scholarly

communication may both improve the pace of discovery and diversify the means of being an active contributor to scientific discourse.

Incentives

Publication is the currency of academic science and increases the likelihood of employment, funding, promotion and tenure. However, not all research is equally publishable. Positive, novel and clean results are more likely to be published than negative results, replications and results with loose ends; as a consequence, researchers are incentivized to produce the former, even at the cost of accuracy[40]. These incentives ultimately increase the likelihood of false positives in the published literature[71]. Shifting the incentives therefore offers an opportunity to increase the credibility and reproducibility of published results. For example, with simulations, Munafò and Higginson developed an optimality model that predicted the most rational research strategy, in terms of the proportion of research effort spent on seeking novel results rather than on confirmatory studies, and the amount of research effort per exploratory study[72]. This showed that, for parameter values derived from the scientific literature, researchers acting to maximize their 'fitness' should spend most of their effort seeking novel results and conduct small studies that have a statistical power of only 10–40%. Critically, their model suggests that altering incentive structures, by considering more of a researcher's output and giving less weight to strikingly novel findings when making appointment and promotion decisions, would encourage a change in researcher behaviour that would ultimately improve the scientific value of research.

Funders, publishers, societies, institutions, editors, reviewers and authors all contribute to the cultural norms that create and sustain dysfunctional incentives. Changing the incentives is therefore a problem that requires a coordinated effort by all stakeholders to alter reward structures. There will always be incentives for innovative outcomes — those who discover new things will be rewarded more than those who do not. However, there can also be incentives for efficiency and effectiveness — those who conduct rigorous, transparent and reproducible research could be rewarded more than those who do not. There are promising examples of effective interventions for nudging incentives. For example, journals are adopting badges to acknowledge open practices (Fig. 2), Registered Reports as a results-blind publishing model (see Box 3) and TOP guidelines to promote openness and transparency. Funders are also adopting transparency requirements, and piloting funding mechanisms to promote reproducibility such as the Netherlands Organisation for Scientific Research (NWO) and the US National Science Foundation's Directorate of Social, Behavioral and Economic Sciences, both of which have announced funding opportunities for replication studies. Institutions are wrestling with policy and infrastructure adjustments to promote data sharing, and there are hints of open-science practices becoming part of hiring and performance evaluation (for example, http://www.nicebread.de/open-science-hiring-practices/). Collectively, and at scale, such efforts can shift incentives such that what is good for the scientist is also good for science — rigorous, transparent and reproducible research practices producing credible results.

Conclusion

The challenges to reproducible science are systemic and cultural, but that does not mean they cannot be met. The measures we have described constitute practical and achievable steps toward improving rigor and reproducibility. All of them have shown some effectiveness, and are well suited to wider adoption, evaluation and improvement. Equally, these proposals are not an exhaustive list; there are many other nascent and maturing ideas for making research practices more efficient and reliable[73]. Offering a solution to a problem does not guarantee its effectiveness, and making changes to cultural norms and incentives can spur additional behavioural changes that are difficult to anticipate. Some solutions may be ineffective or even harmful to the efficiency and reliability of science, even if conceptually they appear sensible.

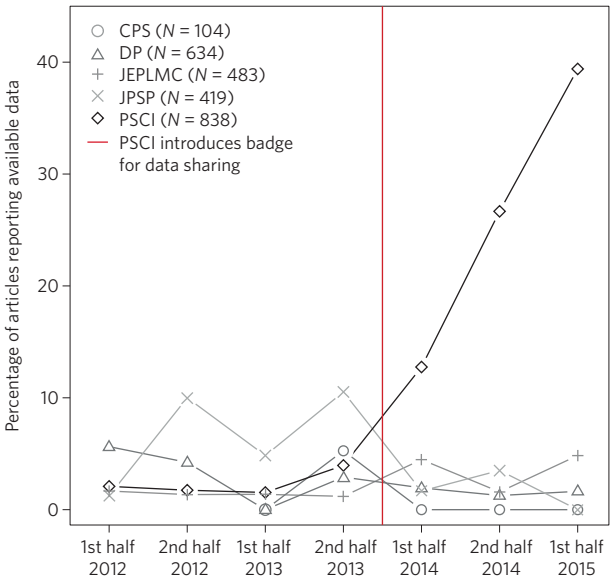

Figure 2 | The impact of introducing badges for data sharing. In January 2014, the journal *Psychological Science* (PSCI) introduced badges for articles with open data. Immediately afterwards, the proportion of articles with open data increased steeply, and by October 2015, 38% of articles in *Psychological Science* had open data. For comparison journals (*Clinical Psychological Science* (CPS), *Developmental Psychology* (DP), *Journal of Experimental Psychology: Learning, Memory and Cognition* (JEPLMC) and *Journal of Personality and Social Psychology* (JPSP)) the proportion of articles with open data remained uniformly low. Figure adapted from ref. 75, PLoS.

The field of metascience (or metaresearch) is growing rapidly, with over 2,000 relevant publications accruing annually[16]. Much of that literature constitutes the evaluation of existing practices and the identification of alternative approaches. What was previously taken for granted may be questioned, such as widely used statistical methods; for example, the most popular methods and software for spatial extent analysis in fMRI imaging were recently shown to produce unacceptably high false-positive rates[74]. Proposed solutions may also give rise to other challenges; for example, while replication is a hallmark for reinforcing trust in scientific results, there is uncertainty about which studies deserve to be replicated and what would be the most efficient replication strategies. Moreover, a recent simulation suggests that replication alone may not suffice to rid us of false results[71].

These cautions are not a rationale for inaction. Reproducible research practices are at the heart of sound research and integral to the scientific method. How best to achieve rigorous and efficient knowledge accumulation is a scientific question; the most effective solutions will be identified by a combination of brilliant hypothesizing and blind luck, by iterative examination of the effectiveness of each change, and by a winnowing of many possibilities to the broadly enacted few. True understanding of how best to structure and incentivize science will emerge slowly and will never be finished. That is how science works. The key to fostering a robust metascience that evaluates and improves practices is that the stakeholders of science must not embrace the status quo, but instead pursue self-examination continuously for improvement and self-correction of the scientific process itself.

As Richard Feynman said, "The first principle is that you must not fool yourself – and you are the easiest person to fool."

References

1. Ioannidis, J. P. A. Why most published research findings are false. *PLoS Med.* **2,** e124 (2005).
2. Button, K. S. *et al.* Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* **14,** 365–376 (2013).
3. Fanelli, D. "Positive" results increase down the Hierarchy of the Sciences. *PloS ONE* **5,** e10068 (2010).
4. John, L. K., Loewenstein, G. & Prelec, D. Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychol. Sci.* **23,** 524–532 (2012).
5. Makel, M. C., Plucker, J. A. & Hegarty, B. Replications in psychology research: how often do they really occur? *Perspect. Psychol. Sci.* **7,** 537–542 (2012).
6. Wicherts, J. M., Borsboom, D., Kats, J. & Molenaar, D. The poor availability of psychological research data for reanalysis. *Am. Psychol.* **61,** 726–728 (2006).
7. Kerr, N. L. HARKing: hypothesizing after the results are known. *Pers. Soc. Psychol. Rev.* **2,** 196–217 (1998).
8. Al-Shahi Salman, R. *et al.* Increasing value and reducing waste in biomedical research regulation and management. *Lancet* **383,** 176–185 (2014).
9. Begley, C. G. & Ioannidis, J. P. Reproducibility in science: improving the standard for basic and preclinical research. *Circ. Res.* **116,** 116–126 (2015).
10. Chalmers, I. *et al.* How to increase value and reduce waste when research priorities are set. *Lancet* **383,** 156–165 (2014).
11. Chan, A. W. *et al.* Increasing value and reducing waste: addressing inaccessible research. *Lancet* **383,** 257–266 (2014).
12. Glasziou, P. *et al.* Reducing waste from incomplete or unusable reports of biomedical research. *Lancet* **383,** 267–276 (2014).
13. Ioannidis, J. P. *et al.* Increasing value and reducing waste in research design, conduct, and analysis. *Lancet* **383,** 166–175 (2014).
14. Macleod, M. R. *et al.* Biomedical research: increasing value, reducing waste. *Lancet* **383,** 101–104 (2014).
15. Baker, M. 1,500 scientists lift the lid on reproducibility. *Nature* **533,** 452–454 (2016).
16. Ioannidis, J. P., Fanelli, D., Dunne, D. D. & Goodman, S. N. Meta-research: evaluation and improvement of research methods and practices. *PLoS Biol.* **13,** e1002264 (2015).
17. Paneth, N. Assessing the contributions of John Snow to epidemiology: 150 years after removal of the broad street pump handle. *Epidemiology* **15,** 514–516 (2004).
18. Berker, E. A., Berker, A. H. & Smith, A. Translation of Broca's 1865 report. Localization of speech in the third left frontal convolution. *Arch. Neurol.* **43,** 1065–1072 (1986).
19. Wade, N. Discovery of pulsars: a graduate student's story. *Science* **189,** 358–364 (1975).
20. Nickerson, R. S. Confirmation bias: a ubiquitous phenomenon in many guises. *Rev. Gen. Psychol.* **2,** 175–220 (1998).
21. Levenson, T. *The Hunt for Vulcan...and How Albert Einstein Destroyed a Planet, Discovered Relativity, and Deciphered the University* (Random House, 2015).
22. Rosenthal, R. *Experimenter Effects in Behavioral Research* (Appleton-Century-Crofts, 1966).
23. de Groot, A. D. The meaning of "significance" for different types of research [translated and annotated by Eric-Jan Wagenmakers, Denny Borsboom, Josine Verhagen, Rogier Kievit, Marjan Bakker, Angelique Cramer, Dora Matzke, Don Mellenbergh, and Han L. J. van der Maas]. *Acta Psychol.* **148,** 188–194 (2014).
24. Heininga, V. E., Oldehinkel, A. J., Veenstra, R. & Nederhof, E. I just ran a thousand analyses: benefits of multiple testing in understanding equivocal evidence on gene-environment interactions. *PloS ONE* **10,** e0125383 (2015).
25. Patel, C. J., Burford, B. & Ioannidis, J. P. Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *J. Clin. Epidemiol.* **68,** 1046–1058 (2015).
26. Carp, J. The secret lives of experiments: methods reporting in the fMRI literature. *Neuroimage* **63,** 289–300 (2012).
27. Carp, J. On the plurality of (methodological) worlds: estimating the analytic flexibility of FMRI experiments. *Front. Neurosci.* **6,** 149 (2012).
28. Simonsohn, U., Nelson, L. D. & Simmons, J. P. *P*-curve: a key to the file-drawer. *J. Exp. Psychol. Gen.* **143,** 534–547 (2014).
29. Nuzzo, R. Fooling ourselves. *Nature* **526,** 182–185 (2015).
30. MacCoun, R. & Perlmutter, S. Blind analysis: hide results to seek the truth. *Nature* **526,** 187–189 (2015).
31. Greenland, S. *et al.* Statistical tests, *P* values, confidence intervals, and power: a guide to misinterpretations. *Eur. J. Epidemiol.* **31,** 337–350 (2016).
32. Sterne, J. A. & Davey Smith, G. Sifting the evidence—what's wrong with significance tests? *BMJ* **322,** 226–231 (2001).
33. Brand, A., Bradley, M. T., Best, L. A. & Stoica, G. Accuracy of effect size estimates from published psychological research. *Percept. Motor Skill.* **106,** 645–649 (2008).
34. Vankov, I., Bowers, J. & Munafò, M. R. On the persistence of low power in psychological science. *Q. J. Exp. Psychol.* **67,** 1037–1040 (2014).
35. Sedlmeier, P. & Gigerenzer, G. Do studies of statistical power have an effect on the power of studies? *Psychol. Bull.* **105,** 309–316 (1989).
36. Cohen, J. The statistical power of abnormal-social psychological research: a review. *J. Abnorm. Soc. Psychol.* **65,** 145–153 (1962).
37. Etter, J. F., Burri, M. & Stapleton, J. The impact of pharmaceutical company funding on results of randomized trials of nicotine replacement therapy for smoking cessation: a meta-analysis. *Addiction* **102,** 815–822 (2007).
38. Etter, J. F. & Stapleton, J. Citations to trials of nicotine replacement therapy were biased toward positive results and high-impact-factor journals. *J. Clin. Epidemiol.* **62,** 831–837 (2009).
39. Panagiotou, O. A. & Ioannidis, J. P. Primary study authors of significant studies are more likely to believe that a strong association exists in a heterogeneous meta-analysis compared with methodologists. *J. Clin. Epidemiol.* **65,** 740–747 (2012).
40. Nosek, B. A., Spies, J. R. & Motyl, M. Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspect. Psychol. Sci.* **7,** 615–631 (2012).
41. Bath, P. M. W., Macleod, M. R. & Green, A. R. Emulating multicentre clinical stroke trials: a new paradigm for studying novel interventions in experimental models of stroke. *Int. J. Stroke* **4,** 471–479 (2009).
42. Dirnagl, U. *et al.* A concerted appeal for international cooperation in preclinical stroke research. *Stroke* **44,** 1754–1760 (2013).
43. Milidonis, X., Marshall, I., Macleod, M. R. & Sena, E. S. Magnetic resonance imaging in experimental stroke and comparison with histology systematic review and meta-analysis. *Stroke* **46,** 843–851 (2015).
44. Klein, R. A. *et al.* Investigating variation in replicability: a "many labs" replication project. *Soc. Psychol.* **45,** 142–152 (2014).
45. Ebersole, C. R. *et al.* Many Labs 3: evaluating participant pool quality across the academic semester via replication. *J. Exp. Soc. Psychol.* **67,** 68–82 (2016).
46. Lenzer, J., Hoffman, J. R., Furberg, C. D. & Ioannidis, J. P. A. Ensuring the integrity of clinical practice guidelines: a tool for protecting patients. *BMJ* **347,** f5535 (2013).
47. Sterling, T. D. Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *J. Am. Stat. Assoc.* **54,** 30–34 (1959).
48. Rosenthal, R. File drawer problem and tolerance for null results. *Psychol. Bull.* **86,** 638–641 (1979).
49. Sterling, T. D. Consequence of prejudice against the null hypothesis. *Psychol. Bull.* **82,** 1–20 (1975).
50. Franco, A., Malhotra, N. & Simonovits, G. Publication bias in the social sciences: unlocking the file drawer. *Science* **345,** 1502–1505 (2014).
51. Simmons, J. P., Nelson, L. D. & Simonsohn, U. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.* **22,** 1359–1366 (2011).
52. Chambers, C. D. Registered Reports: a new publishing initiative at Cortex. *Cortex* **49,** 609–610 (2013).
53. Nosek, B. A. & Lakens, D. Registered Reports: a method to increase the credibility of published results. *Soc. Psychol.* **45,** 137–141 (2014).
54. Nosek, B. A. *et al.* Promoting an open research culture. *Science* **348,** 1422–1425 (2015).
55. Begg, C. *et al.* Improving the quality of reporting of randomized controlled trials: the CONSORT statement. *JAMA* **276,** 637–639 (1996).
56. Moher, D., Dulberg, C. S. & Wells, G. A. Statistical power, sample size, and their reporting in randomized controlled trials. *JAMA* **272,** 122–124 (1994).
57. Schulz, K. F., Altman, D. G., Moher, D. & Group, C. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMJ* **340,** c332 (2010).
58. Grant, S. *et al.* Developing a reporting guideline for social and psychological intervention trials. *Res. Social Work Prac.* **23,** 595–602 (2013).
59. Liberati, A. *et al.* The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *PLoS Med.* **6,** e1000100 (2009).
60. Shamseer, L. *et al.* Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. *BMJ* **349,** g7647 (2015); erratum **354,** i4086 (2016).
61. van 't Veer, A. & Giner-Sorolla, R. Pre-registration in social psychology: a discussion and suggested template. *J. Exp. Soc. Psychol.* **67,** 2–12 (2016).
62. Franco, A., Malhotra, N. & Simonovits, G. Underreporting in psychology experiments: evidence from a study registry. *Soc. Psychol. Per. Sci.* **7,** 8–12 (2016).
63. Alsheikh-Ali, A. A., Qureshi, W., Al-Mallah, M. H. & Ioannidis, J. P. Public availability of published research data in high-impact journals. *PloS ONE* **6,** e24357 (2011).
64. Iqbal, S. A., Wallach, J. D., Khoury, M. J., Schully, S. D. & Ioannidis, J. P. Reproducible research practices and transparency across the biomedical literature. *PLoS Biol.* **14,** e1002333 (2016).
65. McNutt, M. Taking up TOP. *Science* **352,** 1147 (2016).

66. Park, I. U., Peacey, M. W. & Munafò, M. R. Modelling the effects of subjective and objective decision making in scientific peer review. *Nature* **506,** 93–96 (2014).
67. Button, K. S., Bal, L., Clark, A. G. & Shipley, T. Preventing the ends from justifying the means: withholding results to address publication bias in peer-review. *BMC Psychol.* **4,** 59 (2016).
68. Berg, J. M. *et al.* Preprints for the life sciences. *Science* **352,** 899–901 (2016).
69. Nosek, B. A. & Bar-Anan, T. Scientific utopia: I. Opening scientific communication. *Psychol. Inq.* **23,** 217–243 (2012).
70. Walsh, E., Rooney, M., Appleby, L. & Wilkinson, G. Open peer review: a randomised trial. *Brit. J. Psychiat.* **176,** 47–51 (2000).
71. Smaldino, P. E. & McElreath, R. The natural selection of bad science. *R. Soc. Open Sci.* **3,** 160384 (2016).
72. Higginson, A. D. & Munafò, M. Current incentives for scientists lead to underpowered studies with erroneous conclusions. *PLoS Biol.* **14,** e2000995 (2016).
73. Ioannidis, J. P. How to make more published research true. *PLoS Med.* **11,** e1001747 (2014).
74. Eklund, A., Nichols, T. E. & Knutsson, H. Cluster failure: why fMRI inferences for spatial extent have inflated false-positive rates. *Proc. Natl Acad. Sci. USA* **113,** 7900–7905 (2016).
75. Kidwell, M. C. *et al.* Badges to acknowledge open practices: a simple, low-cost, effective method for increasing transparency. *PLoS Biol.* **14,** e1002456 (2016).
76. Munafò, M. *et al.* Scientific rigor and the art of motorcycle maintenance. *Nat. Biotechnol.* **32,** 871–873 (2014).
77. Kass, R. E. *et al.* Ten simple rules for effective statistical practice. *PLoS Comput. Biol.* **12,** e1004961 (2016).
78. Schweinsberg, M. *et al.* The pipeline project: pre-publication independent replications of a single laboratory's research pipeline. *J. Exp. Psychol. Gen.* **66,** 55–67 (2016).
79. Stevens, A. *et al.* Relation of completeness of reporting of health research to journals' endorsement of reporting guidelines: systematic review. *BMJ* **348,** g3804 (2014).
80. Kilkenny, C. *et al.* Survey of the quality of experimental design, statistical analysis and reporting of research using animals. *PloS ONE* **4,** e7824 (2009).
81. Baker, D., Lidster, K., Sottomayor, A. & Amor, S. Two years later: journals are not yet enforcing the ARRIVE guidelines on reporting standards for pre-clinical animal studies. *PLoS Biol.* **12,** e1001756 (2014).
82. Gulin, J. E., Rocco, D. M. & Garcia-Bournissen, F. Quality of reporting and adherence to ARRIVE guidelines in animal studies for Chagas disease preclinical drug research: a systematic review. *PLoS Negl. Trop. Dis.* **9,** e0004194 (2015).
83. Liu, Y. *et al.* Adherence to ARRIVE guidelines in Chinese journal reports on neoplasms in animals. *PloS ONE* **11,** e0154657 (2016).
84. Gotzsche, P. C. & Ioannidis, J. P. Content area experts as authors: helpful or harmful for systematic reviews and meta-analyses? *BMJ* **345,** e7031 (2012).
85. Morey, R. D. *et al.* The Peer Reviewers' Openness Initiative: incentivizing open research practices through peer review. *R. Soc. Open Sci.* **3,** 150547 (2016).
86. Simmons, J. P., Nelson, L. D. & Simonsohn, U. A 21 word solution. Preprint at http://dx.doi.org/10.2139/ssrn.2160588(2012).
87. Eich, E. Business not as usual. *Psychol. Sci.* **25,** 3–6 (2014).

Acknowledgements

M.R.M. is a member of the UK Centre for Tobacco Control Studies, a UKCRC Public Health Research Centre of Excellence. Funding from the British Heart Foundation, Cancer Research UK, Economic and Social Research Council, Medical Research Council, and the National Institute for Health Research, under the auspices of the UK Clinical Research Collaboration, is gratefully acknowledged. This work was supported by the Medical Research Council Integrative Epidemiology Unit at the University of Bristol (MC_UU_12013/6). D.V.M.B. is funded by a Wellcome Trust Principal Research Fellowship and Programme (grant number 082498/Z/07/Z). N.P.d.S. is employed by the NC3Rs, which is primarily funded by the UK government. J.P.A.I. is funded by an unrestricted gift from S. O'Donnell and B. O'Donnell to the Stanford Prevention Research Center. METRICS is supported by a grant by the Laura and John Arnold Foundation. The authors are grateful to Don van den Bergh for preparing Fig. 2.

Additional information

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence should be addressed to M.R.M.

How to cite this article: Munafò, M. R. *et al.* A manifesto for reproducible science. *Nat. Hum. Behav.* **1,** 0021 (2017).

Competing interests

M.R.M, together with C.D.C and D.V.M.B., has received funding from the BBSRC (grant number BB/N019660/1) to convene a workshop on advanced methods for reproducible science, and is chair of the CHDI Foundation Independent Statistical Standing Committee. B.A.N. is executive director of the non-profit Center for Open Science with a mission to increase openness, integrity and reproducibility of research. N.P.d.S. leads the NC3Rs programme of work on experimental design, which developed the ARRIVE guidelines and Experimental Design Assistant. J.J.W. is director, experimental design, at CHDI Management/CHDI Foundation, a non-profit biomedical research organization exclusively dedicated to developing therapeutics for Huntington's disease. The other authors declare no competing interests.