Directing scientists away from potentially biased publications: the role of systematic reviews in health care

Rossella Salandra [1, *], Paola Criscuolo [2], Ammon Salter [1]

[1] School of Management, University of Bath, Bath, UK

[2] Imperial College Business School, South Kensington Campus, London, UK

[*] Corresponding author E-mail address: r.salandra@bath.ac.uk (R. Salandra)

Abstract

Despite increasing concerns about the validity of published research, the issue of how the scientific community can maintain a high-quality body of research is not well understood. We consider the case of systematic reviews in health care, and explore whether risk of bias ratings communicated within these reviews may help shift scientists' attention towards published research that is at a low risk of bias. We focus on publications deemed at risk of bias due to selective reporting; that is, scientific articles with high chances of systematic errors in the published research findings due to flaws in the reporting. Using a matched-sample control group we find that, after potential bias is signalled in systematic reviews, publications at a high risk of bias attract less attention – as indicated by fewer follow-on citations – when compared to a control group of low risk of bias publications. We extend our analysis by considering those cases where risk of bias is unclear, and by examining how different features of the rating system may affect the magnitude of the main effect. The findings provide evidence about whether systematic reviews can play a role in signalling biases in the scientific literature, over and above their established role of synthesising prior research.

1 Introduction

Consideration of prior research, as indicated by the citation of relevant scientific publications, is essential to the conduct of new research (e.g., Dasgupta and David, 1994, Merton, 1973). Building new research cumulatively on previous works is made difficult by high quality research being diluted in the larger body of existing articles, which include less than perfect studies. As reviewed in a recent Special Issue in this journal, the quality of research can be compromised at various points in the production and dissemination of research results (Biagioli et al., 2018). Cases of deliberate fraud and data manipulation are often the most visible, yet more elusive or debateable practices have been proliferating. Although these 'liminal' forms of misrepresentation have not been classified as misconduct, they can be as problematic as the most severe practices (e.g., Hall and Martin, 2019, Biagioli et al., 2018).

Recent studies have drawn attention to systematic errors specifically introduced in the publication phase, and provided evidence that the dissemination of research in scientific journals is often incomplete (e.g., Fanelli, 2011, Fanelli et al., 2017, Franco et al., 2014). The occurrence of unscientific publication practices – including, but not limited to, the misreporting of true effect sizes ('p-hacking'), and hypothesizing after results are known (HARKing) – has been lamented across a number of fields e.g., psychology, management, economics and innovation studies (Necker, 2014, Fanelli, 2009, Bergh et al., 2017, Murphy and Aguinis, 2019, Head et al., 2015, Bettis, 2012, Harrison et al., 2017, Bruns et al., 2019, Halevi, 2020, Craig et al., 2020). Among others, these practices include selective reporting, which consists of including in the final publication only part of the findings originally recorded during a research study, on the basis of their direction or significance (Higgins and Green, 2011, Higgins et al., 2019). In clinical research, estimates suggest that at least half of studies are incompletely reported, in turn becoming unusable for follow-on science and clinical guidelines (Chalmers and Glasziou, 2009). Since reporting rules are often vague, selective reporting may be regarded more as an 'inappropriate' practice than a 'questionable' one (Hall and Martin, 2019). Yet, scientists rate it as the top factor contributing to irreproducible research (Baker, 2016), and members of the public see selective reporting as being immoral and deserving of punishment (Pickett and Roche, 2018).

In addition to obstructing scientists' ability to replicate past research findings (Allison et al., 2016, Collaboration, 2015, Fanelli, 2018), poor reporting is believed to seriously distort science (Young et al., 2008) and generate research waste (Glasziou et al., 2014, Chalmers and Glasziou, 2009). The risks carried by incomplete reporting are particularly high in the biomedical field, for example, in terms of the potential harm for human health.

While prior research has thoroughly documented the proliferation of selective reporting and other poor publication practices, little is known about how scientists can go about detecting them. Work on knowledge governance systems, notably retractions, suggests that the scientific community can detect signals and redirect research efforts away from inadequate or biased publications. For example, retracted papers are cited less than their counterparts following retraction notices (Furman et al., 2012, Lu et al., 2013, Azoulay et al., 2015b). However, retractions are not always fit-for-purpose, particularly in those cases where errors are not necessarily the outcome of deliberate misconduct, or where mistakes are not serious enough to lead to the invalidation of an entire article (Fang et al., 2012, Neale et al., 2010). Therefore, it is useful to investigate other systems that can provide signals about the quality of published research.

We focus here on the role of systematic reviews in health care, which summarise prior medical knowledge, such as randomised controlled trials of medical drugs, with the primary aim of informing clinical and policy decisions (e.g., Cook et al., 1997). Systematic reviews represent a particularly interesting case because most studies on publication and reporting bias have been conducted in the biomedical sciences, reflecting a high awareness for bias in this field (Bekelman et al., 2003, Lexchin et al., 2003, Dwan et al., 2008, Lee et al., 2008b, Dwan et al., 2013, Ross et al., 2008, Fleming et al., 2015). In addition, although the synthesis of research findings – for example, in literature reviews and meta-analyses – has enjoyed a growing interest in several fields (e.g., Aguinis et al., 2011a), systematic reviews in health care are unique because they routinely incorporate systems to appraise the quality of the included studies. An example of such systems is the assessment of a study's risk of bias due to selective reporting.

These features allow us to explore whether systematic reviews in health care can play a secondary role – over and above their established role in summarizing existing medical evidence – by providing a signal to detect biases, and flag them to the scientific community. We explore whether publications deemed at high risk of bias from selective reporting attract less relative attention (as measured by follow-on citations), compared to their low risk of bias counterparts, after potential biases are flagged in systematic reviews. We also examine whether the key features of this signal, or rating in this case, shape its effect on scientists' attention.

To tackle this question, we leverage evidence ratings presented in the reviews compiled by Cochrane, the most authoritative and comprehensive source of systematic reviews in health care (e.g., Jadad et al., 1998). We consider the publication of Cochrane reviews, which include a risk of bias assessment for all appraised studies. The risk of bias for a given study is judged

as being high, low, or unclear, and it is summarised making use of a traffic-light system – i.e. a red, green, or amber flag.

In line with prior literature on retractions, we employ a matched-sample control group, pairing articles deemed at high risk of bias due to selective reporting to similar papers deemed at low risk of bias, and quantify the impact of risk of bias ratings by comparing citation patterns for articles at high risk of bias to those of the matched publications. Our results indicate that systematic reviews provide key signals to follow-on researchers: in the main model specification, high risk of bias articles receive on average 7.9% fewer annual citations relative to their low risk of bias counterparts, following the publication of a Cochrane review. The investigation of citation dynamics after the treatment indicates that the citation effect is strongest in year 3 and 4 after the publication of a review. While our main sample compares the two furthermost categories – that is, high vs. low risk of bias ratings – in additional analyses, we also consider the cases flagged as being at unclear bias. We observe no significant citation effect for high risk of bias publications when compared to those at unclear risk. Low risk of bias publications, instead, receive on average 4.1 % more citations relative to their unclear risk of bias counterparts.

We also examine whether the effect of this quality signal is shaped by the modes of presentation of the risk of bias rating within a review, and by contextual factors such as the timing and subject area of the publication. We find that the effect is stronger when the risk of bias for the focal paper is deemed to be high along other bias domains considered by Cochrane, when the bias judgment is accompanied by a long explanatory comment, and when the review appraises a low number of articles. The effect is also concentrated among the most recent papers, and for papers reviewed within subject areas in which selective reporting bias is prevalent. These results suggest that the form and nature of the rating system's signal shape scientists' response to it.

By exploring the role of systematic reviews in health care in influencing the way scientists place new research in the context of prior publications, this study builds on prior research that has examined other knowledge governance systems, such as the system of retractions (Furman et al., 2012, Lu et al., 2013, Azoulay et al., 2015b). By helping us to investigate practical ways to improve the way scientists can detect publication errors and build upon prior work that is at least well reported, our findings inform the conversation on the detection and possible remedies for academic misconduct, misrepresentation and gaming particularly for the less easily defined, yet alarming, practices (Biagioli et al., 2018). Speaking to the ongoing debate regarding the quality of published research, these results also have

repercussions for important matters such as research waste (Glasziou et al., 2014, Chalmers and Glasziou, 2009), and the crisis of replication of research across various science fields (Allison et al., 2016, Aguinis et al., 2017).

2 Systems to govern the validity of published research

The proliferation of academic misconduct, misrepresentation and gaming, and the resulting publication errors, are posing growing threats to the reliability of the scientific literature (e.g., Biagioli et al., 2018). These issues beg the question of how should scientists detect biases and build new studies upon robust evidence, while trying to navigate across an overwhelming volume of scientific publications (e.g., Bornmann and Mutz, 2015).

From the standpoint of prevention, various initiatives have been put forward to reduce questionable research practices. These proposals, fundamentally aimed at increasing research transparency and reducing researchers' 'degrees of freedom', include, among several others, the promotion of reporting guidelines, data sharing, and the use of publication checklists (e.g., Nature, 2018, Ioannidis, 2014). The biomedical field appears to be more advanced than others in adopting initiatives to counter biases, as reflected in the use of research protocols and the preregistration of clinical trials – e.g., in ClinicalTrials.gov, a registry of clinical studies funded by the US National Institutes of Health (NIH). Journal editors have circulated and endorsed various reporting standards, such as CONSORT, a set of recommendations for the reporting of randomized trials (Moher, 2001). Yet, the implementation of reporting standards remains low (Turner et al., 2012, Péron et al., 2013). Adherence to data repository initiatives has also shown itself to be challenging (Viergever et al., 2014, Tang et al., 2015).

As far as the detection of publication flaws is concerned, the accuracy of published research is generally checked pre- as well as post-publication. The role of ensuring the validity of scientific publishing has traditionally been served by peer review, with checks and corrections before publication – e.g., at the submission or review stage – likely to be preferred to adjustments to published reports. However, the increase in number of submissions, together with the diffusion of open access and online journals (Arns, 2014), are placing an increasing burden on the peer review system, which largely relies on voluntary collaboration (Kovanis et al., 2016). The limits and vulnerability of peer review have also been highlighted by a surge in the number of retracted papers (Van Noorden, 2011, Fang et al., 2012).

In truth, even when the pre-submission and review processes are efficient, errors are still likely to be detected following publication, leading to the general acceptance that 'To err is human, to correct divine', as suggested in the title of a recent editorial in the *Journal of the*

American Medical Association (Christiansen and Flanagin, 2017). As such, monitoring and correction in the post-publication stage are increasingly recognised as being as important as pre-submission checks to improve reporting quality (Glasziou et al., 2014).

Various mechanisms are in place to correct research after publication, for example errata and retractions. Studies of retractions suggest that retraction notices are effective in signalling partial or incomplete science: publications that have been retracted receive less citations than similar papers following retraction notices (Furman et al., 2012, Lu et al., 2013, Azoulay et al., 2015b). While these results offer some reassurance on the ability of the scientific community to govern the validity of publications and correct mistakes, the current mechanisms may not always be appropriate. The system of retractions does not always work smoothly; for example some journals issue undetailed retraction notices – e.g., too vague for future researchers to assess possible effects on their work –, and some retractions are underused or poorly linked to the retracted publications (Neale et al., 2010). Also, retractions may not be suitable for all circumstances. Some errors, possibly those deriving from less severe publication practices such as selective reporting, may not be considered serious enough to invalidate a study's conclusions, in turn requiring that the publication is retracted. In addition, retraction notices are typically subject to legal overview with respect to the liabilities of the publishers and the authors, which means that notices are often written to avoid further legal complications arising from the retraction decision-making process itself, rather than to further the cause of science.

Given the limitations of the current systems to check the quality of published research, the question of whether other systems – e.g., other than retractions – can provide some signals regarding the validity of published research is worthy of investigation. This is particularly important to assist the scientific community in its attempts to deal with errors resulting from proliferating 'liminal' practices in research and publication e.g., selective reporting.

3 Systematic reviews in health care and bias assessment tools

Given that the search costs for scientists are invariably high, signals about the reliability and credibility of research may play an important role in shaping scientific efforts. Signalling theory suggests that signals can lower the uncertainty associated with selection where there is incomplete and asymmetric information about quality (Spence, 1973, Spence, 2002). In science, there are often information asymmetries between what scientists need to know, and what information is available to them in publications (Pavitt, 1987). In theory, scientists could lower this asymmetry by reading carefully each paper and making a judgement about its reliability – or by contacting the author for further information or data –, but this would

significantly increase their search costs. Given the volume of research is high, and scientists' time and attentional resources are limited, scientists tend to rely on credible signals, such as citations (Merton, 1973), journal impact factors (Baum, 2011, Seglen, 1992) or journal rankings (Drivas and Kremmydas, 2020), to lower their search efforts. They may also turn to systematic literature reviews for further help.

Systematic reviews are integrative publications that evaluate unambiguously formulated research questions, and use pre-planned methods to summarise and interpret data from the included studies (e.g., Cook et al., 1997). In health case, based on the rationale that clinical decisions should be based on the totality of evidence rather than a single study, review authors comprehensively search for appropriate articles and make use of reproducible criteria in the selection of publications to synthesise all prior available research on a given topic. Systematic reviews are a key building block of evidence-based medicine, a system aimed at grounding clinical decision-making in prior medical knowledge (Sackett et al., 1996, Guyatt et al., 2004). Over time, these documents have developed to become highly impactful, for example they are used to inform clinical recommendation, policy decisions, and potentially regulatory issues (Barbui et al., 2017).

At present, the most comprehensive repository of up-to-date systematic reviews of health care interventions is the Cochrane Database of Systematic Reviews (CDSR), which is the main source of data used in this study. More broadly, the use of literature reviews and meta-analyses to synthesize existing literature and draw evidence-based recommendations is widespread in many fields, and is enjoying increased popularity (e.g., in management studies, Aguinis et al., 2011a, Aguinis et al., 2011b).

Systematic reviews in health care are distinctive, as careful consideration of the potential limitations of the included studies – for example, flaws in design, conduct, analysis, and reporting – is one of their essential components. The rationale beyond this is that "the Achilles' heel of systematic reviews lies in publication and reporting bias" (Bouter, 2015, p.53); in other words, the extent to which meta-analytic estimates can reach reliable conclusions – e.g., making causal inferences about the effects of a medical intervention – naturally depends on the soundness of each of the individual studies included in the review.

Thus, historically, the improvement of methods for the integration of prior evidence – e.g., systematic reviews – has been accompanied by a need to also develop systems to rate the quality of the evidence that is collected (Guyatt et al., 1992, Guyatt et al., 2008, Atkins et al., 2004). One of the first evidence assessment tools, proposed in 1979 by the Canadian Task Force Periodic Health Examination, was based solely on study design, with randomised

controlled trials deemed of higher quality than case studies and expert opinion (Fletcher and Sackett, 1979). Since then, a number of methodologies have been suggested and used to classify clinical evidence (for a review, see Atkins et al., 2004), and many tools for assessing the quality of primary research are available, including scales, checklists, and the risk of bias assessment tool developed by Cochrane, which we leverage in this study.

The importance of systematic reviews to inform primary research – e.g., to identify research gaps, as well as to generate knowledge – has been widely acknowledged in medical research (Clarke et al., 2010, Bunn et al., 2015). For example, the medical journal *The Lancet* officially asks authors for reports of new research to place the results into the context of the whole body of evidence (Clark and Horton, 2010). Prior works looking at the role of systematic reviews to inform primary research have considered issues such as how many citations reviews can attract, and whether new trials set their conclusions in the context of a systematic review (Clarke et al., 2010, Bunn et al., 2015). Insights into the impact of evidence appraisal systems generally consider the effect of evidence appraisal documents as a whole – e.g., whether the recommendations from a certain systematic review have had an impact on practice – rather than the impact of bias assessment at the level of individual publications, such as whether a piece of research that has been judged as biased continues to attract research interest (e.g., Bunn et al., 2014). Little is known about whether quality ratings communicated in systematic reviews impact referencing.

Cochrane defines bias as a systematic error, "meaning that multiple replications of the same study would reach the wrong answer on average" (Higgins and Green, 2011). Thus, the identification of a high risk of bias indicates that a publication is subject to non-random errors that can potentially affect the extent to which the findings "should be believed" (Higgins and Green, 2011). As such, it is reasonable to ask whether bias identification in systematic reviews provides a signal of reporting quality and accuracy to future researchers.

In this study, we are interested in exploring whether and to what extent, over and above their established role – i.e., summarizing and interpreting existing evidence –, systematic reviews could be used as a practical tool to assist scientists in discriminating between publications that are well reported and those that cannot be fully relied upon due to reporting flaws. Drawing upon Connelly et al. (2001), we suggest that risk bias ratings in systematic reviews may act as a *pointing signal* that helps scientists to identify research of lower (reporting) quality, and separate it from other types of research.

4 Empirical strategy

We leverage a unique dataset of clinical research publications matched to expert-driven assessments of bias derived from the Cochrane Database of Systematic Reviews. Cochrane is a global independent network of researchers, and the leading provider of systematic reviews in health care. Cochrane reviews, which appraise the extant research on international public health priority themes, are recognised as the highest standard in evidence-based health care (e.g., Grimshaw, 2004), and are influential in shaping public health as well as scientific research.[1]

Cochrane authors search for all the existing primary research on a chosen topic and assess it using stringent criteria to establish whether there is convincing evidence – for example, to support the adoption of a certain medical intervention. Included studies are appraised based on various characteristics that may introduce a risk of bias in the published results, such as 'adequate sequence generation', blinding, and selective reporting. Risk of bias judgments (high, unclear, and low risk) relative to each of the assessed domains are reported in a table within each Cochrane review and visually displayed using a 'traffic–light' system, where green indicates a low risk of bias, amber is an unclear risk and red is a high risk of bias (Higgins and Green, 2011, Higgins et al., 2019). Section I of the Online Appendix reports an example of such summary tables.

To mitigate potential concerns that any effect on citations that we observe may be confounded by various factors influencing the scientific attention received by a publication (e.g., Tahamtan et al., 2016), we identify a carefully matched treatment and control group of articles. To define a precise estimate of the impact of bias detection on citations, we apply advanced econometric methods already established in the literature on retractions (e.g., Lu et al., 2013, Furman et al., 2012, Azoulay et al., 2015b)

4.1 Dataset

In the final months of 2019, we submitted a review data request form to the Cochrane Editorial and Methods department. We received N=7,717 Extensible Markup Language (XML) data files, containing detailed information on all Cochrane reviews published up until November 2019. If a review had been updated during its lifetime, we received information for its latest published version. We processed only the XML files for the reviews published from 2008 to 2016. We set this time period because the risk of bias tool was adopted as the recommended

[1] As of 2016, 90% of the World Health Organization guidelines contained Cochrane evidence. In 2018, the Cochrane Database of Systematic Reviews had an impact fact of 7.75 and was ranked 11th of the 160 journals in the Medicine, General & Internal category (Sources: https://www.cochrane.org/news/use-cochrane-reviews-inform-who-guidelines, https://www.cochrane.org/news/2018-journal-impact-factor-cochrane-database-systematic-reviews-7755, Accessed in September 2020).

method throughout the Cochrane Collaboration only after February 2008 (Higgins et al., 2011), and because we wanted to be able to observe at least three full years of citations following the publication of a review. The remaining 4,659 XML files were parsed and combined into a single database using Excel. 527 reviews had no quality ratings information and for 78 reviews we could not extract the references of the appraised studies, leaving us with 4,054 reviews.

We then extracted the references for all the papers included in the remaining reviews (N=56,553 publications). Given that the same publication may be evaluated in more than one Cochrane review we removed any duplicates and, in case of multiple publications associated to a certain study, we linked each trial to the reference identified by Cochrane as primary. Following these procedures, we were left with 48,562 unique references. We were able to find 69% (33,612) of these references in Elsevier's SCOPUS, a comprehensive database of peer-reviewed literature, which we used to obtain bibliometric information. We accessed the data using the package developed by Rose and Kitchin (2019). A comparison of these publications against the full dataset provided by Cochrane revealed no considerable differences across the two sets.[2]

To perform our difference-in-difference analyses based on the above sample, we had to ensure that at least some publications in a review had a selective reporting rating. This meant we had to exclude 748 reviews and the corresponding publications because none of the publications had received a rating for selective reporting. We also had to drop all those reviews where all publications had a low risk of selective reporting bias, and those with all publications with a high risk of bias. These additional exclusions resulted in a sample of 22,928 unique publications assessed in 2,020 Cochrane reviews, prior to the matching exercise discussed later.

4.2 Measures

Our main outcome variable is annual citations received by the included papers, as reported in SCOPUS. Citation data were extracted at the beginning of 2020, so for all articles we have

[2] To assess whether the drop in the number of articles might have introduced some selection bias, we tested whether the articles in this sample were different from those extracted from the main Cochrane dataset with respect to the year of publication, the impact factor of the journal, the Cochrane Editorial group and, most importantly, risk of bias due to selective reporting. We found that on average the articles in the final sample were slightly more recent (mean=2000.3 vs 1999.8, t-test=-4.86, p-value<0.01) and they were published in journals with lower impact factor (mean = 5.5 vs 6.5, t-test= 9.46, p-value<0.01). These findings are consistent with the notion that older publications may be less likely to be retrieved in Elsevier's SCOPUS or to be indexed in the Clarivate Analytics' Journal Citation Reports (JCR). Reassuringly, there was a similar proportion of articles with a high risk of bias due to selective reporting in the final sample as in the dataset obtained from Cochrane (proportion 13.6 vs 14.9, z-test=1.07, p-value=0.28). This suggests that articles with a high risk of bias are not over or under-represented in our sample.

annual citations received till the end of 2019. To identify risk of bias we relied on the Cochrane data, focussing on the risk of bias introduced by selective reporting. Cochrane risk of bias tool is structured into a set of domains of bias, investigating different aspects of trial design, conduct, and reporting (see, for example, Section I of the Online Appendix). For each study we extracted the Cochrane risk of bias judgments for the selective reporting item, which can take the values low risk, high risk, or unclear risk. [3]

We focused on selective reporting for many reasons. First, considering the broad range of misconduct and misrepresentation practices, the heterogeneity in their definitions, and the difficulties in detecting them (e.g., Biagioli et al., 2018), we limit our attention to questionable practices occurring specifically in the *publication* stage (as opposed to the research *conduct* or *review*). Cochrane's selective reporting domain tackles squarely these issues. The impact on follow-on citations driven by potential biased introduced during the design or the conduct of the trials (e.g., whether a trial is blinded) is outside the scope of our analysis because these domains are quite narrowly defined and rather specific to clinical research. A focus on selective reporting allows us to explore a phenomenon that goes beyond clinical trials. In addition, ample evidence indicates that selective reporting can influence study's findings e.g., statistically significant study outcomes are more likely to be published (Smyth et al., 2010). This is worrying, particularly when the outcomes being misreported concern potentially serious drug side effect. While important, other sources of bias may not have equally serious implications. Indeed, they may not indicate that a study is unreliable. As illustrated in the Cochrane Handbook (Higgins et al., 2011), blinding may simply not be feasible for certain trials – for example, in surgical procedures, participants would know if an operation has been performed. While we focus on selective reporting in the main analysis, we linked information on the risk of bias from the remaining domains in the heterogeneity analysis discussed later.

Leveraging information from SCOPUS, we added various paper characteristics such as year of publication, authors count, affiliations count, citations received on the first year and average yearly citations received. We also considered the authors' affiliations to identify any affiliated institutions. At the level of the journal, we attached the journal impact factor from Clarivate Analytics' Journal Citation Reports (JCR), using the average impact factor from 1997 (the first available year in the JCR dataset) to 2018. We used this information to assess the quality of our matching, and in the heterogeneity analysis.

[3] Reasons for a judgment of a high risk of bias include for example that an outcome of interest is partially reported (so that it cannot be entered in a meta-analysis), or that a study's publication does not report results for a key outcome that would be expected to have been measured for such study.

4.3 Statistical methods

Our core analysis focusses on the two categories of bias that are the farthest apart. Thus, our main sample compares scientific articles deemed at a high risk of bias due to selective reporting – identified by a 'red' flag in a Cochrane review – with a sample of control articles deemed at low risk of bias – those identified by a 'green' flag.

The control sample was based on a set of articles appraised in the same Cochrane review, published in same year of the focal article, and with a maximum journal impact factor difference of +/-3. While prior works on the effect of scientific institutions and governance control for citation trajectories by matching articles that appeared in the same journal (e.g., Furman et al., 2012), we focus on articles that appeared in the same Cochrane review. Two main reasons motivate this choice. First, studies appraised in the same review are highly comparable one another in terms of their underlying research issue. This is because Cochrane reviews address narrow research questions – e.g., ranging from examining clinical trials for the same drug, the same class of drugs, or all drugs used to treat a certain disease –, and review authors perform searches of the literature using precise search terms, for example, the chemical or code name of a drug. Second, Cochrane reviewers screen the studies identified by the search based on their methods or quality criteria. For example, they may select for inclusion only trials that are longer than two weeks, or only randomised studies. Taken together, matching publications within the same systematic review allows us to control for any differences in citation patterns and likelihood of bias that may derive from either the subject or methodological characteristics of studies. By selecting papers published in the same year as the treated paper, we implicitly control for factors associated with changes in citation patterns and bias propensity over time (Dechartres et al., 2011). Matching papers published in journals with a similar impact factor, we control for any differences in citations and reporting quality that may be associated to the impact of the journal the focal publication appears in. We matched the treated publications to up to two controls with replacement using a using a Mahalanobis distance measure (based on the journal impact factor, as all pairs would have the same year).[4]

Next, we compared citation patterns for the treated articles to those of the matched control sample following the publication of a Cochrane review. We employ a difference-in-difference analysis, allowing us to assess how the publication of a Cochrane review changes

[4] Replacement means that each control could be used as a control for several treatments. We excluded all instances in which a control was used more than two times. When a control was used more than twice, we included the treatment/control pairs with the smallest relative difference in the journal impact factor, and removed the other pairs.

the rate at which treated articles (i.e. papers at high risk of bias) are cited relative to those in the control group (those at low risk). We estimate the average impact of bias on the forward citations of an article by comparing yearly citations to article before and after the publication of a Cochrane review, controlling for article age. In line with prior work on retractions (Azoulay et al., 2015b, Lu et al., 2013), we tested the model using conditional quasi-maximum likelihood estimates based on the fixed-effect Poisson model (Hausman et al., 1984) clustering the standard errors around 'families' of treatment and control papers.

The estimator identifies the average change in citations that article i received in year t, resulting from the publication of a Cochrane review:

$$(1) \qquad CITES_{it} = f(\gamma_i, \delta_t, \beta POST_t, \psi POST_t \times TREATED_i, \varepsilon)$$

where γ_i is a fixed effect for each article, δ_t reflects the age of the publication, $POST_t$ is a dummy indicating whether year t is strictly after the year of publication of a Cochrane review, and $TREATED_i$ is a dummy indicating whether i is a treated paper – i.e. in our main analysis, a paper at high risk of bias. $POST_t \times TREATED_i$ is equal to one for treated articles only in the years after they have appeared in a Cochrane review. So, for an article published in 2005, but not rated at high risk of bias due to selective reporting until the publication of a Cochrane review in 2012, $POST_t \times TREATED_i$ equals zero in years 2005-2012, and one in the following years. As article fixed effects identify the mean number of annual citations received by each article over its lifetime, δ_t captures the average citation pattern over years, β captures aggregate factors that would cause changes in citations in the second time period even in the absence of the treatment, and ψ reflects the effect on follow-on citations induced by a 'high risk of bias' rating in a Cochrane review.

5 Results

The final sample for our main set ('HIGH vs LOW'), consists of 8,726 unique papers prior to the matching exercise. Out of these, 2,675 (30.7%) were deemed at a high risk of bias due to selective reporting. In Table 1, we report descriptive statistics for the key variables for this sample of publications. It is worth noting that, *before the matching*, the treated and the control samples are statistically significant different along several characteristics. Among others, treated papers have on average less cumulative citations prior to the publication of the Cochrane review, relative to the controls, indicating that high risk of bias publications may attract less citations even before the treatment. Altogether, the pre-matching descriptive

statistics indicate the need to carry out a careful matching process to find control papers that share the same citation pattern with the treatment papers in the pre-treatment period.

[Table 1 approximately here.]

The matching procedure left us with 1,067 articles, 468 at high risk of bias and 599 at low risk (the other 2,207 high risk publications had no counterfactual based on Cochrane review and year of publication). We further processed these data to make them amenable to statistical analysis. First, we excluded all articles published outside a time window running from ten years to one year before the publication of a Cochrane review. This procedure, also adopted in prior work on retractions (Azoulay et al., 2015b), was necessary because for many articles the publication of a Cochrane review occurred several years after the year the papers were published. It is also necessary to remove papers that were rated by Cochrane just after being published. This procedure reduced the sample substantially (from 1,067 to 625 articles). Second, we removed any citations outliers, defined as all treated papers that were on the top 10% percentile of citations prior to the publication of a Cochrane review.

After these procedures, we were left with 230 treated articles matched to 301 controls. As shown in Table 1, there are no considerable differences between the two groups in the post-matching sample. These papers were identified in 145 Cochrane reviews, published between 2008 and 2016, and covering a range of questions related to medical interventions. In Section II in the Online Appendix, we report the distribution of the publications in the matched sample across the different Cochrane Editorial groups.

5.1 Impact of bias ratings on citations

The results of the difference-in-difference analysis, aimed at assessing how the publication of a Cochrane review changes the rate at which articles at high risk of bias are cited relative to those in the low risk group, are shown in Table 2. We employ a quasi-ML Poisson estimator, and for each model we calculate robust standard errors, clustered by article 'family' – i.e. the treated paper plus up to two controls. The table reports the Incidence Rates Ratio (IRR) associated with each coefficient to allow for a direct interpretation of the effect: an IRR of one implies no effect on follow-on citations, whereas a coefficient equal to 0.90 implies a 10% less follow-on citations for the treated articles relative to the controls.

Column 2-1 reports the results for a model including all observations (i.e. where papers outside the [-10;-1] period and outliers are not excluded). Column 2-2 reports the results for a sample in which we exclude all observations outside the time period running from ten years to one year before the publication of a review. In Column 2-3, we report the results for our main sample, which excludes any papers published outside the selected time period and citations outliers. The coefficient on POST, which captures aggregate factors that would cause changes in citations after the publication of a Cochrane review even in the absence of the treatment, is lower than 1 (IRR=0.917). This is line with the expectation of a declining level of citations over time. We are interested in the coefficient on POST x TREATED, which in the 'HIGH vs LOW' sample reflects the effect on follow-on citations induced by a high risk of bias rating in a Cochrane review, relative to the low risk controls. The coefficient (IRR=0.921) implies that following a risk of bias rating, annual citations of high risk of bias articles are 7.9% less than those received by papers rated at low risk.

[Table 2 approximately here.]

While the coefficient on POST x TREATED in Table 2 indicates the average impact of risk bias ratings across all years after the treatment, it is also worth understanding whether this impact occurs as a discontinuity, or whether it induces declining levels of citation. To address these issues, Figure 1 reports the results of a regression in which we included separate interactions between TREATED and dummy variables for each year preceding and following the publication of a Cochrane review. All coefficients (IRRs) are relative to a window period running from one year prior to one year following the publication of a review. In the years prior to the publication of a Cochrane review, the coefficients are decreasing and not significantly different from one. Thus, in the pre-treatment period, we do not observe any significant effect on citations. After the publication of a Cochrane review, citations decrease by 12% by the fourth year for publications reported to have a high risk of bias, compared to the controls. The coefficients are significantly lower than one in year 3 and 4, and the uptick in year 5 suggests that the effect may fade out i.e. treated articles may not experience a long-lasting decline in the rate of citations received.[5]

[5] The fact that the citation penalty is transient may also be explained by negative citations. It could also be the case that citations to older papers "care less" about the specific findings and instead "care more" about the topical relevance of the paper. We thank one of the anonymous reviewers for suggesting these additional possibilities.

[Figure 1 approximately here.]

5.2 Robustness checks

We conducted additional tests to examine the stability of our main results. In line with prior work on retractions (e.g., Furman et al., 2012), we ran an OLS specification and conditional fixed effects negative binomial regression, with bootstrapped standard errors clustered by article. We find (see Section III in the Online Appendix) that OLS estimation and negative binomial regressions with bootstrapped standard errors yield qualitatively similar findings. Our main results are also unchanged when we remove self-citations from the dependent variable (this check is reported in Section IV of the Online Appendix).

A selection effect may confound our exploration of the impact of Cochrane ratings on citations if knowledge of lower intrinsic importance is endogenously embedded in papers that eventually received a high risk of bias rating. In our main analysis, we use article fixed effects to account for heterogeneity across matched article pairs: this allows us to precisely identify a treatment effect but not a potential selection effect. To disentangle the treatment from the selection effect, we developed a specification that includes "family" fixed effects (identifying the selection effect). The results for this model, reported in Section V in the Online Appendix, indicate that in this sample both the selection effect and the marginal treatment effect (over and above the selection effect) are negative and not statistically significant at conventional level.[6]

5.3 Citation effect relative to unclear risk of bias ratings

Our main sample compares scientific articles deemed at a high risk of bias (i.e. red) due to selective reporting with a sample of control articles deemed at low risk of bias (green). However, Cochrane reviews also make use of an unclear risk of bias rating category (amber), which indicates that there is insufficient information to permit judgement of low or high risk.

[6] To understand more how Cochrane review sends quality signal to scientist (e.g., whether the majority of citation changes is induced by those who actually read and cite a review) we checked whether articles citing the focal publication appraised in a review were also citing the review itself. After excluding citations to our main papers occurring before the year of publication of a review, we were left with 688,243 citations to a sample of 14,381 cited publications. Of these citations, only 59,001 (8.6%) were citing the focal Cochrane review. We were not surprised to see a relatively small number of citations to Cochrane reviews, because not all authors would add a reference to a systematic review, even if they might have read it. For those papers that did cite Cochrane reviews, we also checked the risk of bias rating of the cited studies. The expectation was that more of these instances would occur for citing articles drawn from the low risk of bias pool. Consistent with this scenario, out of the papers also citing the focal Cochrane review, the majority (53%) were citing papers that the review found to be at low risk of bias. We thank one of the reviewers for suggesting us to explore this.

Examining unclear risk of bias publications allow us to attempt to separate the citation effect that a favourable versus an unfavourable risk of bias rating may have relative to a common reference point (i.e. unclear risk). While it is reasonable to expect that high and low risk of bias ratings may provide clear-cut signals, there may be some ambiguity around how to interpret an unclear rating.

To explore this issue, we generated two additional matched samples that compare: i) articles deemed at a high risk of bias due to selective reporting with articles deemed at unclear risk ('HIGH vs UNCLEAR' sample), and ii) articles deemed at low risk of bias with a sample of control articles deemed at unclear risk ('LOW vs UNCLEAR' sample). As we do in the main set, we chose the controls among articles appraised in the same Cochrane review, published in same year of the treated article, and with a maximum journal impact factor difference of 3. Descriptive statistics for these two additional samples, before and after the matching procedure, are reported in Section VI of the Online Appendix.

[Table 3 approximately here.]

Table 3 shows the results of the difference-in-difference analysis across the three matched samples. Column 3-1 reports for convenience the results for our main sample ('HIGH vs LOW'). The results in Column 3-2 ('HIGH vs UNCLEAR') indicate that there is no significant citation effect for high risk of bias papers when compared to those at unclear risk (IRR=1.050, not statistically significant). The coefficient on POST x TREATED in 3-3 ('LOW vs UNCLEAR') implies that the effect on citations is significant and positive for low risk of bias papers, when compared to those at unclear risk. The IRR is equal to 1.041, thus we observe a 4.1% increase in citations for low risk of bias papers relative to those at unclear risk, after the treatment.

The lack of a net effect in the 'HIGH vs UNCLEAR' sample suggests that high risk and unclear risk of bias papers may not be seen as noticeably different. To the extent that a high risk of bias rating identifies papers that are associated to more severe forms of selective reporting relative to the unclear risk of bias ones, the results are directionally in line with prior work on retractions (Azoulay et al., 2015b) showing that the citation penalty is more severe when the focal articles retracted because of fraud or misconduct, relative to cases where the retraction occurred because of honest mistakes.

While it is the case that this sample is not contrasting the two extreme categories, the finding that the effect for the 'LOW vs UNCLEAR' sample is lower than that recorded in our

main sample is consistent with prior work (e.g., in consumer research) showing that negative information tends to be over emphasised and is more influential in forming impressions than positive information (e.g., Fiske, 1993). In the context of online holiday reviews, for example, negative reviews have been shown to have a greater impact than positive ones (Papathanassis and Knolle, 2011).

5.4 Effect of features of the rating system on citation behaviour

The results discussed in the previous section indicate that Cochrane ratings provide an important signal to follow-on scientists, and variations in the signal can shape the impact of the rating on follow-on citations – e.g., a 'green' rating has a positive effect, while a 'red' rating has a negative effect. Indeed, research on signals arising from rating systems in consumer research suggests that salient features in these rating systems shape peoples' response to these signals (e.g., Moore and Lafreniere, 2020). Building on this logic and evidence, we explore how the features of the Cochrane rating system itself shape the net effect on scientists' attention, suggesting that the way in which the information is presented in the rating system may amplify or dampen the effect on scientists' citation behaviour (see Table 4).

We start by focussing on three key salient features: i) the prevalence of 'red' flags for a given paper, ii) the length of the text commentary provided for each paper, and iii) the number of papers appraised in a review. We then examine the environment in which the focal paper appears in the literature, focusing on iv) whether the paper was published before or after the requirement to deposit clinical trial protocols, and v) the average prevalence of bias in the review area.

In the context of product ratings, consumers often evaluate a given product attribute depending on the values and availability of other attributes in the choice set (e.g., Watson et al., 2018). In our case, for each paper, we counted the number of high risk of bias ratings across all domains – e.g., recalling the format of the risk of bias table, how many 'red' flags there were in each row of the table. We then divided the sample based on whether the count was above or below the sample median. We expected that having many high risk ratings across all domains would reinforce the main citation effect linked to the rating on selective reporting bias. In the case of online consumer reviews, for example, the higher the proportion of negative reviews, the higher the perception of purchasing risk and the less favourable the product attitude (Lee et al., 2008a). Consistent with this scenario, the results in Table 4-1 and 4-2 indicate that the citation effect is higher among those papers where risk of bias is high across many bias domains (IRR=0.906, corresponding to a 9.4% decrease in citations).

Cochrane authors are encouraged to support their risk of bias judgements with a 'narrative explanation' (i.e. a text comment) of the evidence-based features known to increase the risk of bias. These comments are reported in Cochrane reviews and can, for example, include quotes from the paper that the reviewers have used to inform their judgements. While the interpretation of the length of a review comments is not straightforward – longer reviews may indicate more effort on the part of the reviewer, or that longer explanations are needed to support a 'mixed' review –, customers have been shown to read and respond to review content (e.g., when purchasing books online, Chevalier and Mayzlin, 2006). In the case of negative online consumer reviews, consumer attitudes toward a product become less favourable when that quality of a review – as measured by its relevance, reliability, understandability and sufficiency – is high (Lee et al., 2008a). Our findings in 4-3 and 4-4 indicate that in our sample, papers with a comment longer than the sample median were subject to a larger drop on citations compared to the others (IRR=0.854, corresponding to a 14.6% drop). This suggests that longer explanatory comments may increase the level of detail (e.g., regarding which study outcomes were missing in the publication) and the degree of ease with which the rating can be understood. Longer comments may also indicate more problematic cases.

Consumer research also shows that processing capacity can become cognitively overloaded if consumers attempt to process too much information, and this can result in the inability to locate what is relevant, and overlooking of critical aspects (Malhotra, 1984). In the case of Cochrane ratings, which are summarised in a table, a high number of lines in the table may also increase visual complexity, which has been found to correlate negatively with affective valence (e.g., in websites, Tuch et al., 2009). Accordingly, we tested whether the number of papers appraised in the review had any effect on the impact of a high-risk rating on citations. We divided the reviews based on the number of articles they included (below and above the median). Consistent with our expectations, the results in 4-5 and 4-6 indicate that most of the negative effects occurred in the low traffic review (14.6%), while papers in the high traffic reviews experienced no significant decrease in citations.[7]

[Table 4 approximately here.]

[7] A large number of papers could also characterise reviews in research areas that are more crowded. As such, these results may also reflect the degree of scientific competition within a subject area. Interestingly, Azoulay et al. (2015b) found that the impact of retraction on the citations received by related papers was stronger in "hot fields" (defined as those in which a high proportion of related articles appear with the retracted articles).

Looking at the environment in which the paper was published, we first explored whether the impact on citations varied for most recent versus older papers. For this analysis, we set as a threshold the year 2005, the start year of the International Committee of Medical Journal Editors (ICMJE) policy aimed at promoting registration of all clinical trials. The policy, stating that editors would consider a trial for publication only if it had been registered before the enrolment of the first patient, is considered as an historical turning point for clinical research transparency. In Table 5, the results in columns 5-1 and 5-2 indicate that recent articles in our sample experienced a higher citation penalty (IRR = 0.887, corresponding to a 11.3% drop), whereas older articles were relatively immune to the rating. These findings indicate that scientists might expect papers published after 2005 to achieve a higher "bar" for the underlying quality and reproducibility, and in turn apply a greater citation penalty when bias is detected.

Second, we considered whether the review belonged to a Cochrane Editorial group where bias is prevalent. For this analysis we considered the percentage of high risk of bias (due to selective reporting) publications by Editorial group – this percentage varied broadly across groups, ranging from 1.2% for studies in the Fertility Regulation Group, to 37.7% for studies appraised in the Schizophrenia Group. We then distinguished the papers from groups where prevalence of bias was higher than the sample median, and the others. Consistent with the idea that a stigma may attach to research areas in which bias is known to be prevalent, we expected that scientists would apply a great citation penalty in those fields where bias is more prevalent. Our results in 5-3 and 5-4 show that the citation effect is stronger for papers published in those subject areas where selective reporting is highly prevalent (IRR= 0.869).

[Table 5 approximately here.]

6 Discussion and conclusion

The proliferation of a range of poor research practices is widely seen to be seriously affecting the scientific literature, leading to research waste across many disciplines. Despite these issues, the current systems intended to counter biases are subject to several important limitations. Against this backdrop, we explored the role of systematic reviews in health care in signalling bias from selective reporting. In our main sample, we found a 7.9% relative decrease in annual citations for articles at high risk of bias due to selective reporting, following the publication of bias ratings in a Cochrane review, and compared to a control group of low risk of bias papers. Post-treatment citation dynamics suggest that the effect fades out relatively quickly, with most of the impact being recorded in year 3 and 4 following the publication of a review. In further

analyses, papers rated positively (i.e. as being at low risk of bias) were found to experience a 4.1% increase in citations relative to papers where risk of bias was unclear.

We also investigated when this signalling effect may be weaker or stronger due to different features of the rating system. Drawing from consumer research (e.g., on behavioural responses to online ratings), we found that the citation effect was strengthen when a paper had a high bias of risk along many bias domains, when the text comment attached to the rating was long, and when the number of papers included in the review was low. We found evidence that the context of the rating, such as the vintage of the article and the prevalence of bias in the Editorial group the review belongs, also matters in modulating the main effect.

6.1 Implications for research and policy

This study's findings have implications for countering bias resulting from poor publication practices, increasing replicability of research, and reducing research waste (e.g., Chalmers and Glasziou, 2009, Baker, 2016, Biagioli et al., 2018). Past research has examined various institutional mechanisms in place to correct published research, especially retractions (Furman et al., 2012, Lu et al., 2013). This study builds upon these works but it broadens the focus by examining (i) 'borderline' practices, and (ii) the role of systemic reviews in detecting biases. While retractions largely deal with cases of academic misconduct, we consider here the case of errors introduced by selective reporting. Although the prevalence of poor reporting and risk of bias has been extensively evaluated in prior work, particularly so in the context of clinical research (Hutton and Williamson, 2000, Chan et al., 2004, Chan and Altman, 2005, Dechartres et al., 2017), to our knowledge this study is the first attempt to investigate how scientists react, via the use or avoidance of citations, to the information that a publication has a high or low risk of bias due to selective reporting. Our investigation of potential solutions to detect and counter bias arising from selective reporting contributes to recent works reflecting upon the current status of the academic literature, which highlight that selective reporting – and other questionable or inappropriate research practices – are on the rise, potentially highly damaging, and difficult to tackle (e.g., Hall and Martin, 2019, Biagioli et al., 2018). Our research suggests that ratings of bias might act as a signal to redirect and shift the attention of scientists.

A second novel element of our study is examining the working of systematic reviews in health care. By focusing on whether and how risk of bias assessments reported in these reviews can direct citations away from work that may be subject to reporting errors (i.e. articles at high risk of bias) or towards well-reported research (i.e. articles at low risk of bias), our study goes

beyond prior works that have largely been looking at the role of reviews to inform clinical decisions and primary research (Clarke et al., 2010, Bunn et al., 2015).

From a policy perspective, the finding that systematic reviews in health care, originally devised as tools to inform clinical decision-making, could also serve the purpose of directing researchers' attention towards publications that are better reported, is promising, suggesting a secondary role for these reviews in shaping scientific attention away from biased towards reliable research. Other fields might benefit from developing similar post-publication review mechanisms to signal to scientists where problems lie in past research. However, it must be said that the citation effect associated with a bias rating is rather modest when compared to that observed for retractions, where the effect has been estimated to be large, immediate and persistent: citations decrease by more than 50% in year 2 after retraction and as much as 72% in year 10 (Furman et al., 2012). This might have been expected, for many reasons: retractions invalidate the content of the article, are much more targeted (e.g., they focus on one publication only), and happen relatively quickly after the publication of an article. Scientists would need to read a Cochrane review to became aware of a risk of bias rating. Our findings also indicate that the effect is contingent: in our supplementary analysis, we can only observe an effect when the risk of bias rating has selected characteristics, and when certain contextual circumstances are in place (e.g., in certain disease domains). As such, the risk of bias rating appears to be a relatively weak signal to the scientific community and might need to be complemented with other mechanisms to ensure that scientists' build on reliable research.

Despite these limitations, it is promising that the use of data already collected in systematic reviews can potentially assist scientists in building follow-on research relying more on robust science, and less so on biased publications. Unlike the systems of retractions, systematic reviews were not formally designed to signal bias, or correct the literature. As such, our study shed light into the detection of bias as a supplementary, somewhat unlooked-for, function of systematic reviews. Our findings indicate that the effort and expertise of review authors, which we agree are "probably the most critical readers of scientific articles" (Bouter, 2015), can be leveraged to contribute to the detection of poor reporting. The examination of systematic reviews as a potential solution to tackle mild forms of misrepresentation is especially interesting in light of recent suggestions that policies aimed at less severe practices, such as selective reporting, may also reduce the appeal of resorting to the most serious forms, such as outright fraud (Gall and Maniadis, 2019).

Transferring into as many scientific disciplines as possible research practices that have worked efficiently when applied elsewhere has been suggested as a viable option to make

science better (e.g., Ioannidis, 2014). Although systematic reviews are mainly applied in health care, their principles can be extended in any field of research; for example, the Campbell Collaboration (https://campbellcollaboration.org) is endorsing the use of systematic reviews in policy-making. Consistent with the results of our study, one possible suggestion would be to integrate appropriate bias judgments in meta-analyses across various disciplines. Review articles and meta-analyses are just two types of integrative publication: guidelines and economic evaluations are others, which could include quality ratings of primary research. Thus, our results also beg the question of whether these other integrating documents may also play a role in signalling research quality. The findings from our heterogeneity analysis suggest that the modes of presentation of the rating matter, and that carefully thinking should go into the design of the rating (e.g., how summary characteristics are displayed). The danger is that weak signals may not provide sufficient warning to scientists to divert their attention away from unreliable science.

Finally, our findings have potential implications for clinical practice. Physicians' ratings of the clinical relevance of publications have been shown to correlate with their citation counts (Lokker et al., 2008); so it may be that studies that are poorly reported, after being flagged in Cochrane reviews, may receive less citations and also, in turn, receive less attention from physicians. As commented by Biagioli et al. (2018), our results may also suggest that "the damage to ongoing science may not be that severe, unless the retracted publications involved medical and therapeutic claims that have been adopted prior to the retraction". It is indeed the case that many of the publications included in our sample were appraised by Cochrane years after the market launch of the drug investigated in the underlying trial. Dedicating resources to systematic reviews to ensure they are conducted promptly and continuously in fast-moving and complex research areas might help to ensure that researchers are given an 'early warning' to studies at high risk of bias, before they become embedded in the literature and into clinical practice. Indeed, one option would be to treat post-publication bias scores from systemic reviews as a quality 'kite mark', stamped on the research outputs themselves as retraction notices currently are. This might ensure that when these publications are read, scientists and medical professionals are adequately warned about the potential biases involved in that study.

6.2 Limitations and future research

This study has some limitations that call for caution in the interpretation of results. Our study is unique in that we use data extracted from Cochrane reviews. The inclusion of Cochrane expert-driven assessments of bias allowed us to overcome potential issues around the

identification of bias and the heterogeneities in its definition (e.g., Dechartres et al., 2011). Considering articles appraised within the same review enabled us to adopt matching criteria that are precise, and stricter than looking at the broad scientific field alone, or at the journal of publication. However, with our difference-in-difference estimation, we observe the citation pattern of papers at high risk of bias relative to the citation trajectory of papers with low risk of bias. Since the latter also receive a treatment, more work is needed to understand exactly what drives the net citation effect that we observe.

While we start with a large dataset, the three samples used in our analysis are relatively modest, due to the relatively low incidence of selective reporting in the sample and the desire to ensure a high quality match. In effect, we traded-off power in our statistical tests with precision by only including in our control sample articles that were not only published in the same systematic review, but also in the same year of the treated article, and in journals with a similar impact. Other studies might relax these criteria to include a wider range of studies in the comparison set. Although our main model specification – which, using article fixed effects, accounts for heterogeneity across matched article pairs – allows us to identify the marginal effect of Cochrane rating on follow-on citations precisely, future research could explore potential selection effects (e.g., the selectivity of Cochrane reviews) in more detail.

Several other questions that require investigation emerge from this study. First, in line with past work exploring how the scientific community's perception of a scientist's prior work changes when one of their articles is retracted (Azoulay et al., 2015a, Lu et al., 2013), it would be useful to investigate the impact of bias detection on scientists' careers. It may be that the citation penalty extends beyond the authors, as it is the case in the study of retractions by Hussinger and Pellens (2019). Second, to gain more insight into the reasons why scientists continue to cite papers at high risk of bias, prospective studies could complement our analysis with qualitative analyses of citation behaviour. For example, given that reviews are used to inform regulatory recommendations (Barbui et al., 2017, Bunn et al., 2015), whether potentially biased studies continue to underpin clinical guidelines and continue to be cited in policy documents (e.g., documents produced by organisations such as the WHO) may be a question worthy of investigation. It would be interesting to identify in our sample selective citations (e.g., Duyx et al., 2017), as well as 'negative' citations (e.g., references made to point out limitations or to question the validity of previous results, Catalini et al., 2015). Third, to complement the investigation of traditional metrics such as citations, future research could look at the impact of bias on other measures of impact. For instance, one could look at alternative metrics (Altmetric), which track attention to research outputs from online sources such as news

outlets, social media, and policy documents. Fourth, while this study focuses on the role of systematic reviews in signalling the presence of flaws in the reporting of research, future work could investigate the citation impact deriving from other bias domains (e.g., methodological issues introduced in the conduct of research). In addition, as the effect of questionable research practices can affect a range of stakeholders (Hall and Martin, 2019), examining the impact of bias on medical practice (for example, on drug prescriptions) may be an interesting additional dimension in this area of research.

We hope that despite these limitations, this paper will help to spur research on the effect of systemic reviews on the direction of science, helping to inform our understanding of how science can reform itself to ensure its integrity and reliability.

Acknowledgements

Early versions of this paper were presented at DRUID18, at the 12th Workshop on the Organization, Economics and Policy of Scientific Research, and at Evidence Live 2018. We are thankful for comments received at these events, and grateful to Cochrane Editorial and Methods Department for providing access to the data. We are also indebted to Stefano Baruffaldi, Ruxandra Luca and John Walsh. R. Salandra gratefully acknowledges the financial support received by the UK Engineering and Physical Sciences Research Council (EPSRC) [grant number EP/K502856/1].

Tables

Table 1: Descriptive statistics for treated and control articles: 'HIGH vs LOW' sample

	Before matching						After matching					
	High risk, n=2,675		Low risk, n=6,051		t-test (z-test)	p-value	High risk, n=230		Low risk, n=301		t-test (z-test)	p-value
Variables	Mean	Std. Dev.	Mean	Std. Dev.			Mean	Std. Dev.	Mean	Std. Dev.		
Root article characteristics												
Year of publication	1999.7	10.46	2002.2	9.13	11.18	0.00	2008.4	2.95	2008.5	2.95	0.49	0.62
Authors count	6.28	5.18	6.74	5.67	3.64	0.00	6.23	3.32	6.47	3.10	0.82	0.41
Affiliations count	2.58	2.81	2.88	3.08	4.36	0.00	2.64	2.24	2.94	2.65	1.38	0.17
Citations characteristics												
Average yearly citations received	4.80	8.82	7.11	20.13	5.69	0.00	3.29	2.40	3.48	2.19	0.97	0.33
Citations received on the first year	4.61	9.26	6.99	20.24	5.80	0.00	3.49	3.82	3.58	3.41	0.31	0.76
Cumulative citation prior to Cochrane review	69.78	140.53	97.72	544.45	2.61	0.01	18.09	18.53	19.25	18.00	0.73	0.47
Journal characteristics												
Journal JCR (mean over 1997-2018)	4.68	6.60	5.91	8.67	6.39	0.00	2.63	1.98	2.77	1.78	0.83	0.41
Authors characteristics												
% authors affiliated to university	0.67	1.00	0.72	1.00	2.09	0.04	0.74	1.00	0.81	1.00	0.75	0.45
% authors affiliated to top 100 pharma firm (incl. sponsorship) (i)	0.10	1.00	0.08	1.00	-0.94	0.35	0.08	1.00	0.06	1.00	-0.25	0.80
Trial characteristics												
Blinding bias (Share of high and unclear risk)	0.60	1.00	0.53	1.00	-3.04	0.00	0.62	1.00	0.58	1.00	-0.39	0.70
Other bias (Share of high and unclear risk)	0.63	1.00	0.44	1.00	-7.04	0.00	0.60	1.00	0.50	1.00	-0.93	0.35
Other domains - excl. sel. reporting (Share of high and unclear risk)	0.67	1.00	0.52	1.00	-6.27	0.00	0.60	1.00	0.52	1.00	-0.86	0.39
Other domains - excl. sel. reporting (Share of high risk)	0.21	1.00	0.16	1.00	-2.04	0.04	0.20	1.00	0.16	1.00	-0.39	0.70

(i)Articles affiliated to private firms were identified by looking at whether any of the listed affiliations included suffixes such as Inc., Corp., LLC, Ltd., GmbH, etc. or were found in Informa Pharma Intelligence's Scrip 100 list, which includes the top 100 pharma companies by drug sales. Sponsorship information was derived from SCOPUS.

Table 2: Average impact of bias (detection) on follow-on citations: Main analysis

	(2-1)	(2-2)	(2-3)
	Entire sample	Excludes papers outside the [-10;-1] period	Further excludes outliers
POST	0.850***	0.922**	0.917**
	(0.0336)	(0.0376)	(0.0376)
POST x TREATED	1.024	0.953	0.921*
	(0.0556)	(0.0408)	(0.0410)
Number of Article-Year Obs.	17,200	7,326	6,135
Number of articles	1,067	625	531
Paper FEs	Yes	Yes	Yes
Age FEs	Yes	Yes	Yes

Incidence-rate ratios obtained using conditional quasi-maximum likelihood Poisson models. Dependent variable: number of citations received by each treated article in a particular year. Treated equal to 1 if publication is deemed at a high risk of selective reporting bias, 0 if at low risk of bias. Standard errors in parentheses clustered by article family.

*** Significant at the 1 percent level.

** Significant at the 5 percent level.

* Significant at the 10 percent level.

Table 3: Average impact of bias (detection) on follow-on citations: Different bias ratings

	(3-1) 'HIGH vs LOW' sample	(3-2) 'HIGH vs UNCLEAR' sample	(3-3) 'LOW vs UNCLEAR' sample
POST	0.917**	0.854***	0.879***
	(0.0376)	(0.0469)	(0.0131)
POST x TREATED	0.921*	1.050	1.041**
	(0.0410)	(0.0523)	(0.0167)
Number of Article-Year Obs.	6,135	3,904	21,037
Number of articles	531	325	1,815
Paper FEs	Yes	Yes	Yes
Age FEs	Yes	Yes	Yes

Incidence-rate ratios obtained using conditional quasi-maximum likelihood Poisson models. Dependent variable: number of citations received by each treated article in a particular year. In all samples, we exclude articles published outside a time period running from ten years to one year before the publication of a Cochrane review. We also remove any citations outliers, defined as all treated papers that were in the top 10% percentile of citations prior to the publication of a Cochrane review.

*** Significant at the 1 percent level.

** Significant at the 5 percent level.

* Significant at the 10 percent level.

Table 4: Average impact of bias (detection) on follow-on citations: Heterogeneity analysis (Features of the rating)

		Risk of bias		Comment length		Review traffic	
	(4-0)	(4-1)	(4-2)	(4-3)	(4-4)	(4-5)	(4-6)
	Entire sample	High risk across other dimensions	Low risk across other dimensions	Long comment	Short comment	High number of papers	Low number of papers
POST	0.917**	0.885**	0.918	0.968	0.879**	0.908*	0.905
	(0.0376)	(0.0521)	(0.0527)	(0.0654)	(0.0486)	(0.0528)	(0.0574)
POST x TREATED	0.921*	0.906*	0.943	0.854*	0.950	0.973	0.854***
	(0.0410)	(0.0489)	(0.0724)	(0.0738)	(0.0482)	(0.0633)	(0.0521)
Number of Article-Year Obs.	6,135	3,907	2,228	1,849	4,286	3,143	2,992
Number of articles	531	330	201	164	367	273	258
Paper FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Age FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Incidence-rate ratios obtained using conditional quasi-maximum likelihood Poisson models. Dependent variable: number of citations received by each treated article in a particular year. Treated equal to 1 if publication is deemed at a high risk of selective reporting bias, 0 if at low risk of bias. Standard errors in parentheses clustered by article family. We exclude articles published outside a time period running from ten years to one year before the publication of a Cochrane review. We also remove any citations outliers, defined as all treated papers that were in the top 10% percentile of citations prior to the publication of a Cochrane review.

*** Significant at the 1 percent level.

** Significant at the 5 percent level.

* Significant at the 10 percent level.

Table 5: Average impact of bias (detection) on follow-on citations: Heterogeneity analysis (Context of the rating)

		Pre vs Post 2005		Prevalence of bias in the Editorial group	
	(5-0)	(5-1)	(5-2)	(5-3)	(5-4)
	Entire sample	Article published after December 2005	Article published before December 2005	High prevalence of bias	Low prevalence of bias
POST	0.917**	0.915**	0.770*	0.860**	0.923
	(0.0376)	(0.0361)	(0.110)	(0.0516)	(0.0545)
POST x TREATED	0.921*	0.887**	1.105	0.869*	0.927
	(0.0410)	(0.0421)	(0.129)	(0.0667)	(0.0474)
Number of Article-Year Obs.	6,135	4,739	1,396	2,454	3,681
Number of articles	531	444	87	213	318
Paper FEs	Yes	Yes	Yes	Yes	Yes
Age FEs	Yes	Yes	Yes	Yes	Yes

Incidence-rate ratios obtained using conditional quasi-maximum likelihood Poisson models. Dependent variable: number of citations received by each treated article in a particular year. Treated equal to 1 if publication is deemed at a high risk of selective reporting bias, 0 if at low risk of bias. Standard errors in parentheses clustered by article family. We exclude articles published outside a time period running from ten years to one year before the publication of a Cochrane review. We also remove any citations outliers, defined as all treated papers that were in the top 10% percentile of citations prior to the publication of a Cochrane review.

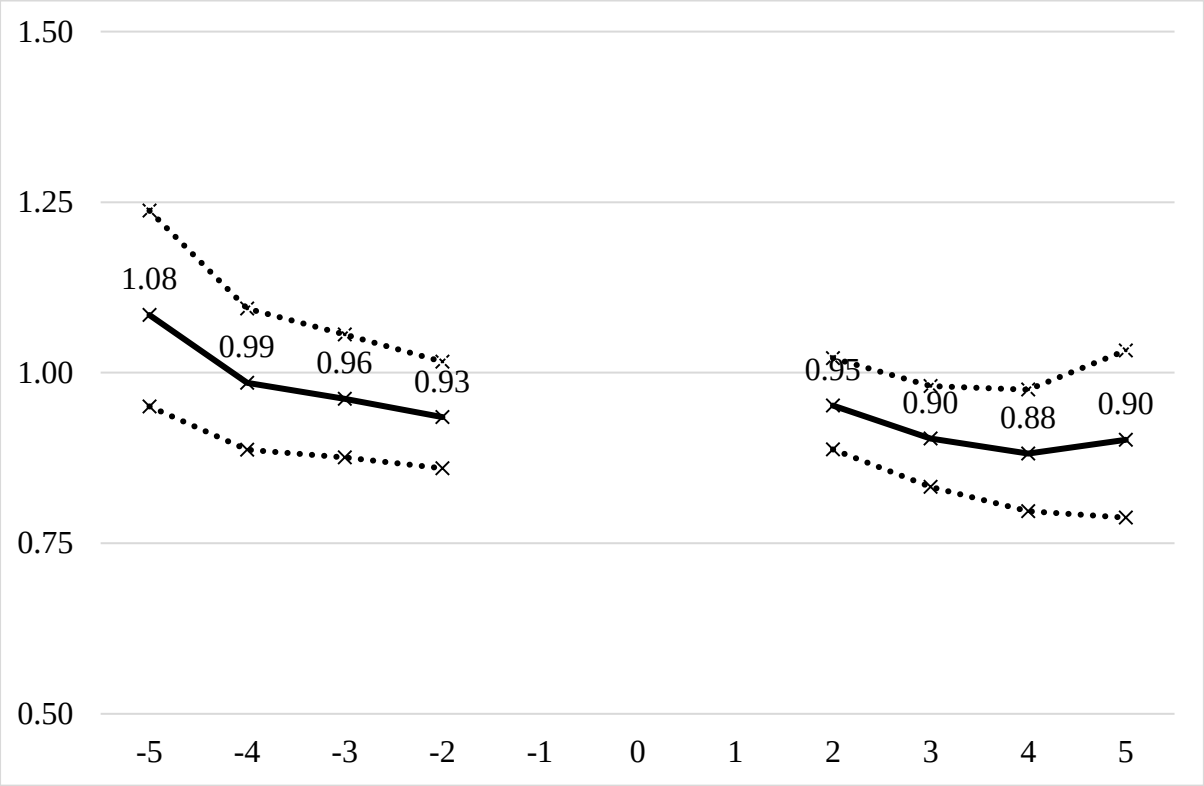
*** Significant at the 1 percent level.

** Significant at the 5 percent level.

* Significant at the 10 percent level.

Figures

Figure 1 - Pre- and Post-Bias detection effect on follow on citations ('HIGH vs LOW' sample)



The solid line reflects IRR from regression (conditional quasi-maximum likelihood Poisson specification) containing separate interactions between TREATED and dummy variables for each year preceding and following the publication of a Cochrane review, along with age fixed effects. All effects are computed relative to the window period (-1,0,1). The dotted lines represent 90% confidence intervals, based on robust standard errors, adjusted for clustering by article family.

References

AGUINIS, H., DALTON, D. R., BOSCO, F. A., PIERCE, C. A. & DALTON, C. M. 2011a. Meta-analytic choices and judgment calls: Implications for theory building and testing, obtained effect sizes, and scholarly impact. *Journal of Management,* 37**,** 5-38.

AGUINIS, H., PIERCE, C. A., BOSCO, F. A., DALTON, D. R. & DALTON, C. M. 2011b. Debunking myths and urban legends about meta-analysis. *Organizational Research Methods,* 14**,** 306-331.

AGUINIS, H., RAMANI, R. & ALABDULJADER, N. 2017. What You See is What You Get? Enhancing Methodological Transparency in Management Research. *Academy of Management Annals***,** annals. 2016.0011.

ALLISON, D. B., BROWN, A. W., GEORGE, B. J. & KAISER, K. A. 2016. Reproducibility: A tragedy of errors. *Nature,* 530**,** 27.

ARNS, M. 2014. Open access is tiring out peer reviewers. *Nature News,* 515**,** 467.

ATKINS, D., ECCLES, M., FLOTTORP, S., GUYATT, G. H., HENRY, D., HILL, S., LIBERATI, A., O'CONNELL, D., OXMAN, A. D. & PHILLIPS, B. 2004. Systems for grading the quality of evidence and the strength of recommendations I: critical appraisal of existing approaches The GRADE Working Group. *BMC Health Services Research,* 4**,** 38.

AZOULAY, P., BONATTI, A. & KRIEGER, J. L. 2015a. The career effects of scandal: Evidence from scientific retractions. National Bureau of Economic Research.

AZOULAY, P., FURMAN, J. L., KRIEGER, J. L. & MURRAY, F. 2015b. Retractions. *Review of Economics and Statistics,* 97**,** 1118-1136.

BAKER, M. 2016. REPRODUCIBILITY CRISIS? *Nature,* 533**,** 26.

BARBUI, C., ADDIS, A., AMATO, L., TRAVERSA, G. & GARATTINI, S. 2017. Can systematic reviews contribute to regulatory decisions? *European journal of clinical pharmacology,* 73**,** 507-509.

BAUM, J. A. 2011. Free-riding on power laws: questioning the validity of the impact factor as a measure of research quality in organization studies. *Organization,* 18**,** 449-466.

BEKELMAN, J. E., LI, Y. & GROSS, C. P. 2003. Scope and impact of financial conflicts of interest in biomedical research. *JAMA: the journal of the American Medical Association,* 289**,** 454-465.

BERGH, D. D., SHARP, B. M., AGUINIS, H. & LI, M. 2017. Is there a credibility crisis in strategic management research? Evidence on the reproducibility of study findings. *Strategic Organization,* 15**,** 423-436.

BETTIS, R. A. 2012. The search for asterisks: Compromised statistical tests and flawed theories. *Strategic Management Journal,* 33**,** 108-113.

BIAGIOLI, M., KENNEY, M., MARTIN, B. & WALSH, J. 2018. Academic misconduct, misrepresentation and gaming: A reassessment.

BORNMANN, L. & MUTZ, R. 2015. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology,* 66**,** 2215-2222.

BOUTER, L. M. 2015. Commentary: Perverse incentives or rotten apples? *Accountability in research,* 22**,** 148-161.

BRUNS, S. B., ASANOV, I., BODE, R., DUNGER, M., FUNK, C., HASSAN, S. M., HAUSCHILDT, J., HEINISCH, D., KEMPA, K. & KÖNIG, J. 2019. Reporting errors and biases in published empirical findings: Evidence from innovation research. *Research Policy,* 48**,** 103796.

BUNN, F., TRIVEDI, D., ALDERSON, P., HAMILTON, L., MARTIN, A. & ILIFFE, S. 2014. The impact of Cochrane Systematic Reviews: a mixed method evaluation of outputs from Cochrane Review Groups supported by the UK National Institute for Health Research. *Systematic reviews,* 3**,** 125.

BUNN, F., TRIVEDI, D., ALDERSON, P., HAMILTON, L., MARTIN, A., PINKNEY, E. & ILIFFE, S. 2015. The impact of Cochrane Reviews: a mixed-methods evaluation of outputs from Cochrane Review Groups supported by the National Institute for Health Research. *Health Technol Assess,* 19**,** 1-100.

CATALINI, C., LACETERA, N. & OETTL, A. 2015. The incidence and role of negative citations in science. *Proceedings of the National Academy of Sciences,* 112**,** 13823-13826.
CHALMERS, I. & GLASZIOU, P. 2009. Avoidable Waste in the Production and Reporting of Research Evidence. *Lancet,* 374**,** 86-89.
CHAN, A.-W. & ALTMAN, D. G. 2005. Identifying outcome reporting bias in randomised trials on PubMed: review of publications and survey of authors. *bmj,* 330**,** 753.
CHAN, A.-W., HRÓBJARTSSON, A., HAAHR, M. T., GØTZSCHE, P. C. & ALTMAN, D. G. 2004. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. *Jama,* 291**,** 2457-2465.
CHEVALIER, J. A. & MAYZLIN, D. 2006. The effect of word of mouth on sales: Online book reviews. *Journal of marketing research,* 43**,** 345-354.
CHRISTIANSEN, S. & FLANAGIN, A. 2017. Correcting the medical literature:"To err is human, to correct divine". *Jama,* 318**,** 804-805.
CLARK, S. & HORTON, R. 2010. Putting research into context—revisited. *The Lancet,* 376**,** 10-11.
CLARKE, M., HOPEWELL, S. & CHALMERS, I. 2010. Clinical trials should begin and end with systematic reviews of relevant evidence: 12 years and waiting. *The Lancet,* 376**,** 20-21.
COLLABORATION, O. S. 2015. Estimating the reproducibility of psychological science. *Science,* 349**,** aac4716.
CONNELLY, B. L., CERTO, S. T., IRELAND, R. D. & REUZEL, C. R. 2001. Signaling Theory: A Review and Assessment. *Journal of Management,* 37**,** 39-67.
COOK, D. J., MULROW, C. D. & HAYNES, R. B. 1997. Systematic reviews: synthesis of best evidence for clinical decisions. *Annals of internal medicine,* 126**,** 376-380.
CRAIG, R., COX, A., TOURISH, D. & THORPE, A. 2020. Using retracted journal articles in psychology to understand research misconduct in the social sciences: What is to be done? *Research policy,* 49**,** 103930.
DASGUPTA, P. & DAVID, P. A. 1994. Toward a new economics of science. *Research policy,* 23**,** 487-521.
DECHARTRES, A., CHARLES, P., HOPEWELL, S., RAVAUD, P. & ALTMAN, D. G. 2011. Reviews assessing the quality or the reporting of randomized controlled trials are increasing over time but raised questions about how quality is assessed. *Journal of clinical epidemiology,* 64**,** 136-144.
DECHARTRES, A., TRINQUART, L., ATAL, I., MOHER, D., DICKERSIN, K., BOUTRON, I., PERRODEAU, E., ALTMAN, D. G. & RAVAUD, P. 2017. Evolution of poor reporting and inadequate methods over time in 20 920 randomised controlled trials included in Cochrane reviews: research on research study. *BMJ,* 357**,** j2490.
DRIVAS, K. & KREMMYDAS, D. 2020. The Matthew effect of a journal's ranking. *Research Policy,* 49.
DUYX, B., URLINGS, M. J., SWAEN, G. M., BOUTER, L. M. & ZEEGERS, M. P. 2017. Scientific citations favor positive results: a systematic review and meta-analysis. *Journal of clinical epidemiology,* 88**,** 92-101.
DWAN, K., ALTMAN, D. G., ARNAIZ, J. A., BLOOM, J., CHAN, A.-W., CRONIN, E., DECULLIER, E., EASTERBROOK, P. J., VON ELM, E. & GAMBLE, C. 2008. Systematic review of the empirical evidence of study publication bias and outcome reporting bias. *PLoS One,* 3**,** e3081.
DWAN, K., GAMBLE, C., WILLIAMSON, P. R. & KIRKHAM, J. J. 2013. Systematic review of the empirical evidence of study publication bias and outcome reporting bias—an updated review. *PloS one,* 8**,** e66844.
FANELLI, D. 2009. How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PloS one,* 4**,** e5738.
FANELLI, D. 2011. Negative results are disappearing from most disciplines and countries. *Scientometrics,* 90**,** 891-904.
FANELLI, D. 2018. Opinion: Is science really facing a reproducibility crisis, and do we need it to? *Proceedings of the National Academy of Sciences,* 115**,** 2628-2631.
FANELLI, D., COSTAS, R. & IOANNIDIS, J. P. 2017. Meta-assessment of bias in science. *Proceedings of the National Academy of Sciences***,** 201618569.

FANG, F. C., STEEN, R. G. & CASADEVALL, A. 2012. Misconduct accounts for the majority of retracted scientific publications. *Proceedings of the National Academy of Sciences,* 109**,** 17028-17033.
FISKE, S. T. 1993. Social cognition and social perception. *Annual review of psychology,* 44**,** 155-194.
FLEMING, P. S., KOLETSI, D., DWAN, K. & PANDIS, N. 2015. Outcome discrepancies and selective reporting: impacting the leading journals? *PloS one,* 10**,** e0127495.
FLETCHER, B. & SACKETT, D. 1979. Canadian task force on the periodic health examination: the Periodic Health Examination. *CMAJ,* 121**,** 1193-1254.
FRANCO, A., MALHOTRA, N. & SIMONOVITS, G. 2014. Publication bias in the social sciences: Unlocking the file drawer. *Science,* 345**,** 1502-1505.
FURMAN, J. L., JENSEN, K. & MURRAY, F. 2012. Governing knowledge in the scientific community: Exploring the role of retractions in biomedicine. *Research Policy,* 41**,** 276-290.
GALL, T. & MANIADIS, Z. 2019. Evaluating solutions to the problem of false positives. *Research Policy,* 48**,** 506-515.
GLASZIOU, P., ALTMAN, D. G., BOSSUYT, P., BOUTRON, I., CLARKE, M., JULIOUS, S., MICHIE, S., MOHER, D. & WAGER, E. 2014. Reducing waste from incomplete or unusable reports of biomedical research. *The Lancet,* 383**,** 267-276.
GRIMSHAW, J. 2004. So what has the Cochrane Collaboration ever done for us? A report card on the first 10 years. *Canadian Medical Association Journal,* 171**,** 747-749.
GUYATT, G., CAIRNS, J., CHURCHILL, D., COOK, D., HAYNES, B., HIRSH, J., IRVINE, J., LEVINE, M., LEVINE, M. & NISHIKAWA, J. 1992. Evidence-based medicine. *JAMA: the journal of the American Medical Association,* 268**,** 2420-2425.
GUYATT, G., COOK, D. & HAYNES, B. 2004. Evidence based medicine has come a long way: The second decade will be as exciting as the first. *BMJ: British Medical Journal,* 329**,** 990.
GUYATT, G. H., OXMAN, A. D., VIST, G. E., KUNZ, R., FALCK-YTTER, Y., ALONSO-COELLO, P. & SCHÜNEMANN, H. J. 2008. Rating quality of evidence and strength of recommendations: GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ: British Medical Journal,* 336**,** 924.
HALEVI, G. 2020. Why Articles in Arts and Humanities Are Being Retracted? *Publishing Research Quarterly,* 36**,** 55-62.
HALL, J. & MARTIN, B. R. 2019. Towards a taxonomy of research misconduct: The case of business school research. *Research Policy,* 48**,** 414-427.
HARRISON, J. S., BANKS, G. C., POLLACK, J. M., O'BOYLE, E. H. & SHORT, J. 2017. Publication bias in strategic management research. *Journal of Management,* 43**,** 400-425.
HAUSMAN, J. A., HALL, B. H. & GRILICHES, Z. 1984. Econometric models for count data with an application to the patents-R&D relationship. national bureau of economic research Cambridge, Mass., USA.
HEAD, M. L., HOLMAN, L., LANFEAR, R., KAHN, A. T. & JENNIONS, M. D. 2015. The extent and consequences of p-hacking in science. *PLoS biology,* 13**,** e1002106.
HIGGINS, J. P., ALTMAN, D. G., GØTZSCHE, P. C., JÜNI, P., MOHER, D., OXMAN, A. D., SAVOVIĆ, J., SCHULZ, K. F., WEEKS, L. & STERNE, J. A. 2011. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *Bmj,* 343**,** d5928.
HIGGINS, J. P. & GREEN, S. 2011. *Cochrane handbook for systematic reviews of interventions*, John Wiley & Sons.
HIGGINS, J. P., THOMAS, J., CHANDLER, J., CUMPSTON, M., LI, T., PAGE, M. J. & WELCH, V. A. 2019. *Cochrane handbook for systematic reviews of interventions*, John Wiley & Sons.
HUSSINGER, K. & PELLENS, M. 2019. Guilt by association: How scientific misconduct harms prior collaborators. *Research Policy,* 48**,** 516-530.
HUTTON, J. & WILLIAMSON, P. R. 2000. Bias in meta-analysis due to outcome variable selection within studies. *Journal of the Royal Statistical Society: Series C (Applied Statistics),* 49**,** 359-370.
IOANNIDIS, J. P. 2014. How to Make More Published Research True. *PLOS Medicine,* 11**,** e1001747.
JADAD, A. R., COOK, D. J., JONES, A., KLASSEN, T. P., TUGWELL, P., MOHER, M. & MOHER, D. 1998. Methodology and reports of systematic reviews and meta-analyses: a comparison of Cochrane reviews with articles published in paper-based journals. *Jama,* 280**,** 278-280.

KOVANIS, M., PORCHER, R., RAVAUD, P. & TRINQUART, L. 2016. The global burden of journal peer review in the biomedical literature: Strong imbalance in the collective enterprise. *PLoS One,* 11**,** e0166387.

LEE, J., PARK, D.-H. & HAN, I. 2008a. The effect of negative online consumer reviews on product attitude: An information processing view. *Electronic commerce research and applications,* 7**,** 341-352.

LEE, K., BACCHETTI, P. & SIM, I. 2008b. Publication of clinical trials supporting successful new drug applications: a literature analysis. *PLoS medicine,* 5**,** e191.

LEXCHIN, J., BERO, L. A., DJULBEGOVIC, B. & CLARK, O. 2003. Pharmaceutical industry sponsorship and research outcome and quality: systematic review. *Bmj,* 326**,** 1167-1170.

LOKKER, C., MCKIBBON, K. A., MCKINLAY, R. J., WILCZYNSKI, N. L. & HAYNES, R. B. 2008. Prediction of citation counts for clinical articles at two years using data available within three weeks of publication: retrospective cohort study. *Bmj,* 336**,** 655-657.

LU, S. F., JIN, G. Z., UZZI, B. & JONES, B. 2013. The retraction penalty: Evidence from the Web of Science. *Scientific Reports,* 3**,** 3146.

MALHOTRA, N. K. 1984. Reflections on the information overload paradigm in consumer decision making. *Journal of consumer research,* 10**,** 436-440.

MERTON, R. K. 1973. *The sociology of science: Theoretical and empirical investigations*, University of Chicago press.

MOHER, D. 2001. CONSORT Group (Consolidated Standards of Reporting Trials). The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. *Ann Intern Med,* 134**,** 657-662.

MOORE, S. G. & LAFRENIERE, K. C. 2020. How online word-of-mouth impacts receivers. *Consumer Psychology Review,* 3**,** 34-59.

MURPHY, K. R. & AGUINIS, H. 2019. HARKing: how badly can cherry-picking and question trolling produce bias in published results? *Journal of business and psychology,* 34**,** 1-17.

NATURE 2018. Checklists work to improve science. *Nature.*

NEALE, A. V., DAILEY, R. K. & ABRAMS, J. 2010. Analysis of citations to biomedical articles affected by scientific misconduct. *Science and Engineering Ethics,* 16**,** 251-261.

NECKER, S. 2014. Scientific misbehavior in economics. *Research Policy,* 43**,** 1747-1759.

PAPATHANASSIS, A. & KNOLLE, F. 2011. Exploring the adoption and processing of online holiday reviews: A grounded theory approach. *Tourism Management,* 32**,** 215-224.

PAVITT, K. 1987. The objectives of technology policy. *Science and public policy,* 14**,** 182-188.

PÉRON, J., MAILLET, D., GAN, H. K., CHEN, E. X. & YOU, B. 2013. Adherence to CONSORT adverse event reporting guidelines in randomized clinical trials evaluating systemic cancer therapy: a systematic review. *Journal of Clinical Oncology,* 31**,** 3957-3963.

PICKETT, J. T. & ROCHE, S. P. 2018. Questionable, objectionable or criminal? Public opinion on data fraud and selective reporting in science. *Science and engineering ethics,* 24**,** 151-171.

ROSE, M. E. & KITCHIN, J. R. 2019. pybliometrics: Scriptable bibliometrics using a Python interface to Scopus. *SoftwareX,* 10**,** 100263.

ROSS, J. S., HILL, K. P., EGILMAN, D. S. & KRUMHOLZ, H. M. 2008. Guest authorship and ghostwriting in publications related to rofecoxib: a case study of industry documents from rofecoxib litigation. *Jama,* 299**,** 1800-1812.

SACKETT, D. L., ROSENBERG, W., GRAY, J., HAYNES, R. B. & RICHARDSON, W. S. 1996. Evidence based medicine: what it is and what it isn't. *Bmj,* 312**,** 71-72.

SEGLEN, P. O. 1992. The skewness of science. *Journal of the American Society for Information Science,* 43**,** 628-638.

SPENCE, M. 1973. Job Market Signaling. *Quarterly Journal of Economics,* 87**,** 355-379.

SPENCE, M. 2002. Signaling in retrospect and the informational structure of markets. *American Economic Review,* 92**,** 434-459.

TAHAMTAN, I., AFSHAR, A. S. & AHAMDZADEH, K. 2016. Factors affecting number of citations: a comprehensive review of the literature. *Scientometrics,* 107**,** 1195-1225.

TANG, E., RAVAUD, P., RIVEROS, C., PERRODEAU, E. & DECHARTRES, A. 2015. Comparison of serious adverse events posted at ClinicalTrials. gov and published in corresponding journal articles. *BMC medicine,* 13**,** 189.

TUCH, A. N., BARGAS-AVILA, J. A., OPWIS, K. & WILHELM, F. H. 2009. Visual complexity of websites: Effects on users' experience, physiology, performance, and memory. *International journal of human-computer studies,* 67**,** 703-715.

TURNER, L., SHAMSEER, L., ALTMAN, D. G., SCHULZ, K. F. & MOHER, D. 2012. Does use of the CONSORT Statement impact the completeness of reporting of randomised controlled trials published in medical journals? A Cochrane review a. *Systematic reviews,* 1**,** 60.

VAN NOORDEN, R. 2011. The trouble with retractions. *Nature,* 478**,** 26.

VIERGEVER, R. F., KARAM, G., REIS, A. & GHERSI, D. 2014. The quality of registration of clinical trials: still a problem. *PLoS One,* 9**,** e84727.

WATSON, J., GHOSH, A. P. & TRUSOV, M. 2018. Swayed by the numbers: the consequences of displaying product review attributes. *Journal of Marketing,* 82**,** 109-131.

YOUNG, N. S., IOANNIDIS, J. P. & AL-UBAYDLI, O. 2008. Why current publication practices may distort science. *PLoS medicine,* 5**,** e201.

Online appendix

Supporting information for:

Directing scientists away from potentially biased publications: the role of systematic reviews in health care

Section I Cochrane risk of bias tool

Figure S1 illustrates how risk of bias ratings are reported in a Cochrane review. From November 2018, a new version (Version 2) of the tool has replaced the original version, which was first published in 2008, and updated in 2011.

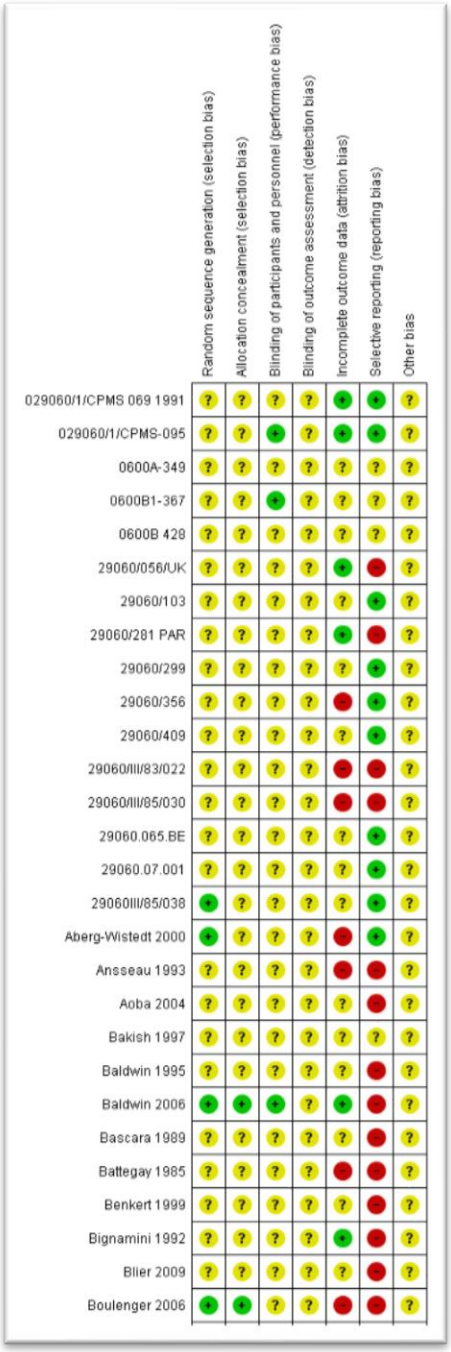


Figure S1 Example of Cochrane risk of bias tool

Source: Reproduced with permission from Purgato, M., Papola, D., Gastaldon, C., Trespidi, C., Magni, L. R., Rizzo, C., ... & Barbui, C. (2014). Paroxetine versus other anti-depressive agents for depression. Cochrane Database of Systematic Reviews, (4). Copyright © 2000 - 2020 by John Wiley & Sons, Inc.

Section II Cochrane Editorial groups included in the analysis

Table S1 Breakdown of articles by Cochrane Editorial group ('HIGH vs LOW' sample)

Cochrane Editorial group	Treatment High risk of bias n=230	Control Low risk of bias n=301	Total n=531
Acute Respiratory Infections	4	5	9
Airways	3	3	6
Anaesthesia	3	3	6
Back and Neck	4	4	8
Bone, Joint and Muscle Trauma	1	2	3
Breast Cancer	2	2	4
Colorectal	3	4	7
Common Mental Disorders	13	15	28
Consumers and Communication	6	9	15
Cystic Fibrosis and Genetic Disorders	1	1	2
Developmental, Psychosocial and Learning	3	3	6
Drugs and Alcohol	7	9	16
ENT	1	1	2
Effective Practice and Organisation	4	5	9
Emergency and Critical Care	3	3	6
Epilepsy	1	1	2
Eyes and Vision	2	3	5
Gynaecological, Neuro-oncology	3	3	6
Gynaecology and Fertility	13	15	28
HIV/AIDS	1	1	2
Heart	14	21	35
Hepato-Biliary	15	21	36
Hypertension	21	37	58
IBD	2	2	4
Incontinence	1	2	3
Infectious Diseases	1	1	2
Injuries	6	7	13
Kidney and Transplant	9	11	20
Multiple Sclerosis	1	1	2
Musculoskeletal	4	6	10
Neonatal	1	1	2
Neuromuscular	1	1	2
Oral Health	11	17	28
Pain, Palliative and Supportive care	8	9	17
Pregnancy and Childbirth	9	10	19
Public Health	4	5	9
Schizophrenia	5	6	11
Skin	21	27	48
Stroke	1	1	2
Tobacco Addiction G..	6	7	13
Upper GI and Pancreatic Disease	5	8	13
Urology	1	1	2
Vascular	2	2	4
Work	1	2	3
Wounds	2	3	5
Total	**230**	**301**	**531**

Section III Alternative modelling approaches.

Our main results are based on a quasi-ML Poisson model, given its robustness. However, the key findings are consistent across different estimation procedures. In Table S3, Column 3–1 reports our main model results using the quasi-ML Poisson estimator, as discussed in the main body of the paper. Column 3–2 reports the results of an ordinary least squares (OLS) model specification, using log (forward citations +1) as the dependent variable. Column 3–3 reports the results of conditional fixed effects negative binomial regression, with bootstrapped standard errors clustered by article family. The coefficient on POST x TREATED, indicating a 5.4% (3-2) and 8.4% (3-3) effect on citations, are in line with or main model results.

Table S3 Average impact of bias (detection) on follow-on citations: Alternative specifications

	(3-1)	(3-2)	(3-3)
	Quasi-ML Poisson	Ordinary Least Squares (OLS)	Conditional fixed effects negative binomial
Dependent variable	Citations count	Log (citations+1)	Citations count
POST	0.917**	0.916***	0.916**
	(0.0376)	(0.0309)	(0.0326)
POST x TREATED	0.921*	0.946*	0.916**
	(0.0410)	(0.0309)	(0.0365)
Constant		1.598***	2.742***
		(0.0435)	(0.266)
Number of Article-Year Obs.	6,135	6,135	6,135
R^2		0.263	
Number of articles	531	531	531
Paper FEs	Yes	Yes	Yes
Age FEs	Yes	Yes	Yes
	SEs, adjusted for clustering by article family, reported in parentheses	SEs, adjusted for clustering by article family, reported in parentheses	Bootstrapped SEs, adjusted for clustering by article, reported in parentheses

Incidence-rate ratios. 'HIGH vs LOW' sample.
We exclude articles published outside a time period running from ten years to one year before the publication of a Cochrane review. We also remove any citations outliers, defined as all treated papers that were in the top 10% percentile of citations prior to the publication of a Cochrane review.
*** Significant at the 1 percent level.
** Significant at the 5 percent level.
* Significant at the 10 percent level.

Section IV Alternative dependent variable operationalisations

In Table S4, Column 4–1 reports our main model results using a quasi-ML Poisson estimator with forward citations as the dependent variable. Column 4–2 reports the results also with a quasi-ML Poisson estimator, but excluding self-citations in the dependent variable. Self-citations were identified by the overlap between any of the authors of the cited article with any of the citing authors. The results of the two models are in line (IRR=0.921 vs IRR=0.920)

Table S4 Average impact of bias (detection) on follow-on citations: Excluding self-citations

	(4-1)	(4-2)
	Quasi-ML Poisson	Quasi-ML Poisson excluding self-citations
POST	0.917**	0.902**
	(0.0376)	(0.0387)
POST x TREATED	0.921*	0.920*
	(0.0410)	(0.0410)
Number of Article-Year Obs.	6,135	6,135
Number of articles	531	531
Paper FEs	Yes	Yes
Age FEs	Yes	Yes
	SEs, adjusted for clustering by article family, reported in parentheses	SEs, adjusted for clustering by article family, reported in parentheses

Incidence-rate ratios obtained using conditional quasi-maximum likelihood Poisson models. Dependent variable: number of citations received by each treated article in a particular year. Treated equal to 1 if publication is deemed at a high risk of selective reporting bias, 0 if at low risk of bias. Standard errors in parentheses clustered by article family. We exclude articles published outside a time period running from ten years to one year before the publication of a Cochrane review. We also remove any citations outliers, defined as all treated papers that were in the top 10% percentile of citations prior to the publication of a Cochrane review.
*** Significant at the 1 percent level.
** Significant at the 5 percent level.
* Significant at the 10 percent level.

Section V Selection effect

Papers that eventually received a worse evaluation (e.g., "high risk of bias") may be less cited than their low risk of bias counterparts even before the treatment.

To investigate this potential 'selection' effect, we estimate an equation that contains fixed effects for each 'family' of a treated and control articles, an article age effect, a dummy of whether *t* is after the year of publication of a Cochrane review (POST), a dummy equal to one for those articles rated at high risk of bias (TREATED), a variable equal to one for the treated papers during the year immediately prior to, the year of, and the year immediately after the publication of a review (WINDOW x TREATED), and a variable equal to one only in the years after the window period for the treated papers (POST x TREATED).

The coefficient on TREATED in Table S5 identifies the selection effect. The results suggest that that articles that are ultimately rated at high risk of bias are associated with a 11 percent lower citation rate relative to the controls (IRR=0.895). A negative selection effect is in line with the expectation that high risk of bias papers may be of lower scientific standing, as proxied by their level of citation regardless of receiving a treatment. The coefficient of TREATED is not statistically significant so we cannot reject the null of no selection in our sample.

The marginal impact of the treatment (controlling for the selection effect) is reflected in the coefficient of POST X TREATED. This is estimated as a 3.1% drop in the citation rate. While the direction of the coefficient is in line with that our main model, the coefficient is not statistically significant. The coefficient of WINDOW X TREATED, accounting for potential pre-announcement effect or lags in the dissemination of the results, is also not significant.

Table S5 Average impact of bias (detection) on follow-on citations: Family FEs specification

	(5-1) Quasi-ML Poisson 'Family' FEs
POST	0.934
	(0.0424)
TREATED	0.895
	(0.0610)
WINDOW X TREATED	1.047
	(0.0577)
POST X TREATED	0.969
	(0.0707)
Number of article 'families'	230
Number of Article-Year Obs.	6,135
Family FEs	Yes
Age FEs	Yes

Incidence-rate ratios obtained using conditional quasi-maximum likelihood Poisson model. Dependent variable: number of citations received by each treated article in a particular year. Treated equal to 1 if publication is deemed at a high risk of selective reporting bias, 0 if at low risk of bias. Robust standard errors in parentheses. We exclude articles published outside a time period running from ten years to one year before the publication of a Cochrane review. We also remove any citations outliers, defined as all treated papers that were in the top 10% percentile of citations prior to the publication of a Cochrane review.

*** Significant at the 1 percent level.

** Significant at the 5 percent level.

* Significant at the 10 percent level

Section VI Descriptive statistics for the 'HIGH vs UNCLEAR' sample and for the 'LOW vs UNCLEAR' sample

Table S6 Descriptive statistics for treated and control articles: 'HIGH vs UNCLEAR' sample

	Before matching						After matching					
	Treated, n=1,984		Controls, n=3,805		t-test (z-test)	p-value	Treated, n=141		Controls, n=184		t-test (z-test)	p-value
Variables	Mean	Std. Dev.	Mean	Std. Dev.			Mean	Std. Dev.	Mean	Std. Dev.		
Root article characteristics												
Year of publication	1999.1	10.68	2000.4	9.92	4.52	0.00	2008.0	3.25	2008.0	3.28	-0.16	0.87
Authors count	6.08	5.10	5.82	4.29	-2.06	0.04	5.84	3.14	5.85	2.93	0.03	0.97
Affiliations count	2.57	2.55	2.43	2.49	-1.87	0.06	2.68	2.00	2.53	1.78	-0.71	0.48
Citations characteristics												
Average yearly citations received	4.65	9.96	5.05	9.99	1.47	0.14	3.73	2.85	3.61	2.64	-0.40	0.69
Citations received on the first year	4.59	11.95	4.45	11.41	-0.45	0.65	3.38	4.06	3.27	3.84	-0.26	0.79
Cumulative citation prior to Cochrane review	61.13	124.19	68.61	191.37	1.57	0.12	18.50	17.61	20.14	20.51	0.76	0.45
Journal characteristics												
Journal JCR (mean over 1997-2018)	4.72	6.99	4.81	7.00	0.45	0.65	2.53	1.45	2.37	1.30	-1.07	0.28
Authors characteristics												
% authors affiliated to university	0.67	1.00	0.71	1.00	1.32	0.19	0.81	1.00	0.82	1.00	0.11	0.91
% authors affiliated to top 100 pharma firm (incl. sponsorship)	0.09	1.00	0.07	1.00	-1.00	0.32	0.05	1.00	0.05	1.00	-0.01	0.99
Trial characteristics												
Blinding bias (Share of high and unclear risk)	0.61	1.00	0.63	1.00	0.39	0.70	0.66	1.00	0.69	1.00	0.25	0.80
Other bias (Share of high and unclear risk)	0.63	1.00	0.59	1.00	-1.33	0.18	0.63	1.00	0.45	1.00	-1.38	0.17
Other domains - excl. sel. reporting (Share of high and unclear risk)	0.68	1.00	0.66	1.00	-0.60	0.55	0.65	1.00	0.64	1.00	-0.14	0.89
Other domains - excl. sel. reporting (Share of high risk)	0.21	1.00	0.18	1.00	-1.06	0.29	0.21	1.00	0.18	1.00	-0.25	0.80

Table S7 Descriptive statistics for treated and control articles: 'LOW vs UNCLEAR' sample

	Before matching						After matching					
	Treated, n=7,319		Controls, n=5,033		t-test (z-test)	p-value	Treated, n=770		Controls, n=1,046		t-test (z-test)	p-value
Variables	Mean	Std. Dev.	Mean	Std. Dev.			Mean	Std. Dev.	Mean	Std. Dev.		
Root article characteristics												
Year of publication	2000.5	9.78	2002.2	9.22	9.39	0.00	2008.5	3.17	2008.3	3.26	-0.95	0.34
Authors count	5.94	3.92	6.81	5.67	9.35	0.00	6.47	3.60	6.40	3.50	-0.43	0.67
Affiliations count	2.43	2.53	2.94	3.17	9.43	0.00	2.81	2.37	2.68	2.03	-1.25	0.21
Citations characteristics												
Average yearly citations received	5.14	10.66	7.65	21.19	7.75	0.00	3.58	2.55	3.20	2.34	-3.33	0.00
Citations received on the first year	4.54	11.43	7.42	19.47	9.43	0.00	3.37	3.54	2.90	3.20	-2.94	0.00
Cumulative citation prior to Cochrane review	76.18	239.54	109.99	966.20	2.43	0.02	18.67	17.30	18.29	17.72	-0.46	0.65
Journal characteristics												
Journal JCR (mean over 1997-2018)	4.83	7.16	6.36	9.28	9.58	0.00	2.82	2.15	2.63	2.00	-1.98	0.05
Authors characteristics												
% authors affiliated to university	0.70	1.00	0.72	1.00	1.03	0.30	0.80	1.00	0.78	1.00	-0.39	0.69
% authors affiliated to top 100 pharma firm (incl. sponsorship)	0.06	1.00	0.07	1.00	0.76	0.45	0.07	1.00	0.04	1.00	-0.61	0.54
Trial characteristics												
Blinding bias (Share of high and unclear risk)	0.63	1.00	0.51	1.00	-6.43	0.00	0.51	1.00	0.63	1.00	2.38	0.02
Other bias (Share of high and unclear risk)	0.59	1.00	0.41	1.00	-8.38	0.00	0.42	1.00	0.57	1.00	2.78	0.01
Other domains - excl. sel. reporting (Share of high and unclear risk)	0.64	1.00	0.50	1.00	-8.18	0.00	0.47	1.00	0.60	1.00	2.75	0.01
Other domains - excl. sel. reporting (Share of high risk)	0.17	1.00	0.14	1.00	-1.31	0.19	0.15	1.00	0.16	1.00	0.32	0.75