Meta-assessment of bias in science

Daniele Fanelli[a,1], Rodrigo Costas[b], and John P. A. Ioannidis[a,c,d,e]

[a]Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Palo Alto, CA 94304; [b]Centre for Science and Technology Studies, Leiden University, 2333 AL Leiden, The Netherlands; [c]Department of Medicine, Stanford University School of Medicine, Stanford, CA 94305; [d]Department of Health Research and Policy, Stanford University School of Medicine, Stanford, CA 94305; and [e]Department of Statistics, Stanford University School of Humanities and Sciences, Stanford, CA 94305

Edited by Susan T. Fiske, Princeton University, Princeton, NJ, and approved February 14, 2017 (received for review November 8, 2016)

Numerous biases are believed to affect the scientific literature, but their actual prevalence across disciplines is unknown. To gain a comprehensive picture of the potential imprint of bias in science, we probed for the most commonly postulated bias-related patterns and risk factors, in a large random sample of meta-analyses taken from all disciplines. The magnitude of these biases varied widely across fields and was overall relatively small. However, we consistently observed a significant risk of small, early, and highly cited studies to overestimate effects and of studies not published in peer-reviewed journals to underestimate them. We also found at least partial confirmation of previous evidence suggesting that US studies and early studies might report more extreme effects, although these effects were smaller and more heterogeneously distributed across meta-analyses and disciplines. Authors publishing at high rates and receiving many citations were, overall, not at greater risk of bias. However, effect sizes were likely to be overestimated by early-career researchers, those working in small or long-distance collaborations, and those responsible for scientific misconduct, supporting hypotheses that connect bias to situational factors, lack of mutual control, and individual integrity. Some of these patterns and risk factors might have modestly increased in intensity over time, particularly in the social sciences. Our findings suggest that, besides one being routinely cautious that published small, highly-cited, and earlier studies may yield inflated results, the feasibility and costs of interventions to attenuate biases in the literature might need to be discussed on a discipline-specific and topic-specific basis.

bias | misconduct | meta-analysis | integrity | meta-research

Numerous biases have been described in the literature, raising concerns for the reliability and integrity of the scientific enterprise (1–4). However, it is yet unknown to what extent bias patterns and postulated risk factors are generalizable phenomena that threaten all scientific fields in similar ways and whether studies documenting such problems are reproducible (5–7). Indeed, evidence suggests that biases may be heterogeneously distributed in the literature. The ratio of studies concluding in favor vs. against a tested hypothesis increases, moving from the physical, to the biological and to the social sciences, suggesting that research fields with higher noise-to-signal ratio and lower methodological consensus might be more exposed to positive-outcome bias (5, 8, 9). Furthermore, multiple independent studies suggested that this ratio is increasing (i.e., positive results have become more prevalent), again with differences between research areas (9–11), and that it may be higher among studies from the United States, possibly due to excessive "productivity" expectations imposed on researchers by the tenure-track system (12–14). Most of these results, however, are derived from varying, indirect proxies of positive-outcome bias that may or may not correspond to actual distortions of the literature.

Nonetheless, concerns that papers reporting false or exaggerated findings might be widespread and growing have inspired an expanding literature of research on research (*aka* meta-research), which points to a postulated core set of bias patterns and factors that might increase the risk for researchers to engage in bias-generating practices (15, 16).

The bias patterns most commonly discussed in the literature, which are the focus of our study, include the following:

Small-study effects: Studies that are smaller (of lower precision) might report effect sizes of larger magnitude. This phenomenon could be due to selective reporting of results or to genuine heterogeneity in study design that results in larger effects being detected by smaller studies (17).

Gray literature bias: Studies might be less likely to be published if they yielded smaller and/or statistically nonsignificant effects and might be therefore only available in PhD theses, conference proceedings, books, personal communications, and other forms of "gray" literature (1).

Decline effect: The earliest studies to report an effect might overestimate its magnitude relative to later studies, due to a decreasing field-specific publication bias over time or to differences in study design between earlier and later studies (1, 18).

Early-extreme: An alternative scenario to the decline effect might see earlier studies reporting extreme effects in any direction, because extreme and controversial findings have an early window of opportunity for publication (19).

Citation bias: The number of citations received by a study might be correlated to the magnitude of effects reported (20).

US effect: Publications from authors working in the United States might overestimate effect sizes, a difference that could be due to multiple sociological factors (14).

Industry bias: Industry sponsorship may affect the direction and magnitude of effects reported by biomedical studies (21). We generalized this hypothesis to nonbiomedical fields

Significance

Science is said to be suffering a reproducibility crisis caused by many biases. How common are these problems, across the wide diversity of research fields? We probed for multiple bias-related patterns in a large random sample of meta-analyses taken from all disciplines. The magnitude of these biases varied widely across fields and was on average relatively small. However, we consistently observed that small, early, highly cited studies published in peer-reviewed journals were likely to overestimate effects. We found little evidence that these biases were related to scientific productivity, and we found no difference between biases in male and female researchers. However, a scientist's early-career status, isolation, and lack of scientific integrity might be significant risk factors for producing unreliable results.

Author contributions: D.F. and J.P.A.I. designed research; D.F. performed research; R.C. contributed new reagents/analytic tools; D.F. analyzed data; D.F. and J.P.A.I. wrote the paper; D.F. conceived the study, sampled meta-analyses, and collected and led the collection of data; and R.C. produced most of the bibliometric data.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. Email: email@danielefanelli.com.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1618569114/-/DCSupplemental.

by predicting that studies with coauthors affiliated to private companies might be at greater risk of bias.

Among the many sociological and psychological factors that may underlie the bias patterns above, the most commonly invoked include the following:

Pressures to publish: Scientists subjected to direct or indirect pressures to publish might be more likely to exaggerate the magnitude and importance of their results to secure many high-impact publications and new grants (22, 23). One type of pressure to publish is induced by national policies that connect publication performance with career advancement and public funding to institutions.

Mutual control: Researchers working in close collaborations are able to mutually control each other's work and might therefore be less likely to engage in questionable research practices (QRP) (24, 25). If so, risk of bias might be lower in collaborative research but, adjusting for this factor, higher in long-distance collaborations (25).

Career stage: Early-career researchers might be more likely to engage in QRP, because they are less experienced and have more to gain from taking risks (26).

Gender of scientist: Males are more likely to take risks to achieve higher status and might therefore be more likely to engage in QRP. This hypothesis was supported by statistics of the US Office of Research Integrity (27), which, however, may have multiple alternative explanations (28).

Individual integrity: Narcissism or other psychopathologies underlie misbehavior and unethical decision making and therefore might also affect individual research practices (29–31).

One can explore whether these bias patterns and postulated causes are associated with the magnitude of effect sizes reported by studies performed on a given scientific topic, as represented by individual meta-analyses. The prevalence of these phenomena across multiple meta-analyses can be analyzed with multilevel weighted regression analysis (14) or, more straightforwardly, by conducting a second-order meta-analysis on regression estimates obtained within each meta-analysis (32). Bias patterns and risk factors can thus be assessed across multiple topics within a discipline, across disciplines or larger scientific domains (social, biological, and physical sciences), and across all of science.

To gain a comprehensive picture of the potential imprint of bias in science, we collected a large sample of meta-analyses covering all areas of scientific research. We recorded the effect size reported by each primary study within each meta-analysis and assessed, using meta-regression, the extent to which a set of parameters reflecting hypothesized patterns and risk factors for bias was indeed associated with a study's likelihood to overestimate effect sizes.

Each bias pattern and postulated risk factor listed above was turned into a testable hypothesis, with specific predictions about how the magnitude of effect sizes reported by primary studies in meta-analyses should be associated with some measurable characteristic of primary study or author (Table 1). To test these hypotheses, we searched for meta-analyses in each of the 22 mutually exclusive disciplinary categories used by the Essential Science Indicators database, a bibliometric tool that covers all areas of science and was used in previous large-scale studies of bias (5, 11, 33). These searches yielded an initial list of over 116,000 potentially relevant titles, which through successive phases of screening and exclusion yielded a final sample of 3,042 usable meta-analyses (Fig. S1). Of these, 1,910 meta-analyses used effect-size metrics that could all be converted to log-odds ratio (n = 33,355 nonduplicated primary data points), whereas the remaining

Table 1. Summary of each bias pattern or risk factor for bias that was tested in our study, parameters used to test these hypotheses via meta-regression, predicted direction of the association of these parameters with effect size, and overall assessment of results obtained

Hypothesis type	Hypothesis tested	Specific factor tested	Variables measured to test the hypothesis	Predicted association with effect size	Result
Postulated bias patterns	Small-study effect		Study SE	+	S
	Gray literature bias		Gray literature (any type) vs. Journal article	–	S
	Decline effect		Year order in MA	–	P
	Early extremes		Year order in MA, regressed on absolute effect size	–	N
	Citation bias		Total citations to study	+	S
	US effect		Study from author in the US vs. Any other country	+	P
	Industry bias		Studies with authors affiliated with private industry vs. Not	+	P
Postulated risk factors for bias	Pressures to publish	Country policies	Cash incentive	+	N
			Career incentive	+	N
			Institutional incentive	+	N
		Author's productivity	(First/last) author's total publications, publications per year	+	N
		Author's impact	(First/last) total citations, average citations, average normalized citations, average journal impact, % top10 journals	+	N
	Mutual control		Team size	–	S
			Countries/author, average distance between addresses	+	S
	Individual risk factors	Career level	Years in activity (first/last) author	–	S
		Gender	(First/last) author's female name	–	N
		Research integrity	(First/last) author with ≥1 retraction	+	P

Symbols indicate whether the association between factor and effect size is predictive to be positive (+) or negative (–). Conclusions as to whether results indicate that the hypothesis was fully supported (S), partially supported (P), or not supported (N) are based on main analyses as well as secondary and robustness tests, as described in the main text.

1,132 meta-analyses (n = 15,909) used a variety of other metrics, which are not readily interconvertible to log-odds ratio (Table S1). In line with previous studies, we focused our main analysis on the former subsample, which represents a relatively homogeneous population, and we included the second subsample only in robustness analyses.

On each included meta-analysis, the possible effect of each relevant independent variable was measured by standard linear meta-regression. The resulting regression coefficients were then summarized in a second-order meta-analysis to obtain a generalized summary estimate of each pattern and factor across all included meta-analyses. Analyses were also repeated using an alternative method, in which primary data are standardized and analyzed with multilevel regression (see *SI Multilevel Meta-Regression Analysis* for further details).

Results

Bias Patterns. Bias patterns varied substantially in magnitude as well as direction across meta-analyses, and their distribution usually included several extreme values (Fig. S2; full numerical results in Dataset S1). Second-order meta-analysis of these regression estimates yielded highly statistically significant support for the presence of small-study effects, gray literature bias, and citation bias (Fig. 1 *A* and *B*). These patterns were consistently observed in all secondary and robustness tests, which repeated all analyses not adjusting for study precision, standardizing meta-regression estimates and not coining the meta-analyses or coining them with different thresholds (see *Methods* for details and all numerical results in Dataset S2).

The decline effect, measured as a linear association between year of study publication and reported effect size, was not statistically significant in our main analysis (Fig. 1*B*), but was highly significant in all robustness tests. Moreover, secondary analyses conducted with the multilevel regression approach suggest that most or all of this effect might actually consist of a "first-year" effect, in which the decline is not linear and just the very earliest studies are likely to overestimate findings (*SI Multilevel Meta-Regression Analysis*, *Multilevel Analyses*, *Secondary Tests of Early Extremes, Proteus Phenomenon and Decline Effect*).

The early-extreme effect was, in most robustness tests, marginally significant in the opposite direction to what was predicted, but was measured to high statistical significance in the predicted (i.e., negative) direction when not adjusted for small-study effects (Dataset S2). In other words, it appears to be true that earlier studies may report extreme effects in either direction, but this effect is mainly or solely due to the lower precision of earlier studies.

The US effect exhibited associations in the predicted direction and was marginally significant in our main analyses (Fig. 1*B*) and was significant in some of the robustness tests, particularly when meta-analysis coining was done more conservatively (Dataset S2; see *Methods* for further details).

Industry bias was absent in our main analyses (Fig. 1*B*) but was statistically significant when meta-analyses were coined more conservatively (Dataset S2).

Standardizing these various biases to estimate their relative importance is not straightforward, but results using different methods suggested that small-study effects are by far the most important source of potential bias in the literature. Second-order meta-analyses of standardized meta-regression estimates, for example, yield similar results to those in Fig. 1 (Dataset S2). Calculation of pseudo-R^2 in multilevel regression suggests that small-study effects account for around 27% of the variance of primary outcomes, whereas gray literature bias, citation bias, decline effect, industry sponsorship, and US effect, each tested as individual predictor and not adjusted for study precision, account for only 1.2%, 0.5%, 0.4%, 0.2%, and 0.04% of the variance, respectively (see *SI Multilevel Meta-Regression Analysis*, *Multilevel Analyses*, *Relative Strength of Biases* further details).

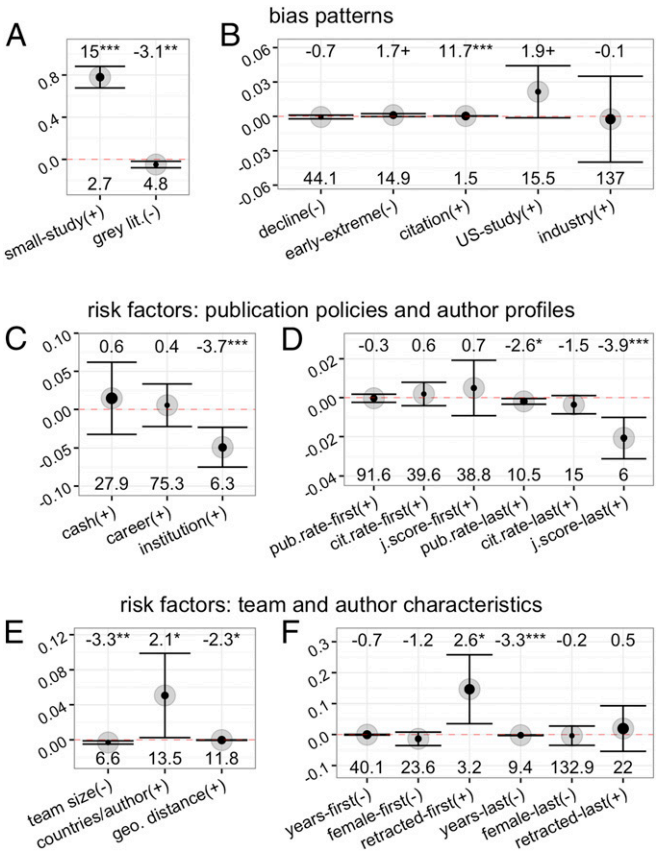

Fig. 1. (*A–F*) Meta-meta–regression estimates of bias patterns and bias risk factors, adjusted for study precision. Each panel shows second-order random-effects meta-analytical summaries of meta-regression estimates [i.e., $b \pm 95\%$ confidence interval (CI)], measured across the sample of meta-analyses. Symbols in parentheses indicate whether the association between factor and effect size is predicted to be positive (+) or negative (–). The gray area within each circle is proportional to the percentage of total variance explained by between–meta-analysis variance (i.e., heterogeneity, measured by I^2). To help visualize effect sizes and statistical significance, numbers above error bars display t scores (i.e., summary effect size divided by its corresponding SE, b/SE) and conventional significance levels (i.e., $+P < 0.1$, $*P < 0.05$, $**P < 0.01$, $***P < 0.001$). Numbers below each error bar reflect the cross–meta-analytical consistency of effects, measured as the ratio of between–meta-analysis SD divided by summary effect size (i.e., τ/b; the smaller the ratio, the higher the consistency). A few of the tested variables were omitted from *D* for brevity (full numerical results for all variables are in Dataset S2). See main text and Table 1 for details of each variable.

Risk Factors for Bias. The pressure to publish hypothesis was overall negatively supported by our analyses, which included tests at the country (i.e., policy) as well as the individual level. At the country level, we found that authors working in countries that incentivize publication performance by distributing public funding to institutions (i.e., Australia, Belgium, New Zealand, Denmark, Italy, Norway, and the United Kingdom) were significantly less likely to overestimate effects. Countries in which publication incentives operate on an individual basis, for example via a tenure-track system (i.e., Germany, Spain, and the United States), and countries in which performance is rewarded with cash-bonus incentives (China, Korea, and Turkey), however, exhibited no significant difference in either direction (Fig. 1*C*). If country was tested separately for first and last authors, results were even more conservative, with last authors working in countries from the latter

two categories being significantly less likely to overestimate results (Dataset S2). At the individual level, we again found little consistent support for our predictions. Publication and impact performance of first and last authors, measured in terms of publications per year, citations per paper, and normalized journal impact score (Fig. 1*D*) as well as additional related measures (Dataset S2), were either not or negatively associated with overestimation of results. The most productive and impactful last authors, in other words, were significantly less likely to report exaggerated effects (Fig. 1*D*).

The mutual control hypothesis was supported overall, suggesting a negative association of bias with team size and a positive one with country-to-author ratio (Fig. 1*E*). Geographic distance exhibited a negative association, against predictions, but this result was not observed in any robustness test, unlike the other two (Dataset S2).

The career level of authors, measured as the number of years in activity since the first publication in the Web of Science, was overall negatively associated with reported effect size, although the association was statistically significant and robust only for last authors (Fig. 1*F*). This finding is consistent with the hypothesis that early-career researchers would be at greater risk of reporting overestimated effects (Table 1).

Gender was inconsistently associated with reported effect size: In most robustness tests, female authors exhibited a tendency to report smaller (i.e., more conservative) effect sizes (e.g., Fig. 1*F*), but the only statistically significant effect detected among all robustness tests suggested the opposite, i.e., that female first authors are more likely to overestimate effects (Dataset S2).

Scientists who had one or more papers retracted were significantly more likely to report overestimated effect sizes, albeit solely in the case of first authors (Fig. 1*F*). This result, consistently observed across most robustness tests (Dataset S2), offers partial support to the individual integrity hypothesis (Table 1).

The between–meta-analysis heterogeneity measured for all bias patterns and risk factors was high (Fig. 1, Fig. S2, and Dataset S2), suggesting that biases are strongly dependent on contingent characteristics of each meta-analysis. The associations most consistently observed, estimated as the value of between–meta-analysis variance divided by summary effect observed, were, in decreasing order of consistency, citation bias, small-study effects, gray literature bias, and the effect of a retracted first author (Fig. 1, bottom numbers).

Differences Between Disciplines and Domains. Part of the heterogeneity observed across meta-analyses may be accounted for at the level of discipline (Fig. S3) or domain (Fig. 2 and Fig. S4), as evidenced by the lower levels of heterogeneity and higher levels of consistency observed within some disciplines and domains. The social sciences, in particular, exhibited effects of equal or larger magnitude than the biological and the physical sciences for most of the biases (Fig. 2) and some of the risk factors (Fig. S4).

Trends over Time. We assessed whether the magnitude of bias patterns or risk factors measured within meta-analyses had changed over time, measured by year of publication of meta-analysis. Because individual-level bibliometric information might be less accurate going backwards in time, this analysis was limited to study-level parameters. Across seven tested effects, we observed a significant increase in the coefficients of collaboration distance and small-study effects (respectively, $b = 0.012 \pm 0.005$/y, $b = 0.020 \pm 0.010$/y; Dataset S3). Analyses partitioned by domain suggest that the social sciences, in particular, may have registered a significant worsening of small-study effects, decline effect, and social control effects (Fig. S5; see Dataset S4 for all numerical results). These trends, however, were small in magnitude and not consistently observed across robustness analyses (Dataset S4).

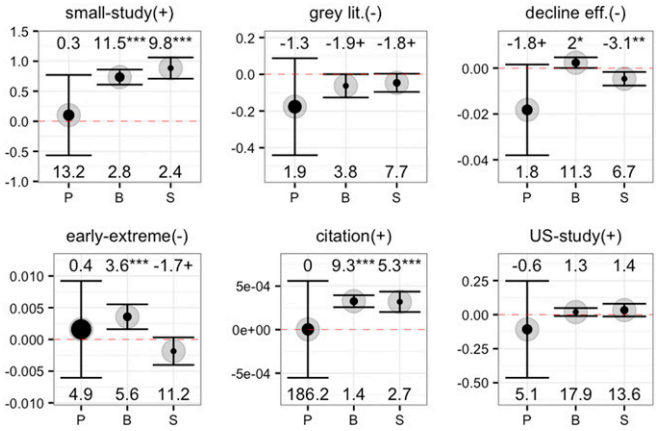

Fig. 2. Bias patterns partitioned by disciplinary domain. Each panel reports the second-order random-effects meta-analytical summaries of meta-regression estimates ($b \pm 95\%$ CI) measured across the sample of meta-analyses. Symbols in parentheses indicate whether the association between factor and effect size is predicted to be positive (+) or negative (–). The gray area within each circle is proportional to the percentage of total variance explained by between–meta-analysis variance (i.e., heterogeneity, measured by I^2). To help visualize effect sizes and statistical significance, numbers above error bars display *t* scores (i.e., summary effect size divided by its corresponding SE, *b*/SE) and conventional significance levels (i.e., $+P < 0.1$, $*P < 0.05$, $**P < 0.01$, $***P < 0.001$). Numbers below each error bar reflect the cross–meta-analytical consistency of effects, measured as the ratio of between–meta-analysis SD divided by summary effect size (i.e., τ/b; the smaller the ratio is, the higher the consistency). The sample was partitioned between meta-analyses from journals classified based on discipline indicated in Thompson Reuters' Essential Science Indicators database. Using abbreviations described in *Methods*, discipline classification is the following: physical sciences (P), MA, PH, CH, GE, EN, CS; social sciences (S), EB, PP, SO; and biological sciences (B), all other disciplines. See main text and Table 1 for further details.

Robustness of Results. Similar results were obtained if analyses were run including the noninterconvertible meta-analyses, although doing so yields extreme levels of heterogeneity and implausibly high regression coefficients in some disciplines (Dataset S5). We also reached similar conclusions if we analyzed the sample using multilevel weighted regression analysis. This method also allowed us to test all variables in a multivariable model, which confirmed that measured bias patterns and risk factors are all independent from one another (see analyses and discussion in *SI Multilevel Meta-Regression Analysis*, *Multilevel Analyses*, *Relative Independence of Biases*). Country of activity of an author might represent a significant confounding factor in our analyses, particularly if country is a surrogate for the way research is done and rewarded. For example, countries with incentives to publish may also have higher research standards, on average. Moreover, the accuracy of individual-level data may be lower for authors from many developing countries that have limited presence in English-speaking journals. However, very similar results were obtained when analyses were limited to studies from authors based in the United States (Fig. S6).

Discussion

Our study asked the following question: "If we draw at random from the literature a scientific topic that has been summarized by a meta-analysis, how likely are we to encounter the bias patterns and postulated risk factors most commonly discussed, and how strong are their effects likely to be?" Our results consistently suggest that small-study effects, gray literature bias, and citation bias might be the most common and influential issues. Small-study effects, in particular, had by far the largest magnitude, suggesting that these are the most important source of bias in meta-analysis, which may be the consequence either of selective reporting of results or of genuine differences in study design between small and large

studies. Furthermore, we found consistent support for common speculations that, independent of small-study effects, bias is more likely among early-career researchers, those working in small or long-distance collaborations, and those that might be involved with scientific misconduct.

Our results are based on very conservative methodological choices (*SI Limitations of Results*), which allow us to draw several conclusions. First, there is a good match between the focus of meta-research literature and the actual prevalence of different biases. Small-study effects, whether due to the file-drawer problem or to heterogeneity in study design, are the most widely understood and studied problem in meta-analysis (1). Gray literature bias, which partially overlaps but is distinct from the file-drawer problem, is also widely studied and discussed, and so is citation bias (1, 34). Moreover, we found at least partial support for all other bias patterns tested, which confirms that these biases may also represent significant phenomena. However, these biases appeared less consistently across meta-analyses, suggesting that they emerge in more localized disciplines or fields.

Second, our study allows us to get a sense of the relative magnitude of bias patterns across all of science. Due to the limitations discussed above, our study is likely to have underestimated the magnitude of all effects measured. Nonetheless, the effects measured for most biases are relatively small (they accounted for 1.2% or less of the variance in reported effect sizes). This suggests that most of these bias patterns may induce highly significant distortions within specific fields and meta-analyses, but do not invalidate the scientific enterprise as a whole. Small-study effects were an exception, explaining as much as 27% of the variance in effect sizes. Nevertheless, meta-regression estimates of small-study effects on odds ratio may run a modest risk of type I error (35) and were measured by our study more completely and accurately than the other biases (*SI Multilevel Meta-Regression Analysis*, *Reliability and Consistency of Collected Data*). Therefore, whereas all other biases in our study were measured conservatively, small-study effects are unlikely to be more substantive than what our results suggest. Moreover, small-study effects may result not necessarily from QRP but also from legitimate choices made in designing small studies (17, 36). Choices in study design aimed at maximizing the detection of an effect might be justified in some discovery-oriented research contexts (37).

Third, the literature on scientific integrity, unlike that on bias, is only partially aiming at the right target. Our results supported the hypothesis that early-career researchers might be at higher risk from bias, largely in line with previous results on retractions and corrections (16) and with predictions of mathematical models (26). The reasons why early-career researchers are at greater risk of bias remain to be understood. Our results also suggest that there is a connection between bias and retractions, offering some support to a responsible conduct of research program that assumes negligence, QRP, and misconduct to be connected phenomena that may be addressed by common interventions (38). Finally, our results also support the notion that mutual control between team members might protect a study from bias. The team characteristics that we measured are very indirect proxies of mutual control. However, in a previous study these proxies yielded similar results on retractions and corrections (16), which were also significantly predicted by sociological hypotheses about the academic culture of different countries (a hypothesis that this study was not designed to test). Therefore, our findings support the view that a culture of openness and communication, whether fostered at the institutional or the team level, might represent a beneficial influence.

Even though several hypotheses taken from the research integrity literature were supported by our results, the risk factors that were not supported included phenomena that feature prominently in such literature. In particular, the notion that pressures to publish have a direct effect on bias was not supported and even contrarian evidence was seen: The most prolific researchers and those working in countries where pressures are supposedly higher were significantly less likely to publish overestimated results, suggesting that researchers who publish more may indeed be better scientists and thus report their results more completely and accurately. A previous study testing the risk of producing retracted and corrected articles, with the latter assumed to represent a proxy of integrity, had similarly falsified the pressures to publish hypothesis as conceptualized here (16), and so did historical trends of individual publication rates (39). Therefore, cumulating evidence offers little support for the dominant speculation that pressures to publish force scientists to publish excessive numbers of articles and seek high impact at all costs (40–42). A link between pressures to publish and questionable research practices cannot be excluded, but is likely to be modulated by characteristics of study and authors, including the complexity of methodologies, the career stage of individuals, and the size and distance of collaborations (14, 39, 43). The latter two factors, currently overlooked by research integrity experts, might actually be growing in importance, at least in the social sciences (Fig. S5).

In a previous, smaller study, two of us documented the US effect (14). We did measure again a small US effect but this may not be easy to explain simply by pressures to publish, as previously speculated. Future testable hypotheses to explain the US effect include a greater likelihood of US researchers to engage in long-distance collaborations and a greater reliance on early-career researchers as first authors.

Other general conclusions of previous studies are at least partially supported. Systematic differences in the risk of bias between physical, biological, and social sciences were observed, particularly for the most prominent biases, as was expected based on previous evidence (5, 11, 33). However, it is not known whether the disciplinary and domain differences documented in this study are the result of different research practices in primary studies (e.g., higher publication bias in some disciplines) or whether they result from differences in methodological choices made by meta-analysts of different disciplines (e.g., lower inclusion of gray literature). Similarly, whereas our results support previous observations that bias may have increased in recent decades, especially in the social sciences (11), future research will need to determine whether and to what extent these trends might reflect changes in meta-analytical methods, rather than an actual worsening of research practices.

In conclusion, our analysis offered a "bird's-eye view" of bias in science. It is likely that more complex, fine-grained analyses targeted to specific research fields will be able to detect stronger signals of bias and its causes. However, such results would be hard to generalize and compare across disciplines, which was the main objective of this study. Our results should reassure scientists that the scientific enterprise is not in jeopardy, that our understanding of bias in science is improving and that efforts to improve scientific reliability are addressing the right priorities. However, our results also suggest that feasibility and costs of interventions to attenuate distortions in the literature might need to be discussed on a discipline- and topic-specific basis and adapted to the specific conditions of individual fields. Besides a general recommendation to interpret with caution results of small, highly cited, and early studies, there may be no one-size fits-all solution that can rid science efficiently of even the most common forms of bias.

Methods

During December 2013, we searched Thompson Reuters' Web of Science database, using the string ("meta-analy*" OR "metaanaly*" OR "meta analy*") as topic and restricting the search to document types "article" or "review." Sampling was randomized and stratified by scientific discipline, by restricting each search to the specification of journal names included in each of the 22 disciplinary categories used by Thompson Reuters' Essential Science Indicators database. These disciplines and the abbreviations used in Fig. 2 and Figs. S1 and S3 are the following: AG, agricultural sciences; BB, biology and biochemistry; CH, chemistry;

CM, clinical medicine; CS, computer science; EB, economics and business; EE, environment/ecology; EN, engineering; GE, geosciences; IM, immunology; MA, mathematics; MB, molecular biology and genetics; MI, microbiology; MS, materials science; MU, multidisciplinary; NB, neuroscience and behavior; PA, plant and animal sciences; PH, physics; PP, psychiatry/psychology; PT, pharmacology and toxicology; SO, social sciences, general; and SP, space sciences. Studies retrieved from MU were later reclassified based on topic.

In successive phases of selection, meta-analyses were screened for potential inclusion (Fig. S1), based on the following inclusion criteria: (*i*) tested a specified empirical question, not a methodological one; (*ii*) sought to answer such question based on results of primary studies that had pursued the same or a very similar question; (*iii*) identified primary studies via a dedicated literature search and selection; (*iv*) produced a formal meta-analysis, i.e., a weighted summary of individual outcomes of primary studies; and (*v*) the meta-analysis included at least five independent primary studies (see *SI Methods* for details).

For each primary study in each included meta-analysis we recorded reported effect size and measure of precision provided (i.e., confidence interval, SE, or *N*) and we retrieved available bibliometric information, using multiple techniques and databases and attempting to complete missing data with hand searches. Further details of each parameter collected and how variables tested in the study were derived are provided in *SI Methods*.

Each dataset was standardized following previously established protocols (14). Moreover we inverted the sign of (i.e., multiplied by −1) all primary studies within meta-analyses whose pooled summary estimates were smaller than zero—a process known as "coining." All analyses presented in the main text refer to meta-analyses coined with a threshold of zero, whilst results obtained with uncoined data and data with a more conservative coining threshold are reported in Datasets S1–S5.

Each individual bias or risk factor was first tested within each meta-analysis, using simple meta-regression. The meta-regression slopes thus obtained were then summarized by a second-order meta-analysis, weighting by the inverse square of their respective SEs, assuming random variance of effects across meta-analyses.

We also analyzed data using weighted multilevel regression, a method that yields similar but more conservative estimates (30). A complete discussion of this method and all results obtained with it can be found in *Supporting Information*.

ACKNOWLEDGMENTS. Data were collected primarily by research assistants Sophy Ouch, Sonia Jubinville, Mélanie Vanel, Fréderic Normandin, Annaelle Winand, Sébastien Richer, Felix Langevin, Hugo Vaillancourt, Asia Krupa, Julie Wong, Gabe Lewis, and Aditya Kumar. This work was entirely supported by the US Office of Research Integrity (Grant ORIIR140006).

1. Song F, et al. (2010) Dissemination and publication of research findings: An updated review of related biases. *Health Technol Assess* 14(8):iii, ix–xi, 1–193.
2. Chavalarias D, Ioannidis JPA (2010) Science mapping analysis characterizes 235 biases in biomedical research. *J Clin Epidemiol* 63(11):1205–1215.
3. Young NS, Ioannidis JPA, Al-Ubaydli O (2008) Why current publication practices may distort science. *PLoS Med* 5(10):e201.
4. Dwan K, et al. (2008) Systematic review of the empirical evidence of study publication bias and outcome reporting bias. *PLoS One* 3(8):e3081.
5. Fanelli D (2010) "Positive" results increase down the Hierarchy of the Sciences. *PLoS One* 5(4):e10068.
6. Dubben HH, Beck-Bornholdt HP (2005) Systematic review of publication bias in studies on publication bias. *BMJ* 331(7514):433–434.
7. Van Bavel JJ, Mende-Siedlecki P, Brady WJ, Reinero DA (2016) Contextual sensitivity in scientific reproducibility. *Proc Natl Acad Sci USA* 113(23):6454–6459.
8. Sterling TD, Rosenbaum WL, Weinkam JJ (1995) Publication decisions revisited - The effect of the outcome of statistical tests on the decision to publish and vice-versa. *Am Stat* 49(1):108–112.
9. de Winter J, Dodou D (2014) A surge of p-values between 0.040 and 0.049 in recent decades (but negative results are increasing rapidly too). *PeerJ PrePrints* (2):e447v443.
10. Pautasso M (2010) Worsening file-drawer problem in the abstracts of natural, medical and social science databases. *Scientometrics* 85(1):193–202.
11. Fanelli D (2012) Negative results are disappearing from most disciplines and countries. *Scientometrics* 90(3):891–904.
12. Fanelli D (2010) Do pressures to publish increase scientists' bias? An empirical support from US States Data. *PLoS One* 5(4):e10271.
13. van Dalen HP, Henkens K (2012) Intended and unintended consequences of a publish-or-perish culture: A worldwide survey. *J Am Soc Inf Sci Technol* 63(7):1282–1293.
14. Fanelli D, Ioannidis JPA (2013) US studies may overestimate effect sizes in softer research. *Proc Natl Acad Sci USA* 110(37):15031–15036.
15. Ioannidis JPA, Fanelli D, Dunne DD, Goodman SN (2015) Meta-research: Evaluation and improvement of research methods and practices. *PLoS Biol* 13(10):e1002264.
16. Fanelli D, Costas R, Larivière V (2015) Misconduct policies, academic culture and career stage, not gender or pressures to publish, affect scientific integrity. *PLoS One* 10(6): e0127556.
17. Sterne JAC, Gavaghan D, Egger M (2000) Publication and related bias in meta-analysis: Power of statistical tests and prevalence in the literature. *J Clin Epidemiol* 53(11):1119–1129.
18. Schooler J (2011) Unpublished results hide the decline effect. *Nature* 470(7335):437–437.
19. Ioannidis JPA, Trikalinos TA (2005) Early extreme contradictory estimates may appear in published research: The Proteus phenomenon in molecular genetics research and randomized trials. *J Clin Epidemiol* 58(6):543–549.
20. Jannot AS, Agoritsas T, Gayet-Ageron A, Perneger TV (2013) Citation bias favoring statistically significant studies was present in medical research. *J Clin Epidemiol* 66(3):296–301.
21. Lexchin J, Bero LA, Djulbegovic B, Clark O (2003) Pharmaceutical industry sponsorship and research outcome and quality: Systematic review. *Br Med J* 326(7400):1167–1170B.
22. Martinson BC, Crain AL, Anderson MS, De Vries R (2009) Institutions' expectations for researchers' self-funding, federal grant holding and private industry involvement: Manifold drivers of self-interest and researcher behavior. *Acad Med* 84(11):1491–1499.
23. Qiu J (2010) Publish or perish in China. *Nature* 463(7278):142–143.
24. Lee C, Schrank A (2010) Incubating innovation or cultivating corruption? The developmental state and the life sciences in Asia. *Soc Forces* 88(3):1231–1255.
25. Fanelli D (2012) When East meets West…does bias increase? A preliminary study on South Korea, United States and other countries. *8th International Conference on Webometrics, Informetrics and Scientometrics and 13th COLLNET Meeting*, eds Ho-Nam C, Hye-Sun K, Kyung-Ran N, Seon-Hee L, Hye-Jin K, Kretschmer H (KISTI, Seoul, Korea), pp 47–48.
26. Lacetera N, Zirulia L (2011) The economics of scientific misconduct. *J Law Econ Organ* 27(3):568–603.
27. Fang FC, Bennett JW, Casadevall A (2013) Males are overrepresented among life science researchers committing scientific misconduct. *mBio* 4(1):e00640–e12.
28. Kaatz A, Vogelman PN, Carnes M (2013) Are men more likely than women To commit scientific misconduct? Maybe, maybe not. *mBio* 4(2):2.
29. Antes AL, et al. (2007) Personality and ethical decision-making in research: The role of perceptions of self and others. *J Empir Res Hum Res Ethics* 2(4):15–34.
30. Davis MS, Wester KL, King B (2008) Narcissism, entitlement, and questionable research practices in counseling: A pilot study. *J Couns Dev* 86(2):200–210.
31. Bailey CD (2015) Psychopathy, Academic accountants' attitudes toward unethical research practices, and publication success. *Account Rev* 90(4):1307–1332.
32. Fanelli D, Ioannidis JPA (2014) Reply to Nuijten et al.: Reanalyses actually confirm that US studies overestimate effects in softer research. *Proc Natl Acad Sci USA* 111(7):E714–E715.
33. Fanelli D, Glanzel W (2013) Bibliometric evidence for a hierarchy of the sciences. *PLoS One* 8(6):e66938.
34. Pfeffer C, Olsen BR (2002) Editorial: Journal of negative results in biomedicine. *J Negat Results Biomed* 1(1):2.
35. Macaskill P, Walter SD, Irwig L (2001) A comparison of methods to detect publication bias in meta-analysis. *Stat Med* 20(4):641–654.
36. Ioannidis JPA (2008) Interpretation of tests of heterogeneity and bias in meta-analysis. *J Eval Clin Pract* 14(5):951–957.
37. Djulbegovic B, Hozo I (2007) When should potentially false research findings be considered acceptable? *PLoS Med* 4(2):211–217.
38. Pimple KD (2002) Six domains of research ethics. A heuristic framework for the responsible conduct of research. *Sci Eng Ethics* 8(2):191–205.
39. Fanelli D, Larivière V (2016) Researchers' individual publication rate has not increased in a century. *PLoS One* 11(3):e0149504.
40. Hayer C-A, et al. (2013) Pressures to publish: Catalysts for the loss of scientific writing integrity? *Fisheries* 38(8):352–355.
41. de Vries R, Anderson MS, Martinson BC (2006) Normal misbehavior: Scientists talk about the ethics of research. *J Empir Res Hum Res Ethics* 1(1):43–50.
42. Tijdink JK, Verbeke R, Smulders YM (2014) Publication pressure and scientific misconduct in medical scientists. *J Empir Res Hum Res Ethics* 9(5):64–71.
43. Vale RD (2015) Accelerating scientific publication in biology. *Proc Natl Acad Sci USA* 112(44):13439–13446.
44. Caron E, Van Eck, NJ (2014) Large scale author name disambiguation using rule-based scoring and clustering. *Proceedings of the 19th International Conference on Science and Technology Indicators*, ed Noyons E (Centre for Science and Technology Studies, Leiden University, Leiden, The Netherlands), pp 79–86.
45. Tijssen RJW, van Leeuwen TN, van Wijk E (2009) Benchmarking university-industry research cooperation worldwide: Performance measurements and indicators based on co-authorship data for the world's largest universities. *Res Eval* 18(1):13–24.
46. Waltman L, van Eck NJ, van Leeuwen TN, Visser MS, van Raan AFJ (2011) Towards a new crown indicator: Some theoretical considerations. *J Informetrics* 5(1):37–47.
47. Horton J, et al. (2010) Systematic review data extraction: Cross-sectional study showed that experience did not increase accuracy. *J Clin Epidemiol* 63(3):289–298.
48. Iqbal SA, Wallach JD, Khoury MJ, Schully SD, Ioannidis JPA (2016) Reproducible research practices and transparency across the biomedical literature. *PloS Biol* 14(1)e1002333.
49. Bates D, Mächler M, Bolker BM, Walker SC (2015) Fitting linear mixed-effects models using lme4. *J Stat Softw* 67(1):1–48.
50. Thompson SG, Sharp SJ (1999) Explaining heterogeneity in meta-analysis: A comparison of methods. *Stat Med* 18(20):2693–2708.
51. Johnson PCD (2014) Extension of Nakagawa & Schielzeth's R(2)GLMM to random slopes models. *Methods Ecol Evol* 5(9):944–946.