HHS Public Access

Author manuscript

Curr Epidemiol Rep. Author manuscript; available in PMC 2021 April 26.

Published in final edited form as:

Curr Epidemiol Rep. 2018 June ; 5(2): 175–183. doi:10.1007/s40471-018-0148-x.

The replication crisis in epidemiology: snowball, snow job, or winter solstice?

Timothy L. Lash, DSc, MPH[1], **Lindsay J. Collin, MPH**[1], **Miriam E. Van Dyke, MPH**[1]

[1]Department of Epidemiology, Rollins School of Public Health, Emory University

Abstract

Purpose of review: Like a snowball rolling down a steep hill, the most recent crisis over the perceived lack of reproducibility of scientific results has outpaced the evidence of crisis. It has led to new actions and new guidelines that have been rushed to market without plans for evaluation, metrics for success, or due consideration of the potential for unintended consequences.

Recent findings: The perception of the crisis is at least partly a snow job, heavily influenced by a small number of centers lavishly funded by a single foundation, with undue and unsupported attention to preregistration as a solution to the perceived crisis. At the same time, the perception of crisis provides an opportunity for introspection. Two studies' estimates of association may differ because of undue attention on null hypothesis statistical testing, because of differences in the distribution of effect modifiers, because of differential susceptibility to threats to validity, or for other reasons. Perhaps the expectation of what reproducible epidemiology ought to look like is more misguided than the practice of epidemiology. We advocate for the idea of "replication and advancement." Studies should not only replicate earlier work, but also improve on it in by enhancing the design or analysis.

Summary: Abandoning blind reliance on null hypothesis significance testing for statistical inference, finding consensus on when pre-registration of non-randomized study protocols has merit, and focusing on replication and advance are the most certain ways to emerge from this solstice for the better.

Keywords

Epidemiologic Methods; Reproducibility of Results

INTRODUCTION

The scientific and lay communities have recently raised concerns about poor reproducibility of scientific results [1–3]. To answer these concerns, a committee has proposed guidelines for journals to achieve transparency and openness [4], editors have suggested changes to reporting principles [5], the United States National Institutes of Health has asked grant applicants to address rigor and reproducibility [6], and a long list of coauthors has proposed lowering the conventional Type 1 error rate from $\alpha=0.05$ to $\alpha=0.005$ [7] while others have

Correspondent: Timothy L. Lash, Department of Epidemiology, Emory University, 1518-002-3BB, 1518 Clifton Rd. NE, Atlanta, GA 30033, (t) 404.712.1270, tlash@emory.edu.

proposed eliminating null hypothesis significance testing altogether [8–11]. These proposals have not included plans to measure program effectiveness or the extra effort required by authors, editors, grant writers, and reviewers. A potential unintended consequence might be, reduced creativity in the scientific enterprise resulting from the increased requirements for compliance [12–14].

Epidemiologic evidence influences almost every aspect of daily life, at least in societies with a higher human development index. We bathe, brush our teeth, eat vitamin-fortified foods, fasten our safety belts, use protective equipment at work, pay attention to hours of screen time and calories consumed, avoid tobacco and excessive alcohol, decry inequities in the distribution of wealth, desire to breath clean air and to drink clean water, put infants to sleep on their backs, and practice safe sex. These daily routines, and many others, are substantially influenced by epidemiologic research. Despite its many successes, concerns about the credibility of epidemiologic research are long-standing. Epidemiologists work perpetually in the long shadows of the shortest day of the year [15–17]. always striving for brighter days ahead when our work will be more highly regarded.

Many of the most critical appraisals of epidemiology have focused on the real or perceived abundance of false-positive associations purportedly found in our research papers [18, 16, 17]. More than twenty years ago, Taubes famously suggested that the practice of epidemiology had reached its limit [16]. In that paper, Trichopoulos said that epidemiology studies will inevitably generate false-positive and false-negative results. That was the only instance in the paper where the concept of a false-negative result received any mention; the remainder of the paper focused entirely on false-positive results. This emphasis on false-positive associations, with almost no attention to false-negative associations, is characteristic of most of the literature that criticizes patterns of results from epidemiologic research [17, 19, 20, 18, 21], and is the foundational literature at the core of the snowball that began rolling downhill at the outset of the most recent epoch (we have been here before [22]).

THE SNOWBALL: RECENT MOMENTUM OF THE REPRODUCIBILITY CRISIS

This most recent epoch traces to Ioannidis' 2005 paper titled "Why most published research findings are false" [17]. This paper has been cited more than 850 times; one well-reasoned response has been cited only 8 times [23]. Momentum grew with a 2009 report from a workshop sponsored by European Centre for Ecotoxicology and Toxicology of Chemicals (ECETOC) [24], which was titled "Enhancement of the Scientific Process and Transparency of Observational Epidemiology Studies." According to its website, the ECETOC is an independent, non-profit, non-commercial and non-governmental organization established to provide a scientific forum through which the extensive specialist expertise of manufacturers and users of chemicals could be harnessed to research, review, assess and publish studies on the ecotoxicology and toxicology of chemicals. It is financed by its membership, which is comprised of the leading companies with interests in the manufacture and use of chemicals, biomaterials and pharmaceuticals. The introductory paragraph of the summary and recommendations section of the workshop report reads:

> Among the workshop participants there was general consensus that current practice in observational epidemiology research is often far from the ideal and that improvements need to be made in particular with regard to enhancing transparency and credibility of observational epidemiological research. The issues of publication and other biases along with undocumented deviation from original study designs were highlighted. The workshop recognized several key points on how to enhance the transparency and credibility of observational epidemiology studies.

First among the recommendations was compulsory preregistration of observational epidemiology studies in a public database following the model of clinical trials. This recommendation was widely discussed in clinical and epidemiology journals (see Lash and Vandenbroucke [25], for example, as well as citations therein). Concerns included a flawed analogy to compulsory registration of clinical trials, philosophical objections, and practical barriers to successful implementation for which there is now empirical support [26]. The original motivation for registration of trials was to assure that null trial results would be known even if never published, so that new trial participants would not be randomized to a treatment thought to be ineffective [27, 28]. There were additional benefits, but the focus was on assuring all evidence came to light and preserving the ethical duty to do no harm to trial participants. The motivation for registration of observational studies, in contrast, was to suppress or down-weight study results that were not preregistered, on the supposition that such studies were more likely to produce false-positive results [29–31]. The lack of evidence to support the contention that preregistered studies were more likely to produce false-positive results was made clear in the workshop report. The second recommendation was to encourage research on the need for preregistering observational epidemiologic studies: “more research is needed to demonstrate that registration can indeed help to obtain the most un-biased estimate of an association from observational epidemiologic studies.”

Support for the idea of preregistration of observational epidemiology studies declined between 2009 to 2012, only to re-emerge as one element of the Transparency and Openness Promotion (TOP) Guidelines [4]. These guidelines were proposed by a self-appointed committee “to encourage greater transparency and reproducibility of research in the published record” with the goal of improving “both research practices and the credibility and reputation of our field.” The guidelines promulgate eight standards to be implemented by journals. Journals that adopt the guidelines choose the level of their implementation for each standard, ranging from level zero—connoting no implementation of the standard—to level three—connoting full implementation and active policing of the standard by the journal editors. Guideline six addresses preregistration of studies. For level 0, journal guidelines would say nothing or encourage preregistration of studies without requiring authors to state whether preregistration occurred and without providing links to the preregistration. For Level 1, journals would encourage preregistration, and require links in text to preregistrations of studies if they exist.

For Level 2, journals would verify that the preregistration follows standards and indicates certification of meeting those standards. In addition to Level 2 guidelines, Level 3 would require all reported studies to be preregistered. The invitation to participate in the TOP guidelines was declined by EPIDEMIOLOGY (N.B., TLL is the current Editor-in-Chief and

wrote this editorial) [12], and a search of the Center for Open Sciences web site does not list the *American Journal of Epidemiology* or *International Journal of Epidemiology* among the signatories as of March 2018.

THE SNOW JOB: "SOLUTIONS" TO THE REPRODUCIBILITY CRISIS

One organization has led the charge to support studies of the reproducibility crisis and to suggest solutions: the Laura and John Arnold Foundation. The Foundation has invested more than $28 million dollars, 52% of their investment in "Research Integrity" grants, to the Center for Open Science and the Meta-Research Innovation Center at Stanford (METRICS). The Center for Open Science has received more than $19 million, amounting to 73% of the direct financial support they list [32]. The Board of Trustees of the Leland Stanford Junior University, largely in support of METRICS, was awarded up to $9,899,739 from the foundation. These two centers have been very high-profile advocates for changing the practice of science, and particularly for the idea of preregistering protocols to improve reproducibility. Stakeholders ought to be aware of the outsized influence of this one foundation, which is funded from the wealth of two individuals, and whose Research Integrity awards appear to be granted without a competition of ideas reviewed by an external panel.

The TOP guidelines were drafted by a committee organized by the Center for Open Science. Stuart Buck, who was Vice-President of Research Integrity at the Laura and John Arnold Foundation, was a member of the committee. He also coauthored the announcement of the TOP guidelines in *Science* [4] and solely authored a commentary [33] titled "Solving Reproducibility," which appeared in the same issue and which included the assertion that demanding more preregistration was an obvious solution to the reproducibility problem. *Science* is published by the American Association for the Advancement of Science, which has received a grant of $98,889 from the Laura and John Arnold Foundation [34]. The publication of the TOP guidelines disclosed support from the Laura and John Arnold Foundation. The commentary by Stuart Buck did not disclose the grant or the potential conflict of interest created when an officer of the awarding foundation authors a commentary in the journal published by the recipient organization.

METRICS is a "research to action center focused on transforming research practices to improve the quality of scientific studies in biomedicine and beyond". It is co-directed by John P.A. Ioannidis, who is a prominent critic of biomedical research [17, 35–37, 1, 38], and was launched by "a founding grant from the Laura and John Arnold Foundation." Brian Nosek, the co-founder and Executive Director of the Center for Open Science, has similarly recognized the central role of the Foundation in its work, saying that the award from the Laura and John Arnold Foundation "completely transformed what we could imagine doing" [39]. Work from both centers emphasizes the importance of preregistration to improve reproducibility, despite the dearth of evidence supporting its effectiveness. The Center for Open Science has offered $1000 each to the first 1000 investigators to preregister their research with its Center, a program suggested by John Arnold to the Center. This million dollars might have been used to implement an intervention with demonstrated effectiveness, as reported in a paper co-authored by Nosek [40], and that does not require preregistration.

The larger point is not who is supporting the work, or why, but that one organization has awarded an unprecedented amount of support to prominent centers in the name of research integrity to improve reproducibility. The awards to the Center for Open Science and METRICS were apparently made without open calls for competitive applications and without external review [39]. In fact, awards were initiated by the foundation contacting the center founders upon learning of the views they held. Both centers suggest that their work is possible at its current level largely because of these awards, and that previous efforts to obtain support at equivalent levels by application to competitive extramural funders had been unsuccessful. It is important for the affected community of scientists to realize that a single foundation is supporting the mission of these centers, and that they are selected for support because of espousing views consistent with the views and priorities of the foundation. It seems unlikely that a second foundation would support similar centers that hold opposing views.

THE WINTER SOLSTICE: REPRODUCIBILITY IN A WIDER CONTEXT

The link between reproducibility and consistency

The underlying concept of reproducibility is familiar because it evokes “consistency,” which is the label for one of the Bradford Hill viewpoints for causal inference [41]. In describing this labeled viewpoint, Bradford Hill posed the question: “Has it (an association) been repeatedly observed by different persons, in different places, circumstances and times?” This description of the question to be asked when evaluating consistency certainly evokes the idea of replication or reproducibility. Asking for consideration of whether consistent results have been reported by different persons suggests that Bradford Hill was concerned about competing interests affecting associations, and presumes that these competing interests would have different influences on different persons. Asking for consideration of whether consistent results have been reported in different places, circumstances, and times suggests that Bradford Hill was concerned about modification of associations by population characteristics that vary with geography, calendar period, or other circumstances. Most importantly, Bradford Hill provided no guidance on how to evaluate whether two or more associations were “consistent.”

Bradford Hill did not intend these viewpoints to be “criteria” for evaluation of causality, which may explain why he avoided describing exactly how to evaluate each viewpoint. In fact, he explicitly disavowed that his viewpoints could be used as a checklist, singly or collectively:

> None of my nine viewpoints can bring indisputable evidence for or against the cause and-effect hypothesis and none can be required as a sine qua non. What they can do, with greater or less strength, is to help us to make up our minds on the fundamental question — is there any other way of explaining the set of facts before us, is there any other answer equally, or more, likely than cause and effect?

Despite this admonition, aspects of these nine viewpoints are frequently used as causal criteria [42, 43]. As anticipated by Bradford Hill (“greater or less strength”), the nine viewpoints are weighed differently by different analysts, with the choice of which viewpoint

to weigh heavily more “a matter of personal preference than of careful inquiry” [44]. Nonetheless, strength of association, consistency, biological plausibility, and dose response are the viewpoints most commonly mentioned in causal inferences that invoke the Bradford Hill viewpoints [44]. Consistency is not as influential on judgements of causation as statistical significance or strength of association, and the influence of consistency reaches saturation when the number of studies reporting a consistent association reaches double digits [45]. This marginal influence of consistency on causal inference has logical underpinnings. Lack of “consistency” across a collection of study results does not rule out a causal relation, because some causal mechanisms only produce an effect when combined with other component causes [46]. Furthermore, a pattern of consistent results across studies cannot be viewed as a clear signal for a causal effect, since biases may be as easily reproduced as causal effects [47]. For these reasons, and in agreement with Bradford Hill’s original manuscript, “consistency” or “reproducibility” of associations should not be viewed as a *sine qua non* when considering whether an association is causal.

Factors contributing to the perception of a “Reproducibility Crisis”

Given the recent attention to the reproducibility crisis, one might think that there would be widespread agreement about whether the result of an initial research study has been reproduced by a second study. There is not. One approach is to assess whether the two studies have results that are concordant with respect to their categorizations as statistically significant or not. This approach has been used in the literature that assesses reproducibility [48], although the idea that two results—one statistically significant and the other not—are necessarily different from one another is a well-known fallacy [49, 50]. Nonetheless, examples of claims of irreproducible results based on p-values falling on opposites of the commonly accepted Type 1 error rate are easy to find. In one example, subcutaneous heparin was reported to reduce the risk of deep vein thrombosis compared with intravenous heparin (OR=0.62, 95% CI 0.39, 0.98); a reanalysis disagreed with the conclusion of a protective effect (OR=0.61, 95% CI 0.30, 1.25) [51]. In a second example, authors concluded that their results (OR=0.75, 95% CI 0.48, 1.17) did not support previous sparse evidence of a protective effect of statins use against glioma (previous study results were reported to be OR=0.72; 95% CI 0.52, 1.00 and OR= 0.76; 95% CI 0.59, 0.98) [52]. Finally, in a study of the association between antidepressant use during pregnancy and autism spectrum disorder in offspring, the authors reported a meta-analysis of earlier studies (OR=1.7; 95% CI 1.1, 2.6), a multivariate adjusted hazards ratio in their study (1.59; 95% CI 1.17, 2.17), and an inverse probability of treatment weighted hazards ratio in their study (1.61; 95% CI 0.997, 2.59) [53]. They concluded “antidepressant exposure compared with no exposure was not associated with autism spectrum disorder in the child.” Examples of this type of misinterpretation abound, and likely contribute to the perception that epidemiologic results are poorly reproducible when, at least in these examples, the evidence base is entirely consistent. It is impossible to estimate the degree to which this common misinterpretation has misguided the impression of a reproducibility crisis.

A more sensible approach is to compare whether the estimated parameter in the two studies are similar [54], which has also been implemented in the reproducibility literature [48]. This approach is more defensible, but some allowance must be made for the expectation that

results will differ due to different external influences, different internal influences, and the role of chance.

In epidemiologic research, the strength of association is partly determined by the distribution of complementary component causes [46, 55]. Thus, an association measured in one population would not be expected to have the same strength of association when measured in a second population, unless the two populations were exchangeable (*i.e.*, had the same distribution of complementary component causes, some of which may be unknown and therefore unmeasured). A second conceptualization of this requirement for expectation of homogenous estimates of strength of association is that the distribution of effect measure modifiers must be the same in the two populations. Recent work has formalized this concept in the realm of understanding generalizability and transportability to a second population [56–58]. If standardization for modifiers is required to generalize or transport an estimate from one population to a second population, then comparison of the same estimate of association obtained from that second population with the estimate from the first population, without standardizing to the distribution of effect modifiers, would lead to an expectation that the two estimates should be different, not to an expectation that the two estimates should be the same. Comparison of estimates of association in two populations without this standardization would lead to the appearance of poor reproducibility, when in fact the estimates may both be valid within their individual populations.

Comparison of the size of estimates of association from two or more studies should also take into account the differential influences of threats to internal validity, which we usually label biases [59, 60]. Studies are susceptible to selection bias because of differential baseline participation or loss to follow-up; to information bias reflecting errors in measuring the exposure, the outcome, and covariates; and to incomplete control of confounding [61]. Studies may also misalign the start of eligibility, the start of follow-up, and the exposure assessment period [62]. Because every study has imperfections [63], two different studies are likely to be differentially susceptible to these biases, which would be expected to yield different estimates of association, even if the source populations were exchangeable. One could argue that this is a source of poor reproducibility in epidemiologic research. Though they are rarely implemented [61, 64, 65], methods exist to quantitatively account for these biases and are commonly referred to as quantitative bias analyses. There are, however, limitations and additional assumptions that need to be made to complete these analyses, which is of concern because of the lack of wide consensus about how to implement these methods [66].

Finally, two studies' estimates of an equivalent parameter may differ from one another by chance, and our intuitions for how much they might differ by chance are not very good [49]. Furthermore, the very strong selection pressures in the publication process that favor statistically significant estimates affects the reproducibility of the magnitude of effect estimates [8]. In a scientific culture that focuses on statistically significant results [67], effects are more likely to be overestimated than underestimated whenever power is less than 100%, as seen in one of the replication projects [48]. In that project, 82 of 99 studies showed a stronger effect size in the original study than in the replication study. This pattern is what should be expected if the original studies were selected because their results were

statistically significant. On average, these studies' results should be overestimates. The selection pressures do not apply to the replication studies, so their results should be expected to regress towards the null [19]. Publication bias [68, 69], p-hacking [70], HARKing [71], and significance questing [67, 72] are manifestations of this problem. By focusing on results that are statistically significant, null hypothesis significance testing has built a machine to overestimate the truth. These pressures cause early studies to have inflated estimates, and then subsequent studies may use the inflated results as the target estimates when designing a replication study, leading to underpowered replication studies that falsely fail to demonstrate reproducibility. One cannot rationally label the resulting poor reproducibility as a crisis; the accumulation of evidence is behaving exactly as expected.

The need for "Replication and Advance"

Given the recent attention to the reproducibility crisis, one might also think that there would be widespread agreement about strategies that ought to yield research results that are more often reproducible. There is not. As noted above, The Center for Open Science's TOP Guidelines [4] have not been adopted by most of the general epidemiology journals, and have been rejected explicitly by one [12]. Compliance at signatory journals has been mixed [73]. The debate over the value, or lack thereof, of pre-registration of observational research studies has reached no consensus [25]. In one study of pre-registration of observational research, few studies had been pre-registered, registration usually occurred after the study had started, and pre-specification of outcomes and statistical analysis rarely occurred [26]. As noted above, a long list of authors has proposed lowering the conventional Type 1 error rate from α=0.05 to α=0.005 [7], a proposal that has been widely criticized by a second long list of coauthors [11] while others have proposed eliminating null hypothesis significance testing altogether [8–11]. In short, while epidemiologic research results may be less reproducible than expected, recommendations for how to address this perceived lack of reproducibility are even less reproducible.

The one set of recommendations to improve reproducibility that has received little criticism are the U.S. National Institute of Health's program to enhance rigor and reproducibility [74]. In keeping with this program, the National Institutes of Health, National Cancer Institute, and National Heart, Lung and Blood Institute have described strategic efforts with respect to the future of epidemiologic research [74]. Although the statements are multifaceted, one concern is related to resource sharing, especially about whether data and analytic code should be available to enhance the reproducibility of results [75]. These concerns have been raised partly in reaction to the era of big data [76]. As big data become more readily available, spurious associations from commonly used statistical tests—leading to false inference and irreproducible results—may become more common. To aid in resource sharing, journals have begun providing authors the option to submit data and code alongside their manuscript submissions. Alternative approaches to open source epidemiology are necessary; application of these methods to publicly available datasets is one safeguard likely to yield benefits [75].

The goal of reproducibility in science should extend beyond the ability to consistently observe comparable estimates of equivalent measures of effect in different populations,

across different points in time, and by different designs. Rather the research community in a particular topic area should move towards a common underlying understanding of a potentially causal effect that is of public health importance [77]. To achieve this goal, each study should contribute more to the state of knowledge than previous reports, ruling out competing explanations for an observed association. If a study reports an important association between an exposure (E) and outcome (D), there could be at least six explanations: E causes D, D causes E, chance, selection bias, information bias, or confounding [78]. Subsequent studies should endeavor to both reproduce the initial association, and to diminish the credibility of the five explanations that compete with the "E causes D" explanation [79, 80]. This goal is in line with an approach that Munafo and Davey Smith refer to as "triangulation" [81]. Advances accrue by the improvement of study design to address the bias [59] and chance [82] explanations, and by the implementation of bias analysis [83] to quantitatively evaluate those biases that remain. This approach would more efficiently allocate scarce research resources compared to merely replicating initial findings without adding further value to the body of literature on a public health question.

Examples of the Role for Reproducibility in Practice

Given that "replication and advancement" are important scientific goals, there is a hazard in emphasizing that the reproducibility of an association is separate from, or more important than, considering its validity, and its population health consequences. In his 1965 Presidential address, Bradford Hill reminded us that "all scientific work is incomplete;" knowledge will always be subject to modification through new advances [41]. Nonetheless, these continuous advancements should not preclude the use of the knowledge already gained to take actions intended to protect public health.

For example, in 2001 Kieler *et al.* estimated the effect of ultrasound exposure during pregnancy on the left-handedness of offspring [84]. The study assessed left-handedness only among men. Handedness had to be measured among Swedish military conscripts, all of whom were men, because one rifle model could not be accurately fired by left-handed soldiers. The excess of left-handedness among men who had undergone ultrasound *in utero* was interpreted by the authors as implicating prenatal ultrasounds as harmful to the fetal brain. Rothman penned a commentary that accompanied this study, suggesting non-causal explanations for the observed association and questioning the importance of left-handedness as a public health problem, unless it really was a surrogate for harm done by ultrasound exposure to the fetal brain [84]. Rothman called for additional studies to "test this finding." Recognize, however, the barriers to direct replication: where else would population-wise (at least among men) measures of left-handedness be available to linkage with histories of ultrasonography *in utero?* Perhaps public health action should have been more immediate. Ten years later, a meta-analysis of data from three randomized trials examining the relation between ultrasound in pregnancy and left-handedness reported similar findings [85]. Although the initial study's association was reproduced—using studies of different design and in different settings—the biological link between excess left-handedness and the health of the fetal brain is still lacking. The American College of Obstetricians and Gynecologists has published guidelines for ultrasound during pregnancy, which includes the statement that "there is no evidence that ultrasound is harmful to a developing fetus" [86].

The topic area of hormone replacement therapy in relation to coronary heart disease provides an example for which reproducibility may have done harm, and whereas "reproduce and advance" helped to move the public health and clinical fields to take warranted and beneficial action. In the early stages of research in this topic area, results from non-randomized studies [87, 88] consistently reported a protective association between hormone replacement therapy and coronary heart disease in menopausal women [87, 88]. It had also been shown that hormone replacement therapy increased the risk of breast cancer, although some viewed the reduction in the risk of coronary heart disease and menopausal symptom relief as outweighing the increased risk of breast cancer. On the strength of the reproduced body of evidence pertaining to both coronary heart disease and breast cancer, many women over the course of decades used hormone replacement therapy. The first advance came when studies of more rigorous design, in this case randomized trials, showed there was no reduction in the risk of coronary heart disease among the women randomized to receive hormone replacement therapy [89, 90], but there was an increased risk of breast cancer [90]. It had been previously suggested that the apparent beneficial impact of hormone replacement therapy on coronary heart disease risk may have been due to uncontrolled confounding [91], such as unmeasured healthy lifestyle factors. For example, Petitti *et al.* reported a reduced risk of cardiovascular mortality among postmenopausal users of estrogen, compared with non-users, and also found a reduced risk of death from accidents, suicide, and homicide. They wrote:

> There is no biologically plausible reason for a protective effect of postmenopausal estrogen use on mortality from accidents, homicide, and suicide. We believe that our results are best explained by the assumption that postmenopausal estrogen users in this cohort are healthier than those who had no postmenopausal estrogen use, in ways that have not been quantified and cannot be adjusted for.

However, this advance, in conjunction with their replication, was largely ignored. A second advance showed that data from a randomized trial and a non-randomized study would have given the same results if analyzed by the same methods, suggesting that the time of onset of hormone replacement use in relation to menopause was an important modifying factor that could explain the discrepant results [92]. In this example, the exposure-outcome relation was consistently reproduced, yet lack of attention to the validity of the consistently reproduced design, analysis, and inference led to inaccurate conclusions that were later changed due to efforts to reproduce and advance, with an emphasis on "advance." In fact, unanimously consistent and reproduced results, such as obtained from the initial non-randomized studies of the relation between hormone replacement therapy and cardiovascular disease, ought to be cause for skepticism, not firm conclusion [93].

CONCLUSION

We began by asking whether the replication crisis in epidemiology was a snowball, snow job, or winter solstice. It is, in fact, all of these. The perception of crisis has outpaced the evidence of crisis. It has led to new actions and new guidelines that have been rushed to market without plans for evaluation, metrics for success, or due consideration of the potential for unintended consequences. In this regard, the current iteration of the

reproducibility crisis is a snowball gathering speed and size as it crashes down a mountain, absorbing or obliterating whatever is in its path. The perception of the crisis is also a snow job, leading practitioners, consumers, policy makers and other stakeholders to believe that epidemiologic research lacks credibility because it is poorly reproducible. Over the long view, epidemiologic research is credible and replicable; it informs all aspects of daily life in many societies, and the public health gains resting in whole or in part on epidemiologic evidence are enormous. The forces that would have us believe otherwise are well organized and have overt or covert alternative motives. Practitioners of our science must be vigilant about the snow job's influence on our self-perceptions or external perceptions. At the same time, careful introspection motivated by concerns of lack of reproducibility may lead to changes that improve the practice of our science. These might include abandoning blind reliance on null hypothesis significance testing for statistical inference and finding consensus on when pre-registration of non-randomized study protocols has merit. If so, then we will emerge from this solstice for the better.

References

1. Ioannidis JP. How to make more published research true. PLoS medicine 2014;11(10):e1001747. doi:10.1371/journal.pmed.1001747. [PubMed: 25334033]
2. Unreliable research: trouble at the lab. Economist 2013 19 10 2013.
3. Collins FS, Tabak LA. Policy: NIH plans to enhance reproducibility. Nature 2014;505(7485):612–3. [PubMed: 24482835]
4. Nosek BA, Alter G, Banks GC, Borsboom D, Bowman SD, Breckler SJ et al. SCIENTIFIC STANDARDS. Promoting an open research culture. Science. 2015;348(6242):1422–5. doi:10.1126/science.aab2374. [PubMed: 26113702]
5. Journals unite for reproducibility. Nature 2014;515(7525):7. doi:10.1038/515007a.
6. US National Institutes of Health. Rigor and Reproducibility 2016. http://grants.nih.gov/reproducibility/index.htm#guidance. Accessed July 6, 2016.
7. Benjamin D, Berger J, Johannesson M, al. E. Redefine Statistical Significance. Unpublished Manuscript. 2017.
8••. Lash TL The Harm Done to Reproducibility by the Culture of Null Hypothesis Significance Testing. American Journal of Epidemiology 2017;186(6):627–35. doi:10.1093/aje/kwx261. [PubMed: 28938715] Demonstrates that null hypothesis significance testing leads to the appearance of poor reproducibility by at least four mechanisms, yet few proposed interventions to improve reproducibility have suggested change to the culture of null hypothesis significance testing.
9. Matthews R, Wasserstein R, Spiegelhalter D. The ASA's p-value statement, one year on. Significance 2017;14(2):38–41. doi:10.1111/j.1740-9713.2017.01021.x.
10. McShane B, Gal D, Gelman A, Robert C, Tackett J. Abandon Statistical Significance. Unpublished Manuscript 2017.
11. Trafimow D, Amrhein V, Areshenkoff C, al. E. Manipulating the alpha level cannot cure significance testing – comments on "Redefine statistical significance". Unpublished Manuscript 2017.
12. Lash TL. Declining the Transparency and Openness Promotion Guidelines. Epidemiology 2015;26(6):779–80. doi:10.1097/ede.0000000000000382. [PubMed: 26348161]
13. Lash TL. Lash Responds to "Is Reproducibility Thwarted by Hypothesis Testing?" and "The Need for Cognitive Science in Methodology". American Journal of Epidemiology 2017;186(6):646–7. doi:10.1093/aje/kwx260. [PubMed: 28938714]
14. Why Crane H. "Redefining Statistical Significance" will not Improve Reproducibility and could make the Replication Crisis Worse. Unpublished Manuscript 2017.

15. Feinstein AR. Scientific standards in epidemiologic studies of the menace of daily life. Science 1988;242(4883):1257–63. [PubMed: 3057627]

16. Taubes G Epidemiology faces its limits. Science. 1995;269(5221):164–9. [PubMed: 7618077]

17. Ioannidis JP. Why most published research findings are false. PLoSMed. 2005;2(8):e124.

18•. Blair A, Saracci R, Vineis P, Cocco P, Forastiere F, Grandjean P et al. Epidemiology, public health, and the rhetoric of false positives. Environ Health Perspect 2009;117(12):1809–13. doi:10.1289/ehp.0901194. [PubMed: 20049197] One of several papers emphasizing the importance of false-positive associations without due consideration to the importance of false-negative associations.

19. Ioannidis JP. Why most discovered true associations are inflated. Epidemiology 2008;19(5):640–8. doi:10.1097/EDE.0b013e31818131e7. [PubMed: 18633328]

20. Ioannidis JP, Tarone R, McLaughlin JK. The false-positive to false-negative ratio in epidemiologic studies. Epidemiology 2011;22(4):450–6. doi:10.1097/EDE.0b013e31821b506e. [PubMed: 21490505]

21. McLaughlin JK, Tarone RE. False positives in cancer epidemiology. Cancer Epidemiol Biomarkers Prev 2013;22(1):11–5. doi:10.1158/1055-9965.EPI-12-0995. [PubMed: 23118145]

22•. Mayes LC, Horwitz RI, Feinstein AR. A collection of 56 topics with contradictory results in case-control research. Int J Epidemiol 1988;17(3):680–5. [PubMed: 3272133] Demonstrates long-standing concerns about the reproducibility of epidemiologic research.

23. Goodman S, Greenland S. Why most published research findings are false: problems in the analysis. PLoS medicine 2007;4(4):e168. doi:10.1371/journal.pmed.0040168. [PubMed: 17456002]

24. Chemicals ECfEaTo. ECETOC Workshop Report No. 18. 2009.

25•. Lash TL, Vandenbroucke JP. Commentary: Should Preregistration of Epidemiologic Study Protocols Become Compulsory?: Reflections and a Counterproposal. Epidemiology. 2012;23(2):184–8. doi:10.1097/EDE.0b013e318245c05b. [PubMed: 22317802] Review of advantages and disadvantages of compulsory preregistration of nonrandomized epidemiologic research.

26. Boccia S, Rothman KJ, Panic N, Flacco ME, Rosso A, Pastorino R et al. Registration practices for observational studies on ClinicalTrials.gov indicated low adherence. J Clin Epidemiol 2016;70:176–82. doi:10.1016/j.jclinepi.2015.09.009. [PubMed: 26386325]

27. De Angelis C, Drazen JM, Frizelle FAP, Haug C, Hoey J, Horton R et al. Clinical Trial Registration: A Statement from the International Committee of Medical Journal Editors. New England Journal of Medicine 2004;351(12):1250–1. doi:10.1056/NEJMe048225.

28. Krleza-Jeric K, Chan AW, Dickersin K, Sim I, Grimshaw J, Gluud C. Principles for international registration of protocol information and results from human trials of health related interventions: Ottawa statement (part 1). BMJ. 2005;330(7497):956–8. doi:10.1136/bmj.330.7497.956. [PubMed: 15845980]

29. Williams RJ, Tse T, Harlan WR, Zarin DA. Registration of observational studies: is it time? CMAJ. 2010;182(15):1638–42. doi:10.1503/cmaj.092252. [PubMed: 20643833]

30. Bracken MB. Preregistration of epidemiology protocols: a commentary in support. Epidemiology 2011;22(2):135–7. doi:10.1097/EDE.0b013e318207fc7c. [PubMed: 21293203]

31. Loder E, Groves T, MacAuley D. Registration of observational studies. BMJ 2010;340. doi:10.1136/bmj.c950.

32. Center for Open Science. Our Sponsors https://cos.io/about/our-sponsors/.

33. Buck S Solving reproducibility. Science 2015;348(6242):1403. doi:10.1126/science.aac8041. [PubMed: 26113692]

34. Laura and John Arnold Foundation. Grants http://www.arnoldfoundation.org/grants/

35. Begley CG, Ioannidis JP. Reproducibility in science: improving the standard for basic and preclinical research. Circ Res 2015;116(1):116–26. doi:10.1161/CIRCRESAHA.114.303819. [PubMed: 25552691]

36. Iqbal SA, Wallach JD, Khoury MJ, Schully SD, Ioannidis JP. Reproducible Research Practices and Transparency across the Biomedical Literature. PLoS Biol 2016;14(1):e1002333. doi:10.1371/journal.pbio.1002333. [PubMed: 26726926]

37. Stodden V, McNutt M, Bailey DH, Deelman E, Gil Y, Hanson B et al. Enhancing reproducibility for computational methods. Science 2016;354(6317):1240–1. doi:10.1126/science.aah6168. [PubMed: 27940837]
38. Munafò MR, Nosek BA, Bishop DVM, Button KS, Chambers CD, Percie du Sert N et al. A manifesto for reproducible science. Nature Human Behaviour 2017;1:0021. doi:10.1038/s41562-016-0021.
39. Apple S John Arnold Made a Fortune at Enron. Now He's Declared War on Bad Science. Wired. 2017.
40. Dreber A, Pfeiffer T, Almenberg J, Isaksson S, Wilson B, Chen Y et al. Using prediction markets to estimate the reproducibility of scientific research. PNAS 2015;112(50):15343–7. [PubMed: 26553988]
41. Hill AB. The Environment and Disease: Association or Causation? Proceedings of the Royal Society of Medicine 1965;58:295–300. [PubMed: 14283879]
42. Lemen RA. Chrysotile Asbestos as a Cause of Mesothelioma: Application of the Hill Causation Model. International Journal of Occupational and Environmental Health 2004;10(2):233–9. doi:10.1179/oeh.2004.10.2.233. [PubMed: 15281385]
43. Degelman ML, Herman KM. Smoking and multiple sclerosis: A systematic review and meta-analysis using the Bradford Hill criteria for causation. Multiple Sclerosis and Related Disorders17:207–16. doi:10.1016/j.msard.2017.07.020.
44. Weed DL. Epidemiologic evidence and causal inference. Hematol Oncol Clin North Am 2000;14(4):797–807, viii. [PubMed: 10949774]
45. Holman CD, rnold-Reed DE, de KN, McComb C, English DR. A psychometric experiment in causal inference to estimate evidential weights used by epidemiologists. 2001. p. 246–55.
46. Rothman KJ. Causes. Am J Epidemiol 1976;104(6):587–92. [PubMed: 998606]
47. Rothman KJ, Greenland S, Poole C, Lash TL. Causation and Causal Inference. In: Rothman KJ, Greenland S, Lash TL, editors. Modern Epidemiology Philadelphia: Lippincott Williams & Wilkins; 2008. p. 5–31.
48. Open Science C PSYCHOLOGY. Estimating the reproducibility of psychological science. Science. 2015;349(6251):aac4716. doi:10.1126/science.aac4716. [PubMed: 26315443]
49••. Gelman A, Stern H. The Difference Between "Significant" and "Not Significant" is not Itself Statistically Significant. The American Statistician 2006;60(4):328–31. doi:10.1198/000313006X152649.Two results, one statistically significant and the other not, are not necessarily different.
50••. Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol 2016;31(4):337–50. doi:10.1007/s10654-016-0149-3. [PubMed: 27209009] Comprehensive review of all the ways that null hypothesis significance testing is misused and misunderstood.
51. Rothman KJ, Lanes S, Robins J. Casual inference. Epidemiology 1993;4(6):555–6. [PubMed: 8268286]
52. Seliger C, Meier CR, Becker C, Jick SS, Bogdahn U, Hau P et al. Statin use and risk of glioma: population-based case–control analysis. European Journal of Epidemiology. 2016;31(9):947–52. doi:10.1007/s10654-016-0145-7. [PubMed: 27041698]
53. Brown HK, Ray JG, Wilton AS, Lunsky Y, Gomes T, Vigod SN. Association between serotonergic antidepressant use during pregnancy and autism spectrum disorder in children. JAMA. 2017;317(15):1544–52. doi:10.1001/jama.2017.3415. [PubMed: 28418480]
54. Utts J. Replication and Meta-Analysis in Parapsychology. Statistical Science 1991;6(4):363–78.
55. Rothman KJ, Poole C. A strengthening programme for weak associations. Int J Epidemiol 1988;17(4):955–9. [PubMed: 3225112]
56. Cole SR, Stuart EA. Generalizing evidence from randomized clinical trials to target populations: The ACTG 320 trial. Am J Epidemiol 2010;172(1):107–15. doi:10.1093/aje/kwq084. [PubMed: 20547574]
57. Lesko CR, Buchanan AL, Westreich D, Edwards JK, Hudgens MG, Cole SR. Generalizing Study Results: A Potential Outcomes Perspective. Epidemiology 2017;28(4):553–61. doi:10.1097/EDE.0000000000000664. [PubMed: 28346267]

58. Westreich D, Edwards JK, Lesko CR, Stuart E, Cole SR. Transportability of Trial Results Using Inverse Odds of Sampling Weights. Am J Epidemiol 2017;186(8):1010–4. doi:10.1093/aje/kwx164. [PubMed: 28535275]
59. Rothman KJ, Greenland S, Lash TL. Design strategies to improve study accuracy. In: Rothman KJ, Greenland S, Lash TL, editors. Modern Epidemiology Philadelphia: Lippincott Williams & Wilkins; 2008. p. 168–82.
60. Greenland S, Lash TL. Bias Analysis. In: Rothman KJ, Greenland S, Lash TL, editors. Modern Epidemiology Philadelphia: Lippincott Williams & Wilkins; 2008. p. 345–80.
61. Lash TL, Fox MP, MacLehose RF, Maldonado G, McCandless LC, Greenland S. Good practices for quantitative bias analysis. Int J Epidemiol 2014;43(6):1969–85. doi:10.1093/ije/dyu149. [PubMed: 25080530]
62. Hernan MA, Sauer BC, Hernandez-Diaz S, Platt R, Shrier I. Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. J Clin Epidemiol 2016;79:70–5. doi:10.1016/j.jclinepi.2016.04.014. [PubMed: 27237061]
63. Maldonado G Adjusting a relative-risk estimate for study imperfections. J Epidemiol Community Health 2008;62(7):655–63. [PubMed: 18559450]
64. Fox MP, Lash TL. On the Need for Quantitative Bias Analysis in the Peer-Review Process. Am J Epidemiol 2017;185(10):865–8. doi:10.1093/aje/kwx057. [PubMed: 28430833]
65. Hunnicutt JN, Ulbricht CM, Chrysanthopoulou SA, Lapane KL. Probabilistic bias analysis in pharmacoepidemiology and comparative effectiveness research: a systematic review. Pharmacoepidemiol Drug Saf 2016;25(12):1343–53. doi:10.1002/pds.4076. [PubMed: 27593968]
66. Greenland S Invited Commentary: The Need for Cognitive Science in Methodology. Am J Epidemiol. 2017;186(6):639–45. doi:10.1093/aje/kwx259. [PubMed: 28938712]
67. O'Boyle EH, Banks GC, Gonzalez-Mulé E. The Chrysalis Effect: How Ugly Initial Results Metamorphosize Into Beautiful Articles. Journal of Management 2014. doi:10.1177/0149206314527133.
68. Sterling TD. Publication Decisions and Their Possible Effects on Inferences Drawn from Tests of Significance--Or Vice Versa. Journal of the American Statistical Association. 1959;54(285):30–4. doi:10.2307/2282137.
69. Begg CB. A measure to aid in the interpretation of published clinical trials. Stat Med 1985;4(1):1–9. [PubMed: 3992068]
70. Motulsky HJ. Common misconceptions about data analysis and statistics. Pharmacol Res Perspect 2015;3(1):e00093. doi:10.1002/prp2.93. [PubMed: 25692012]
71. Kerr NL. HARKing: Hypothesizing After the Results are Known. Personality and Social Psychology Review 1998;2(3):196–217. doi:doi:10.1207/s15327957pspr0203_4. [PubMed: 15647155]
72. Rothman KJ. Significance questing. Ann Intern Med 1986;105(3):445–7. [PubMed: 3740684]
73. Announcement: Transparency upgrade for Nature journals. Nature 2017;543(7645):288. doi:10.1038/543288b.
74. US National Institutes of Health. Rigor and Reproducibility https://www.nih.gov/research-training/rigor-reproducibility.
75. Goldstein ND. Toward Open-source Epidemiology. Epidemiology 2018;29(2):161–4. doi:10.1097/ede.0000000000000782. [PubMed: 29112521]
76. Khoury MJ. Planning for the Future of Epidemiology in the Era of Big Data and Precision Medicine. Am J Epidemiol. 2015;182(12):977–9. doi:10.1093/aje/kwv228. [PubMed: 26628513]
77. Galea S An argument for a consequentialist epidemiology. Am J Epidemiol 2013;178(8):1185–91. doi:10.1093/aje/kwt172. [PubMed: 24022890]
78. von Elm E, Altman DG, Egger M, Pocock SJ, Gotzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Lancet 2007;370(9596):1453–7. [PubMed: 18064739]
79. Lanes SF. Error and uncertainty in causal inference. In: Rothman KJ, editor. Causal Inference Chestnut Hill, MA: Epidemiology Resources Inc.; 1988.

80. Lash TL. Advancing Research through Replication. Paediatr Perinat Epidemiol 2015;29(1):82–3. doi:10.1111/ppe.12167. [PubMed: 25545128]
81. Munafo M, Davey Smith G. Robust research needs many lines of evidence. Nature 2018;553:399–401.
82. Rothman KJ, Greenland S, Lash TL. Precision and statistics in epidemiologic studies. In: Rothman KJ, Greenland S, Lash TL, editors. Modern Epidemiology Philadelphia: Lippincott Williams & Wilkins; 2008. p. 148–67.
83. Lash TL, Fox MP, Fink AK. Applying Quantitative Bias Analysis to Epidemiologic Data. Statistics for Biology and Health, vol Book, Whole. New York: Springer; 2009.
84. Kieler H, Cnattingius S, Haglund B, Palmgren J, Axelsson O. Sinistrality--a side-effect of prenatal sonography: a comparative study of young men. Epidemiology 2001;12(6):618–23. [PubMed: 11679787]
85. Salvesen KA. Ultrasound in pregnancy and non-right handedness: meta-analysis of randomized trials. Ultrasound Obstet Gynecol 2011;38(3):267–71. doi:10.1002/uog.9055. [PubMed: 21584892]
86. The American College of Obstetricians and Gynecologists. Ultrasound Exams 2017. https://www.acog.org/Patients/FAQs/Ultrasound-Exams.
87. Grady D, Rubin SM, Petitti DB, Fox CS, Black D, Ettinger B et al. Hormone therapy to prevent disease and prolong life in postmenopausal women. Ann Intern Med 1992;117(12):1016–37. [PubMed: 1443971]
88. Stampfer MJ, Colditz GA. Estrogen replacement therapy and coronary heart disease: a quantitative assessment of the epidemiologic evidence. Prev Med 1991;20(1):47–63. [PubMed: 1826173]
89. Petitti D Hormone replacement therapy and coronary heart disease: results of randomized trials. Prog Cardiovasc Dis 2003;46(3):231–8. [PubMed: 14685941]
90. Rossouw JE, Anderson GL, Prentice RL, LaCroix AZ, Kooperberg C, Stefanick ML et al. Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results From the Women's Health Initiative randomized controlled trial. 2002. p. 321–33.
91. Lawlor DA, Davey Smith G, Ebrahim S. Commentary: the hormone replacement-coronary heart disease conundrum: is this the death of observational epidemiology? Int J Epidemiol 2004;33(3):464–7. doi:10.1093/ije/dyh124. [PubMed: 15166201]
92. Hernan MA, Alonso A, Logan R, Grodstein F, Michels KB, Willett WC et al. Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. Epidemiology 2008;19(6):766–79. doi:10.1097/EDE.0b013e3181875e61. [PubMed: 18854702]
93. Gunn LJ, Chapeau-Blondeau F, McDonnell MD, Davis BR, Allison A, Abbott D. Too good to be true: when overwhelming evidence fails to convince. Proc Math Phys Eng Sci 2016;472(2187):20150748. doi:10.1098/rspa.2015.0748. [PubMed: 27118917]