A Future for Observational Epidemiology: Clarity, Credibility, Transparency

Sam Harper

December 17, 2018

Abstract

Observational studies are ambiguous, difficult, and necessary for epidemiology. Presently there are concerns that the evidence produced by most observational studies in epidemiology is not credible and contributes to research waste. I argue that observational epidemiology could be improved by focusing greater attention on: 1) defining questions that make clear whether the inferential goal is descriptive or causal; 2) greater utilization of quantitative bias analysis and alternative research designs that aim to decrease the strength of assumptions needed to estimate causal effects; and 3) promoting, experimenting, and perhaps institutionalizing reproducible research standards as well as replication studies to evaluate the fragility of study findings in epidemiology. Greater clarity, credibility, and transparency in observational epidemiology will help to provide reliable evidence that can serve as a basis for making decisions about clinical or population health interventions.

Observational studies are ambiguous, difficult, and necessary. Randomized treatment assignment has important conceptual strengths, especially for reducing bias, but is resource-intensive and often applied to highly selected populations, making generalizations difficult. There are good reasons (feasibility, resources, ethical issues) why observational studies remain the workhorse of epidemiology and constitute the bulk of epidemiologic output. Recent years have seen critiques (and defenses) of observational studies, but the issues are longstanding. In the late 1950s RA Fisher notoriously and mistakenly thought that observational studies of smoking and lung cancer were flawed.(1) Concerns about reliability related to clinical questions were raised in the early 1980s,(2) and were extended to questions about “lifestyle” exposures later in the decade.(3,4) In the early 21st century concerns about conflicting findings from observational studies and large-scale randomized evaluations,(5,6) led some epidemiologists to ask whether it was “time to call it a day” for observational studies,(7) as well as to broad accusations of “junk science” being leveled at the entire discipline in the popular press.(8)

Despite important new efforts to reconcile differences between randomized and non-randomized studies,(9,10) the issue of the reliability of observational evidence isn’t likely to go away (see the recent discrepancy between randomized and observational results from the Illinois workplace wellness evaluation(11)). Exponential growth in the published literature(12) has also increased our reliance on evidence synthesis for observational studies (systematic reviews, meta-analysis, meta-meta-analysis) but such studies also hide many investigator decisions.(13) Moreover, synthesized evidence may reduce the role of sampling error but will do nothing about systematic errors. Consensus, repetition, and synthesis of poor research does not generate strong evidence.

Large scale replication efforts in observational epidemiology are lacking, but there is increasing concern that the considerable flexibility researchers have when analyzing observational data can lead to unreliable results. We know, for example, that if you give multiple authors the same observational dataset and ask them to answer the same (admittedly vague) research question you will see a broad range of estimates, owing largely to different methodological approaches.(14) We also know that when a single investigative team analyzes the same dataset using various regression specifications, coding and inclusion of covariates, and decision rules, they can generate results on either side of the null.(15,16) Such analyses do not suggest any causal hypotheses; rather, they serve to illustrate the myriad ways in which the same data may provide results consistent with many conflicting hypotheses. Such flexibility also makes it extraordinarily difficult to identify true heterogeneity in effects across populations that may be relevant for understanding causal mechanisms, since presently we can’t separate it from the (typically unseen) choices of the investigators.

These problems are not restricted to observational studies in epidemiology; they are pervasive within many scientific disciplines. Distorted incentives for publishing, promotion, and granting that prioritize novelty and quantity rather than quality exacerbate the problem.(17) One goal of observational epidemiology is to provide reliable evidence about causal relationships between exposures and outcomes, such that the evidence could serve as a basis for making decisions about clinical or population health interventions. Unreliable, poor quality epidemiologic research limits scientific understanding of the determinants of health and disease, reduces our

ability to contribute to clinical and policy decisions, and could potentially undermine the public's faith in epidemiologic research.

Clarity: What was the question?

How can we increase the credibility of published observational epidemiological research? Although randomized trials have many potential limitations, one strength is that they often focus clearly on answering a specific causal question for a well-defined treatment. Regardless of whether we are conducting a randomized trial or an observational study, if the purpose is to provide a quantitative estimate of what would happen if we could intervene to change A from a to a^* , the goal is estimating a causal effect.⁽¹⁸⁾ The extent to which the actual study provides good evidence rests on the credibility of the assumptions needed to interpret the estimate as causal (exchangeability, positivity, consistency, correct model specification, absence of measurement error and selection bias). Since these assumptions are unlikely to be met, many observational studies only claim to study associations, but this leads to ambiguity regarding the question, as well as to inaccurate or misleading inference from regression adjustment⁽¹⁹⁾ when there is no hope for causal identification.

Of course, there should be ample room within epidemiology for studies that do not aim to estimate any causal effect, but rather to explore, hypothesize, characterize, interrogate, and describe. Although I disagree with recent arguments that the potential outcomes framework necessarily constrains the kinds of causal questions it is possible to ask,^(20,21) I do think that the emphasis on (and enthusiasm for) formal causal inference may have crowded out many interesting questions that are not explicitly causal. In fact, I would argue for greater role for high quality descriptive studies that do not aim to provide causal estimates, and a lesser role for purportedly “causal” studies that are unlikely to provide robust evidence about causal effects.

For example, consider the recent and deeply disturbing rise in opioid overdose deaths in the United States—presently killing more than 70,000 Americans annually. Not only were descriptive studies crucial in understanding the distribution and determinants of the opioid epidemic,^(22,23) continued surveillance and interest has allowed epidemiologists to identify transitional phases of the epidemic by geography, substance (prescription opioids, heroin, fentanyl) and race-ethnic group,⁽²⁴⁾ as well as contributing to potential interventions. Many of the early reports that found increases in opioid overdose deaths came from epidemiologists conducting basic mortality surveillance in state and local public health agencies.^(25–27) Such analyses are quite explicitly descriptive: Are deaths increasing or decreasing? Who is dying, what are they dying from, and where are they dying? And yet few, if any, of these reports were published in leading epidemiology journals. This is not necessarily an indication of bias (perhaps they were never submitted), but the fact that a massive increase in disability and death due to opioids, with profound implications for society and the economy, has received comparatively little attention in our top journals feels like a lost opportunity.

Similarly, the impact of the disastrous cost-cutting decision to change the city of Flint, Michigan's water supply from Lake Huron to the Flint River was known not only as a

result of citizens voicing their concerns,(28) but from basic descriptive epidemiology.(29) The full impacts of the intervention are still unknown, but routine surveillance of water quality and, subsequently, of blood lead levels in Flint’s children(30) were enormously important in establishing a credible basis for public health action.

Furthermore, in addition to routine surveillance, descriptive observational studies help to reconcile diverging lines of evidence, reflect on the implications of existing theories, and triangulate evidence from multiple sources.(31,32) Combining descriptive analysis with theory and historical context can provide extraordinarily valuable insights into social phenomena with important implications for human health and well being such as climate change,(33) economic crises,(34,35) undernutrition,(36) maternal death,(37) psychosocial stress,(38) or population migration.(39,40) So on one hand I would like to see greater enthusiasm in observational epidemiology for descriptive, historical, and narrative modes of investigation, but on the other I also hope for more reliable and useful causal investigations.

Credibility: Weaker assumptions, by design

We need good observational studies designed to answer causal questions. Having clarified the study purpose (estimating a causal effect) and making clear the assumptions needed for inference, the key question then becomes the credibility of those assumptions. Most authors still rely on intuitions for their assessments of the likely impact of whether these assumptions are reasonable, but quantitative intuitions are notoriously lousy.(41,42) Rather than making strong assumptions about the absence of unmeasured confounding or measurement error, researchers can relax these assumptions by allowing for specific degrees of bias. Methods(43) and guidance(44) for making quantitative bias analysis a routine part of epidemiologic studies have been around a long time, yet few papers conduct any formal quantitative bias analysis,(45) and even fewer take multiple biases(46) into account simultaneously. The upshot is overconfidence in our own findings and likely underestimation of study uncertainty. This cannot be helpful for decision making.

What else can we do? In addition to making quantitative bias analysis standard, another way to potentially improve the credibility of observational epidemiologic studies is to consider alternative designs or analytic choices that aim to account, by design, for some sources of unmeasured confounding or other potential study biases. Quasi-experimental designs(47,48) aim to create exchangeable treatment groups through “acts of chance” or other arbitrary rules or events not under investigators’ control. In the best case scenario this leads to randomized or “as-if” randomized treatment allocation (e.g., a lottery), straightforward analysis, and more robust inference. These designs are not new, but also not extensively taught in most epidemiology programs. However, recent explications in the epidemiologic literature of interrupted time series,(49) difference-in-differences,(50) synthetic controls,(51) instrumental variables,(52) and regression discontinuity(53) should help.

These designs are valuable, and we should encourage greater use of them as ways help to reduce some sources of bias. However, enthusiasm for quasi-experimental designs should be tempered by keeping in mind that they are still observational studies. They require

important assumptions for valid inference, and there is some reason to worry that the label “quasi-experimental” may make researchers less worried about all of the ways that these studies can still produce poor quality evidence. Invalid instruments, poor control groups, and time-varying confounding are serious concerns, in addition to the usual threats of selection bias and measurement error. Two recent “quasi-experimental” analyses of the impact of a national mammography screening program in Switzerland that used identical data came to opposing conclusions about its impact on socioeconomic inequalities in mammography precisely because of important differences in how each set of authors attempted to create exchangeable treatment groups.(54,55)

If it seems like these strategies could make each study considerably more time-consuming and painstaking, so much the better. We should worry deeply about the credibility of our assumptions, and vigorously seek alternative explanations for results. Richard Feynman famously said that we have a responsibility as scientists to each other (and, I might add, to those who fund our work) to bend over backward to show how we might be wrong.(56) This is good advice, and epidemiologists should aspire follow it, though in practice many discussions of study limitations feel more like exercises in hand waving to convince readers of why we are right. But we are all human, and if cognitive science has taught us anything it is that we are unlikely to be able to overcome our biases when evaluating our own work.(57) Confirmation bias is one hell of a drug.

Transparency: Necessary, not sufficient

We can go some way toward more credible observational epidemiology by utilizing better study designs and making quantitative bias analysis standard practice. But we can also go further, and I think greater transparency has some role to play here, not only through better reporting, which has received considerable attention,(58) but woven throughout the entire life cycle of research.(59) For causal questions, this starts with the formulation of a clear research question, ideally specified to represent a target trial,(10) but also with systematically reviewing the extent of prior work to see whether new evidence is needed.(60) From there we need transparency in how the research design was developed, the assumptions that must be credibly believed in order to estimate a causal effect, and how the data analysis was carried out. We also need transparency in reporting conflicts (financial or non-financial) that may bias the authors’ interpretation of results. Finally, I think we need transparency in the preparation and deposition of materials that allow people other than the researchers to reproduce and, possibly, improve upon their results.

Here I think epidemiology can learn from other disciplines that have also been feeling the heat of the so-called “replication crisis”. Large-scale efforts to replicate published results in psychology,(61) genetics,(62) cancer biology,(63) neuroscience,(64) and economics(65) have generally produced results that do not inspire confidence in published work. Quite frankly, I think epidemiology needs to embrace more vigorous challenges of the evidence it produces. To that extent, I hope that the editors and boards of our leading journals may also experiment with strategies for promoting replication studies.

It is important to be clear about the distinction between so-called “pure” replications, independent replications, and reproducible research. “Pure” replications involve independent researchers taking the *same* materials and code developed by the original investigators and trying to reproduce the tables, figures, and estimates of the original report. This is akin to quality control, and it would seem the bare minimum for most published work. Independent replications of the same causal question in another setting are more challenging, since we should not expect exact replications of study findings when there may be good reasons why the effects may demonstrate heterogeneity.(66,67) That being said, if the findings are wildly different in magnitude or in opposing directions, this should probably lead to caution with respect to designing or implementing interventions. Reproducible research more generally means making the materials needed to replicate the work (data, protocols, code) available. Of course, the fact that studies are reproducible does not make them valid. Poorly designed studies are hopeless for producing good evidence, no matter how transparently they are planned, pre-registered, analyzed, and reported. Although much can be done to apply sensitivity analysis to published work without access to the original study materials,(68) having access greatly expands the ability to explore methodological choices that could be consequential for inference (assumptions, data transformations, alternative estimation strategies, effect heterogeneity).

Recent examples in epidemiology of replication studies are instructive. Replication of a high-profile deworming intervention(69) and response by the original authors(70) generated a nuanced discussion of methods and choices for analyzing such interventions. The original authors also provided important discussion about the merits of the replication exercise, ultimately leading to improved estimates. On the other hand, another recent paper(71) replicating a study estimating the impact of anti-gay discrimination on mortality(72) led to considerably different conclusions after accounting for missing data and ultimately to a correction being published. The point here is not to take sides regarding the merits of the particular papers; rather, it is to highlight the enhanced knowledge that comes from a replication exercise. In both cases coding errors were found that surely any author would appreciate, even if it did not change the interpretation. More importantly, the questions were probed by different investigators with different background knowledge, biases, and perhaps alternative causal structures in mind. Interrogating observational studies in such a fashion should help us to be more honest about the fragility of our findings. Re-analyses of published randomized trials with different assumptions that led to different conclusions(73) should make clear that it is not only observational studies that could benefit from greater reproducibility. Clearly there may be limits to such an exercise, since there are far too many papers submitted and published to subject all papers to such scrutiny, but without making the materials behind our studies available such dialogue will not happen.

So I wonder whether epidemiology could benefit from a more structured approach. A recent example from development economics highlights both the promise and some of the perils of replication work.(74) The *Journal of Development Studies* used a structured grant program to fund researchers to replicate 8 high profile publications from a candidate list, enlisted third parties to assist in facilitating contact between original researchers and replication authors, and provided space for the original authors to comment on the replication findings. Errors were discovered and there was some contentious debate, but most findings were generally

robust and several new findings emerged. After all, some degree of non-replication is useful for elaborating alternative theories, exploring, and moving knowledge forward. There are many complications in making such exercises routine, but I think the benefits likely outweigh the costs.

Journals in other disciplines are also actively experimenting with alternative structures for peer review and publishing. More than 120 journals are now supporting “Registered Reports” where initial peer review and publication decisions are made on the basis of specifying the scientific question, research design, and methods.⁽⁷⁵⁾ Publication is effectively guaranteed after successful peer-review (unless there are valid reasons why the research could not be completed, suspicion of fraud, etc.) Of course this won’t help with the problem of null results that are never written up,⁽⁷⁶⁾ but given the potential for results to cloud judgments about the scientific merits of a paper, it’s worth considering. Similar calls for results-free submission⁽⁷⁷⁾ and quantitative bias analysis in peer review⁽⁷⁸⁾ for epidemiologic studies have been suggested but, to my knowledge, never tried in any of our leading journals. Leaving aside the unhelpful dichotomization of results, early research suggests that the proportion of “significant” results among registered reports is considerably lower (~30%)⁽⁷⁹⁾ than the upwards of 90% of published studies that report significant results.⁽⁸⁰⁾ Because null hypothesis significance testing is presently such a strong filter for publication, this isn’t necessarily unexpected,⁽⁸¹⁾ and consistent with clinical trial registration having led to more null reports.⁽⁸²⁾

And yet there are, of course, challenges with conducting replication studies. There are important issues concerning privacy and sharing of sensitive health data. Existing norms and incentives in scientific publishing also continue to prioritize the “new and novel” (e.g., replication studies that “overturn” original studies).⁽⁸³⁾ Care should be taken to make sure that replication studies are themselves conducted transparently, as well as being cognizant of the potential impact of the exercise on young investigators. Guidance on how to make replication studies more mainstream is emerging.⁽⁶⁷⁾

Many of the problems that currently plague observational epidemiology exist because epidemiologic research is done by human beings embedded in social contexts. This has led to misaligned incentives between researchers, funders, and publishers, inadequate peer-review, novelty-chasing, exaggerated claims, and conflicts of interest, all of which contribute to problems with the credibility of published research.⁽⁸⁴⁾ Better training of epidemiology graduate students in how to think critically and formulate well-defined questions, evaluate evidence and adjudicate between competing claims, and maintain humility regarding their own research findings would be valuable. For what is the alternative? Some may argue that epidemiology mostly gets it right over the long haul, but how will we know? What other masses of sliced, prodded, hacked, and unreliably “significant” results might be lurking beneath the surface of the sea of observational epidemiologic papers published every day? Will we bother to wade in and find out?

Acknowledgements

Author affiliations: Department of Epidemiology, Biostatistics & Occupational Health, McGill University and Institute for Health and Social Policy, McGill University, Montreal, Quebec (Sam Harper)

Funding: None

I thank Jay Kaufman, Nick King, and Ari Nandi for constructive comments on an earlier draft of this paper. All errors are mine.

References

1. Fisher RA. Cancer and smoking. *Nature*. 1958;182(4635):596.
2. Byar DP. Why data bases should not replace randomized clinical trials. *Biometrics*. 1980;36(2):337–42.
3. Feinstein AR. Scientific standards in epidemiologic studies of the menace of daily life. *Science*. 1988;242(4883):1257–63.
4. Savitz DA, Greenland S, Stolley PD, et al. Scientific standards of criticism: A reaction to "scientific standards in epidemiologic studies of the menace of daily life," by a.r. Feinstein. *Epidemiology*. 1990;1(1):78–83.
5. Kunz R, Oxman AD. The unpredictability paradox: Review of empirical comparisons of randomised and non-randomised clinical trials. *BMJ*. 1998;317(7167):1185–90.
6. Lawlor DA, Davey Smith G, Kundu D, et al. Those confounded vitamins: What can we learn from the differences between observational versus randomised trial evidence? *Lancet*. 2004;363(9422):1724–7.
7. Davey Smith G, Ebrahim S. Epidemiology—is it time to call it a day? *Int J Epidemiol*. 2001;30(1):1–11.
8. Taubes G. Do we really know what makes us healthy? *New York Times*. 2007;16:2007.
9. Hernán MA, Alonso A, Logan R, et al. Observational studies analyzed like randomized experiments: An application to postmenopausal hormone therapy and coronary heart disease. *Epidemiology*. 2008;19(6):766–79.
10. Hernán MA, Sauer BC, Hernández-Díaz S, et al. Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. *J Clin Epidemiol*. 2016;79:70–75.
11. Jones D, Molitor D, Reif J. What do workplace wellness programs do? Evidence from the illinois workplace wellness study. National Bureau of Economic Research; 2018.(<http://www.nber.org/papers/w24229>)

12. Bowen A, Casadevall A. Increasing disparities between resource inputs and outcomes, as measured by certain health deliverables, in biomedical research. *Proc Natl Acad Sci U S A*. 2015;112(36):11335–40.
13. Ioannidis JPA. The mass production of redundant, misleading, and conflicted systematic reviews and meta-analyses. *Milbank Q*. 2016;94(3):485–514.
14. Silberzahn R, Uhlmann EL, Martin DP, et al. Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*. 2018;1(3):337–356.
15. Madigan D, Ryan PB, Schuemie M. Does design matter? Systematic evaluation of the impact of analytical choices on effect estimates in observational studies. *Ther Adv Drug Saf*. 2013;4(2):53–62.
16. Patel CJ, Burford B, Ioannidis JPA. Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *J Clin Epidemiol*. 2015;68(9):1046–58.
17. Ioannidis JPA, Greenland S, Hlatky MA, et al. Increasing value and reducing waste in research design, conduct, and analysis. *Lancet*. 2014;383(9912):166–75.
18. Hernán MA. The c-word: Scientific euphemisms do not improve causal inference from observational data. *Am J Public Health*. 2018;108(5):616–619.
19. Kaufman JS. Statistics, adjusted statistics, and maladjusted statistics. *American journal of law & medicine*. 2017;43(2-3):193–208.
20. Krieger N, Davey Smith G. The tale wagged by the dog: Broadening the scope of causal inference and explanation for epidemiology. *Int J Epidemiol*. 2016;45(6):1787–1808.
21. Vandenbroucke JP, Broadbent A, Pearce N. Causality and causal inference in epidemiology: The need for a pluralistic approach. *Int J Epidemiol*. 2016;45(6):1776–1786.
22. Fingerhut LA, Cox CS. Poisoning mortality, 1985-1995. *Public Health Rep*. 113(3):218–33.
23. King NB, Fraser V, Boikos C, et al. Determinants of increased opioid-related mortality in the United States and Canada, 1990-2013: A systematic review. *Am J Public Health*. 2014;104(8):e32–42.
24. Alexander MJ, Kiang MV, Barbieri M. Trends in black and white opioid mortality in the united states, 1979-2015. *Epidemiology*. 2018;29(5):707–715.
25. Centers for Disease Control and Prevention (CDC). Unintentional and undetermined poisoning deaths—11 states, 1990-2001. *MMWR Morb Mortal Wkly Rep*. 2004;53(11):233–8.
26. Franklin GM, Mai J, Wickizer T, et al. Opioid dosing trends and mortality in Washington State workers’ compensation, 1996-2002. *Am J Ind Med*. 2005;48(2):91–9.
27. Shah N, Lathrop SL, Landen MG. Unintentional methadone-related overdose death in New Mexico (USA) and implications for surveillance, 1998-2002. *Addiction*. 2005;100(2):176–88.

28. Felton R. Flint residents raise concerns over discolored water. *Detroit Metro Times*. 2014;(<https://www.metrotimes.com/detroit/flint-residents-raise-concerns-over-discolored-water/Content?oid=2231724>)
29. Hanna-Attisha M, LaChance J, Sadler RC, et al. Elevated blood lead levels in children associated with the flint drinking water crisis: A spatial analysis of risk and public health response. *Am J Public Health*. 2016;106(2):283–90.
30. Kennedy C, Yard E, Dignam T, et al. Blood lead levels among children aged <6 years - Flint, Michigan, 2013-2016. *MMWR Morb Mortal Wkly Rep*. 2016;65(25):650–4.
31. Lieberman S. Making it count: The improvement of social research and theory. Univ of California Press; 1987.
32. Lawlor DA, Tilling K, Davey Smith G. Triangulation in aetiological epidemiology. *Int J Epidemiol*. 2016;45(6):1866–1886.
33. Klinenberg E. Denaturalizing disaster: A social autopsy of the 1995 chicago heat wave. *Theory and Society*. 1999;28(2):239–295.
34. Leon DA, Chenet L, Shkolnikov VM, et al. Huge variation in russian mortality rates 1984-94: Artefact, alcohol, or what? *Lancet*. 1997;350(9075):383–8.
35. Venero Fernández SJ, Medina RS, Britton J, et al. The association between living through a prolonged economic depression and the male:Female birth ratio—a longitudinal study from Cuba, 1960-2008. *Am J Epidemiol*. 2011;174(12):1327–31.
36. Jayachandran S, Pande R. Why are indian children so short? The role of birth order and son preference. *American Economic Review*. 2017;107(9):2600–2629.
37. Loudon I. Maternal mortality in the past and its relevance to developing countries today. *Am J Clin Nutr*. 2000;72(1 Suppl):241S–246S.
38. Susser M, Stein Z. Commentary: Civilization and peptic ulcer 40 years on. *Int J Epidemiol*. 2002;31(1):18–21.
39. Fang J, Madhavan S, Alderman MH. The association between birthplace and mortality from cardiovascular causes among black and white residents of new york city. *N Engl J Med*. 1996;335(21):1545–51.
40. Razum O, Twardella D. Time travel with oliver twist—towards an explanation foa a paradoxically low mortality among recent immigrants. *Trop Med Int Health*. 2002;7(1):4–10.
41. Tversky A, Kahneman D. Judgment under uncertainty: Heuristics and biases. *Science*. 1974;185(4157):1124–31.
42. Lash TL. Heuristic thinking and inference from observational epidemiology. *Epidemiology*. 2007;18(1):67–72.
43. Greenland S. Basic methods for sensitivity analysis of biases. *Int J Epidemiol*. 1996;25(6):1107–16.

44. Lash TL, Fox MP, MacLehose RF, et al. Good practices for quantitative bias analysis. *Int J Epidemiol.* 2014;43(6):1969–85.
45. Brakenhoff TB, Mitroiu M, Keogh RH, et al. Measurement error is often neglected in medical literature: A systematic review. *J Clin Epidemiol.* 2018;98:89–97.
46. Greenland S. Multiple-bias modelling for analysis of observational data. *Journal of the Royal Statistical Society: Series A (Statistics in Society).* 2005;168(2):267–306.
47. Shadish WR, Cook TD, Campbell DT. Experimental and quasi-experimental designs for generalized causal inference. Boston: Houghton Mifflin; 2001.
48. Craig P, Katikireddi SV, Leyland A, et al. Natural experiments: An overview of methods, approaches, and contributions to public health intervention research. *Annu Rev Public Health.* 2017;38:39–56.
49. Bernal JL, Cummins S, Gasparrini A. Interrupted time series regression for the evaluation of public health interventions: A tutorial. *Int J Epidemiol.* 2017;46(1):348–355.
50. Ryan AM, Burgess JF Jr, Dimick JB. Why we should not be indifferent to specification choices for difference-in-differences. *Health Serv Res.* 2015;50(4):1211–35.
51. Rehkopf DH, Basu S. A new tool for case studies in epidemiology-the synthetic control method. *Epidemiology.* 2018;29(4):503–505.
52. Hernán MA, Robins JM. Instruments for causal inference: An epidemiologist’s dream? *Epidemiology.* 2006;17(4):360–72.
53. Bor J, Moscoe E, Mutevedzi P, et al. Regression discontinuity designs in epidemiology: Causal inference without randomized trials. *Epidemiology.* 2014;25(5):729–37.
54. Pletscher M. The effects of organized screening programs on the demand for mammography in switzerland. *Eur J Health Econ.* 2017;18(5):649–665.
55. Cullati S, von Arx M, Courvoisier DS, et al. Organised population-based programmes and change in socioeconomic inequalities in mammography screening: A 1992-2012 nationwide quasi-experimental study. *Prev Med.* 2018;116:19–26.
56. Feynman RP. Cargo cult science. *Engineering and Science.* 1974;37(7):10–13.
57. Kahneman D. Thinking, fast and slow. New York: Farrar, Straus; Giroux; 2011.
58. Equator Network. Reporting guidelines for main study types. 2018;(https://www.equator-network.org/)
59. Munafò MR, Nosek BA, Bishop DV, et al. A manifesto for reproducible science. *Nature Human Behaviour.* 2017;1(1):0021.
60. Chalmers I, Bracken MB, Djulbegovic B, et al. How to increase value and reduce waste when research priorities are set. *Lancet.* 2014;383(9912):156–65.
61. Open Science Collaboration. Estimating the reproducibility of psychological science. *Science* [electronic article]. 2015;349(6251):aac4716–aac4716. (<http://dx.doi.org/10.1126/>

science.aac4716)

62. Ioannidis JP, Allison DB, Ball CA, et al. Repeatability of published microarray gene expression analyses. *Nature genetics*. 2009;41(2):149.
63. Errington TM, Iorns E, Gunn W, et al. Science forum: An open investigation of the reproducibility of cancer biology research. *Elife*. 2014;3:e04333.
64. Button KS, Ioannidis JPA, Mokrysz C, et al. Power failure: Why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci*. 2013;14(5):365–76.
65. Camerer CF, Dreber A, Forsell E, et al. Evaluating replicability of laboratory experiments in economics. *Science*. 2016;351(6280):1433–1436.
66. Lash TL, Collin LJ, Van Dyke ME. The replication crisis in epidemiology: Snowball, snow job, or winter solstice? *Current Epidemiology Reports* [electronic article]. 2018;5(2):175–183. (<https://doi.org/10.1007/s40471-018-0148-x>)
67. Zwaan RA, Etz A, Lucas RE, et al. Making replication mainstream. *Behav Brain Sci*. 2017;1–50.
68. Weuve J, Sagiv SK, Fox MP. Quantitative bias analysis for collaborative science. *Epidemiology*. 2018;29(5):627–630.
69. Aiken AM, Davey C, Hargreaves JR, et al. Re-analysis of health and educational impacts of a school-based deworming programme in western kenya: A pure replication. *Int J Epidemiol*. 2015;44(5):1572–80.
70. Hicks JH, Kremer M, Miguel E. Commentary: Deworming externalities and schooling impacts in Kenya: A comment on Aiken et al. (2015) and Davey et al. (2015). *Int J Epidemiol*. 2015;44(5):1593–6.
71. Regnerus M. Is structural stigma’s effect on the mortality of sexual minorities robust? A failure to replicate the results of a published study. *Soc Sci Med*. 2017;188:157–165.
72. Hatzenbuehler ML, Bellatorre A, Lee Y, et al. Structural stigma and all-cause mortality in sexual minority populations. *Soc Sci Med*. 2014;103:33–41.
73. Ebrahim S, Sohani ZN, Montoya L, et al. Reanalyses of randomized clinical trial data. *JAMA*. 2014;312(10):1024–32.
74. Brown AN, Wood BD. Replication studies of development impact evaluations [available online ahead of print september 13, 2018]. *The Journal of Development Studies*.
75. Center for Open Science. Registered reports: Peer review before results are known to align scientific values and practices. 2018;(<https://cos.io/rr/>)
76. Franco A, Malhotra N, Simonovits G. Social science. Publication bias in the social sciences: Unlocking the file drawer. *Science*. 2014;345(6203):1502–5.
77. Lawlor DA. Quality in epidemiological research: Should we be submitting papers before we have the results and submitting more hypothesis-generating research? *Int J Epidemiol*. 2007;36(5):940–3.

78. Fox MP, Lash TL. On the need for quantitative bias analysis in the peer-review process. *Am J Epidemiol.* 2017;185(10):865–868.
79. Allen CP, Mehler DM. Open science challenges, benefits and tips in early career and beyond—manuscript. *PsyArXiv.* 2018;
80. Chavalarias D, Wallach JD, Li AHT, et al. Evolution of reporting p values in the biomedical literature, 1990-2015. *JAMA.* 2016;315(11):1141–8.
81. Lash TL. The harm done to reproducibility by the culture of null hypothesis significance testing. *Am J Epidemiol.* 2017;186(6):627–635.
82. Kaplan RM, Irvin VL. Likelihood of null effects of large nhlbi clinical trials has increased over time. *PLoS One.* 2015;10(8):e0132382.
83. Gertler P, Galiani S, Romero M. How to make replication the norm. *Nature.* 2018;554(7693):417–419.
84. Ioannidis JPA. How to make more published research true. *PLoS Med.* 2014;11(10):e1001747.