Current Environmental Health Reports (2019) 6:8–15
https://doi.org/10.1007/s40572-019-0225-4

METHODS IN ENVIRONMENTAL EPIDEMIOLOGY (AZ POLLACK AND NJ PERKINS, SECTION EDITORS)

Understanding and Mitigating the Replication Crisis, for Environmental Epidemiologists

Scott M. Bartell [1,2,3]

Published online: 22 January 2019

Abstract

Purpose of Review In recent years, investigators in a variety of fields have reported that most published findings can not be replicated. This review evaluates the factors contributing to lack of reproducibility, implications for environmental epidemiology, and strategies for mitigation.

Recent Findings Although publication bias and other types of selective reporting may contribute substantially to irreproducible results, underpowered analyses and low prevalence of true associations likely explain most failures to replicate novel scientific results. Epidemiologists can counter these risks by ensuring that analyses are well-powered or precise, focusing on scientifically justified hypotheses, strictly controlling type I error rates, emphasizing estimation over statistical significance, avoiding practices that introduce bias, or employing bias analysis and triangulation. Avoidance of p values has no effect on reproducibility if confidence intervals excluding the null are emphasized in a similar manner.

Summary Increased attention to exposure mixtures and susceptible subpopulations, and wider use of omics technologies, will likely decrease the proportion of investigated associations that are true associations, requiring greater caution in study design, analysis, and interpretation. Though well intentioned, these recent trends in environmental epidemiology will likely decrease reproducibility if no effective actions are taken to mitigate the risk of spurious findings.

Keywords Reliability · Reproducibility · False positive · Type I error · Family-wise error rate · False discovery rate · p value · Hypothesis testing

Introduction

In recent years, researchers in a variety of fields have investigated whether previously reported scientific findings can be reproduced by repeating the original experiments [1–3], or by comparing meta-analyses to the original study findings [4•]. Findings have been dismal, routinely showing that novel findings have, on average, larger effect sizes than confirmatory studies, and that more than half of initially reported associations are contradicted by later studies. Lack of reproducibility appears to be widespread, affecting all scientific fields of study [5] including risk factor epidemiology [4•].

Although this state of affairs has come as a surprise to some, it is understood by statisticians as a predictable consequence of traditional null hypothesis testing for statistical significance [6••]. This review explains the underlying causes of the crisis, describes some of the implications for environmental epidemiology, and summarizes proposed solutions.

Understanding the Causes of the Replication Crisis

The causes of the replication crisis can easily be understood through the familiar framework of diagnostic testing, and

This article is part of the Topical Collection on *Methods in Environmental Epidemiology*

✉ Scott M. Bartell
sbartell@uci.edu

[1] Program in Public Health, Susan and Henry Samueli College of Health Sciences, University of California Irvine, 2032 Anteater Instruction & Research Building, Irvine, CA 92697-3957, USA

[2] Department of Statistics, Donald Bren School of Information and Computer Sciences, University of California Irvine, Irvine, CA, USA

[3] Department of Epidemiology, School of Medicine, Susan and Henry Samueli College of Health Sciences, University of California Irvine, Irvine, CA, USA

particularly the concept of positive predictive value (*PPV*) [7••, 8•]. *PPV* is a key component of diagnostic test validity, and a very useful metric for imperfect tests because it conditions on a positive diagnostic test result, directly answering the question "how likely is it that I have this disease, given that my test result is positive?" Consider any epidemiological association analysis as a type of diagnostic test, in this case using a cohort study, case-control study, or other sample to "test" for a true disease association in a reference population or study base from which the sample was drawn. In this context, one can consider statistical significance (i.e., a low p value or a confidence interval excluding the null hypothesis) as a "positive" test result, lack of statistical significance as a "negative" test result, and the prevalence as the proportion of investigated disease associations that are true disease associations (causal or non-causal associations that exist in the reference population/study base, but not necessary in the sample being used for epidemiological study). Thus, in this context, the *PPV* answers the question, "how likely is it that there is an association in the reference population, given that there is a statistically significant association in my study sample?"

In this context of the diagnostic reliability of hypothesis testing using p values or confidence intervals, the statistical power of an association analysis is its sensitivity, or the probability of a statistically significant finding given that there is a true disease association in the reference population, and the confidence level, $1-\alpha$, is its specificity, or the probability of a non-significant finding given that there is no disease association in the reference population. By convention, α, the type I error rate, is typically set to 0.05 both as a threshold for declaring statistical significance and for calculation of confidence intervals (i.e., $1-\alpha = 95\%$). The *PPV*, or the probability that there is a true disease association given that a statistically significant result has occurred, has a well-known mathematical relationship to the sensitivity (*sens*), specificity (*spec*), and prevalence (*prev*):

$$PPV = \frac{sens \cdot prev}{sens \cdot prev + (1-spec) \cdot (1-prev)} \quad (1)$$

which, in the context of an epidemiological association analysis, can be written as

$$PPV = \frac{power \cdot prev}{power \cdot prev + \alpha \cdot (1-prev)} \quad (2)$$

From this perspective and a few simple calculations, it is evident that either low statistical power or low prevalence of true disease associations will result in poor diagnostic reliability of a statistically significant finding for predicting a true disease association in the reference population. For example, consider an analysis with 80% power, a level that is often considered adequate for investigating primary aims [9], and that uses $\alpha = 0.05$ as threshold for statistical significance or for comparing confidence intervals to null hypothesis values. If half of such analyses actually investigate true risk factor associations (i.e., $prev = 0.50$), then $PPV = 94\%$ and all is well, because the vast majority of statistically significant findings from such studies will actually reflect true associations in the reference populations for those studies. However, if only 10% of these analyses are for true associations ($prev = 0.10$), then $PPV = 64\%$, indicating that only a little more than half of the study results can be trusted. *PPV* declines rapidly from there, to values of 14% for $prev = 0.01$ and 2% for $prev = 0.001$. Notably, these calculations depend on the actual power of the tests, not the nominal power computed using simplifying assumptions such as no adjustments for measurement error or confounding. Those considerations may reduce the actual power below nominal values, which would decrease *PPV*.

What is an appropriate value for *prev*, the prevalence of true disease associations among those tested in environmental epidemiology? This question is difficult to answer, but relevant published estimates include 0.5 for mundane research on well-understood topics [8•], 0.1 for risk factor epidemiology as a whole [10•], 0.096 for new drug development [11], 0.09 for exploratory epidemiology studies [6••], 0.01 for innovative high-risk research [8•], and 0.001 for discovery-oriented research with "massive testing," such as epigenetics [6••]. These are mostly educated guesses at realistic values of *prev*, but suggest that the example *PPV* calculations presented here are highly relevant to the reliability of epidemiological research using well-powered studies of a priori hypotheses.

It is important to recognize that the above calculations of PPV are not realistic for many secondary analyses, or exploratory analyses. Such analyses are often conducted without any formal power calculations, or with lower than 80% power and the understanding that these more opportunistic analyses have higher risk of failing to detect associations, but may produce more novel findings. Considering the number of confidence intervals or p values contained in a single manuscript, and the number of power calculations in a typical grant application, it seems likely that the majority of analyses in the published epidemiological literature are secondary analyses, with typically less than 80% power. One early study estimated the average power of hypothesis tests to be 50% for the medical literature as a whole [10•], supported by surveys of medical studies published in the 1990s. Secondary analyses may play an even larger role in the literature now, considering the proliferation of statistical methods, software, and computational ability during the last two decades. A more recent analysis of neuroscience topics assessed in meta-analyses reported that the median power of those analyses on those topics was 23% [12], and a similar study of meta-analyses for neurological, psychiatric, and somatic diseases reported a median power of approximately 20% [13••]. Estimates of average power in the published scientific literature routinely fall below 50%, across fields and across decades [14•].

Taking 50% as an optimistic estimate of the average power of all published epidemiological analyses, what are the implications for the reliability of statistically significant results among those findings? Using $\alpha = 0.05$ as the threshold for statistical significance, analyses with 50% power have a *PPV* of 53, 9, and 1% for hypothesis with *prev* values of 0.1, 0.01, and 0.001, respectively. If the average power of published analyses is 23%, similar to the neurosciences, then the *PPV* is 34, 4, and 0.5% for hypothesis with *prev* values of 0.1, 0.01, and 0.001, respectively. Average values of *power* and *PPV* for environmental epidemiology are unlikely to be any higher than these values, and could be even lower.

This framework for evaluating the reliability of statistical significance testing does not account for bias resulting from uncontrolled confounding, selection bias, information bias, p-hacking (fitting various models until a statistically significant result is obtained), variable selection routines, selective reporting, publication bias, or other potential sources of bias. Such biases are often in a direction away from the null hypothesis (e.g., there is a true association in the reference population, but it is not causal), which further decreases *PPV*. In a widely cited paper, discouragingly titled "Why Most Published Research Findings Are False," Ioannidis expanded the framework of Eq. 2 to account for the combined effect of those impacts using a single bias parameter, showing that moderate bias can push *PPV* below 50% even when *power* and *prev* are large [6••].

Implications for Environmental Epidemiology

Although one of the lessons of the replication crisis is that every field of study is at risk, environmental epidemiology has some particular characteristics and trends worth noting. Epidemiologists are fortunate that we already have an established culture of avoiding overemphasis on statistical significance, and of "replication before belief," treating isolated results with skepticism until they are reported in multiple studies. Our journals are generally interested in publishing replication studies with or without statistical significance, limiting publication bias to some extent.

However, some features of our study designs may put our field at higher risk of irreproducible results. In particular, chemical exposures are often difficult to characterize at the individual level, limiting sample sizes that can be obtained with a given budget—particularly using exposure biomarkers or personal exposure monitoring for air pollutants. Because statistical power decreases at smaller sample sizes, obtaining higher quality exposure metrics can hamper reproducibility even as it decreases bias by reducing measurement error. We also have a history of avoiding formal adjustment for multiple comparisons, based on the argument that they overemphasize the joint null hypothesis and discourage scientists from exploring potentially important associations [15]. However, some of our papers contain hundreds or thousands of models [16], which virtually guarantees multiple statistically significant findings when each test is performed at $\alpha = 0.05$, regardless of whether there are any true associations.

Recent years in environmental epidemiology have seen greater emphasis on the study of exposure mixtures, epigenetic mechanisms, and susceptible subpopulations. Although these developments are scientifically important and well intentioned, they should be expected to worsen reproducibility for the reasons described below.

Effect estimation and hypothesis testing for each individual component of a mixture of correlated exposures is difficult enough, due to the need for larger sample sizes to achieve the desired precision/power than would be required for a single exposure variable, but investigation of synergy, antagonism, or other types of effect modification can require a dramatic increase in sample size if more than a few exposure variables are included. Because the toxicological effects of mixtures are often poorly understood, these analyses are typically exploratory, suggesting lower values of *prev* than traditional, hypothesis-driven studies of single exposures. The impact of lower values of *prev* is, of course, lower *PPV* and less trustworthy study results.

In practice, few epidemiology studies achieve the sample sizes necessary to conduct traditional statistical analyses of the individual or interactive effects of all components in an exposure mixture or an omics biomarker; instead, individual exposure models with false discovery rate corrections or variable reduction techniques are routinely employed first to identify a smaller set of predictors, making it possible to fit adequate statistical models [17]. However, because prior data and straightforward methods for power calculation are often lacking for these studies, power calculations tend to be cursory or omitted. Such study designs can be assumed to produce relatively low values for both *power* and *prev*, and thus poor *PPV*. Indeed, a recent simulation study of realistic exposome data for 1200 hypothetical participants investigated several modern statistical methods including dimension reduction that were developed for high-dimensional data, and found that all performed poorly in generating false discoveries for health associations with correlated exposures in the exposome [18•]. The methods examined included including two variations of environmental-wide association study with false discovery rate correction, elastic net, sparse partial least squares, Graphical Unit Evolutionary Stochastic Search, and the deletion/substitution/addition algorithm; false discovery proportions varied from 28 to 86%.

Increasingly, evidence is emerging in support of differences in risk factor susceptibility by individual characteristics such as age, sex, and socioeconomic status [19, 20]. Such investigations often have considerable prior biological support, in which case *prev* is likely to be comparable to other

hypothesis-driven investigations. However, these effect modification analyses will invariably have lower *power* than risk factor analyses for the same outcomes in the entire study, because there are smaller sample sizes in each stratum than in the overall study, and because the difference in effect sizes between strata is typically smaller than the overall effect. Again, lower power results in lower *PPV*. When such investigations are exploratory rather than being supported by prior information, they are more likely to have lower *prev* as well, and thus lower *PPV*.

The limitations of epidemiology for investigating effect modification were discussed extensively in earlier literature [21, 22], which may be worth revisiting as environmental epidemiology struggles with these old problems in the context of its new research priorities.

Proposed Solutions

A variety of solutions to the replication crisis have been proposed. Some are focused on ensuring good laboratory practices, full disclosure of research methods, study registration, and other best practices [23–25] that facilitate replication efforts and may offer some improvements to study precision and bias, but do not directly address the primary causes of the replication crisis elucidated by Eq. 2. This review focuses on potential intervention strategies for the key factors in that equation: high α, low *power*, and low *prev*. Several other proposed strategies that are likely to be effective, including focus on estimation rather than null hypothesis testing, are also discussed.

Choose a Lower Value for α

One of the simplest proposed solutions to the replication crisis is to use a lower value for α. The common choice of $\alpha = 0.05$ is based mostly on tradition rather than principle, and it appears that this somewhat arbitrary choice has allowed countless doubtful conclusions to appear in the scientific literature. Inspection of Eq. 2 reveals that lower values of α can dramatically improve *PPV*. For example, at $power = 0.80$ and $prev = 0.10$, PPV = 95% using $\alpha = 0.005$, much higher than the *PPV* of 64% using $\alpha = 0.05$. Choosing $\alpha = 0.005$ also protects against somewhat lower values for *prev*, but even with 80% power the *PPV* drops below 50% for $prev < 0.006$.

Choosing $\alpha = 0.005$ as the threshold for statistical significance is easily implemented by researchers, grant reviewers, and journals, though it requires about twice the sample size to maintain the same statistical power as $\alpha = 0.05$. Although this solution has been proposed several times over the past few decades, larger numbers of researchers now endorse the routine use of $\alpha = 0.005$ as a threshold for using the term "significant," noting that this approach does not preclude describing results with $0.005 < p < 0.05$ as "suggestive," which might be more appropriate given the likelihood of poor *PPV* for *p* values in that range [26•].

A corollary to choosing a lower value for α is formal adjustment for multiple comparisons. For example, Bonferroni correction multiplies each *p* value in a manuscript, table, or any other group of analyses by *k*, where *k* is the total number of *p* values in that group of analyses. Or, equivalently, the original *p* values can be compared to a threshold of $\frac{\alpha}{k}$ for statistical significance. Although conservative, this approach is easy to implement and highly effective in reducing false discoveries; the effect on *PPV* can be determined by substituting $\frac{\alpha}{k}$ for α in Eq. 2. For example, at $power = 0.80$, $\alpha = \frac{0.05}{k}$, and $prev = \frac{1}{k}$, the *PPV* exceeds 94% for any value of *k*. Thus, if it least one of the *p* values or confidence intervals in a group of analyses is for a true association, then Bonferroni correction ensures a high probability that a statistically significant result in that group actually reflects a true association. Without any correction for multiple comparisons, the PPV can be quite poor for large values of *k*, as shown in the earlier calculations using $\alpha = 0.05$ with $prev = 0.01$ or 0.001.

In practice, researchers often prefer less conservative methods of adjustment for multiple comparisons such as the Holm method or the Benjamini-Hochberg procedure instead of Bonferroni correction. All of these procedures are widely implemented in statistical software packages (e.g., the *p.adjust* function in R) and are reviewed in detail elsewhere [27].

Notably, selection of a more strict threshold for α, whether the value is fixed or dependent on the number of tests, is a reproducibility-enhancing strategy that can be applied to confidence intervals as well as statistical significance testing using *p* values. For example, $\alpha = 0.005$ corresponds to a 99.5% confidence interval, which is computed for approximately normal sampling distributions as the point estimate plus or minus 2.81 times the standard error, rather than 1.96 times the standard error. Similar adjustments to confidence intervals are available to account for multiple comparisons [28]. Of course, any formal adjustment for multiple comparisons requires larger sample sizes to maintain statistical power and precision.

Increase *Power*

It is evident from Eq. 2 that increasing statistical power will increase the *PPV*. Increasing the sample size is not the only viable approach to increasing *power*; epidemiologists can employ a number of efficient study designs that can achieve more precise estimation and larger power with fewer participants than simpler designs, such as pooled sampling, outcome-dependent sampling and other two-phase designs, hybrid designs, or counter-matching [29–33].

Researchers can also identify and prioritize populations with large between-individual disparities in exposure, as greater exposure variation inherently produces more statistical precision and power for exposure effects with monotonic dose-response relationships. For example, PFOA serum concentrations in the C8 Science Panel study population ranged over several orders of magnitude, from background concentrations to occupational levels [34], producing substantially more statistical power for investigating the health effects of PFOA than general population studies with less exposure variation.

Grant reviewers and funding agencies can also play a role by expecting primary aims in confirmatory studies to have high statistical power or precision, justified by clear calculations [35]. Investigators and reviewers face more difficult decisions regarding power expectations for secondary hypotheses and exploratory studies. Innovation and true discovery can be facilitated by funding and conducting studies with low statistical power, but only at the cost of generating more false discoveries that are (or worse yet, are not) later refuted. If nothing else, reviewers should assume that aims not supported by explicit calculations of power or precision most likely will have low power and poor precision, affecting the reliability of those study results.

Increase *Prev*

If scientific knowledge about an exposure or a particular health outcome has any relevance, then informed a priori hypotheses should have higher values of *prev* than hypothesis-free, data-driven analyses [6••]. For example, epidemiological studies of an emerging toxicant affecting a similar system or outcomes in rodent studies have stronger biological plausibility, and likely a higher value of *prev*, than epidemiological studies of a chemical without any supporting toxicological evidence. Preconceived, scientifically supported analyses of potential effect modification by age, sex, race, and other variables are likely to have higher values of *prev* than opportunistic analyses of potential effect modification not predicated on any particular rationale.

A sole focus on well-developed, a priori hypotheses with high *prev* would greatly increase *PPV*, but at a cost of making research less innovative [8•]. Moreover, considering the large expense of conducting a sizable epidemiological study, it seems wasteful to use the resulting data to investigate only preconceived hypotheses with strong prior support. Exploratory analyses should continue to have a role in research, but such analyses have questionable reliability and should be clearly distinguished from confirmatory, hypothesis-driven analyses [36].

Avoid Statistical and Epidemiological Practices That Introduce Bias

Although the solutions highlighted in the previous sections (decrease α, increase power, increase *prev*) can effectively address problems related to random error, they do not address confounding, selection bias, or other types of systematic error. Conducting larger studies can increase power and *PPV* when the studies are performed with the same quality, but if a larger sample size requires other study design changes such as using a different study population with less exposure variability, omitting a key confounding variable, or increasing some other risk of bias, the net effect could be to decrease *PPV* relative to a different study design using a smaller sample size. Thus, it remains important to balance statistical considerations with other principles in epidemiological study design.

Once data are collected, it is always tempting to examine associations in a variety of potentially susceptible subgroups, or to investigate potential effect modifiers, after the primary analyses are completed. These exploratory analyses can be dangerous because they are often post-hoc, without clear plans or sufficient power, and too easily degenerate into a search for something interesting (i.e., statistically significant or with a large effect size) to report. Well intentioned or not, the practice of fitting models until statistical significance is achieved or relatively large effect sizes are found introduces bias away from the null. Even with prior selection of models, reporting bias occurs if "less interesting" results are excluded from publications. Such practices are often referred to as p-hacking or data dredging, and should be avoided [14•].

When many models are fit, such as in air pollution studies with multiple pollutants, outcomes, or lag times, researchers should take pains to include description of all results in a publication, not just those that tell the most interesting story. This is important not only for primary interpretation of the study results, but also because of the potential impact on subsequent systematic reviews and meta-analyses, which can yield biased results if some analyses were preferentially excluded from publication. To put results in context when many models are fit, even without formal adjustment for multiple comparisons, it is useful to focus on patterns and to count the proportion of significant results among any group of related analyses. If that proportion is less than α, the results for the group are indistinguishable from chance, and not particularly persuasive. For example, in an epidemiological analysis with multiple time lags, multiple particulate matter measures, and multiple hourly heart rate variability measures, we noted that our six "statistically significant" associations ($p < 0.05$) constituted only 4.8% of the 126 odds ratios we computed for this pollutant, and therefore did provide support for an association between particulate matter and hourly heart rate variability [37].

Researchers can reduce the potential for bias by developing detailed data analysis plans before statistical analysis is actually performed [14•]. Ideally these plans should include specific statistical models, covariate selection, plans for identifying and handling outliers, and any other necessary details that can be anticipated in advance of data analysis. Although statistical methods may change after diagnostics are used to check assumptions, those contingencies can be included in advance planning. In addition to enhancing reproducibility, detailed data analysis plans can also strengthen grant applications.

Focus on Estimation Instead of Null Hypothesis Testing

Although some investigators and journals advise abandoning *p* values altogether in favor of estimation and confidence intervals, this change will have no impact on reproducibility if confidence intervals excluding the null continue to be emphasized or selectively reported [8•]. This is because hypothesis testing using either a p value < 0.05 or a 95% confidence interval relies on $\alpha = 0.05$, producing the same *PPV* in Eq. 2. Although it is easy to avoid using *p* values and the phrase "statistically significant," it is much more rare to see a paper in epidemiology that does not emphasize confidence intervals that exclude the null hypothesis. It is customary to highlight those findings in the discussion section and abstract, even in journals that ban *p* values. Interpretation of results without emphasizing significant findings seems particularly difficult for papers with dozens, hundreds, or thousands of effect estimates; it might become necessary to interpret each effect estimate and confidence interval in some detail, or perhaps not at all, so as to avoid any emphasis of statistically significant results. Abandonment of hypothesis testing in favor of estimation without null hypothesis comparisons could solve the replication crisis, but this shift would require a cultural revolution in scientific research [8•].

Bias Analysis and Triangulation

As most epidemiologists will recognize, multiple studies on the same topic can yield consistent but incorrect effect estimates if they are all biased due to lack of adjustment for key confounding variables, selection bias, recall bias, or other common threats to validity. The solutions addressing parameters in Eq. 2 will not address those threats, but other tools are available to assess and mitigate them. Some researchers practice quantitative bias analysis, in which probability distributions are used to represent the potential magnitude of major sources of bias, and Monte Carlo sensitivity analysis or Bayesian methods are used to determine their potential impacts on epidemiological effect estimates [38].

With or without quantitative analysis, consideration of the likely direction of bias can be useful in interpreting results, especially when two studies have different sources of bias in opposite directions, in which case the true effect size is more likely to be between the two point estimates [39]. Comparison of results from multiple analyses on the same research question but with different threats to validity has been referred to as "triangulation." It has been argued that triangulation should be a key component in designing replication studies and making science more reliable, considering that systematic errors can not be prevented by repeating previous study designs with larger sample sizes [39]. Examples include comparisons of epidemiological associations for the same outcomes using both measured and modeled serum perfluorooctanoate concentrations in the C8 Science Panel Studies, with measured values being less prone to measurement error for characterizing recent exposures, but more susceptible to reverse causation and physiological confounding than modeled values [40–42]. Instrumental variables and negative controls have also been used to triangulate epidemiological effects, and could be used more widely without much difficulty [43, 44].

Increasing Awareness

Although most of these proposed solutions are focused on steps that researchers can take to increase replicability when designing, conducting, and publishing our studies, it is equally important to combat misconceptions that led to misinterpretation and misuse of published study findings. One common misconception is that $\alpha = 0.05$ indicates that 5% of statistically significant associations are false; the actual proportion of statistically significant associations that are false is $1 - PPV$, and therefore depends not just on α, but on power and *prev* as well. Thus, interpretation of confidence intervals, *p* values, and statistical significance requires understanding the context, such as whether they were computed to address well-powered, plausible a priori hypotheses, or computed for larger sets of underpowered, less plausible exploratory hypotheses. Thus, two numerically identical confidence intervals or *p* values can have very different interpretations.

Perhaps most dangerous is the misconception that "statistically significant" implies that a study result is proven, reliable, clinically relevant, or even true in the reference population. Students, researchers, and the general public should be skeptical of study findings that have not yet been replicated, particularly for findings with low power and *prev*, and those that result from testing many associations without accounting for multiple comparisons. A similar misconception that lack of statistical significance implies no association or relevance in the reference population is also dangerous, as it can result in ignoring strong but non-significant associations that may actually be true and very important in the reference population. Focusing on estimation rather than statistical significance can

help avoid these misunderstandings, but null hypothesis testing is pervasive at present, so it is critical that researchers and science readers become better educated regarding the factors that influence its reliability.

Conclusions

Study results using $\alpha = 0.05$ as the threshold for statistical significance have poor diagnostic reliability for true risk factor associations. This finding is not unique to epidemiology, but appears to be common across all fields using statistical significance testing. However, statistically significant results in environmental epidemiology may warrant even greater skepticism as we increasingly study exposure mixtures, epigenetic mechanisms, and susceptible subpopulations. Fortunately, the causes of the replication crisis are easy to understand, and various interventions can be used to mitigate the problem, including more strict control of type I error rates, using modern study designs to maximize power, conducting hypothesis-driven research, avoiding p-hacking and other practices that introduce bias, focusing on estimation instead of statistical significance, and using study designs with different threats to validity for the same research topic in order to triangulate actual effects.

Compliance with Ethical Standards

Conflict of Interest Scott M. Bartell declares that he has no conflict of interest.

Human and Animal Rights and Informed Consent This article does not contain any studies with human or animal subjects performed by any of the authors.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

Papers of particular interest, published recently, have been highlighted as:
- • Of importance
- •• Of major importance

1. Prinz F, Schlange T, Asadullah K. Believe it or not: how much can we rely on published data on potential drug targets? Nat Rev Drug Discov. September 2011;10(9):712. https://doi.org/10.1038/nrd3439-c1.
2. Begley CG, Ellis LM. Drug development: raise standards for preclinical cancer research. Nature. 2012. https://doi.org/10.1038/483531a.
3. Open Science Collaboration. Estimating the reproducibility of psychological science. Science. 2015;349(6251):aac4716. https://doi.org/10.1126/science.aac4716.
4.• Dumas-Mallet E, Button K, Boraud T, Munafo M, Gonon F. Replication validity of initial association studies: a comparison between psychiatry, neurology and four somatic diseases. PLoS One. 2016;11(6):e0158064. **This study assesses reproducibility by comparing 663 meta analyses of risk factor associations to the initial studies reporting those associations.**
5. Baker M. 1,500 scientists lift the lid on reproducibility. Nat News. 2016;533(7604):452–4.
6.•• Ioannidis JPA. Why most published research findings are false. PLoS Med. 2005;2(8):e124. https://doi.org/10.1371/journal.pmed.0020124. **This study extends the PPV framework to account for bias, and provides example PPV calculations for various types of studies.**
7.•• Browner WS, Newman TB. Are all significant p values created equal? The analogy between diagnostic tests and clinical research. JAMA. 1987;257:2459–63. **This study explains the application of the PPV framework to hypothesis testing in research.**
8.• Lash TL. The harm done to reproducibility by the culture of null hypothesis significance testing. Am J Epidemiol. 2017;186(6):627–35. **This manuscript discusses the poor reproducibility of traditional hypothesis testing, and advocates a change in scientific culture to focus on estimation.**
9. McDonald JH. Handbook of biological statistics. 3rd ed. Baltimore: Sparky House Publishing; 2014.
10.• Sterne JAC, Smith GD. Sifting the evidence—what's wrong with significance tests? Phys Ther. 2001;81(8):1464–9. **This manuscript assesses the impacts of power and type I error rate on the proportion of false positives, and advocates the use of p-values as measures of evidence rather than determining statistical significance.**
11. Mullard A. 2016 FDA drug approvals. Nat Rev Drug Discov. 2017;16:73–6.
12. Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J, Robinson ESJ, et al. Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci. 2013;14(5):365–76.
13.•• Dumas-Mallet E, Button KS, Boraud T, Gonon F, Munafò MR. Low statistical power in biomedical science: a review of three human research domains. R Soc Open Sci. 2017 cited 2018 Aug 5;4(2)160254. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5367316/. **This study assesses reproducibility by comparing 663 meta analyses of risk factor associations to the initial studies reporting those associations.**
14.• Munafò MR, Nosek BA, Bishop DVM, Button KS, Chambers CD, du Sert NP, et al. A manifesto for reproducible science. Nat Hum Behav. 2017;1(1):0021. **This manuscript proposes changes in key elements of the scientific process that could enhance reproducibility.**
15. Rothman KJ. No adjustments are needed for multiple comparisons. Epidemiology. 1990;1(1):43–6.
16. Young SS. Air quality environmental epidemiology studies are unreliable. Regul Toxicol Pharmacol. 2017;86:177–80.
17. Chadeau-Hyam M, Campanella G, Jombart T, Bottolo L, Portengen L, Vineis P, et al. Deciphering the complex: methodological overview of statistical models to derive OMICS-based biomarkers. Environ Mol Mutagen. 2013;54(7):542–57.
18.• Agier L, Portengen L, Chadeau-Hyam M, Basagaña X, Giorgis-Allemand L, Siroux V, et al. A systematic comparison of linear regression-based statistical methods to assess exposome-health associations. Environ Health Perspect. 2016 [cited 2018 Aug 6];124(12)1848–1856. Available from: http://ehp.niehs.nih.gov/EHP172. **This exposome simulation study assesses the false discovery proportion and sensitivity for a variety of common statistical methods addressing multiple comparisons.**

19. Mielke MM, Vemuri P, Rocca WA. Clinical epidemiology of Alzheimer's disease: assessing sex and gender differences. Clin Epidemiol. 2014;6:37–48.
20. van den Berg M, Wendel-Vos W, van Poppel M, Kemper H, van Mechelen W, Maas J. Health benefits of green spaces in the living environment: a systematic review of epidemiological studies. Urban For Urban Green. 2015;14(4):806–16.
21. Greenland S. Tests for interaction in epidemiologic studies: a review and a study of power. Stat Med. 1983;2(2):243–51.
22. Thompson WD. Effect modification and the limits of biological inference from epidemiologic data. J Clin Epidemiol. 1991;44(3):221–32.
23. Landis SC, Amara SG, Asadullah K, Austin CP, Blumenstein R, Bradley EW, et al. A call for transparent reporting to optimize the predictive value of preclinical research. Nature. 2012;490(7419):187–91.
24. Collins FS, Tabak LA. NIH plans to enhance reproducibility. Nature. 2014;505(7485):612–3.
25. LaKind JS, Goodman M, Makris SL, Mattison DR. Improving concordance in environmental epidemiology: a three-part proposal. J Toxicol Environ Health B Crit Rev. 2015;18(2):105–20.
26.• Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers E-J, Berk R, et al. Redefine statistical significance. Nat Hum Behav. 2018;2(1):6–10. **This manuscript by 72 authors advocates the use of 0.005 instead of 0.05 as the standard threshold for statistical significance.**
27. Shaffer JP. Multiple hypothesis testing. Annu Rev Psychol. 1995;46(1):561–84.
28. Benjamini Y, Yekutieli D, Edwards D, Shaffer JP, Tamhane AC, Westfall PH, et al. False discovery rate: adjusted multiple confidence intervals for selected parameters [with comments, rejoinder]. J Am Stat Assoc. 2005;100(469):71–93.
29. Langholz B, Borgan ØR. Counter-matching: a stratified nested case-control sampling method. Biometrika. 1995;82(1):69–79.
30. Weinberg CR, Umbach DM. Using pooled exposure assessment to improve efficiency in case-control studies. Biometrics. 1999;55(3):718–26.
31. Zhou H, Weaver MA, Qin J, Longnecker MP, Wang MC. A semiparametric empirical likelihood method for data from an outcome-dependent sampling scheme with a continuous outcome. Biometrics. 2004;58:413–21.
32. Haneuse S, Bartell S. Designs for the combination of group- and individual-level data. Epidemiology. 2011;22(3):382–9.
33. Kass PH. Modern epidemiological study designs. In: Handbook of epidemiology. Springer, New York, NY; 2014 [cited 2018 Aug 5]. p. 325–63. Available from: https://link.springer.com/referenceworkentry/10.1007/978-0-387-09834-0_8
34. Steenland K, Jin C, MacNeil J, Lally C, Ducatman A, Vieira V, et al. Predictors of PFOA levels in a community surrounding a chemical plant. Environ Health Perspect. 2009;117(7):1083–8.
35. Rothman KJ, Greenland S. Planning study size based on precision rather than power. Epidemiology. 2018;29(5):599–603.
36. Tukey JW. We need both exploratory and confirmatory. Am Stat. 1980;34(1):23–5.
37. Bartell SM, Longhurst J, Tjoa T, Sioutas C, Delfino RJ. Particulate air pollution, ambulatory heart rate variability, and cardiac arrhythmia in retirement community residents with coronary artery disease. Environ Health Perspect. 2013;121(10):1135–41.
38. 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.
39. Munafò MR, Smith GD. Robust research needs many lines of evidence. Nature. 2018;553(7689):399–401.
40. Watkins DJ, Josson J, Elston B, Bartell SM, Shin H-M, Vieira VM, et al. Exposure to perfluoroalkyl acids and markers of kidney function among children and adolescents living near a chemical plant. Environ Health Perspect. 2013;121(5):625–30.
41. Dhingra R, Winquist A, Darrow LA, Klein M, Steenland K. A study of reverse causation: examining the associations of perfluorooctanoic acid serum levels with two outcomes. Environ Health Perspect. 2017;125(3):416–21.
42. Weisskopf MG, Webster TF. Trade-offs of personal versus more proxy exposure measures in environmental epidemiology. Epidemiology. 2017;28:635–43.
43. Lipsitch M, Tchetgen ET, Cohen T. Negative controls: a tool for detecting confounding and bias in observational studies. Epidemiology. 2010;21(3):383–8.
44. Arnold BF, Ercumen A, Benjamin-Chung J, Colford JMJ. Brief report: negative controls to detect selection bias and measurement bias in epidemiologic studies. Epidemiology. 2016;27(5):637–41.