DATA-DREDGING BIAS

Adrian Erasmus [1,2]
ORCID iD: https://orcid.org/0000-0001-8944-1598
[1] Department of History and Philosophy of Science, University of Cambridge, Cambridge, UK
[2] Institute for the Future of Knowledge, University of Johannesburg, Johannesburg, South Africa

Bennet Holman [3,4] (Corresponding Author)
ORCID iD: https://orcid.org/0000-0001-9038-2644
[3] Underwood International College, Yonsei University, Seoul, South Korea
[4] Faculty of Humanities, University of Johannesburg, Johannesburg, South Africa
(POSTAL ADDRESS)
bholman@yonsei.ac.kr

John P.A. Ioannidis [5,6,7,8,9]
ORCID iD: https://orcid.org/0000-0003-3118-6859
[5] Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, CA, USA
[6] Department of Epidemiology and Population Health, Stanford University School of Medicine, Stanford, CA, USA
[7] Stanford Prevention Research Center, Department of Medicine, Stanford University School of Medicine, Stanford, CA, USA
[8] Department of Biomedical Data Science, Stanford University School of Medicine, Stanford, CA, USA
[9] Department of Statistics, Stanford University School of Humanities and Science, Stanford, CA, USA

Word Count:1491

Statements:

Funding Statement: This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Competing Interests Statement: All authors declare no competing interests

Contributorship Statement: AE and BH cowrote a first draft of the article. Which was subsequently corrected and revised by all authors.

Background: What is Data Dredging Bias?

Data-dredging bias encompasses a number of more specific questionable practices (e.g., fishing, p-hacking, etc.) all of which involve probing data using unplanned analyses and then reporting salient results without accurately describing the processes by which the results were generated. Almost any process of data analysis involves numerous decisions necessary to complete the analysis (e.g. how to handle outliers, whether to combine groups, including/excluding covariates). Where possible, it is best practice for these decisions to be guided by a principled approach and prespecified in a publicly available protocol. When it is not possible, authors must be transparent about the open-ended nature of their analysis.

While a set of choices may well be methodologically defensible and reliable had the specifications been made prior to the analysis. However, probing the data and selectively reporting an outcome as if it were always the intended course of analysis dramatically increases the likelihood of finding a statistically significant result when there is in fact no effect (i.e., a false positive).

As an intuitive example, consider a hypothesis that a given coin is unfairly biased to heads. Suppose I flip the coin twenty times each day for a week and assume that I am allowed to (1) eliminate data from any given day; (2) consider only the first 10 flips in a day, the last 10 flips in a day, or all 20 flips; and (3) restrict my consideration to only flips that were preceded by a heads or flips that were preceded by a tails. I would be conducting a fair trial if I prespecify that I will consider only the results when the prior flip was "tails" and the flip was one of the last 10 in the series for the day, but I will be excluding the results from Wednesday. This is because none of these factors actually influence the probability of a flip coming up heads. However, if I am allowed to dredge through the data and freely examine various combinations of these restrictions after the results are known, I am virtually certain to find some set of specifications that makes it appear that the coin is biased. Importantly, the bias depends on analyzing and reporting the data *as if* the analytic choices were specified ahead of time. Once the curtain is pulled back on how the results were generated, no one should be convinced by the evidence, least of all the analyst.

Though the intuitive example suggests that data-dredging is obviously unreliable, many common forms of data-dredging arise from a lack of knowledge rather than an intent to deceive. As [1] note, "it can seem entirely appropriate to look at the data and construct reasonable rules for data exclusion, coding, and analysis that can lead to statistical significance" (p. 461). Because researchers often remain unaware of the ways in which their biases shape their decisions, there is no substitute for a prespecified course of analysis. The absence of a study protocol outlining the planned methods and statistical analyses is red flag that the results may be a result of data dredging.

To be clear, unplanned analyses are appropriate for hypothesis generation. Similarly, planned interim analyses are a standard part of adaptive trials. However, such analyses must be reported accurately, and different statistical analyses are often required. In contrast, data dredging has occurred if the results are reported as if they were the result of a single preplanned analysis, but were actually generated by: collecting the data first and then systematically adjusting data analysis approaches until the researcher finds a set of choices that produces a statistically

significant result ("p-hacking");[2] assessing models with multiple combinations of variables and selectively reporting the "best" model ("fishing");[3] making decisions about whether to collect new data on the basis of interim results; or generating a hypothesis to explain results which have already been obtained but presenting it as if it were a hypothesis one had prior to collecting the data (HARKing ("hypothesizing after the results are known")).[4]

Examples

The use of progestogens to prevent pregnancy loss was supported by numerous randomized control trials and meta-analyses of those trials; however, this apparent benefit disappears when the analysis is restricted to trials reporting preregistered primary outcomes.[5] When meta-analyses included trials that were not preregistered, 19 of 29 meta-analyses found significant benefits for progestogens. In contrast, a meta-analysis constrained to the 22 trials which reported preregistered primary outcomes provided substantial evidence that progestogens were ineffective (RR = 1.00, 95% CI 0.94-1.07). Since for publication bias to account for such a difference, there would have to be an implausibly large number of unpublished trials, the difference is likely caused by various forms of data-dredging in the unregistered trials.[5]

Another well-known and extreme example of data dredging are the studies conducted at the Cornell Food and Brand Lab, under the supervision of Brian Wansink.[6] Scrutiny was drawn to his work after a November 2016 blog post in which Wansink described how to succeed in academia. In the post Wansink recounts a study which was intended to replicate his earlier finding that patrons would eat more at an all-you-can-eat buffet if they paid more. Instead, the results contradicted his earlier work. Rather than publish the results of the trial as designed, Wansink passed the data from the "failed study" to a visiting scholar in his lab. In his original recounting, Wansink valorizes the visiting scholar who, at his direction, continually reanalyzed the data looking for some result they could "salvage." Indeed, she found several.

In response to subsequent criticism, Wansink wrote: "With field studies, hypotheses usually don't 'come out' on the first data run. But instead of dropping the study, a person contributes more to science by figuring out when the hypo (sic) worked and when it didn't. This is Plan B. Perhaps your hypo worked during lunches but not dinners, or with small groups but not large groups. You don't change your hypothesis, but you figure out where it worked and where it didn't. Cool data contains cool discoveries."[7] Again, had the article described all of the unplanned analyses that ultimately culminated in the published result, such an attitude may be defensible. However, the results were presented as if they were testing a hypothesis with a preplanned course of analysis, see [8], and represent an egregious form of data-dredging.

Impact

While the exact frequency of data-dredging has not been determined, the unusually large number of published studies that just pass the threshold $p < .05$ has been offered as evidence for its prevalence.[9, 10] While others have suggested that p-hacking is unlikely to have a significant effect on areas of research where there are enough studies to conduct a meta-analysis,[11] such a conclusion depends on those meta-analyses containing multiple studies with large sample sizes.

P-curve analyses suggest that the distribution of p-values indicates that most research is investigating real effects; however, some forms of data-dredging could also produce a similar distribution.[12, 13] Despite remaining uncertainties, there is no dispute that the bias introduced by data dredging will be most severe where the effect size is small, the dependent measures are imprecise, research designs are flexible, and studies are conducted on small populations.[14]

Preventative steps

Because they do not understand such practices as flawed, researchers often unabashedly endorse data-dredging (e.g., inspecting interim results to determine whether to continue collecting data, deciding whether to exclude outliers after assessing the effects of doing so, etc.). Accordingly, the most promising preventive step is education and change in disciplinary norms. Best practice includes designing a trial with sufficient power to detect meaningful effects, prespecifying rules for stopping data collection, and plans for data analysis, including criteria for excluding outliers, expected variable transformations, whether covariates will be controlled for, etc. Such plans should be registered before data collection commences.[14] Published articles should identify the complete list of variables studied in the trial, report all planned analyses (indicating which results were prespecified as primary and which as secondary), and conduct robustness analyses for methodological choices.[2]

Some researchers have advocated p-curve analysis as a formal means of correcting p-hacking;[15, 16] however, more research is needed to validate the procedure empirically. Because confounding variables can render p-curve analysis unreliable, its application beyond randomized trials is particularly uncertain.[13]

Yet the above solutions fail to capture the extent to which data-dredging is a problem of bad barrels rather than bad apples.[18] Wansink may have been wrong about the reliability of the procedures he used, but he was not wrong about their ability to increase academic output. So long as disciplinary incentives reward data dredging, researchers will face pressures to engage in it. Accordingly, the most successful interventions require changes to disciplinary standards of designing, implementing, and publishing scientific research.[19] For example, journals that accept articles based on their prospective designs mitigate the need and ability to dredge through data to produce "attractive" results.[20] Given the current standards of inquiry, such interventions would require fundamental changes to most disciplines.

References

[1] Gelman, A., & Loken, E. (2014). The Statistical Crisis in Science. American Scientist, 102(6), 460. https://doi.org/10.1511/2014.111.460

[2] Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. Psychological Science, 22(11), 1359–1366. https://doi.org/10.1177/0956797611417632

[3] Selvin, H., & Stuart, A. (1966). Data-dredging procedures in Survey Analysis. The American Statistician, 20(3), 20-23.

[4] Kerr, N. (1998). HARKing: Hypothesizing After the Results are Known. Personality and Social Psychology Review, 2(3), 196-217.

[5] Prior, M., Hibberd, R., Asemota, N., & Thornton, J. (2017). Inadvertent p-hacking among trials and systematic reviews of the effect of progestogens in pregnancy? A systematic review and meta-analysis. BJOG: An International Journal of Obstetrics & Gynaecology, 124(7), 1008–1015. https://doi.org/10.1111/1471-0528.14506

[6] van der Zee, T., Anaya, J., & Brown, N. (2017). Statistical heartburn: an attempt to digest four pizza publications from the Cornell food and brand lab. BMC Nutrition, 3, 1-15.

[7] https://web.archive.org/web/20170312041524/http:/www.brianwansink.com/phd-advice/the-grad-student-who-never-said-no

[8] Just, D. R., Sığırcı, Ö., & Wansink, B. (2014). Lower buffet prices lead to less taste satisfaction. Journal of sensory studies, 29(5), 362-370.

[9] Perneger, T. V., & Combescure, C. (2017). The distribution of p-values in medical research articles suggested selective reporting associated with statistical significance. Journal of Clinical Epidemiology, 87, 70–77. https://doi.org/10.1016/j.jclinepi.2017.04.003

[10] Ioannidis, J. P. A. (2019). What Have We (Not) Learnt from Millions of Scientific Papers with p-Values? The American Statistician, 73(sup1), 20–25. https://doi.org/10.1080/00031305.2018.1447512

[11] Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The Extent and Consequences of p-hacking in Science. PLOS Biology, 13(3), e1002106. https://doi.org/10.1371/journal.pbio.1002106

[12] Bishop, D. V. M., & Thompson, P. A. (2016). Problems in using p-curve analysis and text-mining to detect rate of p -hacking and evidential value. PeerJ, 4, e1715. https://doi.org/10.7717/peerj.1715

[13] Bruns, S. B., & Ioannidis, J. P. A. (2016). P-Curve and p-Hacking in Observational Research. PLOS ONE, 11(2), e0149144. https://doi.org/10.1371/journal.pone.0149144

[14] Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. PLoS Medicine, 2(8), e124. https://doi.org/10.1371/journal.pmed.0020124

[15] Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An Agenda for Purely Confirmatory Research. Perspectives on Psychological Science, 7(6), 632–638. https://doi.org/10.1177/1745691612463078

[16] Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). p-curve: A key to the file-drawer. Journal of Experimental Psychology: General, 143(2), 534–547. https://doi.org/10.1037/a0033242

[17] Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). Better p-curves: Making p-curve analysis more robust to errors, fraud, and ambitious P-hacking, a Reply to Ulrich and Miller (2015). Journal of Experimental Psychology. General, 144(6), 1146–1152. https://doi.org/10.1037/xge0000104

[18] Banks, G. C., O'Boyle, E. H., Pollack, J. M., White, C. D., Batchelor, J. H., Whelpley, C. E., Abston, K. A., Bennett, A. A., & Adkins, C. L. (2016). Questions About Questionable Research Practices in the Field of Management: A Guest Commentary. Journal of Management, 42(1), 5–20. https://doi.org/10.1177/0149206315619011

[19] Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. P. A. (2017). A Manifesto for Reproducible Science. Nature Human Behaviour, 1(1). https://doi.org/10.1038/s41562-016-0021

[20] Nosek, B. A., & Lakens, D. (2014). Registered Reports: A Method to Increase the Credibility of Published Results. Social Psychology, 45(3), 137–141. https://doi.org/10.1027/1864-9335/a0