Journal of Clinical Epidemiology 70 (2016) 258–260

Journal of Clinical Epidemiology

Anticipating consequences of sharing raw data and code and of awarding badges for sharing

John P.A. Ioannidis[a,b,*]

[a]*Stanford Prevention Research Center (SPRC), Stanford University, Stanford, CA, USA*
[b]*Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, CA, USA*

Accepted 22 April 2015; Published online 8 July 2015

West argues [1] that on an elective basis, journals could encourage authors of scientific publications to make data and statistical analysis command files available and get a ''transparency'' quality marker in reward. The concept is interesting, and the suggestion of data and code sharing is not new. Some journals have long made this a prerequisite for publication for some study types, for example, microarrays [2]. Across the scientific literature, the proportion of journals that adopt sharing data and code policies is increasing [3,4], although these editorial policies are not strongly enforced [3]. The Center for Open Science (COS) has a process where it attributes to articles any of the three following rewarding ''badges'' [5]: open data, open materials, and preregistered. What West proposes is practically equivalent to the open data badge, which requires the following two prerequisites to be awarded [5]:

1. Digitally shareable data are publicly available on an open-access repository (eg, university repository or one at www.re3data.org or www.databib.org).
2. A codebook is included with sufficient description for an independent researcher to reproduce the reported analyses and results. Data from the same project that are not needed to reproduce the reported results can be kept private without losing eligibility for the open data badge.

Assuming data and code do become available, an obvious question arises: Now what? What are we going to do with them? West [1] offers two potential uses: checking for and correcting errors and performing additional analyses to improve interpretation and identify new findings. Although I value these possibilities, it is important to probe empirically how fruitful they are likely to be. First, errors in data and computations are likely to be very common, probably almost ubiquitous in any scientific project that has any volume of work behind it [6]. Some of these errors are easily suspected even from the published reports [7,8]. The question is how frequently errors are big enough to make a profound difference in the inferences and conclusions of a article. We have more limited evidence to answer this question. The frequency of such errors may increase with the complexity of data set and analytical issues and the lack of expertise of the researchers. However, researchers who take the step to share their raw data and analysis codes in public are likely to be among the most sophisticated and most sensitized to error issues. They are also likely to take extra steps to check that what they submit in public squares with what they report in the article, so as to avoid major embarrassments. These extra steps may require some extra resources and effort spent on preparing the publicly viewable versions of the data [9] because data are often used in chaotic formats [10]. Even then, some errors would be unavoidable. In that case, the paradox may arise that the most meticulous and sophisticated and method-savvy and careful researchers may become more susceptible to criticism and reputation attacks by reanalyzers who hunt for errors, no matter how negligible these errors are. This means that we need to find ways to incentivize researchers to share but also avoid unnecessary aggressiveness or defensiveness when shared data and code have errors or do not get replicated [11]. In a new paradigm of openness, we should accept that error is ubiquitous, as opposed to the dominant current paradigm where reviewers cannot find essential flaws in most articles simply because they have no means to really check the work that was done. It is not surprising that most peer reviews focus on conceptual and philosophical issues, presentation aesthetics (eg, how to best present tables or figures), or grammar and syntax rather than the core of the science, with some exceptions [12].

Another disturbing corollary is that, if the only request to get an open data badge is the availability of data and

Conflicts of interest: The author reports no conflicts of interest. The author alone is responsible for the content and writing of the article.

* Corresponding author.1265 Welch Road, Medical School Office Building, Room X306, Stanford, CA 94305, USA. Tel.: 650-7045584.
E-mail address: jioannid@stanford.edu

http://dx.doi.org/10.1016/j.jclinepi.2015.04.015

code, there is absolutely no guarantee that what is deposited reflects what really was done. Often a multitude of analyses are conducted, and only the ''best'' results or products of ''double-dipping'' are written up for publication [13–15]. ''Best'' could mean most-significant largest effects (and thus probably grossly inflated [15]), or most congruent with whatever bias is held by the researcher, sponsor, or other stakeholders involved [16]. In this case, awarding an open data badge sadly makes the situation more misleading because it can masquerade as a marker of quality, although the presented analyses may represent the extreme of selection bias processes.

The notion that open availability of data and code will allow the performance of more analyses that may change the interpretation of the original findings and offer new insights has merit but is also fraught with some difficulties. These analyses add to the multiplicity burden. When data become widely available, in theory, millions of scientists could now use them in billions of different ways, leading to lower poststudy odds of derived significant results [17]. Moreover, these analyses may add multiplicity that is mostly nontransparent, if most of these scientists do not disclose all their attempts to find something that would change the interpretation of the original findings. The credibility of such analyses may be pretty low. West identifies [1] conditions that would make the credibility suffer even further, such as analyses by authors with vested interests and those with poor analytical credentials. One hopes that with maximized transparency, transparency rules would apply also to these additional analyses. Unfortunately, currently scientists are incentivized to claim new discoveries so as to publish and get funded. It is thus not surprising that when reanalyses of raw data are performed, even for randomized trials and even for the same hypothesis as in the original article, in 35% of the cases, the conclusions change about whether patients should be treated or not and which patients should be treated [18]. Interestingly, authors from the original articles almost always coauthor these reanalyses that lead to different interpretations. This reinterpretation challenge is likely to become even greater with wider availability of raw data, unless the incentive structure changes and reward are given for getting high-quality results rather than ''new,'' ''significant,'' and ''different'' results.

Another important potential use of raw data is not in performing additional analyses using just the same data set, but in integrating raw data from multiple similar data sets. The advantages and caveats of such meta-analyses using raw data are beyond the scope of this commentary. However, when raw data are available from only a small subset of the relevant studies, perhaps a biased subset, there can be major disadvantages compared with a more comprehensive meta-analysis of group-level data.

In all, I think that data that have been collected and code that has been written and implemented with effort and with expenditure of resources are worthwhile saving for further use—by default. However, the way these tools are going to be used and the interpretation of such uses need careful consideration. More sophisticated approaches are currently available that allow capturing every step in the design, conduct, and implementation of a study and its computational analyses. For example, the entire design and computational process may be summarized in a barcode that can then be re-expanded to show all the steps that were taken [19], including the false leads, the hidden multiplicity of dead-end analyses, and all the persistent plain errors that remain to be discovered by future reanalysts. Such detailed disclosure and documentation would make most sense in the setting where it is coupled with some sort of study pre-registration [20]. Preregistration is one of the three badges mentioned previously, and to get one, the following criteria must be met [5]:

1. A public date-time stamped registration is in an institutional registration system (eg, ClinicalTrials.gov, open science Framework)
2. Registration predates realization of the outcomes
3. Registered design and analysis plan corresponds directly to reported design and analysis
4. Full disclosure of results following the registered plan

Although I believe that preregistration and full documentation is a good idea, it probably applies only to a fraction of conducted research. Much research is and should continue to be exploratory. It would be misleading to try to make exploratory research seem more preconceived than it really is. Scientists should be allowed to ponder ideas and analyze data in nonpredetermined, unexpected, and unconventional ways, provided that they also fully acknowledge that this is exactly what they have done. Softwares that track their tortuous process through exploratory ideas and analyses may be useful for being able to back track what was done, but the exact role of such back tracking remains to be determined. Eventually, one should specify what exactly worked, if any, among zillions of performed exploratory analyses, so that independent hypothesis-validating studies could test the very same analytical sequence for validation. However, we should not create an artificial environment where investigators are incentivized to seemingly preregister studies that are not really preconceived, simply because carrying a ''preregistered'' badge carries value. Although it may sound simple, creating the optimal research environment that fosters disclosing what we did and what we found and sharing the evidence is not straightforward.

References

[1] West R. Promoting greater transparency and accountability in clinical and behavioural research by routinely disclosing data and statistical commands. J Clin Epidemiol 2015. http://dx.doi.org/10.1016/j.jclinepi.2015.06.015. [Epub ahead of print].

[2] Ioannidis JP, Allison DB, Ball CA, Coulibaly I, Cui X, Culhane AC, et al. Repeatability of published microarray gene expression analyses. Nat Genet 2009;41:149–55.

[3] Alsheikh-Ali AA, Qureshi W, Al-Mallah MH, Ioannidis JP. Public availability of published research data in high-impact journals. PLoS One 2011;6:e24357.
[4] Stodden V, Guo P, Ma Z. Toward reproducible computational research: an empirical analysis of data and code policy adoption by journals. PLoS One 2013;8:e67111.
[5] Center for Open Science, Open Science Framework: Badges to acknowledge open practices. Available at https://osf.io/tvyxz/wiki/home/. Accessed December 24, 2014.
[6] Dewald WG, Thursby JG, Anderson RG. Replication in empirical economics. The Journal of Money, Credit and Banking Project. Am Econ Rev 1986;76:587−603.
[7] Garcia-Berthou E, Alcaraz C. Incongruence between test statistics and P values in medical papers. BMC Med Res Methodol 2004;4:13.
[8] Bakker M, Wicherts JM. The (mis)reporting of statistical results in psychology journals. Behav Res Methods 2011;44:666−78.
[9] Khokhar RH, Chen R, Fung BC, Lui SM. Quantifying the costs and benefits of privacy-preserving health data publishing. J Biomed Inform 2014;50:107−21.
[10] Doshi P, Jefferson T, Del Mar C. The imperative to share clinical study reports: recommendations from the Tamiflu experience. Plos Med 2012;9:e1001201.
[11] Bohannon J. Replication effort provokes praise—and 'bullying' charges. Science 2014;344:788−9.
[12] Hopewell S, Collins GS, Boutron I, Yu LM, Cook J, Shanyinde M, et al. Impact of peer review on reports of randomised trials published in open peer review journals: retrospective before and after study. BMJ 2014;349:g4145.
[13] Kriegeskorte N, Simmons WK, Bellgowan PSF, Baker CI. Circular analysis in systems neuroscience: the dangers of double dipping. Nat Neurosci 2009;12:535−40.
[14] Dwan K, Altman DG, Clarke M, Gamble C, Higgins JP, Sterne JA, et al. Evidence for the selective reporting of analyses and discrepancies in clinical trials: a systematic review of cohort studies of clinical trials. Plos Med 2014;11:e1001666.
[15] Ioannidis JP. Why most discovered true associations are inflated. Epidemiology 2008;19:640−8.
[16] Ioannidis JP. How to make more published research true. Plos Med 2014;11:e1001747.
[17] Ioannidis JP. Why most published research findings are false. Plos Med 2005;2:e124.
[18] Ebrahim S, Sohani ZN, Montoya L, Agarwal A, Thorlund K, Mills EJ, et al. Reanalyses of randomized clinical trial data. JAMA 2014;312:1024−32.
[19] Gavish M, Donoho D. A universal identifier for computational results. Procedia Comput Sci 2011;4:637−47.
[20] Nosek BA, Lakens D. Registered reports. Soc Psychol 2014;45:137−41.