Journal of Clinical Epidemiology 70 (2016) 254–255

Journal of Clinical Epidemiology

COMMENTARIES

Data and statistical commands should be routinely disclosed in order to promote greater transparency and accountability in clinical and behavioral research

Robert West*

Department of Epidemiology and Public Health, Health Behaviour Research Centre, University College London, Gower Street, London WC1E 6BT, UK

Accepted 26 June 2015; Published online 9 July 2015

This commentary argues for clinical, public health, and health science journals routinely to invite authors to make data and statistical analysis command files underlying the findings reported in articles available as supplementary files and signal prominently articles for which this is done using a ''transparency'' quality marker.

The background to this is a need to make the conduct of science more efficient and effective. The Lancet recently published a series of articles on waste in science which drew attention to the many ways in which public money is wasted unnecessarily because of the way we undertake science [1]. There is waste throughout the process from awarding of research funds, conduct of research, writing up of findings, publishing findings, and even their dissemination and the use to which they are/are not put.

One part of the process that requires special attention is transparency with regard to the data and its statistical analysis. Calls for greater transparency have a long history [2], and there are now guidelines concerning data sharing and availability [3]. Apart from providing some protection against fraud and misrepresentation of findings, there are at least two important ways in which transparency could improve our science. One is in reducing the error rate, both in the data itself and in its analysis. The other is in facilitating additional analyses that can help in the interpretation of reported findings or establishing new findings that were not included in the original published report.

Considering first the issue of data errors, mistakes in recording and handling of data could account for part of the high rate of failure to replicate findings in clinical and health research. There are numerous potential entry points for errors in the data including misrecording and mistranscribing, and incorrect commands for recoding variables and computing new variables. It is all too easy for mistakes to creep in, and in most cases, these will not be checked by an independent source. If the data and the commands are available, it will be possible for someone after publication to check them. It will also provide a greater incentive to ensure that they are checked before publication.

When it comes to the opportunities provided for additional analyses, it is common when reading an article to want to know more about the analyses and what results would have been obtained had somewhat different analyses or coding been undertaken. It is often not realistic for all the plausible different ways of analyzing data to answer a research question to be presented in an article, but if the data and commands were available, it would be open to others to undertake the analyses and either reassure themselves that the findings were robust or identify weaknesses. For example, how one groups a quantitative variable can make a substantial difference to one's findings, as can the way in which one combines variables into a composite score. It can also make a difference whether one uses one underlying statistical model or another. Besides establishing the robustness or otherwise of findings, making the data available offers the opportunity for readers of an article to answer questions that might otherwise never be addressed. If additional analyses lead to new insights and substantially change the interpretation of findings, this could be communicated to the original authors in the first instance and then, if appropriate, form the basis for public correspondence. It is not expected that anyone other than the owner(s) of the data set would have rights to publish findings beyond this limited quality control process. If a substantive new finding was to emerge from further analysis, publication would normally have to be agreed with the owners of the data. An exception would be data sets that were explicitly designated as being in the public domain.

Possible negative consequences need to be considered. One is that issues of intellectual property will need to be clarified. It is important to note that what is being proposed is data ''disclosure,'' not data sharing. Thus, ownership and

Conflict of interest: None.

* Corresponding author.
E-mail address: robert.west@ucl.ac.uk

http://dx.doi.org/10.1016/j.jclinepi.2015.06.015

intellectual property would clearly remain with the primary authors. However, if another researcher identifies a flaw in the analysis, it will be necessary to ensure that he or she has the right to publish this, having first alerted the primary authors. It will also be necessary to inculcate routine annotation of data sets so that they can be used by others. Again, this should not be a major issue as it is already established as good practice. A further issue is the opportunity for vested interests such as the tobacco industry to make use of the data sets. This is a serious problem and could be grounds for more limited availability so that researchers wanting to look at the data have to state their credentials. There may occasionally be issues of privacy for patients or participants. If the data set could not be anonymized, this would probably override the disclosure principle.

Bearing in mind the possible drawbacks, as a first step, it may be the best to encourage rather than require authors to make data sets and command files available as supplementary files. Those who do could have their article ''kitemarked'' as meeting that particular quality standard. Depending on how this goes, one could then move to making data disclosure compulsory.

References

[1] Glasziou Paul, Altman Douglas G, Bossuyt Patrick, Boutron Isabelle, Clarke Mike, Julious Steven, et al. Reducing waste from incomplete or unusable reports of biomedical research. Lancet 2014;383:267–76.

[2] Fienberg SE, Martin ME, Straf ML, editors. Sharing research data. Washington, D.C: National Academy Press; 1985.

[3] Inter-university Consortium for Political and Social Research (ICPSR). Research transparency, data access, and data citation: a call to action for scholarly publication. Available at http://datacommunity.icpsr.umich.edu/research-transparency-data-access-and-data-citation-call-action-scholarly-publications. Accessed July 23, 2015.