If you have ever had the pleasure of leafing through old journals you would appreciate that the journals of yesteryear were quite different. A typical article from a scientific journal of the early 20th Century was much longer than those of today. Many articles in old journals contained lots of tables but few graphs. (The graph is a remarkably recent innovation.) And, surprisingly to the modern reader, it was not uncommon for researchers to publish a complete record of all of the data from a particular experiment. That is, researchers often reported data from each subject.

These days journal articles rarely report data from individual subjects. Instead, it is more usual to see summary statistics reported. For example, researchers may report the mean and standard deviation of a distribution. Descriptive statistics such as means and standard deviations, when used appropriately, provide a concise summary which substitutes for a tedious enunciation of each datum.

The convention of reporting summary statistics, rather than ‘raw’ data, is a pragmatic one. Modern readers are faced with unmanageably large amounts of research data so they prefer to read concise research reports. Moreover, contemporary clinical studies are very large. Obviously it would be impossible to provide, in hard copy, data for each of the 38 050 participants in the observational study of low back pain reported by Smith and colleagues in this journal (Smith et al 2006). The sensible short-cut is to report only summaries of data.

There are, however, reasons why some readers might want access to raw data. Access to raw data makes it possible to:

1. Scrutinise data. By inspecting raw data readers can ascertain how complete the data set is and identify anomalies in the data such as outliers. This provides an indication of data quality that may not be apparent in summary statistics.

2. Re-analyse data. Some published statistical analyses (perhaps most often the simplest analyses) are performed incorrectly. Even when the analysis is conducted correctly, it may be suboptimal. When raw data are available it is possible to check the accuracy of an analysis or to subject the data to better analyses.

3. Incorporate data in meta-analyses. Ideally most quantitative research data would eventually be incorporated in a meta-analysis. But meta-analysis is often thwarted by incomplete reporting of data. This problem could be circumvented if meta-analysts routinely had access to raw data. Access to raw data also opens up the possibility of conducting analyses on individual patient data rather than on summary statistics. Where possible, meta-analyses on individual patient data are strongly preferred to meta-analyses on summary data (Higgins and Green 2005).

It is now a simple matter to make data available electronically. Authors can easily make large data files instantly accessible to others over the internet. There is no practical impediment to making all research data freely available to anyone who wants it.

Now that it is an easy matter to make data freely available this should become an expected practice. Only when researchers make their data freely accessible can we achieve the goal of making researchers fully accountable for what they publish. And only when data are freely accessible will it be possible to extract all of the potential information from published research.

We should recognise that there is almost always a large degree of public investment in any piece of research. Often the investment is financial; many research projects are funded by government and most researchers’ salaries are paid by government. But the investment is usually also personal because members of the public volunteer to participate as research subjects in clinical research. These public contributions to research mean that most researchers are not entitled to think of a particular data set as belonging to them alone. Instead, most research data should be seen as a public commodity that is to be made available for the common good.

Some researchers are reluctant to make their data publicly available. This may stem from a legitimate concern about confidentiality, but such concerns are usually not warranted: usually it is a simple matter to ‘de-identify’ data sets. I suspect that researchers’ coyness stems, more often, from a desire to avoid having their data subject to scrutiny. For this reason, journals should be reluctant to publish articles from researchers who are not prepared for their data to see the light of day.

Australian Journal of Physiotherapy has led the research community in its efforts to make research data publicly available. Authors of papers published in the journal are able to make their data available electronically on the journal web site. Many authors have already done so. For example, Smidt and colleagues provided 109 pages of supplementary data to their landmark review of systematic reviews of therapeutic exercise (Smidt et al 2005). The data were published as an electronic eAddendum (http://www.physiotherapy.asn.au/AJP/51-2/AustJPhysiother51i2Addendum2.pdf). These data are tremendously useful to other researchers in the field of therapeutic exercise, and were the starting point for a further review conducted by Taylor and colleagues in 2007 (Taylor et al 2007).

Researchers should make their research data freely available to anyone who wants it. This will go some way to providing much needed transparency and accountability in research.

References


