

Chances and problems of doing science online

Martin Fenner, Gobbledygook

March 15, 2010

Last month (shortly after ScienceOnline2010) David Crotty wrote in a blog post *Science and Web 2.0: Talking About Science vs. Doing Science*:

Nearly all of the more visible attempts (of science and Web 2.0) so far have focused on talking about science, rather than tools for actually doing science.

Flickr picture from Ivan Walsh.

The blog post is required reading for everybody interested in science and Web 2.0 and has attracted a lot of thoughtful comments (on the blog and on FriendFeed). In another discussion Thomas Söderquist from the Medical Museion in Copenhagen reminded me that there are limitations of what can be done online. My blog focusses on talking about science rather than doing science, but this post is about doing science online. My research focusses on clinical cancer research, and in this field the advantages and limitations of doing science online are obviously different from other subject areas (bioinformatics for example obviously looks very different). It probably makes sense to ask yourself the following questions:

- Are your research data collected in (or easily converted into) digital form?
- Are standard data formats and standard tools (preferably as open source software) available?
- Do you regularly collaborate with scientists in other locations?
- Is information about ongoing research projects publicly available?
- Do journals have policies regarding the publication of your primary research data?
- Are there objections to make the research data freely available?

Are your research data collected in (or easily converted into) digital form

Electronic medical records (EMR) have the potential to improve patient care and reduce costs. But for now, often only some clinical information (particularly lab and radiology results) is available electronically and paper-based patient records are still commonly used. And both electronic and paper-based records

have to be adapted to be useful for clinical research, e.g. by allowing a detailed documentation of adverse events.

The raw clinical data of a patient in a trial (called source data in clinical research) are entered into a case-report form (CRF). The purpose of this two-step process is to make sure that all required data are collected and that they are entered correctly. Many clinical trials now use electronic CRFs or electronic data capture (EDC). But these tools are still surprisingly difficult to use and more expensive than paper-based solutions, so that many trials stick to paper CRFs and enter the data into a computer at a later stage. It also doesn't help that the EDC market is very fragmented, so that institutions have to learn to use several different tools.

Are standard data formats and standard tools (preferably as open source software) available?

Clinical Data Interchange Standard (CDISC) is the standard format for clinical research data. OpenClinica and Clinical Trials Management System of CaBIG are two examples of Open Source tools for clinical research.

Do you regularly collaborate with scientists in other locations?

Most clinical trials are multi-center trials that are conducted in different locations, often even in different countries or continents. The coordination of the different trial locations uses email and web conferencing, but often relies more on human resources than on modern Web 2.0 tools.

Is information about ongoing research projects publicly available?

Clinical research is one of only a few research areas where information about (almost) all ongoing research project is publicly available. Clinical trial registries serve two purposes. They make it much easier for patients and their treating physicians to find relevant clinical trials. And they allow clinical researchers to understand what clinical research is going on in their field, and to avoid publication bias. ClinicalTrials.gov at the US National Institutes of Health is the largest clinical trial registry. For reasons that are difficult to understand, the European Clinical Trials database (EudraCT) is not available to the public, but work is in progress to change that.

Do journals have policies regarding the publication of your primary research data?

An article in the BMJ last month by Iain Hrynaszkiwicz and colleagues¹ tries to provide guidance on how to provide raw clinical data for publication. The main focus of the paper is patient privacy. Publication of raw clinical data either as dataset or as part of a research paper is still very uncommon. The meta-analysis

of individual patient data² requires the raw clinical data of several clinical trials, and because of the required effort is probably underused.

Are there objections to make the research data freely available?

Patient privacy is a major concern when publishing raw clinical data, and it's therefore critical to remove all identifying information from the dataset. This not only includes direct identifiers such as patient names, birthdates, unique identifying numbers or facial photographs, but also indirect identifiers such as place of treatment, rare disease or treatment, occupation or place of work, etc. It is the consensus of the authors of the BMJ paper that datasets with three or more indirect identifiers should be evaluated for the risk that individuals might be identifiable before they are made available,

In contrast to Open Notebook Science, it is impossible to make the results of a clinical trial publicly available before the trial is completed. The statistical design of the clinical trial is based on the number of patients needed to show a significant difference – looking at interim data could influence patient recruitment. Double-blind designs (where neither patient nor treating physician now which treatment arm the patient is in) are based on the same principle.

In clinical research there is often more at stake than the well-being of patients and the careers of the scientists involved. Pfizer and Roche last week each lost \$1 billion in stock market value after they both announced negative results of large phase III cancer trials. Drug companies therefore have a great interest in whether and when research findings are published, and that includes the raw clinical data. The selective publication of positive research findings is called publication bias and the mandatory reporting of clinical trial results in ClinicalTrials.gov was introduced by the FDA to reduce publication bias.

Summary

Online tools can help with doing clinical research and there is probably a lot of untapped potential. From the perspective of an individual researcher or a small research group, it probably makes the most sense to develop and/or use tools that solve specific problems. I use the online project management tool Basecamp to coordinate one clinical research project. And I have created a web-based clinical trials registry for our university hospital. The internet version of this registry helps patients and referring physicians to find clinical trials at our institution. The intranet version helps us manage our clinical trials, e.g. by keeping all required documents in one place, keeping track of the patients registered in clinical trials, and serious adverse event reporting. And I don't see why clinical researchers can't adopt the Panton Principles – which endorse that data related to published science should be explicitly placed in the public domain – whenever possible.

References

Hrynaskiewicz et al. Preparing raw clinical data for publication: guidance for journal editors, authors, and peer reviewers. *BMJ* 2010 <https://doi.org/10.1136/bmj.c181>

Simmonds et al. Meta-analysis of individual patient data from randomized trials: a review of methods used in practice. *Clin Trials* 2005 <https://doi.org/10.1191/1740774505cn087oa>