Published by Oxford University Press on behalf of the International Epidemiological Association
International Journal of Epidemiology 2007;36:677–678
doi:10.1093/ije/dym112

Commentary: Assessing the quality of observational studies—or a lesson from Mars

Erik von Elm

Accepted 25 April 2007

Keywords Observational study, quality assessment tool, checklist, quality of reporting

In 1999, the Mars Climate Orbiter (MCO), a spacecraft full of sophisticated scientific instruments, was sent to our neighbouring planet to explore its atmosphere. On entering the Martian orbit, complex calculations had to be performed by teams in different countries in order to stabilize the trajectory of this first interplanetary weather satellite. But the MCO failed to reach its planned altitude and disappeared from the screens, most likely burned in the Martian atmosphere. During later examination, it turned out that one of the teams provided data essential for course correction in English units (Pound second) while it was specified and expected in metric units (Newton second).[1] The difference by a factor of 4.45 could not be reconciled by MCO's board computer; the calculated altitude was much too low. Consequently, the unique opportunity for Martian weather forecasts (and some $125 million) was lost.

This story from Mars may seem a bit obscure, but there is an analogy with measuring study quality. Jüni *et al*. showed that the use of scales to decide whether a randomized clinical trial should be regarded as of high or low quality is problematic.[2] Depending on which scale was used, estimated treatment effects in high quality trials varied considerably. Many different quality scales exist for randomized clinical trials.[3] For both clinical trials and observational studies we would be glad if we had a single system to measure study quality. However, in their comprehensive overview, Sanderson and colleagues identified not less than 86 different checklists and scales for the assessment of observational studies.[4]

We basically do not know very well how to capture the 'amorphous concept'[4] that is study quality. The diversity of tools identified by Sanderson and colleagues illustrates the difficulty. Wisely, the authors resist the temptation to recommend one tool as the 'gold standard'. Instead, they provide useful information about them all, which may help potential users to choose the most suitable one for their own purpose. When assessing the quality of observational studies, one may also prefer to set ready-made checklists and scales aside and to describe the methodological strengths and weaknesses of each published article individually.

Institute of Social and Preventive Medicine (ISPM), University of Bern, Finkenhubelweg 11, CH-3012 Bern, Switzerland.
E-mail: vonelm@ispm.unibe.ch

When using tools for study quality we rely on what was *reported* in published articles. As with most research manuscripts, those on observational studies undergo a long and iterative process of editorial peer review.[5] However, the effectiveness of peer-review to improve the quality of research articles is not well established.[6] Sometimes, longer manuscripts may be stripped of information, in particular in their Methods sections, that later may turn out to be important to assess study quality. A reader's perspective on published articles is quite different from an author's point of view since the reader can only appreciate the final product. It is important to keep this in mind when scrutinizing articles using a checklist. If an essential piece of information is missing from an article, we cannot usually give the authors 'the benefit of the doubt' and assume that, for instance, they accounted for loss to follow-up if they did not say so in their article. Of course, it is advisable to contact the investigators and ask about unclear methodological issues in such situations. But reviewers also need to be prepared that many of them will not respond.[7]

In the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) initiative we are elaborating a set of recommendations for the *reporting* of epidemiological articles, rather than producing another quality assessment tool.[8] The STROBE statement is intended to help those who write up epidemiological research to report completely and transparently what was done and what was found. The group has been asked several times for advice by researchers who planned to assess study quality in bibliographic studies. The list of available tools compiled by Sanderson and colleagues is helpful in these situations but should be used with great caution. Thirty per cent of the reviewed tools were devised for the single purpose of assessing a defined set of literature. But even with the tools that are intended for future use in systematic reviews or empirical studies, one should bear in mind that they usually come with limitations and may not be applicable to the situation at hand. Only about half of the reviewed sources included a description of the tool's development process. Certainly, the present collection should not be seen as a 'laundry list' from which to pick a suitable template for elaboration of yet another tool.

Sanderson and colleagues analysed to what extent domains, which they deemed important for an assessment of study quality, were covered in the tools. Their data clearly show how these tools differ in the relative weight they give to these quality domains. A study like the one by Jüni and colleagues

would probably show that scoring the quality of epidemiological studies using different scales produces equally inconsistent results.[2] Nevertheless, it is reassuring that a great majority of tools uses appropriate methods in the following five areas as criteria: selecting study participants, measuring exposures and outcomes, addressing design-specific sources of bias, control of confounding and analysing data.

The co-existence of metric and English units of measure will probably remain a source of confusion in the future, hopefully with less impact generally than what happened in space. In the earthly case of biomedical research the consequences of the uncertainty around assessments of study quality may also be important, and perhaps more important than weather forecasts from Mars (or the lack thereof). In the abundance of biomedical research articles published each year, it is hard to separate the wheat from the chaff, even when quality assessment tools are employed. Unfortunately, we will probably continue to use studies of poor quality in some instances, and as a consequence, use biased results as the basis for aetiological reasoning and medical decision making.

References

[1] Mars Climate Orbiter Mishap Investigation Board. Phase I Report. November 10, 1999. Available at: http://sunnyday.mit.edu/accidents/MCO_report.pdf. (Accessed June 15, 2007).

[2] Juni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. *Jama* 1999;**282:**1054–60.

[3] Moher D, Jadad AR, Nichol G, Penman M, Tugwell P, Walsh S. Assessing the quality of randomized controlled trials: an annotated bibliography of scales and checklists. *Control Clin Trials* 1995;**16:**62–73.

[4] Sanderson S, Tatt ID, Higgins JP. Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: a systematic review and annotated bibliography. *Int J Epidemiol* 2007;**36:**670–80.

[5] Hall SA, Wilcox AJ. The fate of epidemiologic manuscripts: a study of papers submitted to epidemiology. *Epidemiology* 2007;**18:**262–65.

[6] Jefferson T, Rudin M, Brodney Folse S, Davidoff F. Editorial peer review for improving the quality of reports of biomedical studies. *Cochrane Database Syst Rev* 2007;**2:**MR000016.

[7] Gibson CA, Bailey BW, Carper MJ *et al*. Author contacts for retrieval of data for a meta-analysis on exercise and diet restriction. *Int J Technol Assess Health Care* 2006;**22:**267–70.

[8] STROBE Initiative. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE). Available at: www.strobe-statement.org (Accessed April 23, 2007).