Reviewer's report

Title: Harmonisation of datasets structures prior to conducting statistical analyses with multiple datasets: an automated approach.

Version: 1 Date: 16 February 2011

Reviewer: Marc Chadeau-Hyam

Reviewer's report:

In this paper, the author presents an algorithm to automatically harmonize data from a large set of studies. The rationale together with the usefulness of such a work is clearly and convincingly described in the paper. The step-by-step decomposition of the algorithm to achieve that task is very clearly presented, and surely represents a considerable amount of work. Reading this article has convinced me that such a tool could make a significant contribution to Epidemiological scientific community. However in its current form, and due to the points listed below, it does not provide a satisfactory answer to the interesting issues it raises, which is even more regrettable since most of the underlying conceptual work seems to have been done.

First of all, it is difficult to have an overview of the whole algorithm, some apparently important steps are only mentioned and not detailed. The whole focus of the paper is on the search engine, while these side programs are of critical importance (e.g. p. 5, two ‘programs’ are being alluded to and are neither detailed, nor is their source code provided as part of the SI). For instance it is difficult to understand how data from the different datasets are merged and used by the STATA program. This might be improved by:
- Providing all the source code files required to run the program, including pre-processing routines
- Giving example input and output files, the user could play with and follow the step-by-step run of the program. This would also help to understand the structure of the input files required together with the types of results generated at each step.
- Providing a extensive documentation, describing each of these routines, their usage and the way they are used in the context of this paper
- Providing an overall representation of the algorithm, and detail for each part the code involved (for instance as was done for the 3 searches in the current manuscript).

The choice to write the program in STATA is highly questionable. Indeed, in order to increase the number of potential users and subsequently define a community-based library of input files (e.g. a library of thesaurus files), this work would benefit from being written using a free/open-source software (e.g. R) and maybe precompiled as a package.
The use of iterative runs is only presented in the results section, while they appear as being an important part of the algorithm itself. It would therefore be better to add to the methods section a first sub-section providing an overview of the algorithm (including its iterative structure), and then detail each of the (i) pre-processing steps currently not described, (ii) each the searches involved and (iii) the outcome combination, as done in the current manuscript.

The description of how variables re-coding is performed remains quite cryptic and should be detailed, as it might be critical for subsequent statistical analyses. For example in case of continuous variables, will there be a unit consistency check?

Although the description of searches A, B, and C provides an insight into how the program works, some essential questions remain unanswered. For instance, it seems that the way to make the most of all available datasets and therefore not to loose valuable information from some datasets with specific variables labels strongly relies on the definition of the thesaurus file. It is difficult to assess how much work would be required to define an appropriate/general thesaurus file for several variables of interest? The question would then be how much effort would it require compared to manual selection of the candidate lists of variables. This should be quantified and clearly stated in the text. Is is also unclear if it is possible to miss a better candidate variable due to only variable labeling inconsistencies between datasets.

Examples presented in the paper do not provide a striking evidence of how general (and therefore useful) this program is. Testing it on a several sets of variables would definitely improve this point.

Finally there are many formatting issues throughout the manuscript: some of the author’s comments are still in the text, the references are not included in the main text, the tables/figures numbering is either absent or does not correspond the order of appearance in the main text.

Dr Marc Chadeau-Hyam.

**Level of interest:** An article of insufficient interest to warrant publication in a scientific/medical journal

**Quality of written English:** Acceptable

**Statistical review:** No, the manuscript does not need to be seen by a statistician.

**Declaration of competing interests:**

I declare that I have no competing interests