Reviewer’s report

Title: Integrating Historical Clinical and Financial Data for Pharmacological Research

Version: 2 Date: 26 July 2011

Reviewer: James J Cimino

Reviewer's report:

1. Is the question posed by the authors well defined?
   Yes.

2. Are the methods appropriate and well described?
   Yes.

3. Are the data sound?
   Yes.

4. Does the manuscript adhere to the relevant standards for reporting and data deposition?
   Yes.

5. Are the discussion and conclusions well balanced and adequately supported by the data?
   Yes.

6. Are limitations of the work clearly stated?
   Yes.

7. Do the authors clearly acknowledge any work upon which they are building, both published and unpublished?
   Yes.

8. Do the title and abstract accurately convey what has been found?
   Yes.

9. Is the writing acceptable?
   Yes.

- Major Compulsory Revisions (which the author must respond to before a decision on publication can be reached)
  None.

- Minor Essential Revisions (such as missing labels on figures, or the wrong use of a term, which the author can be trusted to correct)
- Discretionary Revisions (which are recommendations for improvement but which the author can choose to ignore)

1) The term "semantic enrichment" bothers me a bit. It seems a bit grandiose for what is essentially mapping one set of codes to another. I guess, in a sense, the records as a whole are being enriched, but the codes themselves are simply being mapped. Perhaps a clearer definition could be added to make clear what "wealth" is being added, exactly.

2) While I understand the translation process in general, I found it hard to follow all the systems, codes and transformations. I think that the description of the example on Page 9 could be enhanced with the addition of a figure that shows, graphically, an entry in, say, the PM with its attendant codes and names, and links to entries in other systems that contain the mapping from the codes/names to RxNorm, with its codes and names. Then a reader could visually see how the links were being established.

3) I am not clear on how name matches were accomplished. Was exact match required?

4) While the exercise carried out by the authors seems like it paid off, and might encourage readers to do this with their own sources, I don't think the methodology will be useful to them unless either they are using the same databases or they can figure out what characteristics of their own databases map to the characteristics of the authors' databases and then figure out which method is uses for each particular set of available characteristics. Table 1 almost does this, but not quite. I am thinking about something that characterizes the specific data types into more generic types. For example, local system-specific codes, foreign system-specific codes (that is, those that are local in some other system), standard codes, local names, local abbreviations, standard names, etc. For example, maybe the Allegra system has local codes only, and no names. And maybe some other system has the foreign (Allegra) codes, its own codes, and NDC codes or names. And then some third system has NDC codes and RxNorm codes. Then the translation process can be seen as mapping from a local-code-only system, through some bridging system with certain kinds of data available, to the target terminology. This would make the authors' methods more easily reproduced by the readers of the paper.

**Level of interest:** An article whose findings are important to those with closely related research interests

**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.