Reviewer’s report

Title: Assessing Accuracy of Record Linkage between Administrative Data and Vital Statistics

Version: 1 Date: 26 August 2005

Reviewer: robert D pates

Reviewer’s report:

General
1) As this paper is written and presented, I must say I have difficulties identifying the value of this work to the health evaluation community.
2) The problem of achieving a rapid, cost-effective merge of acceptable quality has been dealt with extensively in the past by Roos and Wajda (as the authors point out) but the work presented here does not add very much to that debate.
3) Perhaps the authors would be best advised to switch the focus of this work entirely. For example, given the best deterministic linkage combination of identifiers – why not validate the results of using such a linkage by examining the bias introduced into the resultant datasets? If I understand this work correctly, the authors address this issue in Tables 1-4. It might be useful to other investigators to be presented with some detailed notions of how they would be impacting their data by following the example of Li et al using a “quick and dirty” deterministic linkage methodology.
4) Even given 3) above, it would be necessary – in my opinion – for the authors to characterize in much greater detail the data they have chosen for this study. I suspect that the hospital data they have used is of significantly higher quality than routine discharge abstract type data obtained in the US. It would be helpful, too, if they could somehow relate their results to a well-defined geographic region such as an entire Canadian Province (rather than an ill-defined local region). This would give the reader a decent assessment of whether or not they would be able to use the methods outlined in the paper and still get acceptable results.

-------------------------------------------------------------------------------

Major Compulsory Revisions (that the author must respond to before a decision on publication can be reached)
5) See 3) and 4) above.
6) Although the best-case linkage and accuracy rates the authors report are quite good (most studies require at least 95% of the rows to be accurately linked) -- I remain unconvinced that the authors have hit upon quite the best combination of identifiers deterministic linkage. Did the authors try surname, first initial, sex and birthdate – and what about middle initials? In my experience, the middle initials have been helpful in breaking ties, and I suspect that for people using data of lower quality than the authors this might be helpful.

-------------------------------------------------------------------------------

Minor Essential Revisions (such as missing labels on figures, or the wrong use of a term, which the author can be trusted to correct)
7) Background -- I found it distracting that the databases were not named and defined clearly early in the paper and referred to succinctly by acronym thereafter.
8) Methods – Many questions occur to the uninitiated reader – What is the total population of the Calgary Health Region? (is it 1.1 million?) Which organizations assemble and distribute the population / vital statistics / hospital databases? Under what conditions did the authors get access to the data? (i.e. what confidentiality arrangements were made – if any? Did they have to pay cash for the data?). A brief mention of the way in which death data are exchanged between the Canadian
Provinces might be helpful, too (in the US I believe approximately 5% people die out of state – this may be significant in some studies).

9) Regarding the hospital database – exactly how many hospitals contribute to this database? What percentage of Alberta does that represent? As a reader I was constantly trying things in perspective – many more details on the hospital data would be important.

10) What software / hardware were employed by the authors to carry this work out? A brief description would be helpful. Would the authors be prepared to share their code with others to help them get stated in similar linkage efforts?

11) Just how “unique” is the PHN? It’s fair enough that this is a gold standard, of course – but some idea of the known limitations of this identifier might be helpful. (I would imagine that it is excellent owing to the nature of the Canadian health system – superior to the SSN used in the US – but in the US family members may share a single SSN.)

12) In many places the language could do with some clarification/ work (e.g. “Both of studies didn’t take surname and first name … “ should be improved to something along the lines of “Neither of these studies employed surname and first name …”

Discretionary Revisions (which the author can choose to ignore)

13) The 2 figures might be contracted into one.

14) The 4 tables are a bit excessive, and repetitive too. From the point of view of the reader, data like these might be better displayed graphically.

15) Instead of linkage/accuracy rates the authors might consider calculating some more statistically generic measure to present their results – like positive predictive value, for example.

What next?: Reject because too small an advance to publish

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

Quality of written English: Needs some language corrections before being published

Statistical review: No

Declaration of competing interests:

I declare that I have no competing interests.