Reviewer’s report

Title: Event linkage of health and aged care services data: comparing event-based and name-based methods

Version: 2 Date: 4 February 2008

Reviewer: Michael J Goldacre

Reviewer’s report:

Major Compulsory Revisions

1. This is a complicated paper that needs a lot of determination to read. Users of linked data as distinct from technical linkers generally only want to know the key principles behind the linkage of a dataset, and the results of assessments of likely error rates. This paper, in my view, is unlikely to appeal to anyone other than those involved in the technicalities of linkage using these Australian datasets. I do not think that the findings are necessarily generalisable (see below). A judgement for the editors of BMC Health Services Research is whether this journal is the right vehicle for the paper. Can the authors give it appeal both in terms of attractiveness of presentation and in terms of generalisability of findings?

Minor essential revisions

2. The utility of postcode/region in aiding reliable matching depends on the size of the population covered by the postcode (as the authors say). Put simply, the likelihood that two records with the same date of birth and postcode relate to the same person depends on how many people share the same postcode. Postcodes in England typically cover about 14 households, and the likelihood that two (or more) people share the same date of birth and postcode is small (leaving aside twins). The authors describe postcode areas with hundreds or thousands of people. As they imply, it is not surprising that there are fairly high false positive linkage rates. This is one reason why I doubt that the findings are of generalisable value to investigators in other populations.

Nonetheless, the authors say that narrowing the geographical matching criteria would result in dropping true links. This is not what one would expect with reliably recorded geographical data. It raises questions about data quality, inconsistency in recording addresses, or movement between addresses. If people change address between one event and the next, of course, this may render the small-area geographical identifier almost useless. (This, too, depends a bit on local geography. If, say, a patient is admitted to hospital from a home in a small remote settlement and is discharged to stay with a relative nearby, the location may be similar. In a large city, it may be very different. Hard to generalise). The authors need to be convincing on these points.

3. The tables are very hard to follow, in terms of key messages. Table 3 is
probably the most important, in that it shows some headline findings on quality. The importance of false positive links and missed links depends to some extent of the importance, in real terms, of the outcome. If the authors were considering linkage to death (rather than to another type of care), I do not think that it would usually be acceptable to have a false positive linkage rate of 2.3% (ie 2.3% of those thought to be dead weren’t really dead) or a missed match rate of 8.5% (ie 8.5% of those who died are thought to still be alive; Table 3, line 2)). But much depends on context.

4. It could be argued that this level of error in matching does not matter if the outcomes under study are the same in those record pairs that accurately match and those that do not. However, I am unconvinced by the examples. The authors show that the median length of stay in the N linkages and in the E linkages are similar (text and Figure 3). One could, I think, get similar results by taking the whole hospital dataset and splitting it into two samples by using random sampling numbers. All this means is that the reliability of linkage does not vary according to length of stay in the first episode of care. Do the authors think that this an important finding? Perhaps it is, but the readers need to be persuaded.

5. A few small points:

a. Abstract, Results, second sentence: What does the sentence starting However, mean? Is the less than 10% part of the over 95%, does it overlap with it, or are the points made in the two sentences fundamentally different?

b. Conclusions, line 3: the many can be used to investigate many aspects seems a bit sweeping.

c. Methods, first line: I think I know what the authors mean by unidirectional but an explanation would help.

d. Page 11 and elsewhere: I think I know what the authors mean by match passes but an explanation would help.

e. Page 11, para 3, first line: confident

f. Page 12, quality of the N linked dataset: checks have indicated a very low error rate what is it and how does it compare with the findings in this study?

g. Discussion: the reader thinks that the results from the previous sections will be discussed, but in fact gets new results (Figures 2, 3, Tables 6,7)

6. Positive comments to, and about, colleagues:

I am sure that the details of methods, and findings, will be very useful to investigators who use these Australian datasets (and therefore to Australian users of these linked data. I have gone through the paper very carefully and confirm that it seems technically sound.
Michael Goldacre

What next?: Unable to decide on acceptance or rejection until the authors have responded to the major compulsory revisions

Level of interest: An article of limited interest

Quality of written English: Acceptable

Statistical review: Yes, and I have assessed the statistics in my report.

Declaration of competing interests:

I declare that I have no competing interests