Reviewer's report

Title: Event based record linkage in health and aged care services data: a methodological innovation

Version: 1 Date: 17 February 2007

Reviewer: Tim Churches

Reviewer's report:

General

Major Compulsory Revisions (that the author must respond to before a decision on publication can be reached)

The type of matching which the authors are considering is generally referred to in the literature (perhaps incorrectly, because the data are often eminently re-identifiable, as they are in this paper) as "anonymous record linkage". The Background section needs to mention and briefly consider at least some of the studies which have examined this type of linkage in the past - a PubMed search on "anonymous record linkage" revealed several relevant papers in the first few pages of hits.

The authors also need to briefly discuss the ethical and privacy considerations and frameworks in which such record linkage might take place. Individual consent to link these records has not been given, nor is such linkage part of the primary purpose for which the data are collected. Thus, in the privacy regimes operating in many jurisdictions, a properly constituted ethics committee or institutional ethics review board needs to give its assent that such linkage is in the public interest before it can be allowed to take place. The authors need to mention these issues, not just in the context of their particular study, but also for linkage studies of this type in general.

However, the greatest weakness of the paper as it stands is its purely theoretical basis. As the authors note in the background section, false links may be due to chance agreement between the data items used to match records, or "...errors or inconsistencies in the matching data may cause false links to be made...". They then explain that their paper focuses purely on the former. That is a forgiveable limitation given the difficulty and resource-intensiveness of estimating data errors (which requires going back to source records in hospitals and nursing homes, and/or to the patients or their carers themselves, at least for a sample of records). However, the implications of such a purely theoretical approach must be clearly acknowledged. As it stands, the authors fail to mention the likely size of the effect of data errors on the false linkage rate anywhere else in the paper. This is a major omission, because it is wrong to assume that chance agreement is the only source of false links in the scenario in question, or even that it is the major cause of such false links. A study of Canadian hospital discharge abstracts, which are similar to one of the data sources considered in this paper, estimated that date of birth and date of admission were wrongly abstracted from the medical record in about 0.1% of records (see http://secure.cihi.ca/cihiweb/dispPage.jsp?cw_page=GR_1107_E ) - and this estimate does not include any estimate of the proportion of dates of birth which are in error in the source medical record, merely errors in transcribing or transferring that information to the hospital discharge abstract record. Furthermore, it is likely that errors in one data item are not independent of errors in other data items: some patients (or their carers) are simply much less reliable providers of demographic information than others. Thus it is essential that the authors revise their discussion and conclusion to be far more circumspect about their false match estimates, and to make it very clear that there is another source of false links (that is, data error as opposed to chance agreement) which they have not considered.

The paper would be greatly strengthened by the testing of the theoretical false match rates against empirically derived false match rates using a gold-standard linkage mechanism, or some other external source of verification of the matches - that is, by a validation study, such as those reported in [1] or [2]. Again, it is understandable that the authors are unable to carry out such a study due to limited time or resources, but the desirability of such empirical validation should be mentioned.

Finally, at the very end of the paper the authors acknowledge the trade-off between false match rates and missed match rates i.e. between Type I and Type II linkage errors. This point needs to be made much earlier in the paper, perhaps in the Background section. Indeed, it makes little sense to consider false
match rates in isolation, because false match rates can always be lowered simply by tightening the matching criteria, but this comes at the expense of an increased missed match rate. It seems very incomplete to consider one (false match rates) without simultaneously considering the other (missed match rates), in the same way that the sensitivity of a diagnostic test cannot be considered in isolation of its specificity, and vice versa. It would perhaps be better if the authors combined this paper with the subsequent paper on missed match rates which they mention in their conclusion, even if the resulting paper were rather long (but note that BioMed central has no space limitations).


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 the author can choose to ignore)

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

Declaration of competing interests:
I declare that I have no competing interests.