Author's response to reviews

Title: Validation of de-identified record linkage to ascertain hospital admissions in a cohort study

Authors:

Alison Beauchamp (Alison.Beauchamp@monash.edu)
Andrew M Tonkin (Andrew.Tonkin@monash.edu)
Helen Kelsall (Helen.Kelsall@monash.edu)
Vijaya Sundararajan (Vijaya.Sundararajan@health.vic.gov.au)
Dallas R English (Dallas.English@cancervic.org.au)
Lalitha Sundaresan (Lalitha.Sundaresan@health.vic.gov.au)
Rory Wolfe (Rory.Wolfe@monash.edu)
Gavin Turrell (g.turrell@qut.edu.au)
Graham G Giles (Graham.Giles@cancervic.org.au)
Anna Peeters (Anna.Peeters@monash.edu)

Version: 2 Date: 7 February 2011

Author's response to reviews: see over
Re: Response to reviewer’s comments for MS: 4926256324749679 - Validation of de-identified record linkage to ascertain hospital admissions in a cohort study.

Dear Editors

Please find below our response to the reviewer’s comments for the above article submitted to BMC Medical Research Methodology.

We believe that we have adequately addressed the reviewer’s concerns in our response, and thank the editors for the opportunity to do so. We are most willing to make further amendments should this be required.

Thank you for your consideration of this re-submitted article.

Yours sincerely

Alison Beauchamp

Reviewer 1 comments

This is a very important issue but from a policy perspective as well as data management issue (cohort databases from differing sources), I’m not convinced that the authors defined the issue sufficiently nor offered up advice for policy makers and regulators on how to get better data. They more or less assumed that you’ll have ‘so-so’ data and went from there. OK but surely there is some interest in making the data more valid from the ‘get-go’.

Response: We agree that these are important issues and provide our response later in the referee’s report.

From strictly a cohort study question, they may have left out one methodological consideration. It appears that they worked until they found a set of records that were valid using a triangulation of approaches including examination of actual medical records and so
forth. So far, so good. However, as I read the manuscript, they then checked their algorithms against the entire common cohort database that had resulted from this data validation exercise. I would have been happier if they had used 60% or even 75% of the common cohort database and then tested their final “best” algorithm against the remaining 25-40% to see how it performed. I don’t consider myself a basically a methodologist so other reviewers should weigh in on this methods issue.

Response: The algorithms were developed using the entire pilot sample of 2000, with sensitivity tested using only a small part of this sample. To clarify this, the text has been amended slightly to read: “The pilot sample of 2000 used in this study included 101 participants with a confirmed hospital admission…..” (page 5, paragraph 3).

We consider that having two separate ‘development’ and ‘validation’ datasets within this pilot sample would not be the most appropriate methodological approach for achieving the aims of this study in developing and assessing the accuracy of the linkage algorithm. Ultimately, the final algorithm will be applied to linkage of the entire MCCS with the VAED to obtain data on CVD events occurring in the cohort during 19 years of follow-up (as described in current manuscript page 4, Background, final paragraph, first sentence), and we have made a minor wording change in this paragraph to clarify further for the reader what this study relates to.

Policy Dimension.

From the perspective of policy, at least in the USA where we are trying to demonstrate the value of having a Unique Personal Health Identifier (UPHI), their report is only helpful in part. Obviously, part of the utility, or decay of value, of a unique identifier apparently relates to the nature of how Australia assigns unique identifiers, e.g., unique to families and then members within families which can be confusing and, apparently, a source for slippage in validity as well as how well the data are entered in the first place. These issues being the case, I was a bit puzzled why they ‘declared victory’ with their algorithm rather than also suggesting in the discussion how Australians might improve their system of identifiers both from the perspective of precision as well as administration.

Response: We appreciate the referee’s comments and in the revised manuscript have expanded our discussion relating to implications of our findings for improvement of identifiers and relevant policies and have amended the text and included new references (27-29 and 30) to read:

"Limited identifying variables and advantages of Unique Personal Health Identifiers

This study addresses an important aspect of record linkage, that is how to link without using names and addresses or unique health identifiers. Our findings are relevant for custodians of existing research or administrative datasets who seek to increase their value through record linkage but do not have access to such identifying information. In the future, it is likely that
record linkage with hospital and health administration datasets will be much enhanced by use of a Unique Personal Health Identifier (UPHI). These electronic numbers will be used to uniquely identify healthcare providers and individuals, and aim to both improve communication between health care providers and support the delivery of health services thus enhancing the quality of patient care. UPHIs have also been advocated as a way of accurately linking records in a privacy preserving way. While UPHIs are not currently used in Australia, legislation was recently passed authorising the issue of individual identifier numbers. However, it will be some time before uptake of the UPHI is widespread enough to allow for its use in linkage of health records, and for some existing datasets, this will never be possible. As such, for those datasets with limited identifying variables, our study demonstrates that record linkage to other datasets is achievable, further increasing their importance as a valuable source of health-related data.” (page 11, paragraph 2)

Another question of importance relates to the impact of anonymization and what methods are used to anonymize the data. This may have been clear to other reviewers but it was not sufficiently clear for me.

Response: We have amended the text to read: “The above identifiers were joined to create a single linkage variable which was then encrypted.” (page 6, paragraph 3)

Having made the comment relating to assessing the value of having a unique personal health identifier for improved cohort analyses, I’m not positive how I could ‘cure’ this issue for this study or for any other future study for that matter. It appears that their analysis finds that one gets a ‘false sense’ that one has not lost any value through anonymization by not having a totally unique personal health identifier but that this represents ‘false security’ due to a high ‘false positive’ rate. Writing as just a single reviewer, what I’m most interested in knowing is how much ‘loss of signal’ results from not having 1) a totally UPHI and 2) accurate recording. I gather that this is asking the wrong question at least from the researcher’s perspective.

Where does this leave me? Happily, I’m not the editor. There may be sufficient value to justify publication of the paper based solely upon what they accomplished. Even if this is the case, I’d be much happier to see discussion of the implications for policy and administration and not just how researchers address such slippery slopes encountered in their work.

Response: In addition to the paragraph on UPHI above, we have also included discussion of the implications of missing data and data entry error in the relevant sections of the discussion to include

“While use of a UPHI or other unique identifier in combination with demographic details is likely to provide the most accurate linkage for health-related datasets, the datasets may be subject to data entry errors. We found for example, that errors for day of birth ranged from between 2 to 10 days. We were unable to quantify the impact of data entry error in our study.
However, others have found that such errors can be significant. It may be possible for researchers using administrative or data sources to reduce the impact of data entry error by varying the combinations of variables used, as shown in our final algorithm. In addition, errors could be minimised by performing regular audits of data quality.” (Paragraph 1, page 12)

Both when UPHIs, names and addresses or when only demographic details are used as linkage variables, the degree of missing data must first be quantified and its impact reduced by creating linkage algorithms that allow for missing data. Our final linkage algorithm allows for missing health insurance numbers in either dataset by including abbreviated first name in addition to demographic variables. However, for current or ongoing data collections, researchers and data custodians will benefit from ensuring practices and protocols are in place to minimise the risk of missing data. (Paragraph 2, page 12)

Reviewer 2

We thank the reviewer for his positive comments regarding the manuscript and acknowledge that no revisions were recommended.