Author's response to reviews

Title: Assessing Record Linkage between Health Care and Vital Statistics Databases Using Deterministic Methods

Authors:

Bing BL Li (LIB@UCALGARY.CA)
Hude HQ Quan (hquan@ucalgary.ca)
Andrew AF Fong (andrew.fong@calgaryhealthregion.ca)
Mingshan ML Lu (lu@ucalgary.ca)

Version: 3 Date: 25 January 2006

Author's response to reviews: see over
January 24, 2006
To The BioMed Central Editorial Team

Dear Editors;

RE: Manuscript #1883117543723368 - Assessing Record Linkage between Health Care and Vital Statistics Databases Using Deterministic Methods

Thank you for your ongoing interest in our paper. In response to reviewers’ comments, we have again revised the paper. We list reviewer’s comments below in bold followed by our responses. As before, additions to the manuscript itself are bolded in one file.

Reviewer 2

1. I don't like the hyphenated term "correct-linkage" which the authors have used throughout the work and I would either delete the term or alter it appropriately. For example in the first paragraph of "Background", perhaps the second question would read better as "What is the accuracy of the record linkages obtained?"

This is a controversial comment with the comment from a reviewer in the previous letter from your journal. In the previous letter, a reviewer did not like the term we used in the first version of manuscript (i.e. accurate rate) and suggested several alternatives. Following the recommendation, we used term of ‘correct linkage’ replacing ‘accurate linkage’ in the first revision as suggested by the previous comment. In the previous response letter, we stated: “The term ‘accuracy’ quantifies the preciseness of the linkage. Considering that we are reporting the proportion of records that were correctly linked, we employed the term of ‘linkage’ and ‘correct-linkage’ in this revision.” Therefore, in this revision, we did not change the term of ‘correct linkage’. However, we are willing to change it to ‘accurate linkage’ if the editor feels it reads better than ‘correct linkage’.

2. The background section is now much better and explains why the work is being done and what's new about it. However, it remains a little disjointed and I think it lacks a clear statement of methodological strategy. For example the authors compare 3 deterministic linkage approaches and use an unique identifier (Personal Health Number) to validate the
linkages. I think that this needs to be added to this introductory section.

We enlarged the introduction through specifying three linkage approaches on page 4 to 5, reading as:

“In our study, we selected four common identifiers (surname, first name, sex, and date of birth) from a set of personal identifiers and composed three different combinations of four identifiers: (1) surname (i.e. surname at birth or marital surname as recorded in these three databases), sex and date of birth (i.e. month, day, and year of birth); (2) first name, sex, and date of birth, and (3) surname, first name, sex and date of birth. Then we assessed the deterministic linkage between the Vital Statistics registry and the population registry in one scenario, and between the Vital Statistics registry and the in-hospital death records in hospital discharge data in the second scenario for the fiscal years 1998/99 through 2001/02 for each combination. We assessed these three databases because they are widely used in population and health services research to determine death status, cause of death and medical history.”

3. Some of the wording needs some work and I would make the following suggestions: Page 5: ... 1358 were also excluded because the individuals were not CHR residents, ... I would delete or clarify the sentence: “No postal code is missing and 6762 were used.” Apart from my no understanding this, it contains a mixture of tenses, for which the authors might want to check the entire manuscript for the next draft.

The manuscript was grammatically edited again to keep consistent tenses. We followed the suggestion and deleted the following sentence.
“No postal code is missing and 6762 were used.”

Page 6: I would change ... inpatient separations (by discharge or death) ... to inpatient discharges (dead or alive) ...

On the page 6, we revised the sentence as the following:
“Abstracts are filed for all inpatient discharges from all hospitals in the CHR”

Page 7 (top): I don't understand the following ... The extracted data are stored in the region as well as ....

We deleted the sentence “The extracted data are stored in the region as well as submitted to the Alberta provincial discharge database”.

Page 8: Change ... Although this number is complete ... to ... Although this identifier is complete ...

We revised the sentence on the page 7 as:
“Although this identifier is complete in both of the hospital discharge data and population registry, ”

4. The sections added to the discussion are helpful. On page 12, though, the paragraph on
those VS records without PHN is a little tentative. Is there no way of taking those data and characterizing them more definitely?

We revised the paragraph and made the following revisions on page 12:

“Vital Statistics records without PHN are likely to be persons who died out of hospital. Personal information for those deaths is from various sources, resulting in inconsistencies in personal information between Vital Statistics and the Alberta population registry. For persons who are not eligible for Alberta Health Insurance plan (such as inmates, travelers, visitors, expatriates, armed forces, and Royal Canadian Mounted Police), their PHNs would not be recorded in the Vital Statistics if they died in Alberta. However, such cases account for a small proportion of all deaths in Alberta.”

5. The figures are still a little troublesome -- are they really necessary? They take up a lot of space and are repetitive. I would think about sharpening up the methods to describe the processes in detail, and delete them.

We appreciate for reviewer’s suggestion. However, we think the figures would help readers to understand the data linkage flow better although it takes space. Description of the process is complicated and very wordy. Therefore, we prefer to keep these figures in the paper.

6. The same comments could be made about the tables. There are a lot of numbers to look at / interpret (!). Have the authors tried reducing the number of age categories? Could the same points be made by reducing the number of categories to four: (1-9) / (10-19) / (20-64) / (65+)

The reason that we presented several age groups is that we attempted to provide detail information to readers since their study populations may vary by age.

7. I would say that the title is inadequate and needs some work. I would suggest something along the lines of: "Assessing Record Linkage Between Health Care and Vital Statistics Databases Using Deterministic Methods".

We followed the reviewer’s suggestion and changed the manuscript title.

Reviewer 3

1. It is disappointing that the authors have ignored my suggestion (which was provided under the heading "Major Cumpulsory Revisions") to adopt, or at least refer to, terminology which is more commonly used in the record linkage literature. That notwithstanding, it is still essential that the authors a) mention the concept of missed links (or Type 1 errors) and b) the fact that the rate of missed links is not addressed in the study and c) the implications of this. Epidemiologists and other data analysts will typically want to know (or have an estimate of) how many links have been missed by a given record linkage process, as well as what proportion of links produced by a record linkage process are
actually correct. The study reported in the manuscript only covers the latter aspect. Acknowledgement of the importance of the former aspect - of missed links - is necessary. Furthermore, it is unclear why the authors have not attempted to estimate the proportion of missed links. Such an estimation would appear to be feasible, at least for the (biased) subset of records in the Vital Statistics Registry which contain a PHN.

Thanks for pointing out this matter again. We did not ignore previous comments and this research team discussed the concept extensively. The reviewer suggested us to select Vital Statistics records with valid PHNs and link the data with population registry using PHN as identifier. In those subsets, three deterministic approaches will be applied to re-linking records. Finally a Cartesian product of records will be calculated. The Cartesian approach is to address linkage from four aspects. However, we did not apply this approach in our databases. The major reason is that all Vital Statistics Registry records are expected to appear in the Population Registry, and all in-hospital death records are expected to be present in the Vital Statistics Registry. Conceptually, the linkage rate should be 100% while personal information is perfectly accurate. Because of selection of records with PHN for the analysis, we are more likely to over-estimate the linkage with PHN only. This is applicable to our reported results (selection bias) as we acknowledged in the limitation. Although we did not report the Cartesian products in our paper, we acknowledge this in our limitation as the reviewer recommended. We hope to address the Cartesian products in the future using a better database.

The limitation is stated on page 14 to 15 as the following:

“Fourthly, our linkage rate may be applicable to linking nested databases; one database contains all records of other database. In our study, all Vital Statistics Registry records are expected to appear in the Population Registry, and all in-hospital death records are expected to be present in the Vital Statistics Registry. Correct linkage could be assessed by four measures: true-link, false-link, true-nonlink and false-nonlink. Assessment of these four types requires a unique identifier present in both databases to establish the “gold standard”. Our study addressed one question: what is correct linkage rate among links through deterministic record linkage?”

2 and 3) With the most recent revisions, it is now much clearer what quantities are being calculated by the authors. However, what is being calculated is the proportion of "correct-links", as defined by a matching PHN, within the subset of Vital Statistics Registry records which have a PHN - which is only about 70% of all Vital StAtistics Registry records. It seems likely that this is not a random subset of records, but may incorporate potential biases which may have improved the correct linkage rate in the subset of records which have a PHN. For example, it is not unreasonable to assume that if the PHN is missing from a Vital Statistics Registry record, then the accuracy or completeness of other data items on that record may also be sub-standard. The authors need to discuss this potential bias and acknowledge that it may have caused an overestimate of the correct-linkage rate produced by this study.

This matter has been addressed in previous revision. Again, we agreed with reviewer’s opinion and made revision on page 14:
“Secondly, records without PHN may have less complete and accurate information on common identifiers than those with PHN. Therefore the higher the rate of missing PHNs is, the more likely the linkage rate is to be lower. We selected linked records with PHNs only to assess correct linkage rate. The potential selection bias may cause the correct linkage rate to be overestimated, particularly for children aged 1 to 9 since they have more missing PHNs than those aged 10 or older in Vital Statistics data.”

Thank you again for giving us the opportunity to revise the paper again. We greatly appreciate the care that your reviewers took in reviewing the paper, as they have clearly improved our final product through their comments. We hope that you will be satisfied with our responses and revisions.

We look forward to hearing from you with a decision on our paper.

Yours sincerely,

Bing Li and Hude Quan  
(on behalf of all co-authors)