Author's response to reviews

Title: Assessing Correct Record Linkage between Administrative Data and Vital Statistics

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: 2 Date: 8 November 2005

Author's response to reviews: see over
November 8, 2005

To The BioMed Central Editorial Team

Dear Editors;

RE: Manuscript # 1883117543723368 - Assessing Accuracy of Record Linkage between Administrative Data and Vital Statistics

Thank you for giving us the opportunity to resubmit the above manuscript. We greatly appreciate the comments provided by the reviewers and have revised the manuscript in response to the comments that you provided us. Below, we provide an itemized summary of the changes made to the paper. Reviewer comments are shown in bold, followed by our responses. Additions to the manuscript itself are bolded.

Reviewer 1

1. To clarify the meaning of “the combination of these two approaches” : does this mean that the deterministic and the probabilistic linkage approaches were successively applied on the same database or that pairs selected by the deterministic approach were excluded before application of the probabilistic linkage? What are the consequences of this potential exclusion on the estimation of the parameters (weights for each variable) of the probabilistic linkage? To give a precise definition of the variables used for linkage (to be able to compare them to those proposed in the paper), give more information on the discrepancies between the two studies conducted by Roos and Wadja (is it the same methodology, does the only difference concern the data collection periods?).

We have expanded the description of the previous two Roos et al. papers clarifying methods and identifiers used for the linkage. We added the following passage on Page 4.

“Using both deterministic and probabilistic linkage approaches, Roos, Wajda and Nicol [8] assessed the linkage rate between Manitoba Health Services Commission and Canadian Vital Statistics data to verify Manitoba deaths aged 25 or older during 1970 and 1979. The deterministic linkage approach was first used to link the records in the two databases using eight personal identifiers (i.e. sex, death year, death month, death day, birth year, birth month, initial and location). Then, the probabilistic record linkage approach was used to further match the remaining unlinked records after deterministic linkage in these two databases. The linkage rate increased
from 80.7% in 1970 to 95.8% in 1979. Roos and Wajda [9] repeated the linkage assessment but used specific weights model methodology. The model was based on weights for eight personal identifiers (i.e., sex, death month, death day, birth year, birth month, location, initial and marital status). A numerical weight for each variable was calculated by logarithm of agreement or disagreement frequency in linked pairs divided by that in unlinked pairs. They found that with the improvement in data quality, the record linkage rate improved significantly. The deterministic linkage approach matched 98.6% of Vital Statistics records with 1987 Manitoba Health Services Commission data in contrast to 93.7% for 1973 data. Neither study took surname as an identifying variable because researchers did not have access to individual or family names.”

2. The definition of the study population should be clarified, in particular in the case of the differences in the residential postal codes recorded in the three databases: are individuals included in the study population on the basis of the postal code in only one database corresponding to the Calgary Health Region?

We used Standard Geographical Classification in Vital Statistics and postal codes in the hospital discharge data to define Calgary Health Region (CHR) residents. Because there were two linkages, we defined two study populations. This is clarified on page 5.

“In Vital Statistics individuals were defined as CHR residents based on Standard Geographical Classification (SGC) for assessment of the linkage between Vital Statistics and the population registry. The SGC is used to classify residential areas based address or community [14]. Of 21679 deaths occurred in the CHR during 1998/99 to 2001/2002, 99 were excluded due to unknown SGC, 1358 were also excluded because of non-CHR residents, and 20222 were finally analyzed. Residential postal codes recorded in the hospital discharge data were used to define CHR residents for assessment of the linkage between Vital Statistics and hospital discharge data. No postal code is missing and 6762 were used.”

4. The authors state that the registry is complete: does this mean that residents of the first nations are included in this registry?

The Alberta provincial health insurance plan covers permanent residents of the province. However, Registered First Nations, prison inmates, and members of the military and the Royal Canadian Mounted Police are covered by the federal health plan. It is not mandatory for the Alberta registry to register those populations. We added this piece information in the Method section on page 6.

“Canada has a government-financed universal health insurance system. All permanent residents are covered by the provincial health insurance plan except for Registered First Nations, prison inmates, and members of the military and the Royal Canadian Mounted Police all of which are the responsibility of the federal government. All eligible Alberta residents are assigned a unique lifetime Personal Health Number (PHN). Therefore, PHN is an ideal variable for performing record linkage. The Alberta insurance registry is nearly complete and consistent, and is used as a proxy for the population of Alberta.”

5. This sentence helps in understanding how information on deaths of Alberta residents that occurred outside of Alberta can be obtained. However, the authors should precise how long does it take to obtain this information and whether it is complete for the relevant period.
Alberta Vital Statistics capture deaths occurring in the province only. In our study, we determined Calgary Health Region residents from the Alberta Vital Statistics. This region is the largest in the province, accounting for one third of provincial population. Therefore, we identified those who died in the region or other places in Alberta. We missed residents who died out of Alberta. We clarified this on page 6 as:

“The registry captures deaths that occurred within Alberta but misses Alberta residents who died out of Alberta. Provincial and territorial Vital Statistics registries on all deaths are submitted to Statistics Canada annually.”

6. The use of the PHN for assessing accuracy amongst the linked records may be questioned: only 70% of the records in the Vital Statistics registry have a valid PHN, ie the population with a valid PHN may differ from the rest of the population on the quality of the variables considered for the linkage. This information may be useful for assessing positive links. The authors should explain whether it will be used or not for assessing negative links as well.

We agree with this reviewer’s comment. Data quality may vary by presence of PHD in Vital Statistics. To assess whether records are accurately matched, we included records with PHN and excluded those without PHN. Then, we determined correct-link based on identical PHN in matched records. Therefore, positive linkage rate was assessed. The negative link was not be evaluated. The following sentence was added on page 7-8:

“We excluded matched records without PHN in both files.”

7. The authors should better explain why they used the Vital Statistics registry as the master in the process of linkage between this registry and population registry, although the in-hospital death records were accepted as the master in the linkage between these records and Vital Statistics files (page 8, line 6 from the bottom). It may have been more convenient, for the comparison of linkage rates, to use the same master.

We used the different files as the master in the two linkages. In the linkage between Vital Statistics and the population registry, Vital Statistics data are accepted as the master file. In the linkage between hospital discharge and Vital Statistics, we used hospital discharge data as the master file. The reason for choosing discharge data is that we could not distinguish precisely whether deaths recorded in Vital Statistics occurred in hospital or out of hospital. Therefore, we were forced to choose discharge data as the master file rather than Vital Statistics.

8. As mentioned above, a selection bias may lead to an over-estimation of the accuracy rate, especially if the quality of the variables included in the combination identifiers is poorer for the sub-population without a valid PHN. This potential selection bias is mentioned in the Discussion section. However, more information on the sub-population without a valid PHN (native populations, expatriates, etc…) would have helped the reader in the interpretation of this bias.

This is an excellent comment. In the discussion (page 12), we stated the possible reason as the
“Vital Statistics records without PHN are likely to be persons who died out of hospital. Personal information for those deaths might be from various sources, resulting in inconsistencies in personal information between Vital Statistics and the Alberta population registry. Death records without PHN might also be for persons who are not eligible for Alberta Health Insurance plan (such as inmates, travelers, visitors, expatriates, armed forces, and Royal Canadian Mounted Police). Those individuals were not present in the population registry but were recorded in the Vital Statistics if they died in Alberta. However, such cases account for a small proportion of all deaths in Alberta.”

9. Table 1: the authors should provide the reader with some help in the interpretation of this table. For instance, they could explain that this table gives an estimation of the percentage of deaths (among those included in the Vital Statistics registry) retrieved in the Population registry. It is quite surprising to observe that, for period 1999/00, only 22% to 26% (depending on the combination) of deaths, for children aged from 1 to 9 years, are retrieved in the Population registry. If we admit that the quality of the linkage is high, as shown in Table 2 (accuracy rate of 100% for all combinations for this age group), would that mean that only 1 child out of every 4 would have benefited from health care before death? If the above is not the right explanation, would that confirm our concerns about the potential over-estimation of the accuracy rate due to a selection bias of the sub-population, with a valid PHN? The authors should provide the missing PHN rate for “1 to 9 years children, which may even be higher for this specific age group.

One may wonder if the increase in missing PHN rate (from 31% to 48%) is not higher for the period 1999-00 which could explain why the linkage rate decrease (especially for “1 to 9 years” group) was not associated with a decrease in the accuracy rates, for the same period.

In the Results section (page 9), we stated:

“Table 1 presents the percentage of deaths in the Vital Statistics registry which can be linked with the population registry.”

We reported the missing of PHN on page 13.

“In fact, we found the linkage rate between the population registry and Vital Statistics was higher for Vital Statistics records with PHN than for records without PHN (89.5% versus 81.0% in 2001/2002). There were more PHNs missing for children aged 1 to 9 years (42.9% in 2001/2002) than for individuals aged 10 or older (13.8% in 2001/02).”

In the limitation (page 14), we also stated:

“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. However, we assess correct-linkage rate only among records with PHNs. Thus the correct-linkage rate found in this study is likely 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.”

10. As mentioned above, it may have been more convenient, for the comparison of linkage rates given in Tables 1 and 3, to use the same master. The linkage rates are much higher in Table 3 than in Table 1, probably, because Table 3 describes the linkage between two death records files. The authors should discuss these discrepancies more thoroughly in the Discussion section. For instance, hospitalized patients may have a better identification. Moreover, when they die during their hospitalization stay, the identification items of the death certificates may be filled in by the hospital, facilitating the linkage between Vital Statistics and hospital records. It would be interesting to discuss why, contrary to the results of Table 1, Table 3 reveals an increase in the linkage rate, for “1 to 9 years group”, between 1998-99 and 1999-00.

This was added in the discussion (page 11-12).

“The linkage and correct-linkage rates depend on the databases being linked. The linkage between the population and Vital Statistics registries produced lower linkage and correct-linkage rates than those between hospital discharge data and the Vital Statistics registry. One possible reason for this is that while completing death certificates for in-hospital deaths, hospital charts were consulted and personal information recorded in the chart was copied. Personal information in the chart is generally from the health insurance card, on which the Alberta health insurance plan prints PHN, full name, sex and date of birth. Thus more complete and accurate information might be included on death certificates for in-hospital deaths, thereby improving the linkage between Vital Statistics and hospital discharge data.”

11. Authors should explain how the comparison of the results of the three combinations of the four identifiers reveals that “first name is the least reliable”. This result is thus rather unexpected as other authors (Quantin et al, Methods of Information in Medicine, 2005) have found that gender, because of a low specificity, was poorly reliable, when comparing the same (or very close) identifiers through their likelihood ratios. I admit that the use of the Soundex method may have increased the sensitivity of the first name. But the same authors have shown that the increase in sensitivity due to the soundex, was accompanied by a decrease in specificity, with globally a marginal improvement in the likelihood ratio.

Thanks for pointing out this previous paper. We agree that gender was the least reliable indicator for linkage since it lacks specificity. In our study, we found adding first name to a combination of surname, sex and date of birth did not improve the linkage and correct-linkage rates much. Based on this finding, we intend to state that the combination of first name and other identifiers does not enhance linkage. We clarified this statement on page 13 as:

“In our study, a combination of surname, sex and date of birth has the highest linkage rate and second highest correct-linkage rate. The combination of first name, sex and date of birth generated relatively low linkage and correct-linkage rates. The combination of all four identifiers (first, surname, sex and date of birth) resulted in an even lower linkage rate with little increase in the correct-linkage rate. One possibility is that alternate spellings, initials, abbreviations and shortened forms of first names are common in the database. Therefore we recommend the use
surname, sex and date of birth as common identifiers in linking databases deterministically when a unique identifier is not available.”

12. The authors should have given the rate of missing information on postal codes (4% to 16%) in the Results section. As this information is required to select the study population, the authors should discuss the potential selection bias in the Discussion section. One may wonder if the quality of identifiers would not be poorer in the sub population with missing data on postal codes, resulting to a decrease in the linkage rate in this sub-population and to an over-estimation of this rate after exclusion of this sub-population.

We used Standard Geographical Classification in Vital Statistics and postal codes in the hospital discharge data to define Calgary Health Region residents. No postal codes are missing in hospital discharge data. We reported the missing Standard Geographical Classification codes in Vital Statistics in the Results section (on page 5) and also acknowledged this as a limitation on page 14.

“Of 21679 deaths occurred in the Calgary Health Region during 1998/99-2001/2002, 99 were excluded due to unknown SGC, 1358 were also excluded because of non-Calgary Health Region residents, and 20222 were finally analyzed.”

“Thirdly, we excluded deaths with unknown residence area information and missed residents of the region who died out of Alberta, possibly leading to overestimates of our linkage rate if personal information on these deaths was less complete than those we analyzed.”

13. The conclusion (at the end of page 13), recommending the use of the three indicators for the linkage, seems consistent with the recommendation given in page 10, (last line) but not with that given in page 12, second paragraph (3rd line : “it is recommended to use surname, first name, sex and date of birth...”). The authors should clarify their choice, and in particular revise the redaction of the 2nd paragraph page 12.

It is corrected on page 13 as:

“Therefore we recommend the use surname, sex and date of birth as common identifiers in linking databases deterministically when a unique identifier is not available.”

14. The authors should give more explanation on “acute hospitals”, does this mean “tertiary care hospitals”.

In the region, we included all hospitals regardless of type of hospital (tertiary or not). Therefore, we revised the sentence on page 6 as:

“Abstracts are filed for all inpatient separations (by discharge or death) from all hospitals in the CHR.”

15. Does the surname correspond to the family name at birth or may it refer to the marital name, for women?
In our databases, it is not distinguished for women. We clarified this on page 7:

“We assessed 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; (3) surname, first name, sex and date of birth.”

16. It would be preferable to discuss separately the results for each age group. For instance, contrary to the “over 40 years” age groups, Table 1 shows a decrease in the linkage rate for “1 to 9 years” group over time, which seems inconsistent with the statement about “improvements in data quality over the years”.

In Result section page 10, we pointed out this finding:

“The linkage rate across fiscal years varied for age group of 1 to 39 years old and tended to increase for age groups of 40 years or older.”

Reviewer 2

1. The authors have used the terms "linkage rate" and "accuracy". These terms are not generally used in the record linkage literature - the terms "sensitivity" and "specificity" or the complements of these, the missed-linkage and incorrect-linkage error rates are more commonly used (or "Type 1" and "Type 2" error rates).

In our study, we considered ‘PHN’ as the reference standard rather than to assess whether linked records are correct. The terms of ‘sensitivity’ and ‘specificity’ are generally used when the standard is accepted as ‘GOLD’. The terms ‘missed linkage’ and ‘incorrect linkage’ address record linkage from a negative angle. We reported our data from a positive perspective, linked and correct-linked. 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.

2. (We grouped Reviewer 2 comments 2, 3 and 5 in the file sent to us into one since they are related to linkage rate.) The definition for "linkage rate" appears to be the number of records (n) from the two data sources which are linked to each other using a combination of identifiers, divided by the total number of records (N) in one of the data sources. However, there does not appear to be any check that the linked records comprising n do in fact refer to the same person. Surely the numerator for the "linkage rate" measure should be nb (n subscript b)? As it stands, n would appear to contain a mixture of true links and false links and thus the ratio n/N overstates the completeness of the linkage process as measured (by convention) as the number of true-links found divided by the number of true-links possible (i.e. the sensitivity of the linkage process). Similarly, the denominator N should be the number of records in one data source which are expected to be present in the other data source - in other words the number of true-links which are possible. Presumably all Vital Statistics Registry records are expected to appear in the Population Registry, and all
In-hospital death records are expected to appear in the Vital Statistics Registry. If so, this should be stated explicitly. Further to the comments above, it is recommended that the authors think of the results of a the linkage process as a set containing pairs of records representing the Cartesian product of records in both data sources. This set then contains four types of pairs of records: true-links, false-links, true-nonlinks and false-nonlinks. These basic record linkage concepts are explained in more detail in the works of Howard Newcombe and others, some of which the authors cite in their reference list. It is recommended that the authors refer to these or other introductory works on medical record linkage before revising the paper.

This reviewer suggested excellent comments for calculation of correct or incorrect (i.e. true/false) linkage rate using the number of records in the master file as the denominator. In our analysis, we specifically aimed to address what the linkage rate is between two databases and then what the correct linkage rate among those linked records is. Therefore, we did not employ the true/false link although some, not all, literature uses these terms, for example the two papers published by Roos reported in our paper did not use the term of true/false link. The reviewer’s comment on “true/false-linkage” is addressed in our paper among matched records, but not in non-linked records. We reported the correct (“true”) linkage in the table 2 and 4 but did not report the incorrect linkage (“false”) linkage since the false linkage rate can be calculated by 100% – true linkage rate. We included deaths with PHN to check whether the same persons were linked and missed those without PHN. In our databases, we do not have other unique personal identifier except for PHN to address true/false linkage among deaths without PHN. This limitation is acknowledged. Reviewer 1 has suggested many excellent comments on this issue.

In health services or population studies, researchers frequently conduct record linkage to obtain enhanced information without confirmative knowledge of whether the same persons are contained in both databases. For example, researchers have defined a cohort of myocardial infarction patients using hospital discharge data which will be linked with Vital Statistics data to obtain 1 year mortality. Unlinked patients are assumed to be alive. Then, factors of interested are assessed for identify risk factors associated with the poor outcome. In this process, researchers are assuming the mortality obtained from Vital Statistics is complete and correct. In such a study, two major questions are unanswered: 1. How many records could be linked? 2. Which method is the best to produce the highest rate of correct linkage (i.e., linked records are referring the same patients)? Following these practical questions, we designed our study by looking at three combinations of identifiers in three databases.

3) After linking, duplicate records have been eliminated using the "gold standard" linkage reference mechanism of the Vital-PHN. Unfortunately, it is not valid to do this -one cannot alter the results of the trail of deterministic linkage on a combination of partial identifiers by using a mechanism which removes the need for such linkage in the first place. In other words, if Vital-PHN numbers are available, why bother with using combinations of partial identifiers as the authors have done? If, however, there are circumstances in which the Vital-PHN number is not available, and the study is attempting to simulate such circumstances, then it is not acceptable to use the Vital-PHN number to prune out duplicates.
from the linked records -it may only be used as a "gold standard" to establish the true linkage or non-linkage status of records and thus evaluate the performance of the deterministic linkage using the combinations of identifiers.

This is a great comment. Following the recommendation, we re-analyzed the data after grouping duplicates (unresolved records, one record from a database linked to more than one record from another) into incorrect-link group since it is uncertain which link is correct. We revised tables 2 and 4 and the Results section following this new analysis. However, changes between the previous and the new analysis are very minor and did not alter our interpretation and conclusion.

Reviewer 3

1. 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.

Impacts of correct linkage on applied studies vary by study purposes. We could analyze the data for assessment of the impact using an example. However, we feel this will increase the length of the paper greatly and distract the reader’s focus from the methodology of record linkage. Therefore, we decided not to do further analysis. However, we would take this reviewer’s suggestion if the editor feels strongly that additional analysis is necessary.

2. 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.

Use of Canadian national data would increase the generalizability of our study findings. Hospital discharge data and Vital Statistics are centralized in Canada but the population registry is held by each province. It is not feasible to our team to obtain all highly confidential information, often protected by provincial privacy legislation from all Canadian provinces/territories. Therefore, we used available data only (i.e. a Canadian large health region database).

3. 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 birth date – 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.

We did not use middle name as an identifier since this is not consistently recorded in the population registry and Vital Statistics.

4. 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.

We provided more detail of these databases in the Methods section rather than in the Introduction because these are the materials used for analysis. We just listed databases in the Introduction to highlight what we are analyzing.

5. 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).

We specified the population as the following on page 5:

“As of March 2002, CHR’s population was approximately 1.1 million.”

The Alberta provincial government collects deaths and births that occur in the province and submits the data to Statistics Canada. However, the national Vital Statistics data are not shared with provinces except special request.

We described the population registry in more detail. See responses to Reviewer 1 Comment 4 above.

5. 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.

Hospital discharge data are widely used for research purposes. Coders extract demographic and clinical information from each inpatient chart at the hospital. The data are submitted to the provincial government and then centralized at the Canadian Institute of Health Information. We added the following information on page 6-7:

“Abstracts are filed for all inpatient separations (by discharge or death) from all hospitals in the CHR. Coders review inpatient charts and extract data on PHN, demographics, diagnoses, procedures and physician specialty. The extracted data are stored in the region as well as submitted to the Alberta provincial discharge database.”
6. 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?

We used SAS 8.1. We would love to share the SAS codes, however we must point out the codes are simple and programmer with minimum SAS skill could do this themselves. The same analysis could be done using other statistical computer software packages, such as STATA.

7. 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.)

See responses to Reviewer 1 Comment 4.

8. 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 ...”

We carefully checked grammar for this revised paper.

Again, thank you for providing us the opportunity to revise and resubmit this manuscript. It has certainly improved through these suggested revisions. 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)