Reviewer’s report

Title: Assessing Accuracy of Record Linkage between Administrative Data and Vital Statistics

Version: 1 Date: 26 July 2005

Reviewer: catherine quantin

Reviewer’s report:

General

-------------------------------------------------------------------------------

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

Page 4 – 1st paragraph : this paragraph should be rewritten in order 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 ?

- 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 ?).

Page 5 – 1st paragraph :
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 ?

Page 5 – 2nd paragraph : the authors state that the registry is complete : does this mean that residents of the first nations are included in this registry ?

Last sentence of page 5 : this sentence helps in understanding how information on deaths of Ablerta 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.

Page 7 – 1st paragraph : 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.

Page 7 – line 10 (from the bottom)
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.

Page 8 – line 5 : 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.

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.

Page 9 – line 3 from the bottom : 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.

Page 10 – line 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.

Page 12 – line 7 of the second paragraph : The 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.

Page 12 – last line : 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.

Page 13 – last paragraph : 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.

-------------------------------------------------------------------------------------------------------------------

Minor Essential Revisions (such as missing labels on figures, or the wrong use of a term, which the author can be trusted to correct)

Page 6, line 3 : the authors should give more explanation on “acute hospitals”, does this mean “tertiary care hospitals”.

Page 6, beginning of the second paragraph : does the surname correspond to the family name at birth or may it refer to the marital name, for women?

Page 11 – lines 7-9 : 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”.

-------------------------------------------------------------------------------------------------------------------

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 whose findings are important to those with closely related research interests

Quality of written English: Acceptable

Statistical review: No

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