Reviewer's report

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

Version: 1 Date: 21 March 2007

Reviewer: leicester Gill

Reviewer's report:

General

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

Karmel and Gibson

To review this paper I have used the structure outlined in your email.

Overall, the paper is well structured with few, if any, typographical errors. The abstract and title do represent the content and aim of the paper. Although the paper presents a theoretical approach to the methodology of record linkage and does not contain experimental data, I feel that it is a useful addition to the literature and it should be accepted for publication.

This paper covers a very particular type of event linkage, where there are no names or other personal identifiers, but there are other data items that could be used for matching/linking are common to both datasets. The data items discussed in the paper include a range of administrative dates, date of birth with the errors associated with poor recording or lapse of memory, and place of residence which can be very variable for older members of the population, since they may stay with relatives or move to other types of accommodation.

The selection of a matching/linking strategy depends on the availability and quality of the matching items, and the number of events entering the system. The larger the population to be linked the greater the need for a wide selection of high discrimination items. There are two problems that need to be addressed in any record matching and linking exercise, firstly to determine the percentage of potential match pairs that are not linked together (false negative or missed matches) and the percentage of pairs that are wrongly linked together (false negative matches). The strategy normally chosen for linkage is a compromise in which these two measures are minimised. These measures are usually determined using sampling of the records on the file, or capture-recapture methods. The determination of the overall quality of a linked file is usually based on a measure of the false positive rate. In this paper the authors have developed and used an innovative statistical approach to estimate this rate.

The statistical method is fully described in the paper, and is based on two assumptions, firstly, the dates of birth are within a 30 year span starting at age 65, are evenly distributed within this span, and the expected numbers can be estimated using a Poisson process. The paper demonstrates that the linking method works where the number of people living in a defined geographical area is less than 10,000. Using these parameters the false positive rate is about 1%. Since the calculation of the false positive rate is a linear function of the number in the matching group, extension of the method to large groups can generate a large false positive rate. The method works for small populations only, and would give unacceptably high rates if it were used nationwide. Taking the line on Fig. 1 and extrapolating it to a population of 500,000 would give a false positive rate of 25%. In the Oxford experience of matching, using date of birth, sex and postcode, the false positive rates determined through sampling and capture/recapture are not as high as those predicted using this methodology. It will be interesting to see in a future paper how the experimental false match rates compare with the theoretical ones.

Examining the two graphs in Figure 2, the figures in the first graph with a population of 3000 are three times higher than the figures in the second graph, thus demonstrating the linear relationship between the false positive rate and the population size.

The data are collected from the hospital separation data and the transfers to and from residential care. The quality of the data relies on the data collection systems, and is little influenced by the research staff. While the date of birth and gender are personal constants, in practice, administrative dates are volatile and errors
do creep in. One problem that can arise is the poor recording of dates, especially on the Friday to Monday of a weekend, hence the reason for relaxing the methodology to allow a two day difference between admission, discharge and transfer dates. This paper shows that even with such medium discriminating variables, the match rate is quite high with a false positive rate of about 1% and does give encouragement to other groups contemplating similar studies.

The paper has been crafted to the standards required of a scientific study. In the results section, no consideration is given to the reduction in the population as the age of the group increases, for example a life table could be used for this purpose. For example would the population of 75 year olds be better matched with lower false positive rates that a group of 65 year olds. The problem with the approach might be that although the population would be smaller, the data may be of poorer quality.

The paper is a useful starting point for other groups contemplating similar studies.

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

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?: Accept without revision

Level of interest: An article whose findings are important to those with closely related research interests

Quality of written English: Acceptable

Statistical review: Yes, and I have assessed the statistics in my report.

Declaration of competing interests:

I declare that I have no competing interests