Reviewer's report

Title: Optimal strategy for linkage of datasets containing a Statistical Linkage Key and datasets with full personal identifiers

Version: 1 Date: 16 February 2014

Reviewer: Griffin Weber

Reviewer's report:

SLK-581 is a statistical linkage key used in a number of aged care datasets in Australia instead of listing patients' full names or other personal identifiers. This study looks at how best to link SLK-581 indexed datasets to other datasets that contain full patient identifiers. They show that by simply calculating SLK-581 values for the other datasets, and matching on those values, there are many missed links. This is due mainly to missing data needed to create a complete SLK-581 value. However, by first using a probabilistic linkage method to group all records belonging to the same patient in the dataset containing full patient identifiers, it becomes more likely that at least one of those records has enough data to form a SLK-581 value that has a match.

The authors were able to conduct this analysis because of access to three things: (a) a database of admitted patient data (APD) for 7.2 million patients; (b) death registration data (RBDM) that contains full patient identifiers and has similar demographics as the aged care data; and (c) a service in Australia called the Center for Health Record Linkage (CHeReL), which contains highly accurate probabilistic linkage data for more than 9 million patients. By first linking APD and RBDM using CHeReL, the authors created a gold standard to compare other algorithms. By showing how best to link APD to RBDM, if only SLK-581 values were available in the RBDM data, they approximate how best to link APD to actual aged care data.

Minor Essential Revisions

There are some typos. In the Methods section, there seems to be one or more missing words in the first sentence under Data Sources. In the Discussion section, the second sentence in the second paragraph is either also missing words or ending incorrectly with a semicolon.

Discretionary Revisions

I think this is a nice study with potential implications far greater than the authors mention in the manuscript. They focus exclusively on matching to SLK-581 encoded identifiers. However, I think the more important message is that by introducing a third dataset (CHeReL in this case), the linkage between two other datasets can be greatly improved. This idea can be generalized to many other linkage examples. It also highlights a privacy/security issue. If a purpose of a statistical linkage key is to protect the identities of patients, then large population
datasets like CHeReL increase risk to patients by making it easier to re-identify the data. I would encourage the authors to discuss these broader benefits and risks of their method.

If possible, the authors might want to investigate why the Basic algorithm performed more poorly with arthritis than other conditions. Are these patients, for example, seen more at private hospitals, which have higher rates of missing data?

**Level of interest:** An article of importance in its field

**Quality of written English:** Acceptable

**Statistical review:** No, the manuscript does not need to be seen by a statistician.

**Declaration of competing interests:**

I declare that I have no competing interests.