Author's response to reviews

Title: Technical challenges of providing record linkage services for research

Authors:

James H Boyd (J.Boyd@curtin.edu.au)
Sean M Randall (sean.randall@curtin.edu.au)
Anna M Ferrante (A.Ferrante@curtin.edu.au)
Jacqui K Bauer (jacqui.bauer@curtin.edu.au)
Adrian P Brown (Adrian.Brown@birchman.com.au)
James B Semmens (James.Semmens@curtin.edu.au)

Version: 4
Date: 2 December 2013

Author's response to reviews: see over
Reviewer’s report

Title: Technical challenges of providing record linkage services for research

Version: 2

Date: 20 October 2013

Reviewer: Griffin Weber

Reviewer’s report:

MAJOR COMPULSORY REVISIONS
None

MINOR ESSENTIAL REVISIONS
None

DISCRETIONARY REVISIONS

1) The authors discuss technical challenges when trying to match patients (patient linkage) in research datasets. This is the first time I've read this manuscript. From the cover letter I learned that there was a previous version that included descriptions of a patient linkage software tool that is used by the authors' organization. All references to this program were removed in the version I read. In my opinion, I think this is unfortunate because the manuscript might have been much stronger if it described the actual experiences the authors had working with research datasets. Also, the manuscript does not go very deep into the issues. For example, they introduce linkage automation as an approach, but don't provide detail on any specific algorithms used for doing this. Basically what this manuscript does is give a high-level classification of methods of patient linkage with research data. That has some scientific merit. However, it would have been of interest to a broader audience if it had either 1) presented a new algorithm/code for doing patient linkage, or 2) provided data about the number of datasets, accuracy, or costs involved in patient linkage at the authors' organization.

Response:

Accepted, unfortunately we are unable to make the patient linkage software available under an open source licence at the moment (as outlined in the journal guidelines). Instead the manuscript explores the technical challenges associated with record linkage. These include the general management of data; handling different linkage scenarios; the management of routine, on-going linkage (and the complexity of storing and handling changing data); the need for automation and the ever present need to accommodate larger sized datasets. We hope the paper provides information on the challenges associated with operating record linkage services.

2) The word "bespoke" was used several times. I had to lookup this word in the dictionary. Apparently this is a British English word that is rarely used in the United
States, where I'm from. I would have used "customized". Though, I don't know if that would look strange to a person in Australia.

Response: “bespoke” has been replaced with “customised” throughout the document
Reviewer's report

Title: Technical challenges of providing record linkage services for research

Version: 2

Date: 7 November 2013

Reviewer: Vassilios S Verykios

Reviewer's report:

Major Compulsory Revisions

The work presented in this study refers to some guidelines that need to be taken into consideration when some kind of incremental linking is in place. Even though different options are presented with respect to the grouping of similar records no evaluation is presented to shed light into this process.

Response:

Accepted, the paper has been updated to reflect the previous research in the area which has explored incremental linking:

METHODS SECTION

Alternative approaches to on-going incremental linkage have been developed in recent year, including those outlined by Kendrick [16, 17] in his description of Best-link matching. Kendrick’s paper expands on the principles outlined by Newcombe [18,19] which describes the factors which could have an effect on the linkage quality, including the likelihood that a record in one file is represented in the matching file.

RESULTS SECTION

On-going Linkage
There are several possible methods for conducting on-going linkage and the linkage output will be influenced by a number of factors. One factor is the overlap of people between the files being matched i.e. how many of the new records have true matches in the existing linked file. Another influence is the size of the existing file, the larger the number of records involved in a probabilistic linkage the greater the likelihood that information will agree ‘by chance’ across records being compared.

These factors have an influence on the number of records brought together for linkage, the matching strategy and in the post-linkage processes that convert pairs of matched records into groups of records that are stored in a linkage map.

The relationship within and between files and the level of confidence in existing links/relationships are important considerations in the design and optimisation of linkage strategies.
For example, one approach is to link all records in the incoming dataset to all other records in the system. This method allows pairs to be created describing the relationships between all records in the system. With this approach, there are no expectations or assumptions made about how records match against each other or how they group together to become ‘sets’ of records that belong to the same individual. In terms of linkage strategy, this scenario represents a relatively unconstrained many-to-many linkage. If, however, the linkage task involves linking records to an authoritative record type (i.e. where only one high-quality record per person is known and maintained), then a one-to-one or many-to-one linkage may be more appropriate and there is opportunity to adapt matching strategies to leverage this knowledge (Ref: Newcombe 1988, Newcombe 1995).

There is also no particular algorithm provided to explain the exact process that should be followed so that the linkage is accomplished, under different scenarios. I am also aware of another older piece of work by Mauricio A. Hernández and Salvatore J. Stolfo: Real-world Data is Dirty: Data Cleansing and The Merge/Purge Problem. Data Min. Knowl. Discov. 2(1): 9-37 (1998) that talks about a similar approach of linking data in phases, that need to be taken into consideration by the current work.

Response:

A variety of techniques have been used to undertake record linkage. These include deterministic rules-based approaches, sort and match algorithms, and the more classical probabilistic techniques that involve multiple iterations of blocking and record matching. Probabilistic techniques have been most popular amongst SLUs and there have been a number of enhancements to take advantage of improvements in technology and to allow processing of large data files. The paper has been updated to reflect these developments while addressing the complexity around managing the associated factors.

METHODS

The role of SLUs has become more prominent in the research infrastructure landscape and the level and complexity of demands placed on them for linkage services has increased. While there are a variety of techniques available to undertake record linkage such as deterministic rules-based methods, sort and match algorithms (Hernández and Stolfo, 1998), and probabilistic techniques (Fellegi and Sunter, 1969; Newcombe and Kennedy, 1962), the tendency for most SLUs has been to implement a probabilistic framework, owing to its robustness, adaptability (particularly in relation to linkage of large datasets – see, for example Clark and Hahn (1995)) and high-quality output (Pinder and Chong, 2002; Gomatam et al., 2002). Probabilistic methods involve sophisticated blocking techniques (to streamline comparisons) and the application of matching methods that incorporate both deterministic and probabilistic comparisons (Roos et al., 1986; Kendrick et al., 1998; Winkler, 1994). In recent times, there has been extensive work on extending probabilistic approaches and to improving efficiency through the use advances in technology (Winkler, 2000; Herzog et al., 2010). However, beyond the complexity of the linkage process per se, there are other technical challenges that present to SLUs. These include the general management of data, handling different linkage scenarios, the management of routine, on-going linkage (and the complexity of
storing and handling changing data), the need for automation and the ever present need to accommodate larger sized datasets. In this section we discuss each of these emerging problems.

References related to updates:


A reference regarding privacy preserving record linkage is missing, since the authors talk about the process of outsourcing the record linkage process where the privacy of the records is in danger.

Response: A reference to privacy preserving record linkage has been added to the paper.