

# International Journal of Population Data Science

Journal Website: [www.ijpds.org](http://www.ijpds.org)



## Unsupervised evaluation of a large-scale historical population linkage

Charini Nanayakkara<sup>1</sup> and Peter Christen<sup>1,2</sup>

<sup>1</sup>Australian National University, Canberra, Australia

<sup>2</sup>University of Edinburgh, Edinburgh, United Kingdom

### Objectives

Record linkage can be viewed as a classification task where pairs of records are classified as matches (same individual) or non-matches (different individuals). Alternatively, clustering methods generate groups of records each referring to one person. We discuss methods to evaluate the quality of large-scale population linkages without ground truth data.

### Conclusion

Evaluating the quality obtained for a linkage outcome is challenging when no ground truth data are available. We have shown how our unsupervised evaluation methods can provide linkage quality estimates close to the actual ground truth results, and how our methods can be applied on large-scale population linkage outcomes.

### Methods

In practical applications of record linkage, ground truth data are often not available, or they can be incomplete or biased, making quality evaluation challenging. To overcome this gap, we present multiple methods to evaluate the quality of a record linkage outcome. These methods are either applicable to one-to-one linkages, or they consider clustering results by taking the similarities (match scores) calculated between records and the structure of a similarity graph into account. We apply our methods on linkage outcomes from a large historical population database where we consider multiple one-to-one, one-to-many, and many-to-many linkage types.

### Results

We apply our methods on both publicly available data sets where the linkage ground truth is available, as well as a large historical population database. With the former data sets, depending upon the structure and quality of the similarity graph they are applied on, our methods can accurately estimate precision (positive predictive value) and recall (sensitivity) results. While precision is sometimes overestimated, the opposite happens with recall. As a result, on average our best methods provide estimated F-measure results within 7% of the corresponding ground truth-based results. We then apply our methods on different linkage types conducted on a large population database containing over 20 million records, and we report on the quality obtained when several different linkage methods are applied on this large database.

