Evaluating Population Data Linkage: Assessing stability, scalability, resilience and robustness across many data sets for comprehensive linkage evaluation

Özgür Akgün[1], Ahmad Alsadeeqi[2], Peter Christen[3], Tom Dalton[1], Alan Dearle[1], Eilidh Garrett[4], Graham Kirby[1], Alice Reid[5]

[1] School of Computer Science, University of St Andrews, UK
{ozgur.akgun, tsd4, alan.dearle, graham.kirby}@st-andrews.ac.uk
[2] Heriot-Watt University, Edinburgh, UK
[3] The Australian National University, Canberra, Australia
[4] University of Essex, UK
[5] University of Cambridge, UK

Data linkage approaches are often evaluated with small or few data sets. If a linkage approach is to be used widely, quantifying its performance with varying data sets would be beneficial. In addition, given a data set needs to be linked, the true links are by definition unknown. The success of a linkage approach is thus difficult to comprehensively evaluate.

This talk focuses on the use of many synthetic data sets for the evaluation of linkage quality achieved by automatic linkage algorithms in the domain of population reconstruction. It presents an evaluation approach which considers linkage quality when characteristics of the population are varied. We envisage a sequence of experiments where a set of populations are generated to consider how linkage quality varies across different populations: with the same characteristics, with differing characteristics, and with differing types and levels of corruption. The performance of an approach at scale is also considered.

The approach to generate synthetic populations with varying characteristics on demand will also be addressed. The use of synthetic populations has the advantage that all the true links are known, thus allowing evaluation as if with real-world 'gold-standard' linked data sets.

Given the large number of data sets evaluated against we also give consideration as to how to present these findings. The ability to assess variations in linkage quality across many data sets will assist in the development of new linkage approaches and identifying areas where existing linkage approaches may be more widely applied.

This work was presented at the following workshop.

Venue	ADRN 2017 - The UK Administrative Data Research Network Annual Research Conference
Location	The Royal College of Surgeons of Edinburgh
Date	June 1-2, 2017
URL	http://www.adrn2017.net/