International Journal of Population Data Science

Journal Website: www.ijpds.org

Finding Maternal Siblings in Birth Registration Data to form a Pregnancy Spine – Data Linkage & Graph Based Methods for Unknown Cluster Sizes

Gammon, S¹ and Morris, C¹

¹Office for National Statistics, UK

Introduction

We have developed an innovative methodology to link maternal siblings within 2000 – 2005 England and Wales Birth Registration data, to form a Pregnancy Spine, a unification of all births to each unique mother.

Key challenges in this many-many linkage scenario:

1. Blocking (reduction of record pair comparisons)
2. Cluster resolution

Objectives and Approach

Probabilistic data linkage (Python) was followed by generation of clusters (using igraph in R) and graph theory community detection techniques.

To optimise geographical blocking and increase accuracy, we incorporated Internal Migration data to map the likely geographic movement of mothers between births.

Maternal sibling clusters were modelled as a graph and the structure of clusters was optimised using community detection methods to link, split and evaluate sibling groups.

Additionally, we incorporated additional childhood statistics data relating to child date of birth to evaluate likely accuracy of sibling pairs and remove false edges (links).

Results

Our development has resulted in a new blocking method and cluster resolution method. In addition, we developed new ways to assess and measure the accuracy of sibling groups, beyond traditional classifier metrics, and infer error rates.

We applied our method to Registration Data used in earlier studies for QA of our methods.

Using this, and by comparing against other statistics on maternal sibling composition we will present results which show that a high degree of accuracy (precision / recall and new checks) was obtained for precision, recall, and other evaluation metrics.

Conclusion/Implications

These methods will improve other linkage projects with unknown clusters sizes; for de-duplicating datasets, linkage of multiple datasets, or incorporation of data from a longer time-period through longitudinal linkage.

To this Spine, researchers can now append and link other data sources to answer questions about maternal and child health outcomes.