International Journal of Population Data Science

Journal Website: www.ijpds.org

Investigation and reporting of Data Quality within and between linked SAIL datasets

Rees, Sarah¹ and Rees, Arfon¹

¹Swansea University

Objectives

The SAIL databank brings together a range of datasets gathered primarily for administrative rather than research processes. These datasets contain information regarding different aspects of an individual's contact with services which when combined form a detailed health record for individuals living (or deceased) in Wales.

Understanding the quality of data in SAIL supports the research process by providing a level of assurance about the robustness of data, identifying and describing where there may be sources of potential bias due to invalid, incomplete, inconsistent or inaccurate data and therefore helping to increase the accuracy of research using these data.

Designing processes to investigate and report on data quality within and between multiple datasets can be a time-consuming task to undertake; it requires a high degree of effort to ensure it is genuinely meaningful and useful to SAIL users and may require a range of different approaches.

Approach

Data quality tests for each dataset were written, considering a range of data quality dimensions including validity, consistency, accuracy and completeness. Tests were designed to capture not just the quality of data within each dataset, but also to assess consistency of data items between datasets. SQL scripts were written to test each of these aspects: in order to minimise repetition, automated processes were implemented where appropriate.

Batch automation was used to call SQL stored procedures, which utilise metadata to generate dynamic SQL. The metadata (created as part of the data quality process) describes each dataset and the measurement parameters used to assess each field within the dataset. However automation on its own is insufficient and data quality process outputs require scrutiny and oversight to ensure they are actually capturing what they set out to do.

SAIL users were consulted on the development of the data quality reports to ensure usability and appropriateness to support data utilisation for research.

Results

The data quality reporting process is beneficial to the SAIL databank as it provides additional information to support the research process and in some cases may act as a diagnostic tool, detecting problems with data which can then be rectified.

Conclusion

The development of data quality processes in SAIL is ongoing, and changes or developments in each dataset lead to new requirements for data quality measurement and reporting. A vital component of the process is the production of output that is genuinely meaningful and useful.

*Corresponding Author:

Email Address: s.e.r.rees@swansea.ac.uk (S. Rees)