International Journal of Population Data Science

Journal Website: www.ijpds.org

Repeatable Research Infrastructure Enabling Administrative Data Analysis

Thayer, D[1*], Elmessary, M[1], Mallory, D[1], Arnold, P[1], Cichowski, M[1], Brooks, C[1], Rees, S[1], Wang, T[1], Collins, H[1], and Ford, D[1]

[1]Swansea University

Background/Rationale

Linked administrative datasets offer great potential for research, but also present major challenges—including the preparation of operational data into a form suitable for efficient research, complex and computationally demanding analysis, and the need to capture and share information about dataset contents and research methods.

Main Aim

The analytical services team in the Secure Anonymised Information Linkage (SAIL) Databank is creating interconnected tools and systems to automate the preparation and analysis of research data and to curate information about datasets and research methods. Our underlying goal is to make linked data research orders of magnitude faster and cheaper, as well as improve its consistency and quality.

Methods

Several key developments are ongoing:

- Automation of data quality checking.
- Management of dataset metadata.
- Processing of raw source datasets into cleaned, research-ready data assets.
- The Concept Library, an application for creating, using, and sharing knowledge about research definitions and methods.
- A suite of R packages for analysis.

Web Application Programming Interfaces will allow these pieces to work together as an integrated system enabling efficient research.

Results

Initial versions of dataset quality checking, cleaned datasets, and R code to implement common tasks are already in day-to-day use by researchers within SAIL. An advisory group has been convened to help guide the work.

For example, shared library code that flags conditions within health data has been used across multiple projects; a cleaned dataset measuring follow-up within primary care has been used by more than 100 projects.

Conclusion

Our proof-of-concept work demonstrates the ability of shared code and cleaned data to meet needs across multiple projects, saving effort and standardizing results. Ongoing work to develop and integrate these tools should further streamline the research process, increasing the output and public benefit of SAIL and other data sources.

*Corresponding Author:
Email Address: d.s.thayer@swansea.ac.uk (D Thayer)

https://doi.org/10.23889/ijpds.v4i3.1268