International Journal of Population Data Science

Journal Website: www.ijpds.org

Developing a Research-Ready-Data-Asset (RRDA) for Welsh primary care data within the SAIL Databank: enhancing data quality and reproducible research.

Hoda Abbasizanjani[1], Stuart Bedston[1], and Ashley Akbari[1]

[1]Swansea University, Swansea, United Kingdom

Objectives

We aimed to develop a high-performance RRDA for the Welsh Longitudinal General Practice (WLGP) data to standardise curation, enhance reproducibility, improve query performance and add additional value/features for research. The RRDA provides a curated normalised asset with a comprehensive clinical code look-up and assigned activities type.

Methods

WLGP data has a long-format event-list structure with potential data quality issues, including duplicates, re-inserted GP-to-GP-transferred records, and missing/invalid entries. To address these, the RRDA involves three steps: data cleaning, data curation using patient's GP registration history from demographic data, and transforming data into a structured, normalised format to eliminate redundancy and support faster, flexible large-scale queries.

The WLGP-RRDA includes a look-up of primary care official/local codes (Read-V2/SNOMED/EMIS/Vision). Additionally, we implemented a four-layer approach for identifying healthcare providers, patient access mode, interaction type, and details of individual codes to capture the complexity of activities, enabling patient-practice interaction analysis.

Results

Curating WLGP data (1990-2024, 4,565m records, 5m people) revealed significant improvements in data quality and completeness over time, with data retaining rates after cleaning/curation increased from 38% to 94%. Similarly, patient inclusion in WLGP-RRDA improved from 43% to 98% during the same period, indicating improved data accuracy and Welsh residents coverage.

The normalisation process resulted in an efficient three-table structure with unique integer keys for clinical codes and events, optimising database performance/scalability. The extensive clinical code look-up improved coverage of events with known descriptions, showing increased local/SNOMED code use since the pandemic.

Additionally, implementing a multi-layered approach to identify interaction types (e.g., face-to-face/remote consultations) using official/local code hierarchies enabled analysis of national trends in GP activities and impact of the pandemic.

Conclusion

The WLGP-RRDA development enhanced data quality and streamlines the research processes through a reproducible, maintainable, standardised curation and a multi-layered approach to extract activity types. This methodology/RRDA benefits SAIL users and wider across other environments with similar data, through our shared resources to promote transparency and collaboration.

https://doi.org/10.23889/ijpds.v10i3.3123

Temporary page!

LATEX was unable to guess the total number of pages correctly. As there was some unprocessed data that should have been ad to the final page this extra page has been added to receive it.

If you rerun the document (without altering it) this surplus page will go away, because LATEX now knows how many pages expect for this document.