

# International Journal of Population Data Science

Journal Website: [www.ijpds.org](http://www.ijpds.org)



Swansea University  
Prifysgol Abertawe

## Emergency department attendance after telephone triage

Denaxas, Spiros<sup>1\*</sup>, Kunz, Holger<sup>1</sup>, Smeeth, Liam<sup>2</sup>, Gonzalez-Izquierdo, Arturo<sup>1</sup>, Boutselakis, Harry<sup>1</sup>, Pikoula, Maria<sup>1</sup>, and Hemingway, Harry<sup>1</sup>

<sup>1</sup>University College London

<sup>2</sup>London School of Hygiene & Tropical Medicine

### Objectives

Electronic health records (EHR) across primary, secondary, and tertiary care are increasingly being linked for research at a population level. The increasing volume, variety, velocity, and veracity of big biomedical data makes research reproducibility challenging. Research reproducibility and replicability is essential for the external validity and generalizability of scientific findings and the lack of standardized approaches and tools and relative opaqueness of data manipulation methods is detrimental to their integrity.

The objective of this study was to explore, evaluate and propose methods, tools and approaches for addressing some of the challenges associated with reproducibility when using linked national electronic health records for research.

### Approach

We systematically searched literature and internet resources for well-established and appropriate methods, tools, and approaches used in related scientific disciplines. The identified techniques were systematically evaluated in terms of their capacity to facilitate reproducible research in routinely collected health data across the life course of a research project: from protocol creation and raw data curation to data transformation and statistical analysis though to finding dissemination and impact. Most importantly, the identified techniques were tested and applied in a contemporary database of linked electronic health records. CALIBER is a research data platform of linked national electronic health records from primary care (Clinical Practice Research Datalink), secondary care (Hospital Episode Statistics), acute coronary syndrome disease registry (Myocardial Ischaemia National Audit Project) and cause-specific mortality (Office for National Statistics) for roughly 2 million adults.

### Results

Firstly, we present the review of methods and approaches which we identified through our search. Secondly, we propose a set of recommendations for applying them within the context of research projects making use of linked routinely collected health data. Focal interests included: a) documentation of data (attributes, relationships, and interpretation), b) data processing (source code, instructions, and parameters), c) results (visualizations, figures), and any supplementary material. Thirdly, we present approaches around a) raw data curation using international metadata standards, b) study protocol encoding, c) provenance and sharing of data transformation and statistical analysis operations, d) public and private data retention, and e) computable EHR-driven phenotypes.

### Conclusion

The complexity and size of routinely collected health data is increasing through linkages across distributed data sources. The scientific community benefits from findings which can be replicated. This study presents a number of methods, tools and approaches across the project life course for ensuring that their research studies are reproducible and replicable from the wider scientific community.

\*Corresponding Author:

Email Address: [s.denaxas@ucl.ac.uk](mailto:s.denaxas@ucl.ac.uk) (D. Spiros)

