International Journal of Population Data Science

Journal Website: www.ijpds.org

Harmonising data from different sources to conduct research using linked survey and routine datasets

Bandyopadhyay, A[1], Tingay, K[1], Borja, MC[2], Griffiths, L[2], Akbari, A[3], Bedford, H[2], Brophy, S[4], Walton, S[2], Dezateux, C[5], and Lyons, R[6]

[1]Swansea University

[2]Institute of Child Health, University College London

[3]Health Data Research UK - Wales and Northern Ireland, Swansea University Medical School

[4]Swansea University Medical School

[5]Centre for Primary Care and Public Health, Barts and The London School of Medicine and Dentistry, Queen Mary University London

[6]Farr Institute, Swansea University Medical School

Introduction

Harmonization of different data sources from various electronic health records across systems enhances the potential scope and granularity of data available to health data research, providing more opportunities for research by improving the generalizability and effective sample size of a range of outcome metrics.

Objectives and Approach

This study describes data harmonisation for a UK longitudinal birth cohort, the Millennium Cohort Study (MCS) which was linked to routine inpatient and emergency department, and, where available, general practice and child health records for 1838 Welsh and 1431 Scottish consenting MCS participants. Datasets requiring harmonisation were: from Wales, Patient Episode Dataset for Wales (PEDW) and Emergency Department Data Set (EDDS) data and from Scotland, Scottish Medical Record 01 (SMR01) and Accident and Emergency dataset (A&E2). Heterogeneous variables were created by transforming variable names, concepts, codes to improve scope for analysis.

Results

A harmonized dataset of 2166 participants and 5747 hospital admissions were derived of cohort members who had at least 1 hospital inpatient or A&E event before their 14th birthday. Harmonisation included: dealing with date granularity by generating random dates of birth; standardising periods of data collection; identifying inconsistencies and then mapping and bridging differences in definitions of periods of care and levels of diagnostic and operational coding across countries and datasets.

Conclusion/Implications

Heterogeneous variables from different data sources were pooled and converted into standardised data for research, extending existing harmonisation work, including curation of a population based anonymously linkable longitudinal cohort. [AA1] These methods are reproducible and can be utilised by other researchers and projects applying to use these routine data sources.

http://dx.doi.org/10.23889/ijpds.v3i4.750