



Swansea University
Prifysgol Abertawe



Cronfa - Swansea University Open Access Repository

This is an author produced version of a paper published in:
International Journal of Population Data Science

Cronfa URL for this paper:
<http://cronfa.swan.ac.uk/Record/cronfa44390>

Paper:

Bandyopadhyay, A., Tingay, K., Borja, M., Griffiths, L., Akbari, A., Bedford, H., Brophy, S., Walton, S., Dezateux, C. et. al. (2018). Harmonising data from different sources to conduct research using linked survey and routine datasets. *International Journal of Population Data Science*, 3(4)
<http://dx.doi.org/10.23889/ijpds.v3i4.750>

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

This item is brought to you by Swansea University. Any person downloading material is agreeing to abide by the terms of the repository licence. Copies of full text items may be used or reproduced in any format or medium, without prior permission for personal research or study, educational or non-commercial purposes only. The copyright for any work remains with the original author unless otherwise specified. The full-text must not be sold in any format or medium without the formal permission of the copyright holder.

Permission for multiple reproductions should be obtained from the original author.

Authors are personally responsible for adhering to copyright and publisher restrictions when uploading content to the repository.

<http://www.swansea.ac.uk/library/researchsupport/ris-support/>

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

Bandyopadhyay, A¹, Tingay, K¹, Borja, MC², Griffiths, L², Akbari, A³, Bedford, H², Brophy, S⁴, Walton, S², Dezateux, C⁵, and Lyons, R⁶

¹Swansea University

²Institute of Child Health, University College London

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

⁴Swansea University Medical School

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

⁶Farr 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.

