

International Journal of Population Data Science

Journal Website: www.ijpds.org



Making using administrative data easier: How To Guides and the Phenotype Code List Repository for ECHILD

Matthew Jay¹, Kate Lewis¹, Difei Shi¹, Rebecca Langella¹, Tony Stone^{1,2}, Sorcha Ní Chobhthaigh¹, Ania Zylbersztejn¹, Ruth Blackburn¹, and Katie Harron¹

¹University College London, London, United Kingdom

²University of Sheffield, Sheffield, United Kingdom

Objectives

Even experienced researchers face difficulties in finding and using administrative data and producing analysis-ready datasets. Here we describe two resources developed to support users of the Education and Child Health Insights from Linked Data (ECHILD) project: How To Guides and the Phenotype Code List Repository.

Methods

We developed two additional websites that host the How To Guides and, separately, the Code List Repository, to supplement the main website and user guide for ECHILD. We used Quarto markdown with source code freely accessible via GitHub. Both resources were designed according to an overarching set of principles that ensure that content and example code is as clear and simple as possible, and fully reproducible. The Code List Repository seeks to make finding and implementing phenotype code lists easy, by drawing on published code lists, ensuring consistent formatting, and providing machine-readable files.

Results

These are updateable resources designed to facilitate transparent research in line with open science principles. The How To Guides take users through a series of steps to produce an analysis-ready dataset: identifying a cohort from hospital birth admissions or school enrolments, extracting data to produce a cohort spine, extracting and cleaning data on exposures, outcomes and covariates, and merging it all together. We provide R scripts explained in detail, with guidance for workspace and code management. The Code List Repository is a fully open and searchable database of phenotype code lists for ECHILD and beyond. It includes a primer on phenotyping in administrative data and a brief introduction to coding systems used. It, too, comes with scripts that users can adapt to their own projects.

Conclusion

The How To Guides and Phenotype Code List Repository are an example of how to provide guidance on using linked administrative datasets for research. We hope that members of the ECHILD community will contribute to improving code and documentation as more researchers gain access to the data.

