

International Journal of Population Data Science

Journal Website: www.ijpds.org



Navigating Data Acquisition and Data Quality Validation of Large Databases: Practical Lessons

Nedeene Hudema¹, Amanda Hutton¹, Nirmal Sidhu¹, Xinya Lu¹, and Xue Feng¹

¹Saskatchewan Health Quality Council

Objective and Approach

This paper discusses the Data Linkage Hub (DLH) which is a centre for data linkage excellence and aims to produce and support the delivery of high quality, linked data assets for researchers in the UK. Here, we will give an overview of the DLH's activity as a Community of Practice for Data Linkage.

Results

We will discuss the different project types conducted in the DLH; the methods used ranging from bespoke linkage development to the adoption of generalised methodologies; and the customers of the DLH.

We will also discuss the tools the DLH have developed, and make available to others, to standardise, streamline and improve quality assessment of linkage. This will include clerical matching tools known as CROW and the CMS, BAT, a bias analysis tool, and a package of PySpark code specifically developed for data linkage.

In addition, we will explore the data linkage training materials we offer for different audiences and our involvement in a Data Linkage Champions Network, both increasing the awareness of data linkage and promoting best practise.

Conclusion

Investigating errors, deciding level of acceptance of errors, managing relationships, and communication are the most critical aspects of the work. Contingency plans are important when it takes longer than expected. Data are never perfect, but a thoughtful approach to validation and documentation of limitations and decisions support using the data appropriately. Also, a kind demeanor, an open mind, and a sense of humour can go a long way!

