

International Journal of Population Data Science

Journal Website: www.ijpds.org



Using administrative datasets to improve linkage rates and quality of PII in the Generation Victoria cohort

Sharon Goldfield¹, Jatender Mohal¹, Anthony Del Monaco¹, Anna Fedyukova¹, Adam Ismail², Philip Witowski², Adam Moss², Sharon Williams², and Melissa Wake¹

¹Murdoch Children's Research Institute, Melbourne, Australia

²Centre for Victorian Data Linkage, Melbourne, Australia

Maintaining accurate personal identifying information (PII) is a challenge for large population cohorts. Over time, identifier quality declines, reducing linkage sensitivity and increasing false matches and manual clerical review. Generation Victoria (GenV) is addressing this by systematically enhancing cohort PII through linkage with trusted administrative sources, in partnership with the Centre for Victorian Data Linkage (CVDL). GenV is a whole-of-state, consented, cell-to-society cohort comprising ~50,000 children born in Victoria (2021–2023) and ~75,000 parents. GenV and CVDL co-developed a staged PII enhancement and linkage framework to improve identifier completeness, accuracy and downstream linkage performance. In Stage 1, GenV PII was validated and updated through CVDL linkage with the Victorian Birth Registry, enabling correction of core identifiers, identification of previous or alternative names and confirmation of vital status. In Stage

2, this enriched PII dataset linked participants to state administrative datasets. Deterministic and probabilistic linkage rates were calculated, applying standard validation and quality assurance to outputs. PII enhancement markedly improved identifier quality and linkage performance. Core identifier completeness increased by 1–10% (eg imprecise name from 89% to 98% for babies). Datasets that previously failed quality thresholds, including pathology providers and hospital systems, were successfully linked. Systematic, registry-based PII enhancement developed collaboratively between a population cohort and a trusted linkage unit can substantially improve linkage quality, efficiency and scalability in large cohorts experiencing PII decay. Improved identifier quality benefits participants, researchers and data custodians by increasing confidence in linkage outputs while reducing operational burden, with our approach transferable to other studies.

