

International Journal of Population Data Science

Journal Website: www.ijpds.org



Gold Standard Identifiers within Linkage

Michael Edwards^{1,2}, Anahita Zamani^{1,2}, and Simon Thompson^{1,2}

¹SeRP, Swansea, United Kingdom

²Swansea University, Swansea, United Kingdom

Explore utilisation of “gold standard” personal identifiers within linkage processes, focusing on quality of NHS numbers within a representative sample of the Welsh population. We provide insight into reliability and impact on metrics of accuracy for linkages produced, including the need to consider merging or splitting such gold standard IDs.

A longitudinal, population-scale sample of Welsh individuals was internally linked using deterministic models, with an NHS number presumed as a “gold standard” personal identifier. 18,906,276 records were analysed, with 5,654,212 distinct “ground truth” individuals. Initial linkage was performed using only the gold standard ID, whilst comparison models required exact matching on name, date of birth, address, and postcode to produce novel links not identified by the gold standard, indicating proposed ID merger based on tight overlap of record-level data. Performance metrics were compared against the presumed ground truth, with cluster quality metrics identifying instances where merging of clusters should occur.

An initial deterministic model built on name, date of birth, and postcode exact match produces 5,651,983 clusters containing a single distinct ID, 1,113 clusters having 2 distinct IDs, and a single cluster containing 3 distinct IDs. Tightening rules to only consider those also matching on the first line of address results in 5,652,125 clusters containing a single distinct ID, 1,042 clusters having 2 distinct IDs, and a single cluster containing 3 distinct IDs. Cases of being allocated secondary IDs, where multiple gold standard IDs should be merged to a singular ID, can be easily rectified by the linkage process. Conversely, individuals being erroneously allocated onto another’s ID, causing linkage to produce strongly weighted bridging links, is more problematic to split even with probabilistic linkage methods.

“Gold standard” IDs still fall foul of classic data quality issues and their impact should be carefully considered depending on the application. Lower metrics against a ground truth often reflect not just model performance but also underlying data collection errors or systemic discrepancies, suggesting areas for further refinement and research.

