

International Journal of Population Data Science

Journal Website: www.ijpds.org



Diversity of record pairs and mental models in clerical review: What reviewers use to decide matches

Yashithi Dharmawimala¹, Peter Christen², Sumayya Ziyad¹, Joseph Lam³, Anushka Vidanage¹, and Rainer Schnell⁴

¹Australian National University, Canberra, Australia

²Australian National University, Canberra, Australia; University of Edinburgh, Edinburgh, United Kingdom

³University College London, London, United Kingdom

⁴University of Duisburg-Essen, Duisburg, Germany

Clerical review is an integral step of many data linkage systems. It employs human expertise to assess record pairs for which a decision model, such as probabilistic record linkage, has not been able to make a match (two records refer to the same individual) or non-match (two records refer to different individuals) decision. While clerical review is used in many practical applications, there is surprising little research that systematically investigates how to best conduct this important step in the data linkage process. In this work, we explore one aspect of clerical review: How does the diversity of the record pairs selected for manual review affect the final linkage quality? To investigate this question, we generated diverse samples of difficult to classify record pairs from a large public population database, where each sample contained 100 pairs

of true matches and 100 pairs of true non-matches. The samples differed in (1) the number of unique similarity patterns they had (calculated over a set of compared quasi-identifiers such as names and addresses), and (2) the actual set of quasi-identifiers available for manual review. All authors conducted a blinded manual assessment of all samples. Our results indicate that the diversity of the agreement patterns is less important for making informed decisions compared to the set of quasi-identifiers available for review. Furthermore, reviewers seem to make decisions based on mental models of how they think a match or non-match looks like. Our results will help design improved approaches for clerical review in practical linkage applications.

