International Journal of Population Data Science

Journal Website: www.ijpds.org

The influence of ethnic name characteristics on string similarities when linking large heterogeneous population databases

Yashithi Dharmawimala[1], Peter Christen[2], Sumayya Ziyad[1], Anushka Vidanage[1], Joseph Lam[3], and Rainer Schnell[4]

[1]Australian National University, Canberra, Australia

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

[3]University College London, London, United Kingdom

[4]University of Duisburg-Essen, Duisburg, Germany

Comparing names is at the core of many data linkage applications. Names can be recorded with variations and errors (typographical, phonetic, or scanning, depending on how names have been captured). Furthermore, names can change over time. Therefore, approximate string comparison functions are commonly employed to calculate similarities between names. There are many such comparison functions, with popular ones including Jaro-Winkler, edit distance, and q-gram based techniques. Which ones are suitable for a given data linkage application is a topic that has not been explored in much detail. Selecting a suitable string comparison function is, however, crucial in the context of bias and fairness when linking population databases that can contain ethnically diverse names, as recent research has shown (for example, see Lam et al,. IPDLN 2024).In this work, we investigate the following question: Do different string comparison functions result in different levels of bias when names from different ethnic groups are being compared? We conducted extensive experiments on a large public population database where the ethnicity of each record is available. By comparing name variations of the same person using multiple string comparison functions, we show that some of these functions exhibit larger similarity ranges than others, while very different similarities are obtained for the same name pair depending upon which comparison function is employed. Our results provide important insights into how practical data linkage of large population-level databases should be conducted in order to limit linkage bias with regard to ethnic groups by employing suitable string comparison functions.

https://doi.org/10.23889/ijpds.v11i5.3756