

International Journal of Population Data Science

Journal Website: www.ijpds.org



Mitigating Ethnoracial-Based Linkage Bias with Feature and Frequency-Aware Methods

Joseph Lam¹, Mario Cortina-Borja¹, Ruth Blackburn¹, and Katie Harron¹

¹UCL Great Ormond Street Institute of Child Health, London, United Kingdom

Record linkage using names as identifiers is central to population data integration yet may systematically disadvantage minoritised ethnoracial groups. We investigate how ethnoracial-based differences in name characteristics and error distributions interact with probabilistic linkage models to produce unequal missed match rates, and whether feature-based methods can mitigate this bias. Using 8.7 million records from the North Carolina Voter Registry, we derive race-specific error profiles and simulate three data corruption scenarios, including a biased setting with higher corruption among minoritised groups. We compare three linkage strategies within a mixed, cluster-based framework: (i) conventional probabilistic linkage using Jaro–Winkler and Levenshtein similarities; (ii) rarity-aware variants with term-frequency (TF) adjustment; and (iii) a proposed name-feature model combining principal-component embeddings of first-name characteristics

with TF-adjusted last-name similarity. Across all settings, we fix the overall false negative rate at 20% and quantify group-specific false match rates differences relative to Non-Hispanic White voters. Baseline probabilistic models exhibit clear ethnoracial bias mechanisms: Non-Hispanic White individuals consistently have the lowest false negative rates, with larger missed-match penalties for Non-Hispanic Black, Asian, and Hispanic groups under race-specific error patterns. Term-frequency adjustment reduces but does not eliminate ethnoracial disparities. The name-feature model yields the most equitable outputs, minimising false negative rate gaps across racial groups in the most realistic corruption scenario, with modest increases in false match rate. Distribution-aware name features and rarity adjustment can be embedded in cluster-based linkage to reduce ethnic bias in name-based linkage without substantial loss of accuracy.

