

International Journal of Population Data Science

Journal Website: www.ijpds.org



Alternative Name Encodings - Using Jyutping or Pinyin as tonal representations of Chinese names for data linkage

Joseph Lam¹, Mario Cortina-Borja¹, Robert Aldridge², Ruth Blackburn¹, and Katie Harron¹

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

²Institute for Health Metrics and Evaluation, University of Washington, Seattle, USA

Objectives

Accurate data linkage across large administrative databases is crucial for addressing complex research and policy questions, yet linkage errors—stemming from inconsistent name representations—can introduce biases, predominantly for names not given in English. We identify three primary issues in processing non-English names not sufficiently considered by data linkers: language-specific variations in romanisation, the loss of tonal information inherent to tonal languages, and discrepancies in name order conventions.

Methods

This work examines the impact of romanisation on linkage accuracy, focusing on Chinese names and comparing standardised systems (Jyutping and Pinyin) with the non-standardised Hong Kong Government Cantonese Romanisation (HKG-romanisation). Using a dataset of 771 Hong Kong student names, we derived Jyutping and Pinyin using the `pinyin_jyutping` package in Python 3.8, which used an online open-source Cantonese dictionary CC-Canto, for each character of the Chinese name. We compared how closely three different systems of romanisation (HKG-romanisation, Jyutping, Pinyin) represented the original Chinese characters, in terms of uniqueness, which provided some information on the utility of these systems for balancing sensitivity and specificity of linkages.

Results

Our analysis reveals that standardised romanisation systems enhance the uniqueness and consistency of name representations, thereby improving linkage precision and recall compared to HKG-romanisation. Specifically, Jyutping and Pinyin achieved over 95% recall in blocking strategies, whereas HKG-romanisation only reached 68.8%. We explored tonal distribution of Jyutping and Pinyin in our sample. Incorporating tonal information further improved recall, and hold potential for character re-encoding when developed using a larger, more representative Chinese name database.

Conclusion

These findings underscore the necessity of adopting standardised, tone-sensitive romanisation systems and flexible database designs to reduce linkage errors and promote data equity for under-represented groups. We advocate for the implementation of phonetic encodings in databases, alongside language-specific pre-processing protocols, to ensure more inclusive and accurate data linkage processes.

