International Journal of Population Data Science

Journal Website: www.ijpds.org

IJPDS International Journal of Population Data Science

Swansea University Prifysgol Abertawe

Impact of linkage quality on inferences drawn from analyses using imperfectly matched data with high rates of linkage errors

Rentsch, C[1], Reniers, G[1], Harron, K[2], Machemba, R[3], Mtenga, B[3], Michael, D[3], Kabudula, C[4], Natalis, R[3], Urassa, M[3], Todd, J[3], and Zaba, B[1]

[1]London School of Hygiene & Tropical Medicine

[2]University College London

[3]National Institute for Medical Research

[4]University of the Witwatersrand

Introduction

Studies based on high-quality linked data in developed countries show that residual linkage errors impact the bias and precision of subsequent analyses. Since 2015, we conducted point-of-contact interactive record linkage (PIRL) between serological survey data and manually digitised medical records with low data quality from three clinics in rural Tanzania.

Objectives and Approach

We sought to determine the impact of the substantial linkage errors made by automated probabilistic linkage (a commonly used, less accurate, but much cheaper alternative to PIRL) on the bias and precision of inferences drawn from Cox regression analyses, comparing time from a positive HIV diagnostic test to registration at a local HIV care and treatment clinic (CTC) by testing modality (sero-survey vs. clinic). Using PIRL links as the gold standard, we quantified false/missed matches, compared characteristics between linked and unlinked data, and evaluated regression estimates at low, medium, and high (25th, 50th, and 75th percentile) match score thresholds.

Results

Between 2015-2017, 297 and 147 individuals with gold standard links received HIV+ test results in sero-surveys and clinics, respectively. Automated probabilistic linkage correctly identified 276 individuals (positive predictive value [PPV]=62%) at the low threshold and 43 individuals (PPV=96%) at the high threshold. At the lowest threshold, false matches were more likely to be clinic testers and less likely to register at CTC. These differences attenuated with increased threshold. Testing modality was significantly associated with time to CTC registration in the gold standard data (adjusted hazard ratio [HR] 6.42, 95%CI 4.45-9.28). Increasing false matches progressively weakened the association (low threshold: HR 4.99, 95%CI 3.45-7.21). Increases in missed matches were strongly correlated with a reduction in the precision of coefficient estimates (R-squared=0.94; p=0.0001).

Conclusion/Implications

While the significance of inferences did not change, a clear direction of bias was identified. High rates of false matches in this setting reduced the magnitude of the association; missed matches reduced precision. Adjusting for these biases could provide more robust results using data with considerable linkage errors.

http://dx.doi.org/10.23889/ijpds.v3i4.605