

International Journal of Population Data Science

Journal Website: www.ijpds.org



Enhancing integration of administrative databases in South Africa's HIV program: Validation of record linkage using non-representative gold standard

Evelyn Lauren^{1,2}, Dickman Gareta^{3,4}, Khumbo Shumba², Dorina Onoya², and Jacob Bor^{2,3}

¹Department of Biostatistics, Boston University School of Public Health

²Health Economics and Epidemiology Research Office

³Africa Health Research Institute

⁴Institute of Social and Preventive Medicine, University of Bern

Introduction

Linked administrative data are widely used in epidemiology to capture patient data across multiple databases. Linkage error rates, critical to measure linkage performance, are rarely reported due to difficulty in obtaining representative gold standard. We propose a training and validation approach for linkage procedures that yield unbiased performance estimates even with a non-representative gold standard.

Implications

Failure to adjust for informative missingness in the gold standard may lead to biased validation metrics and over/underconfidence in linked data.

Methods

We linked patient records from two non-deduplicated databases for HIV monitoring in South Africa, TIER.Net and NHLS laboratory database, using a network-based probabilistic linkage and deduplication approach. National IDs (gold standard) were available for a non-representative minority of records (10%). We calculated sensitivity (Sen, share of true matches identified by the algorithm) and positive predictive value (PPV, share of algorithm-identified matches that were true matches). We adjusted for bias due to informative missingness in National IDs using inverse probability weights to break the link between missingness and match probability.

Results

111,755 record pairs were considered. National IDs were not missing completely at random. Match probabilities for National ID record pairs exhibited substantially less uncertainty (mid-range match probabilities), inflating Sen and PPV. Before bias correction, Sen and PPV were estimated at 97.0% and 97.8% respectively. After bias correction for missing National IDs, Sen and PPV were estimated at 95.7% and 96.6%. Failure to address this bias understated the overlinkage rate (100% - PPV) by 35% and the underlinkage rate (100% - Sen) by 30%.

