

International Journal of Population Data Science

Journal Website: www.ijpds.org



The National HIV Cohort of South Africa's National Health Laboratory Service (NHLS): tracking the first 20 years of SA's HIV response

Jacob Bor¹, Evelyn Lauren², Trevor Bell³, Siyabongal Zondi³, Alana Brennan⁴, William MacLeod⁵, Matthew Fox⁴, Mhairi Maskew⁶, Musa Manganye⁷, and Mazvita Muchengeti³

¹Health Economics and Epidemiology Research Office, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa; Africa Health Research Institute, Somkhele, South Africa; Department of Global Health, Boston University School of Public Health, Boston, USA; Department of Epidemiology, Boston University School of Public Health, Boston, USA

²Health Economics and Epidemiology Research Office, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa; Department of Biostatistics, Boston University School of Public Health, Boston, USA

³National Institute for Communicable Diseases, National Health Laboratory Service, Johannesburg, South Africa

⁴Health Economics and Epidemiology Research Office, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa; Department of Global Health, Boston University School of Public Health, Boston, USA; Department of Epidemiology, Boston University School of Public Health, Boston, USA

⁵Health Economics and Epidemiology Research Office, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa; Department of Global Health, Boston University School of Public Health, Boston, USA

⁶Health Economics and Epidemiology Research Office, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa

⁷HIV/AIDS and STIs Cluster, National Department of Health, Pretoria, South Africa

Background

South Africa's public sector HIV program, the largest globally, relies on the National Health Laboratory Service (NHLS) for all laboratory monitoring. However, the absence of consistent unique identifiers has limited longitudinal analyses. We previously constructed the NHLS National HIV Cohort by probabilistically linking HIV test results through 2018. Here, we report on our updated linkage of >388 million laboratory records from 2004–2023.

Methods

We implemented graph-based probabilistic linkage, using national IDs (available for 15% of records) for training and validation. Record similarity was assessed using a modified Fellegi-Sunter framework based on demographics, with multiple blocking strategies and parallel processing to handle scale. As individuals form cluster of records, we applied a graph-informed thresholding approach that penalizes implausibly large clusters, reducing false merges over traditional pairwise linkage.

Results

Linkage generated 10 billion candidate pairs. We identified 17.2 million unique patient clusters out of 110 million HIV-related records. Performance relative to the national ID subset yielded 97.2% sensitivity (SEN) and 94.4% positive predictive value (PPV). Within the whole dataset, we estimated 90.4%

SEN and 91.9% PPV, accounting for the larger population and the non-representativeness of national IDs. Among patients with viral load (VL) between July 2021–June 2022, 18% had no follow-up VL within 18 months, indicating potential care attrition.

Conclusion

The NHLS National HIV Cohort offers the first longitudinal patient-level analysis covering the first 20 years of SA's national HIV programme. Our findings highlight the promise of probabilistic record linkage for epidemiology and the importance of robust validation strategies.

