

Factors associated with Sexually transmitted infection literacy among men who have sex with men and transgender men and women in Soweto: A machine learning approach

Background: Sexually transmitted infections (STIs) are a big health problem and machine learning algorithms are used to report STI literacy [1]. In SSA, the prevalence of STIs remains alarmingly high, particularly among key populations (KPs) such as sex workers, men who have sex with men, transgender individuals, and people who inject drugs [3]. Key populations are a vulnerable group who can benefit from STI literacy which is essential for promoting informed and empowered sexual choices. Understanding STI literacy levels is essential for addressing STI challenges at the grassroots level, enabling the development of targeted and effective interventions.

Objective: To identify factors associated with STI literacy among MSM and transgender men and women in Soweto using machine learning algorithms.

Methodology: A retrospective observational study involving 1240 MSM and transgender men and women from Soweto, aged 18 years and above was conducted using secondary data collected through self-administered questionnaires by Best Health Solutions. Data cleaning and analysis were conducted using R version 4.4.1 (released on 2024-06-14), employing seven machine-learning algorithms: XGBoost, AdaBoost, Naive Bayes, Decision tree, Random Forest, Logistic regression, and K-Nearest Neighbors. Data were split into training (80%) and testing (20%) sets [3]. The outcome variable was based on the question: "Have you ever received counselling or education about STI diagnosis and results interpretation?" This variable was a binary factor with "Yes" and "No" representing STI literacy. Model evaluation was based on metrics including accuracy, precision, and AUC. We employed repeated 10-fold cross-validation with 100 repetitions to ensure reliable performance estimates [4]. Metrics such as AUC (ROC), accuracy, sensitivity (recall), and precision were used to evaluate the model's classification ability [5]. The final model was assessed based on its accuracy in distinguishing between individuals with and without STI literacy. Variable importance was calculated to highlight the most influential predictors [6].

Preliminary results: Random Forest was the best-performing model due to its high AUC, F1 score, and robustness against overfitting. However, SVM was the second best and can provide the highest separation ability while XGBoost remains a close competitor and may be preferable in applications where predictive precision is a priority. The final model was assessed based on its accuracy in distinguishing between individuals with and without STI literacy.

Education level, employment status and accessibility to STI testing and treatment services emerged to be the most influential factors in determining STI literacy while marital status and age were found to be the least significant.

Conclusions and recommendations: Advanced ML techniques can be used to establish STI literacy among key populations for comprehensive targeted interventions including sexual health education. Education level, employment status and accessibility to STI testing and treatment services, popularity of STIs in the community, sexual orientation, perceived high risk of contracting STI in the community, confidence in STI knowledge, comfortable in discussing STI status, marital status and age were found to be important in predicting STI literacy. Machine learning techniques can be developed to classify and predict STI literacy based on these factors, enabling informed decision-making and improving sexual health outcomes. The findings may help to address critical knowledge gaps and misconceptions about STIs, offering a strong foundation for developing tailored educational interventions [7].

The work reported herein was made possible through funding by the South African Medical Research Council (SAMRC) Project Code #57035 (SAMRC File ref no: HDID8528/KR/202) through its Division of Research Capacity Development under the Mid-Career Scientist Programme through funding received from the South African National Treasury. The content hereof is the sole responsibility of the authors and does not necessarily represent the official views of the SAMRC. This research was funded, in part, by the U.S. National Institute of Allergy and Infectious Diseases under award R01AI170249.

Primary author: Dr SIYAMAYAMBO, Claris (University of Johannesburg)

Co-authors: Prof. PHASWANA-MAFUYA, Refilwe Nancy (1South Africa Medical Research Council/ University of Johannesburg (SAMRC/UJ) Pan African Centre for Epidemics Research (PACER) Extramural Unit, Faculty of Health Sciences, South Africa); Dr PHALANE, Edith (1South Africa Medical Research Council/ University of Johannesburg (SAMRC/UJ) Pan African Centre for Epidemics Research (PACER) Extramural Unit, Faculty of Health Sciences, South Africa)

Presenter: Dr SIYAMAYAMBO, Claris (University of Johannesburg)