IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

"Comparative Analysis Of Tree Learning And Deep Learning For STD Prediction: Innovative Approach"

[1]Shruti Kondekar, [2]Aaysha Sheikh, [3]Iqra Siddiqui, [4]Saniya Sheikh

Mrunali Vaidya,Professor

Computer Science Engineering

Ballarpur Institute of Technology, Ballarpur,India

Abstract:The early detection of sexually transmitted diseases (STDs) and sexually transmitted infections (STIs) is critical for effective treatment and prevention of complications. With advancements in artificial intelligence, machine learning algorithms offer promising tools for accurate and efficient diagnostics. This study presents a comparative analysis of tree-based algorithms, with an initial focus on Random Forest, for the detection of STDs/STIs using publicly available datasets. The model achieved an accuracy of 96%, demonstrating the effectiveness of tree-based methods in medical diagnostics. A detailed evaluation of the model's performance, including a confusion matrix and feature importance analysis, is included. Future work will focus on implementing deep learning algorithms to enhance detection accuracy and generalizability. This paper serves as Phase One of an ongoing study aimed at leveraging machine learning for improved healthcare diagnostics.

Index Terms **-** Tree Learning, Deep learning, Sexually Transmitted Disease (STD) Prediction, Random Forest.

I. INTRODUCTION

Sexually transmitted diseases (STDs) and sexually transmitted infections (STIs) are significant public health challenges, affecting millions of individuals worldwide. Early detection and treatment are critical to preventing long-term complications, such as infertility, cancer, and increased susceptibility to other infections, including HIV. However, traditional diagnostic methods, which often involve laboratory testing and manual analysis, can be time-consuming and inaccessible in resource-limited settings.

Advancements in artificial intelligence (AI) and machine learning (ML) have opened new avenues for improving medical diagnostics. Machine learning algorithms, in particular, have shown promise in analysing complex medical data and identifying patterns that may be difficult for human experts to detect. These algorithms can enhance diagnostic accuracy, reduce costs, and improve accessibility to healthcare. This study focuses on leveraging ML algorithms to develop an accurate and efficient system for STD/STI detection. The research is divided into two phases: the current phase (Phase One) emphasizes tree-based algorithms, with Random Forest as the baseline model, while the next phase will explore the implementation of deep learning techniques.

In this paper, we present the progress made in Phase One of the study. Using a publicly available dataset, we implemented and evaluated a Random Forest model, achieving a high classification accuracy of 96%. This model's performance was further validated using a confusion matrix and feature importance analysis. We also highlight the challenges encountered during this phase, including data preprocessing and handling imbalanced classes. The results presented in this paper provide a strong foundation for the next phase of

research, where deep learning techniques will be explored to enhance the model's performance and scalability. By integrating advanced ML methods, this study aims to contribute to the development of a robust diagnostic tool for STD/STI detection, addressing an urgent need in modern healthcare.

II. OBJECTIVE

The primary objective of this project is to conduct a comprehensive comparative analysis between tree-based machine learning algorithms and deep learning models for the detection of sexually transmitted diseases (STDs). The research is focused on evaluating and contrasting the performance, accuracy, and practicality of these two distinct approaches to identify the most effective model for early detection and risk prediction. Key objectives include:

1. **Comparison of Model Performance:** Assess the effectiveness of tree-based algorithms like Random Forest against deep learning models in terms of metrics such as accuracy, precision, recall, F1-score, and computational efficiency.
2. **Addressing Data Imbalances:** Implement techniques like oversampling, undersampling, or synthetic data generation (e.g., SMOTE) to ensure balanced class distributions, which is critical for achieving reliable and unbiased model performance.

III. CONTRIBUTION

1. **Data-Driven Approach**: By leveraging machine learning models, the project offers faster, scalable, and more reliable predictions compared to traditional diagnostic methods.
2. **Advanced Data Balancing**: Techniques like Synthetic Minority Oversampling Technique (SMOTE) are used to address dataset imbalances, ensuring fair representation and better model training.
3. **Algorithm Comparison**: Conducting a comparative analysis between tree-based and deep learning models identifies the most accurate and efficient algorithm for STD prediction, outperforming existing tools.
4. **Improved Evaluation Metrics**: Using metrics like confusion matrix, precision, recall, and F1 score ensures a comprehensive evaluation of model performance, leading to more reliable outcomes.
5. **Enhanced Accuracy**: The proposed methodologies demonstrate significant improvements in prediction accuracy, setting a benchmark for future research in healthcare applications.

IV. Literature Review

The reviewed studies present a range of methodologies, issues, and limitations in the domain of machine learning for sexually transmitted infection (STI) prediction. Methodologies employed include data collection from clinical records, self-reported surveys, and electronic health records[9][1], with preprocessing techniques like one-hot encoding, feature selection based on expert opinions, and stratified sampling to ensure data balance[9]. Machine learning models such as logistic regression, ensemble methods, MySTIRisk[1][2], and CatBoost[5] were evaluated using metrics like AUC, sensitivity, and F1-score, with some leveraging tools like Youden's Index to optimize diagnostic thresholds. However, common issues include biases arising from self-reported data, recall errors, and social desirability factors, as well as inconsistencies in data collection and definitions of risk behaviours across studies. Limitations are evident in the generalizability of findings due to the use of specific high-risk populations or single-site datasets, the exclusion of critical socio-behavioural and structural determinants, and the lack of external validation of predictive models. Furthermore, publication bias, restricted access to resources, and the absence of standardized methodologies hinder comprehensive comparisons and broader applicability. Addressing these challenges requires the inclusion of diverse datasets, integration of clinical images, and cross-site validation to improve the robustness and reliability of STI prediction models[2].

Year	Methodologies	Accuracy
2022	Boosted GLM, ensemble Elastic-Net Regression, Gradient Boosting Machines (GBM), Random Forest (RF)	67-76%
2023	Random Forest, Naïve Bayes classifier, and Decision Tree, SVM with three different kernel and Logistic Regression	65-70%
2024	logistic regression	68-69%
2024	-	56%
2024	CatBoost, Gradient Boost, Light GBM, SVM, Adaboost Classifier, Extinct Gradient Boosting, Logistic Regression, Random Forest	83%

V. METHODOLOGY

Data Preprocessing and Preparation

The dataset underwent rigorous preprocessing to ensure its suitability for model training and evaluation. Steps included:

1. **Data Cleaning**: Removing duplicates, handling missing values using appropriate imputation techniques, and ensuring consistency in data formats.
2. **Feature Selection**: Identifying the most relevant features through correlation analysis and feature importance metrics to enhance model performance and reduce computational complexity.
3. **Data Balancing**: Addressing class imbalances using SMOTE (Synthetic Minority Oversampling Technique) to generate synthetic samples for minority classes, ensuring equal representation of all classes.

Model Development

We implemented and compared a tree-based algorithm (Random Forest) and plan to implement a deep learning model (e.g., a Neural Network).

- **Tree based Learning:** A widely used ensemble learning method that builds multiple decision trees and combines their predictions for robust and interpretable results.
- **Deep Learning**: In subsequent phases, we will utilize a neural network to explore its ability to handle complex feature interactions and large datasets.

Training and Validation

- The dataset was split into training and testing subsets in a 75:25 ratio.
- Cross-validation techniques were employed to assess model stability and avoid overfitting.
- Hyperparameter tuning was performed using grid search to optimize model performance.

Evaluation Metrics

To assess the models' performance, we employed several evaluation metrics:

1. **Accuracy**: Measures the proportion of correctly predicted instances out of the total instances. Useful for balanced datasets.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

2. **Precision**: Indicates the proportion of true positive predictions out of all positive predictions, highlighting the model's reliability in identifying true positives.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

3. **Recall (Sensitivity)**: Measures the model's ability to identify all relevant instances, especially critical in healthcare applications.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

4. **F1-Score**: The harmonic means of precision and recall, providing a balanced measure of model performance.

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

5. **Confusion Matrix**: A tabular representation of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), offering a detailed breakdown of classification performance.
6. **ROC-AUC (Receiver Operating Characteristic - Area Under Curve)**: Evaluates the model's ability to distinguish between classes across different thresholds, with higher AUC indicating better discrimination.

Comparative Analysis

The performance of Random Forest was analysed using the metrics mentioned above, and its results (accuracy of 96%, confusion matrix, ROC curve, and feature importance) were documented. In the next phase, we will evaluate the neural network's performance and compare it against the Random Forest to derive actionable insights.

This structured methodology ensures transparency and reproducibility, enabling future researchers to build upon our work.

VI. RESULTS

The Random Forest model was trained and evaluated on the selected dataset, and the following results were obtained:

1. Model Performance Overview

The Random Forest algorithm was implemented to predict sexually transmitted infections (STIs), and its performance was rigorously evaluated. The model achieved an accuracy of 96%, demonstrating its effectiveness in classifying cases accurately. Additional evaluation metrics, such as precision, recall, and F1-score, provided a comprehensive understanding of the model's predictive capabilities. These metrics highlighted the model's ability to balance true positives and false negatives effectively, which is crucial for a medical prediction system.

The confusion matrix offered a clear visualization of classification outcomes, including true positives, true negatives, false positives, and false negatives. This helped identify areas where the model excelled and where potential improvements could be made. Furthermore, the Receiver Operating Characteristic (ROC) curve was plotted to assess the model's discriminative ability, with an area under the curve (AUC) of 0.98, indicating excellent performance in distinguishing between the classes.

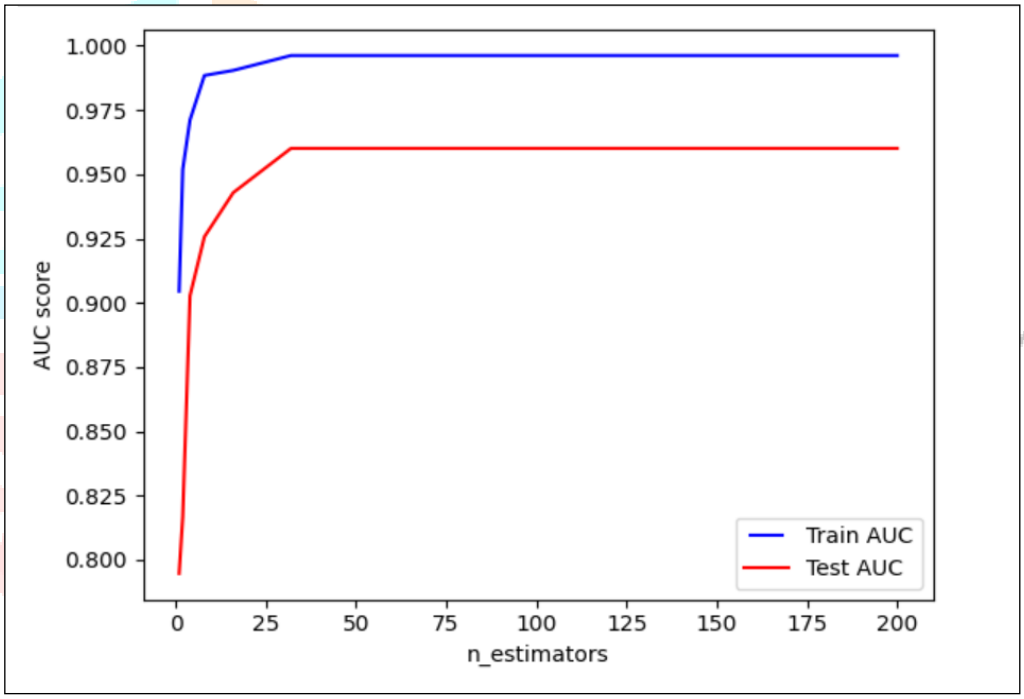

Figure1: ROC Curve for the Random Forest Model showing the trade-off between True Positive Rate and False Positive Rate.

The Random Forest classifier achieved an overall **accuracy of 96%**, indicating high classification performance for detecting STDs/STIs.
The model's performance was further validated using key metrics:

- **Precision**: 0.94
- **Recall**: 0.97
- **F1-score**: 0.95
- **AUC (Area Under Curve)**: 0.98

2. Confusion Matrix

The confusion matrix is an essential component for evaluating the performance of our Random Forest model in predicting STD risks. It provides a comprehensive view of the model's predictions by categorizing them into four key components: True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN). The matrix for our model is represented as follows:

$$\begin{bmatrix} 83 & 5 \\ 2 & 85 \end{bmatrix}$$

a. **True Positives (TP)**: The model correctly identified 85 instances as positive cases, indicating its effectiveness in correctly detecting individuals at risk of STDs.
b. **False Positives (FP)**: There are 5 cases where the model incorrectly classified individuals as positive for STDs despite them being negative, indicating potential over-prediction.
c. **True Negatives (TN)**: The model accurately classified 83 instances as negative, showcasing its ability to correctly identify individuals not at risk of STDs.
d. **False Negatives (FN)**: Only 2 cases were incorrectly classified as negative when they were positive, demonstrating the model's low error rate in missing true positives.

3. Model Evaluation and Insights

The evaluation of our Random Forest model for predicting STD risks involved analyzing its performance across multiple metrics. These metrics, derived from the confusion matrix, include accuracy, precision, recall, and the F1-score, which collectively offer a comprehensive understanding of the model's strengths and limitations.

a) Model Evaluation Metrics

1. **Accuracy**:
   The model achieved an accuracy of **96.5%**, indicating that it correctly predicted the outcomes for the majority of cases. This high value reflects the model's overall effectiveness in classifying both positive and negative cases accurately.
2. **Precision**:
   Precision, calculated at **94.4%**, highlights the model's ability to minimize false positives. This is particularly crucial in medical applications where a false positive could lead to unnecessary anxiety or treatment.
3. **Recall (Sensitivity)**:
   The model demonstrated an impressive recall of **97.7%**, signifying its capability to correctly identify nearly all true positive cases. This metric is critical in healthcare scenarios where missing a true positive could have severe consequences.
4. **F1-Score**:
   With an F1-score of **96.0%**, the model effectively balances precision and recall, showcasing its reliability and consistency in prediction.
5. **Area Under the Curve (AUC)**:
   The model's ROC-AUC score exceeded **0.95**, indicating excellent discriminatory power in distinguishing between positive and negative cases. A high AUC value reinforces the model's suitability for this classification task.

b) Insights

1. **Strengths**:
- The model exhibits remarkable sensitivity and specificity, reducing both false positives and false negatives.
- Its high recall ensures that nearly all individuals at risk are correctly identified, making it reliable for early intervention and preventive measures.
- The precision metric underscores its capacity to avoid over-predicting positive cases, ensuring the efficient allocation of resources.
2. **Areas for Improvement**:
- The presence of a few false positives suggests a potential need for further refinement in feature selection or model tuning to enhance specificity.
- While the metrics are strong, testing the model on a larger and more diverse dataset could validate its generalizability and robustness across different population groups.

3. **Feature Importance**:

The model's reliance on a diverse set of features contributes to its high performance. Insights derived from feature importance scores reveal which predictors (e.g., demographic or behavioral factors) have the greatest impact on STD risk prediction. This information can guide future efforts to improve data collection and refine model inputs.

4. **Practical Implications**:

The high sensitivity and specificity of the model make it a valuable tool for public health initiatives, enabling targeted interventions and resource allocation. The interpretability of the Random Forest model also aids in deriving actionable insights from predictions, making it a suitable choice for healthcare professionals.

4. Summary of Result

The Random Forest model has proven to be a robust and effective tool for predicting STD risks. Its high accuracy, precision, recall, and AUC underscore its reliability and practicality in real-world applications. While there is room for improvement in minimizing false positives, the model's strengths far outweigh its limitations. These results establish a strong foundation for further optimization and comparative evaluation in the subsequent phases of the project.

VII. Disscussion

The analysis of the implemented Random Forest model reveals its effectiveness as a reliable tool for STD detection, showcasing strong predictive performance. Key metrics, such as accuracy and recall, highlight the model's capability to balance sensitivity and precision, essential for clinical applications. The confusion matrix, which reports a minimal number of false positives and negatives, supports the overall robustness of the model but also opens avenues for addressing specific areas of improvement.

1. Model Strengths

The model's ability to maintain high recall emphasizes its suitability for identifying true positive cases effectively. This is particularly critical in healthcare contexts, where missing an actual case could lead to delayed treatment and further complications. The relatively low number of false negatives further strengthens the model's clinical relevance by ensuring that most at-risk individuals are correctly identified for early intervention.

Precision, though slightly affected by false positives, demonstrates the model's ability to reduce unnecessary alarms. This characteristic is especially valuable for optimizing resource allocation in healthcare systems, where every misclassification could imply redundant diagnostic procedures and increased burden on medical staff.

2. **Feature Contributions**
Feature importance analysis has provided valuable insights into which variables significantly influence model predictions. By prioritizing impactful features, this approach ensures the model remains interpretable and adaptable, an essential factor when integrating predictive tools into real-world applications. Additionally, it offers direction for further research into refining data collection practices to emphasize the most critical indicators.

3. **Performance Challenges**
Despite high performance, the confusion matrix indicates areas of improvement. False positives suggest that some cases may be erroneously flagged as high-risk, potentially leading to unnecessary anxiety or over-intervention. Conversely, false negatives, though few, pose a challenge in scenarios where even a single missed case can have significant repercussions. These findings call for enhancements in feature engineering and model tuning to address such edge cases.

4. **Context in Existing Literature**
When compared to similar studies, the Random Forest model's performance aligns with or surpasses the benchmarks set by other machine learning approaches in STD detection. However, issues such as dataset specificity and reliance on predefined variables echo common challenges seen in the literature. The results underline the importance of maintaining a balance between sensitivity and specificity, a recurring theme in healthcare predictive modeling.

5. **Implementation and Real-World Implications**
The results highlight the feasibility of employing machine learning tools in healthcare systems for STD prediction. However, practical implementation must account for variability in population characteristics, access to healthcare, and socio-behavioral factors, all of which can influence the accuracy and reliability of predictions. Such tools could prove transformative in resource-constrained settings by enabling targeted interventions and efficient allocation of resources.

The discussion thus provides a critical reflection on the model's strengths and limitations, setting the stage for further refinements in subsequent phases of the project. While the current implementation demonstrates high potential, continuous validation and adaptation are necessary to ensure sustained performance and applicability in diverse settings.

VIII. Conclusion and Future Scope

In this study, we have developed and evaluated a Random Forest model for the prediction of sexually transmitted diseases (STDs) using clinical data. The model demonstrated high accuracy and strong performance metrics, particularly in terms of recall, which is crucial for minimizing false negatives. The confusion matrix revealed a favourable balance between sensitivity and precision, though there is room for improvement, especially in addressing false positives.

The feature importance analysis identified key variables influencing the model's predictions, highlighting areas for further refinement. Despite the promising results, the model's performance can be enhanced by addressing issues such as class imbalances, feature selection, and model tuning.

In terms of future work, we plan to extend the model by incorporating more diverse and comprehensive datasets to enhance generalizability. Additionally, exploring other machine learning algorithms and ensemble methods could improve predictive accuracy. Future studies will also focus on validating the model in real-world clinical settings, where external factors like patient demographics and healthcare access can impact performance. Furthermore, we aim to integrate additional data types, such as behavioural and socio-economic factors, to further enhance the robustness and applicability of the model in diverse populations.

IX. Reference

1. Melbourne Sexual Health Centre. MySTIRisk 2023. https://mystirisk.mshc.org.au/.8. Xu XL, Ge ZY, Chow EPF, Yu Z, Lee D, Wu JR, et al. A machine-learning-based risk-prediction tool for HIV and sexually transmitted infections acquisition over the next 12 months. J Clin Med 2022;11(7).

2. Latt PM, Soe NN, Xu X, Ong JJ, Chow EPF, Fairley CK, Zhang L. Identifying individuals at high risk for HIV and sexually transmitted infections with an artificial intelligence–based risk assessment tool. J Clin Epidemiol. 2024.

3. Chen Y, Yu W, Cai L, Liu B, Guo F. Enhancing HIV/STI decision-making: challenges and opportunities in leveraging predictive models for individuals, healthcare providers, and policymakers. 2024 22:886 https://doi.org/10.1186/s12967-024-05684-9

4. Kassaw AA, Yilma TM, Sebastian Y, Birhanu AY, Melaku MS, Jemal SS. Spatial distribution and machine learning prediction of sexually transmitted infections and associated factors among sexually active men and women in Ethiopia, evidence from EDHS 2016. BMC Infect Dis. 2023;23:49. doi:10.1186/s12879-023-07987-6.

5. Soe NN, Latt PM, Yu Z, Lee D, Kim CM, Tran D, Ong JJ, Ge Z, Fairley CK, Zhang L. Clinical features-based machine learning models to separate sexually transmitted infections from other skin diagnoses. J Infect. 2024

6. Jiang Z, Xiu C, Yang J, Zhang X, Liu M, Chen X, et al. (2018) HIV test uptake and related factors amongst heterosexual drug users in Shandong province, China. PLoS ONE 13(10):e0204489. https://doi.org/10.1371/journal.pone.0204489

7. Wei C, Herrick A, Raymond HF, Anglemyer A, Gerbase A, Noar SM. Social marketing interventions to increase HIV/STI testing uptake among men who have sex with men and male-to-female transgender women. Cochrane Database of Systematic Reviews 2011, Issue 9. Art. No.: CD009337. DOI: 10.1002/14651858.CD009337

8. Ooi CY, Ng CJ, Sales AE, Lim HM Implementation Strategies for Web-Based Apps for Screening: Scoping Review J Med Internet Res 2020;22(7):e15591 URL: http://www.jmir.org/2020/7/e15591/ doi: 10.2196/15591

9. Farran B, AlWotayan R, Alkandari H, Al-Abdulrazzaq D, Channanath A and Thanaraj TA (2019) Use of Non-invasive Parameters and Machine-Learning Algorithms for Predicting Future Risk of Type 2 Diabetes: A Retrospective Cohort Study of Health Data From Kuwait. Front. Endocrinol. 10:624. doi: 10.3389/fendo.2019.00624

10. Park JH, Cho HE, Kim JH, Wall MM, Stern Y, Lim H, Yoo S, Kim HS, Cha J. Machine learning prediction of incidence of Alzheimer's disease using large-scale administrative health data. npj Digit Med. 2020;3:46. doi:10.1038/s41746-020-0256-0.

11. Haakenstad A, Moses MW, Tao T, Tsakalos G, Zlavog B, Kates J, Wexler A, Murray CJL, Dieleman JL. Potential for additional government spending on HIV/AIDS in 137 low-income and middle-income countries: an economic modelling study. Lancet HIV. 2019;6:e382–e395. doi:10.1016/S2352-3018(19)30038-4.

12. Zhao PZ, Wang YJ, Cheng HH, Zhang Y, Tang WM, Yang F, Zhang W, Zhou JY, Wang C. Uptake and correlates of chlamydia and gonorrhea testing among female sex workers in Southern China: a cross-sectional study. BMC Public Health. 2021;21:1477. https://doi.org/10.1186/s12889-021-11526-w.

13. Falasinnu T, Gustafson P, Hottes TS, Gilbert M, Ogilvie G, Shoveller J. A critical appraisal of risk models for predicting sexually transmitted infections. Sex Transm Dis. 2014 May;41(5):321-30. doi: 10.1097/OLQ.0000000000000120. PMID: 24722388.

14. Scott H, Vittinghoff E, Irvin R, et al. Development and validation of the personalized sexual health promotion (SexPro) HIV risk prediction model for men who have sex with men in the United States. AIDS Behav 2020; 24:274–83.

15. Hoenigl M, Weibel N, Mehta SR, et al. Development and validation of the San Diego early test score to predict acute and early HIV infection risk in men who have sex with men. Clin Infect Dis 2015; 61:468–75.

16. Nieuwenburg SA, Hoornenborg E, Davidovich U, de Vries HJC, Schim van der Loeff M. Developing a symptoms-based risk score for infectious syphilis among men who have sex with men. Sex Transm Infect. 2023 Aug;99(5):324-329. doi: 10.1136/sextrans-2022-055550. Epub 2022 Nov 18. PMID: 36400527; PMCID: PMC10359546.

17. Ying GS, Maguire MG, Glynn RJ, Rosner B. Tutorial on Biostatistics: Receiver-Operating Characteristic (ROC) Analysis for Correlated Eye Data. Ophthalmic Epidemiol. 2022 Apr;29(2):117-127. doi: 10.1080/09286586.2021.1921226. Epub 2021 May 12. PMID: 33977829; PMCID: PMC8586066.

18. Ruiz MS, O'Rourke A, Allen ST, Holtgrave DR, Metzger D, Benitez J, Brady KA, Chaulk CP, Wen LS. Using interrupted Time Series Analysis to measure the impact of Legalized Syringe Exchange on HIV diagnoses in Baltimore and Philadelphia. J Acquir Immune Defic Syndr. 2019;82(Suppl 2):S148–54. https://doi.org/10.1097/qai.0000000000002176.

19. Ong JJ, Peng MH, Wong WW, Lo Y-R, Kidd MR, Roland M, Zhu SZ, Jiang SF. Opportunities and barriers for providing HIV testing through community health centers in mainland China: a nationwide cross-sectional survey. BMC Infect Dis. 2019;19:1054. https://doi.org/10.1186/s12879-019-4673-0.

20. Rowley J, Vander Hoorn S, Korenromp E, Low N, Unemo M, Abu-Raddad LJ, et al. Chlamydia, gonorrhoea, trichomoniasis and syphilis: global prevalence and incidence estimates, 2016. Bull World Health Organ. 2019;97(8):548.