International Journal of Women in Technical Education and Employment
ISSN: 2811-1567. Volume 4 – Issue 1. July 2023
https://fpiwitedjournal.federalpolyilaro.edu.ng

Statistical Analysis of Sexually Transmitted Infections among Patients Treated in a Nigerian Hospital

[1]Rasaki Akinbo & [2]Oluwatobi Ogunnusi
[1,2]Department of Mathematics and Statistics, Federal Polytechnic, Ilaro, Ogun State, Nigeria
rasaki.akinbo@federalpolyilaro.edu.ng; ogunnusioluwatobi@gmail.com

ARTICLE HISTORY

Received: May 30, 2023
Revised: June 9, 2023
Accepted: July 5, 2023

Abstract

Sexually transmitted infection (STI) has been one of the diseases prevalent among youths within society, which has affected the lives of individuals in the area of infertility and morbidity. Several factors have been attributed to sexually transmitted infections of which proper diagnosis has not been carried out as a result of misleading information given by infected personalities. Secondary data consisting of 400 patients treated for STI covering January 2020 to August 2021 was extracted from The Federal Polytechnic Ilaro Medical Centre record. STI status of patients was considered as a response variable while age, gender, vaginal diseases, vaginal itching, foul smelling, dysuria, penile itching, and vaginitis were the explanatory variables. Establishing a model that predicts the odds of each factor's contribution to the prevalence of STIs among the treated patients was achieved using the Logistic Regression Model (LRM), while the Random Forest Algorithm (RFA) was also considered as an alternative method. The area under the curve, accuracy, recall, precision, and F-score were used as evaluation metrics for the two techniques. The Chi-square technique was also used to test the independence of all the symptoms in association with the patient's status. The result showed that the female gender has a higher chance of being infected compared to its male counterpart with a 90.3% chance of being infected. The logistic regression model revealed that the odds of STI positive for patients suffering from penile itching is higher than every other factor considered in this research, while both logistic regression and random forest performances were evaluated with the area under the curve (0.947 and 0.907) accuracy (0.950 and 0.933) recall (0.939 and 0.849), precision (0.886 and 0.849) and F-score (0.912 and 0.875) respectively. Sex, vaginal itching, and foul-smelling were found to be significantly associated with patient status while age, dysuria, penile itching, and vaginitis were not significant at 5% and 10% levels. Based on the five (5) evaluation metrics, the logistic regression model outperformed the random forest model, hence, it is adjudged to be the best of the two models and can be relied upon for the analysis.

Keywords: Chi-Square test, Demographic variables, Logistic regression, Random Forest, sexually transmitted infections.

Citation
Akinbo, R. & Ogunnusi, O. (2023). Statistical Analysis of Sexually Transmitted Infections among Patients Treated in a Nigerian Hospital. *International Journal of Women in Technical Education and Employment*, 4(1), 110 – 119.

Introduction

Sexually transmitted diseases (STIs), which afflict millions of people globally, are a serious threat to global health. The majority of sexually transmitted diseases (STIs) are spread through oral, anal, and vaginal sex. If left untreated, these illnesses, which can be brought on by bacteria, viruses, parasites, and fungi, can have detrimental effects on one's health.

STIs are pervasive and affect people of all sexes, ages, and socioeconomic levels. Globally, the World Health Organisation (WHO) reports that more than a million STIs are contracted each day (WHO, 2021). Chlamydia, gonorrhea, syphilis, and trichomoniasis are the four most prevalent treatable STIs, and there were reportedly 376 million new cases of these in 2016 among adults aged 15 to 49 worldwide (WHO, 2019). It is crucial to remember that these numbers can be understated due to underreporting and limited access to healthcare in some areas.

Sexual intercourse is the most common way for STIs to spread (Cramer et al., 2020; Prah et al., 2018). Sexual intercourse without protection, including vaginal, anal, and oral sex, carries a substantial risk of infection. Some STIs can also be passed from mother to child during childbirth or through blood transfusions (CDC, 2021; Hickson et al., 2021; Prah et al., 2018; Mirzaiel et al., 2022). Multiple sexual partners, inconsistent or inaccurate condom usage, and participating in high-risk sexual behaviors can all increase the risk of transmission.

To battle the transmission and effect of STIs, prevention and timely treatment are essential. Education and awareness programmes play an essential role in promoting safer sexual practices, such as consistent and correct condom usage, regular testing, and seeking early medical attention. Some STIs, such as human papillomavirus (HPV) and hepatitis B, have vaccines that protect against these illnesses (CDC, 2021). Individuals must also talk honestly about their sexual health status with their sexual partners, as well as practice mutual trust and respect, to prevent the transmission of STIs (Aladeniyi et al., 2017; Rönn & Ward, 2022; Tabrizi et al., 2006; Shafir et al., 2009).

The transmission dynamics of several STIs have been studied using mathematical models. Susser et al. (2020) established a dynamic model to examine HIV transmission in a heterosexual community, taking into account characteristics including sexual behaviour, condom use, and antiretroviral medication coverage. Such models aid in understanding the impact of various factors on STI transmission, identifying major drivers of infection, and informing targeted therapies.

The efficiency of various intervention options for STI control can be assessed using modeling. Khan et al. (2019) investigated the impact of pre-exposure prophylaxis (PrEP) on HIV prevention in high-risk populations using a mathematical model. The model anticipated a significant reduction in new infections with PrEP deployment, underlining the drug's potential as a preventative measure. Modeling is also used to evaluate the effectiveness of condom distribution, screening programmes, vaccine campaigns, and behavioural interventions in reducing STI transmission.

Vickerman et al. (2021) utilised models to assess the cost-effectiveness of delivering a syphilis vaccine in Sub-Saharan Africa, hence influencing vaccine implementation policy decisions. Models can also predict how changes in sexual behaviour patterns, such as greater condom use or decreases in high-risk behaviours, will affect the trajectory of STI outbreaks.

The rise of machine-learning approaches may aid in the complete identification of parameters linked with STIs (Uddin et al, 2019). Different machine-learning approaches can be used to examine data from various angles and summarise it into meaningful knowledge. Available research studies in Ethiopia relied on traditional statistical approaches to evaluate the link between variables that are dependent on prior assumptions, limiting the ability to reveal hidden knowledge (Dagnew et al., 2020; Tamrat et al., 2020; Amsale & Yemane, 2012).

Machine learning models, on the other hand, are built to provide the most accurate predictions possible, allowing systems to learn from data rather than prior assumptions (Dhar, 2013). Furthermore, numerous aspects have been incorrectly linked to STIs for which accurate diagnosis has not been performed in the past, owing to deceptive information provided by the affected personality and inappropriate use of statistical methodologies capable of determining individual sufferers. It is essential to create a better STI prediction model. Because STIs are complicated diseases, employing predictive modeling with a novel technique will bring fresh insight into the condition, hence improving population care.

In this research, machine learning prediction methods will be applied to fill this gap. The methods have a mechanism that handles the imbalanced data which makes it biased in the traditional statistical regression model such as Logistic Regression and Random Forest Algorithm (RFA). Therefore, this research aimed to assess predictors of STIs using machine learning techniques and thereby identify the model that best fits the STI data using several prediction metrics.

This paper aimed to compare Logistic Regression and Random Forest Algorithm (RFA), a machine learning technique in the modeling and prediction of sexually transmitted infections status among patients treated in

the Federal Polytechnic Ilaro, Medical Centre, Ogun State, Nigeria. The methods provide a methodology for dealing with the imbalanced data that causes bias in the classic statistical regression model. To achieve this aim, several prediction metrics will be estimated and compared based on the models fitted to identify the best fit for the STI data.

Materials and Methods

Data on sexually infected individuals was collected at The Federal Polytechnic Ilaro, Medical Centre. The dataset includes the demographic information of 400 patients across genders whose ages range between 09 – 60 years. Also collected was information on symptoms recorded for each patient which comprise dysuria, penile itching, vaginal discharge, vaginal itching, candidiasis, vaginitis, and foul-smelling. Gender was scaled as '0' for male and '1' for female, while '0' and '1' also indicates respectively the absence and presence of the diseases. Disease status was considered the dependent variable. The value of the STI status was determined based on the presence or absence of laboratory test results in patients' files during data extraction, as well as the dictate of the test reports. While age, gender, and all seven symptoms were considered independent variables.

Logistic Regression

Suppose we have n independent observations $y_1, y_2, \cdots, y_n$ and that the i^{th} observation can be treated as a realization of a random variable Y_i. Assuming that Y_i is Bernoulli distributed with parameter θ. The logit of the underlying probability θ_i is a linear function of the predictors

$$logit(\theta_i) = x'\beta \tag{1}$$

Exponentiating equation (1), the odds for the i^{th} unit becomes

$$\frac{\theta_i}{1-\theta_i} = \exp(x'\beta) \tag{2}$$

Solving for the probability θ_i in equation (2), we have

$$\theta_i = \frac{\exp(x'\beta)}{1+\exp(x'\beta)} \tag{3}$$

Explicitly written, equation (3) is expressed as

$$\theta_i = \frac{e^{\beta_0+\beta_1X_1+\beta_2X_2+\cdots\beta_kX_k}}{1+e^{\beta_0+\beta_1X_1+\beta_2X_2+\cdots\beta_kX_k}} \tag{4}$$

Chi-Square Test

This was used to carry out the test of independence between the status of patients and all the independent variables considered in this research.

The chi-square model is given as;

$$x^2 = \frac{(0_1-E_1)^2}{E_1} + \frac{(0_2+E_2)^2}{E_2} + \cdots + \frac{(0_K+E_K)^2}{E_K} = \sum_{f=1}^{k} \frac{\left(0_f-E_f\right)^2}{E_f} \tag{5}$$

Where 0_f is the observed frequency and E_f is the expected frequency

Random Forest

This algorithm combines a huge number of distinct decision trees, with the final prediction coming from the class that received the most votes. The logic is that an attribute of a large collection of uncorrelated models should outperform any of the randomly chosen individual constituent models. The technique's ease of implementation and flexibility have fueled its adoption, as it handles both classification and regression problems. Mathematically, a random forest can be expressed as:

$$RF_i = \frac{\sum jnormF_{ij}}{\sum jCall\ features, kCall\ trees^{normFi_{jk}}} \tag{6}$$

Where RF_i is the importance of feature i calculated from all trees in the random forest model and $\sum jnormF_{ij}$ is the normalized feature importance for i in tree j.

The evaluation metrics considered to measure the performances of the two techniques are Area Under Curve (AUC), Classification Accuracy (CA), F1-Score, Precision, and Recall.

The proposed machine learning algorithms were implemented using Jupiter I Python Notepad as a simple tool cache. The toolbox makes use of a wide range of high-performance calculating bundles to handle input data, build elements, and train, and test

models. Grid search with 5fold cross-validation was used for each model, with numerous hyper-parameters. Based on the findings of that analysis, the grid search found the optimal hyper-parameters and the adopted evaluation metrics were generated.

Furthermore, the logistic regression model and its associated odds were carried out on R software to achieve a robust model specification.

Results

The frequency distribution of age variable among STI-infected individuals is summarized in Table 1.

Table 1. Age Distribution of Infected Patients

	Frequency	Percent
Under 19	76	19.0
20-29	295	73.8
30-39	14	3.5
+40-49	10	2.5
50-59	3	.8
60-69	1	.3
70+	1	.3
Total	**400**	**100.0**

Source: Researchers' Self Computation

The age group 20-29 has the highest proportion (73.8%) of infected individuals, while the age group 60 and older has the lowest rate (0.6%). Figure 1 depicts the graphical distribution of the infected individuals' gender, with the female category constituting the largest populace. According to Table 2, the female category is the most endangered species as all the females had complained of at least one symptom for treatment, with vaginal itching being the most rampant among the symptoms.

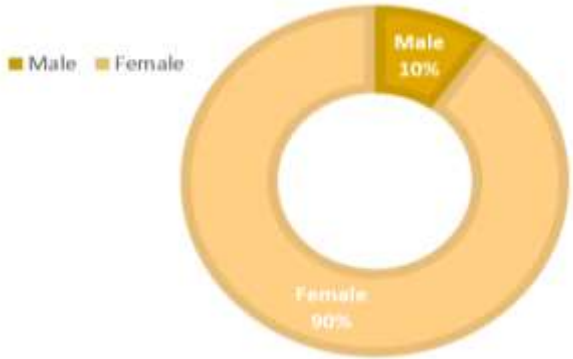

Figure 1: *Doughnut chart*

From the data collected, about 90% of the patients were female while 10% were male. Figure 1 shows **a** **d**oughnut chart showing the Sex distribution of Patients with sexually transmitted Infection

Table 2. Distribution of the Symptoms across Gender

Symptoms	Male	Female
Dysuria	25	63
Peniel Itching	15	2
Vaginal discharge	0	155
Vaginal itching	0	176
Candidiasis	0	20
Vaginitis	0	18
Foul-smelling	0	45

Source: Researchers' Self Computation

From Table 2, the result showed that female patients were majorly affected by dysuria while male patients were affected by penile itching. In addition, only female patients suffer from vaginal discharge, vaginal itching, candidiasis, vaginitis, and foul-smelling. This implies that the female population is majorly contracted with STIs.

Results

Model coefficients, standard error, and p-values for each variable are summarized in Table 3.

Table 3: Logistic Regression Coefficients

	Estimate	Standard error.	Z	Pr(>|Z|)
Constant	-18.105	5.894	-3.071	.002***
Age	.043	.070	0.608	.542
Sex	3.734	5.005	0.746	.456
Dysuria	7.754	1.341	5.782	.000***
Penile itching	9.457	4.980	1.899	.058*
Vagina discharge	7.387	1.165	6.341	.000***
Vaginal itching	7.864	1.290	6.095	.000***
Candidiasis	8.771	2.433	3.605	.000***
Vaginitis	6.300	1.413	4.458	.000***
Foul-smelling	8.285	1.319	6.282	.000***

Source: Researchers' Self Computation

The fitted logit model from the estimates presented in the table is given as:

$$Logit\theta_i = -18.105 + .043(age) + 3.734(Sex) + 7.754(dysuria) + 9.457(penile\ itching) + 7.387(vaginal\ discharge) + 7.864(vaginal\ itching) + 8.771(candidiasis) + 6.300(vaginitis) + 8.285\ (foul\ smelling) \quad (7)$$

Equation (6) fitted from the results presented in Table 3, implies that when all independent variables (symptoms) are equal to zero, the log odds of being positive to STI (θ_i) = -18.105. At 0.05 level of significance, all the symptoms apart from the age and sex of the patients are significant in the model. The coefficient estimate of the variables age and sex is 0.043 and 3.734 respectively, which implies that an increase in age and sex is associated with an increase in the probability of being positive to STI but not statistically significant. Taking the various symptoms into consideration, it can be evidenced that the

coefficient of the symptoms dysuria, penile itching, vaginal discharge, vaginal itching, candidiasis, vaginitis, and foul smelling were found to be 7.754, 9.457, 7.387, 7.864, 8.771, 6.300 and 8.285 respectively, which indicates that an increase in those symptoms was associated with increased probability of being positive as they were found to be highly statistically significant, saves for penile itching that was only significant at 5% level.

The odds ratio in Table 4 is used to interpret the logistic beta coefficients since it measures the association between a predictor variable and an outcome variable. It represents the probability of an event occurring given the presence of the predictor.

Table 4: Odds Ratios

Age	Sex	Dysuria	penile itching	vagina discharge	vaginal itching	candidiasis	Vaginitis	foul-smelling
1.044	41.826	2330.504	12802.21	1615.634	2602.742	6444.846	544.59	3962.451

Overall percentage correctly predicted = 75.0%

Source: Researchers' Self Computation

Taking the odds ratios of Table 4 into consideration, results indicate that a one-unit increase in the age and sex of patients will increase the odds of being STI-positive by 1.044 and 41.826 times respectively. More so, a unit increase in dysuria, penile itching, vaginal discharge, vaginal itching, candidiasis, vaginitis, and foul smelling will increase the odds of being STI-positive by 41.826, 2330.504, 12802.21, 1615.634, 2602.742, 6444.84l6, 544.59 and 3962.451 times respectively. This implies that penile itching increases the odds of being infected with sexually transmitted infections the most, as the odds are higher compared to other identified symptoms.

Table 5: Chi-square test

Interacting variables	Test Statistic*	df	p-value	Remark
Sex in Relation to patient status	10.313	1	0.001321	Significant
age in relation to patient status	27.938	1	0.758600	Not significant
dysuria in relation to patient status	0.000	1	1.000000	Not significant
Peniel itching in relation to patient status	1.0034	1	0.316500	Not significant
Vaginal discharge in relation to patient status	63.989	1	0.000000	Significant
Vaginal itching in relation to patient status	26.757	1	0.000000	Significant
Vaginitis in Relation to patient status	0.000	1	1.00000	Not significant
Foul smelling in relation to patient status	66.112	1	0.00000	Significant

* represents Pearson's Chi-squared test with Yates' continuity correction

Source: Researchers' Self Computation

Table 5 presents a test of independence between variables at a 5% level of significance. The results indicate that there are significant relationships between sex, vaginal discharge, vaginal itching, foul smelling, and patient STI status while age, dysuria, peniel itching, and vaginitis have no significant relationship with STI status.

Table 6: Performance Evaluation Results of the Models

Classifier	AUC	Accuracy	Recall	Precision	F-Score
Logistic Regression	**0.947**	**0.950**	**0.939**	0.886	**0.912**
Random forest	0.907	0.933	0.848	0.848	0.875

Source: Researchers' Self Computation

Table 6 presents the performance results of both logistic regression and random forest techniques. The results show that logistic regression outperforms the random forest model with an AUC of approximately 95%, CA of 95%, recall of 93%, precision of 89%, and F-Score of approximately 91%.

Discussion

Sexually transmitted infections have been one of the most prevalent contagious diseases among youths and adults within society. It has become the most complex problem facing humanity today and has affected the lives of individuals in the area of infertility and morbidity. Several factors have been attributed wrongly to STI of which proper diagnosis had not been carried out as a result of misleading information given by infected personalities. Thus, this study revealed how STIs can be identified on individual sufferers through modeling of the prevalence disease based on various symptoms regardless of their age and sex.

Logistic regression is considered suitable for this research based on the results of the evaluated metrics and this is in tandem with the established convention in literature, which has confirmed the Logistic Regression model as one of the most suitable techniques in healthcare investigations (Adeboye and Adesanya, 2022; Adeniyi et al., 2022). Mirzaie et al. (2022) used a Logistic regression model and genetic algorithm to predict the mortality rate for breast cancer patients. In their comparison, the two (2) methods compared favorably well and gives better prediction based on the evaluation metrics employed; they recommended the use of the two (2) methods for better prediction.

Based on the data analysed, various inferences were drawn from this study. It was discovered that the female gender was more affected when compared to their male counterparts. Furthermore, the male gender has a lesser risk of contracting any sexually transmitted infection. The numerous symptoms of sexually transmitted infection that are common among the targeted gender have an equivalent effect on the status outcome, i.e. infected or not infected (positive or negative). Age and gender have little bearing on whether or not a patient is infected. In addition, findings also indicated from the Chi-square test of linear association that vaginal discharge, vaginal itching, and foul smelling were found to be statistically related to the patient's status while dysuria, penile itching, and vaginitis do not significantly associate with the patient's status.

Conclusion

This study has established awareness, understanding, and modeling of the prevalence of sexually transmitted infections (STIs) among patients. in a Nigerian Hospital. Their causes, sources, and modes of transmission have all been highlighted in this study. The logistic regression model revealed that the odds of STI positive for patients suffering from penile itching is higher than every other factor considered in this research. However, age and sex are not statistically significant in determining the patients' STIs status. The remaining factors graduated in the odds of their ratios of infection are candidiasis, foul smelling, vaginal itching, dysuria, vaginal discharge, and vaginitis. However, the female gender was diagnosed to be the most infected with STIs compared to their male counterparts. The chi-square test of independence revealed that the patient's STI status depended statistically only on vaginal discharge, vaginal itching, and foul smelling while the effects of other symptoms were found to be statistically insignificant. It can as well be concluded that the random forest algorithm does not outperform the logistic regression model in predicting the patients' status for STI based on their respective AUC, accuracy, recall, precision, and F-scores. Knowing all of this, a programme against STIs can be launched with greater success by raising awareness among a large audience through mass media (television and radio). The inclusion of a course on sexual and reproductive health in the educational curriculum will be of utmost significance. We can have a sound social economic planning for the current and future generations if the aim is accomplished. Future research could focus on developing models that capture the complex interactions between different STIs, understanding the impact of emerging drug resistance, and integrating models with real-time surveillance data for more accurate predictions.

References

Adeboye N.O, Adesanya K.K.(2022). On The Survival Assessment of Diabetics Patients using Machine Learning Techniques. *International Journal of Research and Innovation in Applied Science*. 7(1):69-75.

Adeniyi O.I, Afolabi N.B, Akinrefon A.A, Omekam I.V, Olonijolu I.R. (2022). Factors Influencing the Choice of Place of Delivery of a First Child among Nigerian Women. *Tanzania Journal of Science*. 48(2):324-334.

Aladeniyi O.B, Bodunwa O,K, Sonde M. (2017). Statistical Analysis of Reported Cases of Sexually Transmitted Diseases. *International Journal of Statistics and Applications*, 7(3): 186-191.

Amsale C, Yemane B. (2012). Knowledge of sexually transmitted infections and barriers to seeking health services among high school adolescents in Addis Ababa, Ethiopia. *J AIDS Clin Res.* 3(5).

Centers for Disease Control and Prevention (CDC). (2021). Sexually Transmitted Infections (STIs). Retrieved from https://www.cdc.gov/std/default.htm

Cramer R, Leichliter J.S, Stenger M.R, et al. (2020). The role of gonorrhea and syphilis in HIV acquisition: an analysis of heterosexual partnerships in the United States. *Sex Transm Dis.* 47(7):473-479.

Dagnew G.W, Asresie M.B, Fekadu G.A. (2020). Factors associated with sexually transmitted infections among sexually active men in Ethiopia. Further analysis of 2016 Ethiopian demographic and health survey data. *PLoS ONE*. 2020;15(5): e0232793.

Dhar V. (2013). Data science and prediction. *Commun ACM.* 56(12):64–73.

Hickson F, Reid D, Davies P, Weatherburn P, Wayal S, McKeown E. (2021). Sex, sexuality, and sexual health of adolescents and young adults in England: findings from the third National Survey of Sexual Attitudes and Lifestyles (Natsal-3). Lancet Public Health. 6(12):e855-e866.

Khan M.R, Parvez S.M, Mahmud S, et al. (2019). Impact of pre-exposure prophylaxis for HIV prevention among high-risk populations: an update from mathematical modeling. *J Acquir Immune Defic Syndr*. 80(4):e93-e99.

Mishra S, Pickles M, Blanchard J.F, Moses S, Shubber Z, Boily M.C. (2014). Validation of the Modes of Transmission Model as a tool to prioritize HIV prevention targets: a comparative modeling analysis. *PLoS One*. 9(8):e103691.

Prah P, Hickson F and Bonell C (2018). Men who sell sex in London: a qualitative comparison of different sub-groupings in a diverse sexual marketplace. *HIV Med*. 19(8):529-535.

Rönn MM, Ward H. (2011). The association between lymphogranuloma venereum and HIamong men who have sex with men: systematic review and meta-analysis. *BMC Infectious Diseases.* 11:70–78.

Shafir S.C, Sorvillo F.J, Smith L.(2009). Clinical Microbiology Reviews, Current Issues and considerations regarding trichomoniasis and human immunodeficiency virus in African-Americans. A*merican Society for Microbiology*. 22(1):37-45.

Susser S.R, Oh C, Makinde O.A (2020). Modeling HIV transmission dynamics in a heterogeneous population: a case study of South Africa. *Math Biosci Eng.* 17(6):6525-6552.

Tabrizi S. N., Fairley C. S., Bradshaw C. S., Garland S. M. (2006). Prevalence of Gardnerella vaginalis and Atopobium vaginae in virginal women. *Journal of the American Sexually Transmitted Diseases Association.* 33(11):663–665.

Tamrat R, Kasa T, Sahilemariam Z, Gashaw M.(2020). Prevalence and factors associated with sexually transmitted infections among Jimma University students, Southwest Ethiopia. *Int J Microbiol.* 2020.

Uddin S, Khan A, Hossain M.E, Moni M.A.(2019) Comparing different supervised machine learning algorithms for disease prediction. *BMC Med Inform Decis Making*. 19(1):281.

Vickerman P, Devine A, Foss A.M, Delany-Moretlwe S, Mayaud P, Meyer-Rath G (2016). The cost-effectiveness of herpes simplex virus-2 suppressive therapy with daily aciclovir for delaying HIV disease progression among HIV-1-infected women in South Africa. *Sex Transm Infect.* 92(7):527-533.

World Health Organization (WHO) (2019). Sexually transmitted infections (STIs). Retrieved from https://www.who.int/news-room/fact-sheets/detail/sexually-transmitted-infections-(stis)

World Health Organization (WHO). (2021). Sexually transmitted infections (STIs) - Key facts. Retrieved from https://www.who.int/news-room/fact-sheets/detail/sexually-transmitted-infections-(stis)-key-facts