



Predictive model for genital tract infections among men and women in Ghana: An application of LASSO penalized cross-validation regression model


Original Paper

Cite this article: Ntumy MY, Tetteh J, Aguadze S, Swaray SM, Udofia EA and Yawson AE (2024). Predictive model for genital tract infections among men and women in Ghana: An application of LASSO penalized cross-validation regression model. *Epidemiology and Infection*, **152**, e160, 1–9 <https://doi.org/10.1017/S0950268824001444>

Received: 13 September 2023
Revised: 21 July 2024
Accepted: 28 July 2024

Keywords:
genital tract infections; Ghana; LASSO; prediction model; reproductive

Corresponding author:
John Tetteh;
Email: bigjayasamoah@gmail.com

Michael Yao Ntumy¹, John Tetteh² , Stephen Aguadze³, Swithin M. Swaray⁴, Emilia Asuquo Udofia² and Alfred Edwin Yawson²

¹Department of Obstetrics and Gynaecology, University of Ghana Medical School, College of Health Sciences, Accra, Ghana; ²Department of Community Health, University of Ghana Medical School, College of Health Sciences, Accra, Ghana; ³Research Unit, Korle Bu Teaching Hospital, Accra, Ghana and ⁴National Cardiothoracic Centre, Korle Bu Teaching Hospital, Accra, Ghana

Abstract

To enhance the capacity for early and effective management of genital tract infections at primary and secondary levels of the healthcare system, we developed a prediction model, validated internally to help predict individual risk of self-reported genital tract infections (sGTIs) at the community level in Ghana. The study involved 32973 men and women aged 15–49 years from three rounds of the Ghana Demographic Health Survey, from 2003 to 2014. The outcomes were sGTIs. We applied the least absolute shrinkage and selection operator (LASSO) penalized regression with a 10-fold cross-validation model to 11 predictors based on prior review of the literature. The bootstrapping technique was also employed as a sensitivity analysis to produce a robust model. We further employed discriminant and calibration analyses to evaluate the performance of the model. Statistical significance was set at P -value <0.05 . The mean \pm standard deviation age was 29.1 ± 9.7 years with female preponderance (60.7%). The prevalence of sGTIs within the period was 11.2% (95% CI = 4.5–17.8) and it ranged from 5.4% (95% CI = 4.8–5.86) in 2003 to 17.5% (95% CI = 16.4–18.7) in 2014. The LASSO regression model retained all 11 predictors. The model's ability to discriminate between those with sGTIs and those without sGTIs was approximately 73.50% (95% CI = 72.50–74.26) from the area under the curve with bootstrapping technique. There was no evidence of miscalibration from the calibration belt plot with bootstrapping (test statistic = 17.30; P -value = 0.060). The model performance was judged to be good and acceptable. In the absence of clinical measurement, this prediction tool can be used to identify individuals aged 15–49 years with a high risk of sGTIs at the community level in Ghana. Frontline healthcare staff can use this tool for screening and early detection. We, therefore, propose external validation of the model to confirm its generalizability and reliability in different population.

Introduction

Genital tract infections (GTIs) are infectious diseases that often go undetected as epidemics and constitute a huge public health concern [1]. They are mostly ignored, misdiagnosed, or unreported leading to incorrect treatment and undetected transmission [2–5]. Among pregnant women, GTIs can cause spontaneous abortions and where the foetus survives, the risk of congenital diseases adversely affects the quality of life [4].

The WHO estimated that globally, more than one million sexually transmitted infections (STIs) affecting the genitals occur daily, the majority of which are asymptomatic. In 2016, more than 490 million people worldwide had genital herpes [6]. In sub-Saharan Africa (SSA), approximately 6.9% of self-reported STIs exist among young women aged 15–24 years, of which Ghanaian women accounted for 0.3% [7]. For sexually active men, the average prevalence of self-reported STIs in SSA was found to be 3.8%, with Ghana accounting for 5.7%. The highest prevalence was found amongst sexually active men aged 15–24 years [8].

It is important to enhance the capacity of frontline healthcare providers to detect and manage GTIs effectively. This underscores the need to strengthen capacity beginning at the community level to ensure that most infections that otherwise go undetected are managed to curb transmission. This is because genital infections are treatable, provided they are diagnosed early [9, 10]. In low-resource nations, where there are numerous barriers to assessing medical treatment, infections spread quickly and widely compounding amongst other things, the adverse outcomes of reduced fecundability and sterility associated with genital infections in men and women [11, 12]. At the community level, resource constraints in terms of diagnostic capacity and attendant

© The Author(s), 2024. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives licence (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided that no alterations are made and the original article is properly cited. The written permission of Cambridge University Press must be obtained prior to any commercial use and/or adaptation of the article.



costs justify the need to have a robust model to predict GTIs. In order to accurately select the individual risk of GTIs at the community level, this study uses the machine learning (ML) technique, i.e. the least absolute shrinkage and selection operator (LASSO) penalized cross-validation regression approach [17] to generate a prediction model. The model can be used to identify the most significant predictors and offer insights into contextual factors that contribute to the occurrence of sGTIs. We, therefore, integrate the LASSO regression model to find characteristics that predict sGTIs to increase early detection and intervention.

The merits of ML compared to traditional statistical methods for predictive modelling is that the nonlinear nature of many real-world phenomena is better captured by ML techniques such as deep learning, offering superior predictive power [13, 14]. Again, the LASSO ML can simultaneously perform variable selection and regularization (penalization or shrinkage) to constrain or shrink the regression coefficients [15], thereby simplifying models and improving predictive performance.

LASSO regression has emerged as a powerful tool in predicting sexually transmitted infections (STIs), as evidenced by studies conducted both in the United States and Africa. In the United States, Comulada *et al.* [16] utilized LASSO regression to identify predictors of STIs, leveraging its capacity for variable selection and regularization. Similarly, in Africa, Thivalapill *et al.* [17] developed a predictive tool for STIs in Eswatini, employing LASSO regression as a central component of their methodology. Through the application of LASSO, they tackled issues of multicollinearity and overfitting while discerning significant predictors of STIs tailored to the specific demographic and environmental characteristics of Eswatini. These studies underscore the versatility and relevance of LASSO regression in public health research, particularly in identifying critical factors associated with STI transmission. By leveraging the predictive capabilities of LASSO, researchers can enhance their understanding of STIs dynamics and inform targeted interventions and policies aimed at mitigating the burden of STIs globally.

Methods

Description and study design

We utilized five rounds of data from the Ghana demographic health survey (GDHS) conducted from 1993 to 2014. GDHS was a nationally stratified survey conducted across the country employing a multi-stage cluster sampling design. The surveys were supported by the United Nations Population Fund (UNFPA), the United States Agency for International Development (USAID), the World Bank, and other development partners. These demographic health surveys are aimed at providing information on fertility, family planning, infant and child mortality, maternal and child health, and nutrition. The purpose of GDHS is to inform policy decisions, planning, monitoring, and evaluating programmes related to health in general and reproductive health across the country [18].

GDHS employs a cross-sectional study design to a nationally representative sample, using two-stage sampling criteria. The first stage involved selecting clusters consisting of enumeration areas (EAs) independently within the then ten administrative regions in Ghana. This was done considering the rural-urban differential characteristics. The second stage involved systematic sampling from a list of households in all the selected EAs. The number of households enlisted in each EA makes up the EA size. A household was selected, and all men and women aged 15–49 who met the

inclusion criteria were enumerated for the study. Details of the Demographic Health Survey (DHS) sampling design can be found elsewhere [19].

Study participants

This current study merged data on men and women in their reproductive age (i.e. those aged 15–49 years) from 2003 to 2014.

Outcome measures

The main outcome was self-reported genital tract infections (sGTIs). The GDHS assessed sGTIs by asking participants who had ever engaged in sexual intercourse and had experienced STI or symptoms of an STI (a foul-smelling, abnormal discharge from the vagina or penis or a genital sore or ulcer) within 12 months preceding the GDHS survey. The item response theory (IRT) [20] was used to compute the outcome variable. The IRT is a sophisticated statistical framework used to understand how individuals' responses to test items relate to their underlying latent traits, such as abilities or attitudes. At its core, IRT views each test item as a statistical model with its own unique parameters, including difficulty, discrimination, and guessing. These parameters describe the item's characteristics and how it interacts with individuals' trait levels. By modelling the relationship between item responses and latent traits, IRT enables more precise measurement and assessment of individuals' abilities (representing some underlying trait related to the variables for which respondents are providing yes or no responses), allowing for fairer comparisons across different tests and populations [21]. Three items were considered to generate sGTIs composite variable (coded as 0 and 1): abnormal genital discharge, genital sore or ulcer, and any sexually transmitted diseases. These variables are coded as 0 "No" and 1 "Yes". The one-parameter logistic model for the probability sGTI among the participants based on the three items was defined as $p(X_{ni}) = \frac{\exp(\theta_n - b_i)}{1 + \exp(\theta_n - b_i)}$ where $p(X_{ni})$ is the probability of an individual n having sGTI to item i ; θ_n represent the ability of an individual n to report GTIs at community level and b_i

representing the difficulty in reporting GTIs of item i . After applying IRT application, predicted probabilities were generated. Predicted value ≥ 0.5 was classified as having sGTI (coded as 1) and otherwise (coded as 0). [Supplementary Figure 1](#) presents the items characteristics curves for the individual sGTIs, which are simultaneous to each other.

Data analysis

Two approaches to data analyses were employed: descriptive and inferential models. For descriptive analysis, independent variables were described using frequencies and weighted percentages for categorical variables while measures of dispersion involving weighted means \pm standard deviation and median (interquartile range) were adopted for continuous variables. Weighted analysis was adopted because the study design of DHS allows for adjusting for the sampling weight, sampling unit, and strata. The forest plot was used to present the prevalence of sGTIs by GDHS year of study (i.e. 2003, 2008, and 2014) considering the DerSimonian–Laird random-effect meta-analysis [22] to assess differences of sGTIs between the years. This was employed to assess heterogeneity in the prevalence rate of sGTIs across the demographic health survey years. The DerSimonian and Laird model is a widely recognized

method in meta-analysis that effectively incorporates random effects to address variability among studies. This model is particularly useful when dealing with heterogeneous studies where differences in results are anticipated [22, 23]. Unlike fixed-effect models that assume a single common effect size, the DerSimonian and Laird model accounts for varying effect sizes across studies by allowing for random effects. This flexibility helps in producing a more accurate and generalizable summary of the effect [22]. This method employs inverse-variance weighting and considers both within country and between country variability. By effectively managing the inherent variability and heterogeneity in our data, this model offers a more thorough and reliable evaluation of temporal sGTI differences. The test of non-linear simultaneous equality of proportion was adopted to assess the differences in sGTI by socio-demographic categories using the Rao–Scott Wald χ^2 test.

For inferential analysis, we adopted the (LASSO) penalized regression to predict and select the best predictors of sGTI. The DHS sampling weight and clusters were controlled for during estimation. The LASSO is an extension of ordinary least square (OLS) regression, which adds a penalty to the OLS residual sum of squares. When a data value narrows towards a central point, shrinkage occurs. As a result, it is well suited to models with high levels of multicollinearity to identify any potential high correlation between the outcome variable and the predictors [24]. LASSO selects a subset of predictors by shrinking the coefficients of the least dominant variables to zero, thereby excluding them from the model.

The tenfold cross-validation was adopted to determine the amount of coefficient shrinkage. The cross-validation process divides the available data into multiple folds, using one of these folds as a validation set, and training the model on the remaining folds. This process is repeated multiple times and the results from each validation step are averaged to produce a more robust effect size estimate of the model [25]. The performance of the model was assessed using the cross-validation discriminant analysis considering the area under the curve (cvAUC). The value of the AUC indicates the ability of the model to differentiate between individuals with sGTI and otherwise [26].

The bootstrapped Bias-corrected 95% confidence intervals [27] for the AUC were also generated as a sensitivity analysis. The calibration belt plot was additionally adopted to assess the agreement between the model-predicted probabilities and actual observed rates of sGTI. The calibration belt plots examine the variation between expected and observed probabilities (miscalibration) at certain confidence levels [28].

Model building and checking

Factors for model building were considered following a priori review of the literature identifying 11 potential factors associated with STIs. These factors included; wealth quintile [29], number of household members [30], region [29], place of residence [29, 31], sanitation [32], sex [33], age group [34, 35], educational level [34], sexual initiation [31, 35, 36], currently working [35] and staying with partner [34]. These predictors are depicted in the conceptual framework in Supplementary Figure 2. Three models were estimated as presented in the Table 1.

The best-fitted model was selected following the assessment of the AUC. Additionally, the Akaike information criterion (AIC) given by $AIC = -2\ln(L) + 2p$ and Bayesian information criterion (BIC) given by $BIC = -2\ln(L) + \frac{1}{2}p * \ln(n)$ were adopted to assess the best model. The smallest AIC and BIC were considered the best

Table 1. Components involved in the model-building process

Model	Components in the model
Model 1	wealth quintile + number of household members + region + place of residence + sanitation + sex + age group + educational level+ sexual initiation + currently working + and staying with partner
Model 2	Model 1 + interaction term (wealth index * sex)
Model 3	Model 2 + cohort effect (GDHS time trend for year)

GDHS = Ghana demographic and health survey

fit. We implemented a sensitivity analysis to examine the potential influence of GDHS stratification (enumeration areas) on both the discriminant and calibration belt plot analyses. This involved utilizing bootstrapping resampling methods with 1000 replicates for the optimal model. The probability of an individual self-reporting sGTI in Ghana among men and women aged 15–49 years equals the inverse of a logistic regression equation (model) given as;

$$\text{probability} = \frac{e^{\beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \dots + \beta_k * x_k}}{1 + (e^{\beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \dots + \beta_k * x_k})}$$

where β is the coefficient estimate for all the parameters from the LASSO penalty regression analysis. All analyses were conducted using Stata 17 (StataCorp. 2017, Stata Statistical Software, College Station, TX, StataCorp LLC.).

Ethical requirements

Permission to use the secondary data was requested from the DHS Monitoring and Evaluation to Assess and Use Results of Demographic and Health Surveys (MEASURE DHS) Department. Request can be obtained from DHS https://dhsprogram.com/data/dataset_admin/login_main.cfm

Results

The analysis involved 32973 men and women aged 15–49 years with a mean age \pm standard deviation of 29.1 ± 9.7 years. The majority were females (60.7%) with an approximately equal numbers of rural-urban differential. Most of the participants had secondary level education (58.1) and the majority were currently working (Table 2).

Generally, the prevalence of sGTIs within the period (2003–2014) was 11.2% (95% CI = 4.5–17.8) and it ranged from 5.4% (95% CI = 4.8–5.86) in 2003 to 17.5% (95% CI = 16.4–18.7) in 2014 (Figure 1). The differences in sGTIs proportions by socio-demographic characteristics were significantly different within all predictors (p -value<0.05) (Supplementary Table 1).

Model building

In Model 1, the analysis revealed that the mean AUC from tenfold cross-validation discriminant analysis was approximately 70.79% (95% CI = 69.8–71.6). Upon adjusting for an interaction effect in Model 2, the probability increased slightly to 70.81% (95% CI = 69.81–71.59). Further adjustments for cohort effect in Model 3 significantly increased the probability to 73.50% (95% CI = 72.50–74.26). This means that there is approximately a 74% probability that Model 3 will correctly rank a person chosen at random from the community level in Ghana (aged 15–49 years) with sGTIs

Table 2. Socio-demographic characteristics of men and women aged 15–49 years, GDHS 2003–2014

Variable	Frequency	Weighted %	Mean ± SD
Wealth quintile			
Poorest	7466	16.4	
Poorer	5829	17.4	
Middle	5802	19.5	
Richer	6163	22.6	
Richest	6203	24.1	
Household members			
≤3	9597	32.8	5.0 ± 2.9
4–5	9298	30.3	
6–8	8908	27.0	
9+	3660	9.9	
Region			
Western	3161	10.3	
Central	2431	8.7	
Greater Accra	3927	18.0	
Volta	2641	8.4	
Eastern	2993	9.9	
Ashanti	4404	19.4	
Brong Ahafo	3219	8.9	
Northern	3497	9.1	
Upper west	2667	3.8	
Upper east	2523	3.5	
Place of residence			
Urban	14044	49.9	
Rural	17419	50.1	
Sanitation			
Unimproved	19029	56.3	
Improved	12434	43.7	
Sex			
Female	19114	60.7	
Male	12349	39.3	
Age group			
15–19	6818	21.1	30.2 ± 10.8
20–24	5383	17.1	
25–29	5100	16.5	
30–34	4246	13.8	
35–39	3939	12.6	
40–44	3157	10.1	
45–49	2820	8.8	
Educational level			
None	7108	18.2	
Primary	5801	17.6	
Secondary	16794	58.1	

(Continued)

Table 2. (Continued)

Variable	Frequency	Weighted %	Mean ± SD
Higher	1760	6.2	
Sexual initiation			
Late	20005	64.2	
Early	11456	35.8	
Currently working			
No	7880	24.8	
Yes	23583	75.2	
Staying with partner			
No	28279	89.4	
Yes	3184	10.6	

higher than a negative random chosen person. This suggests that Model 3 demonstrates the highest ability to accurately classify individuals with and without sGTIs at the community level in Ghana. This conclusion is supported by the fact that Model 3 has the smallest AIC (21223.99) and BIC (21240.79) values compared to Models 1 and 2 (Table 3 and Figure 2).

The AUC result from the sensitivity analysis was statistically not different from the main model result. The bootstrap resampling adjusting for clustering supported this (Supplementary Figure 1).

Calibration

The calibration belt plot analysis showed that the predicted probabilities of sGTIs from the LASSO penalty regression model are statistically not different from the observed sGTIs rates across all the probabilities (p -value > 0.05). Indicating that all the models performed well by not misclassifying individuals from the LASSO at 10-fold cross-validation (Figure 2).

The best predictive model

The best model penalty prediction from LASSO regression retained all variables as predictors. By default, LASSO regression fits 76 models using different values of lambda. The best model (model 76) had the smallest cross-validation mean prediction error with a mean deviance of 0.6524. The overall mean lambda was approximately 0.01 (Table 4).

The best model equation is presented in the Supplementary Material.

The probability of an individual self-reporting sGTI in Ghana among men and women aged 15–49 years ranged from 0.51% to 53.33%. Figure 3 illustrates that higher probabilities are strongly associated with the occurrence of sGTIs, suggesting that as the predicted probability increases, the likelihood of experiencing sGTIs also rises. This demonstrates that the model effectively captures the relationship between higher probability scores and the presence of sGTIs.

Discussion

This current study generated a prediction model for sGTIs at the community level in Ghana. We derived this model using a machine

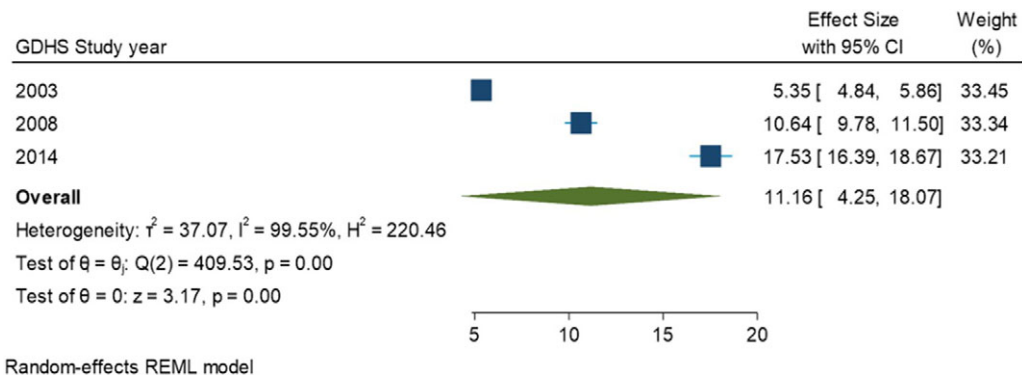


Figure 1. Prevalence of self-reported genital tract infections among men and women aged 15–49 years, GDHS 2003–2014.

Table 3. Predictive model building and checking.

Model	% cvMean AUC	95%CI	AIC	BIC
Model 1	70.79	69.78–71.57	21750.25	21767.05
Model 2	70.81	69.81–71.59	21736.30	21753.11
Model 3	73.50	72.50–74.26	21223.99	21240.79

cv = cross validation; AUC = area under the curve; CI = confidence interval; AIC = Akaike information criterion; BIC = Bayesian information criterion. 95% CI estimates are bootstrap bias-corrected from the LASSO tenfold cross-validation.

learning technique involving objective socio-demographic, and behavioral/environmental indicators. The model performed well in predicting an individual at the community level with sGTI. The AUC from the discriminant analysis was over 70% indicating an acceptable level and the ability of the prediction model to discriminate the true sGTIs at the community level. The calibration belt also showed a good overall performance with no miscalibration indicating that the predicted values were not statistically different from the observed values. We believe our analysis offers a

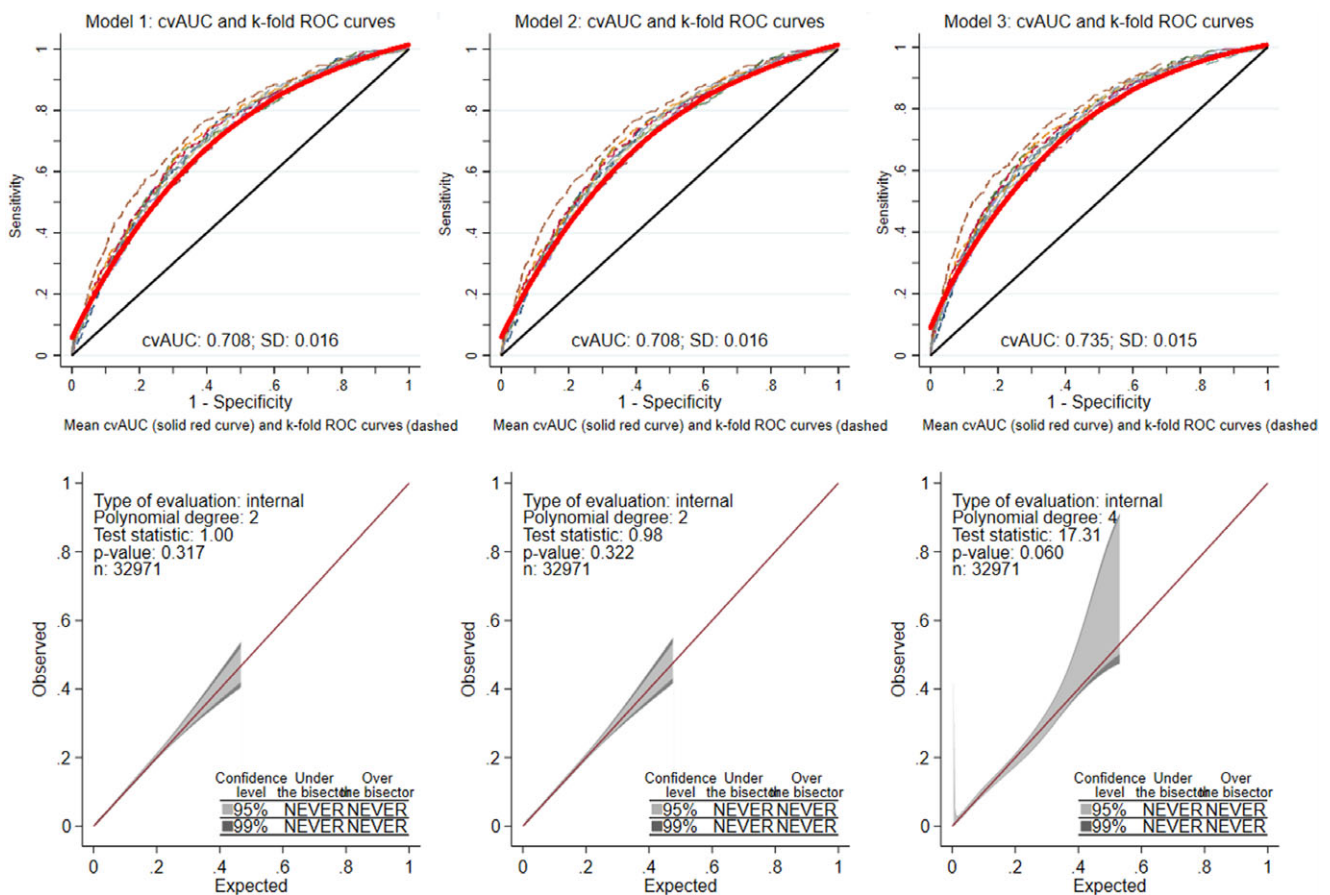


Figure 2. LASSO penalized tenfold cross-validation regression model discrimination and calibration belt: Analysis of receiver operating characteristic (ROC) curve showing area under the curve (AUC) for both cross-validation and calibration belt. The 45° diagonal line represents a model that discriminates by chance (AUC = 50); the x-axis shows the proportion with no sGTI who were incorrectly classified as reporting sGTI (false positive rate or 1 - Specificity); the y-axis shows the proportion with STIs who were correctly classified as reporting STIs (true positive rate or Sensitivity). cvAUC = mean cross-validated area under the curve.

Table 4. Predictors of sexually transmitted infection from LASSO penalized tenfold cross-validation regression model among men and women aged 15–49 years, GDHS 2003–2014

Predictors	Equation parameters	Parameter codes	Lasso regression penalized coefficient
Wealth quintile			
Poorest	p1a	1	0.2012
Poorer	p1b	2	0.1287
Middle	p1c	3	0.1506
Richer	p1d	4	0.0893
Richest	p1e	5	X
Sex			
Male	p2a	1	X
Female	p2b	2	1.4802
Interaction term			
Poorest*Male	p3a	1	0.2496
Poorer*Male	p3b	2	0.4514
Middle*Male	p3c	3	0.1934
Richer*Male	p3d	4	0.5585
Number household members			
≤3	p4a	1	0.1327
4–5	p4b	2	X
6–8	p4c	3	–0.1863
9+	p4d	4	–0.2292
Region			
Western	p5a	1	–0.2307
Central	p5b	2	–0.0336
Greater Accra	p5c	3	X
Volta	p5d	4	0.3024
Eastern	p5e	5	0.0094
Ashanti	p5f	6	0.1797
Brong Ahafo	p5g	7	0.3554
Northern	p5h	8	0.5028
Upper west	p5i	9	–0.3033
Upper east	p5j	10	–0.0300
Place of residence			
Urban	p6a	1	0.1701
Rural	p6b	2	X
Sanitation			
Unimproved	p7a	0	0.0015
Improved	p7b	1	X
Age group			
15–19	p8a	1	–0.1371
20–24	p8b	2	0.4918
25–29	p8c	3	0.2904
30–34	p8d	4	0.0507

(Continued)

Table 4. (Continued)

Predictors	Equation parameters	Parameter codes	Lasso regression penalized coefficient
35–39	p8e	5	X
40–44	p8f	6	–0.3001
45–49	p8g	7	–0.6623
Educational level			
None	p9a	1	–0.2108
Primary	p9b	2	X
Secondary	p9c	3	0.0416
Higher	p9d	4	–0.1887
Sexual initiation			
Late	p10a	0	0.4167
Early	p10b	1	X
Currently working			
No	p11a	0	–0.1874
Yes	p11b	1	X
Staying with partner			
No	p12a	0	0.0845
Yes	p12b	1	X
GDHS year	p13a	linear	0.1008
Intercept			–206.3026
Mean lambda			0.0070

All variables were selected by the LASSO prediction model. Penalized regression coefficients were derived after a penalty was applied which reduces overfitting of the data during model development. Using the regression coefficients from the predictors. X indicates category used for reference.

promising diagnostic tool for screening individual risks of sGTIs in the community.

The overall prevalence of sGTIs during the study period was 11.2%, and this prevalence ranged from 5.4% in 2003 to 17.5% in 2014, showing fluctuations over time. Similarly in sub-Saharan Africa, Thivalapill *et al.*, identified a prevalence of 10.12% among adolescents and young adults in Eswatini [17] and approximately 13.5% in Liberia [37]. This variation might be due to differences in awareness or knowledge about sexually transmitted diseases and differences in measuring the outcome (self-reporting versus diagnosis or screening). Differences in access to health care services, socio-economic/cultural systems that influence health-seeking behavior, and demographic characteristics as well as in sample size, could also account for the variation observed [34].

The high prevalence of sGTIs and the increasing rate at the community level indicate an urgent need for early screening and detection. Screening high-risk individuals at the community level using clinical history will be economically viable for a developing country like Ghana. Self-reporting if encouraged, can result in more testing and diagnosis. Diagnosing cases helps to prevent infected individuals from remaining a reservoir of infection within the community. This study has identified 12 key predictors of sGTIs including; wealth quintile, sex, interaction term (sex and wealth), number of household members, region, place of residence, sanitation, educational level, sexual initiation, currently working, staying with partner and cohort effect. Thivalapill *et al.* also incorporated

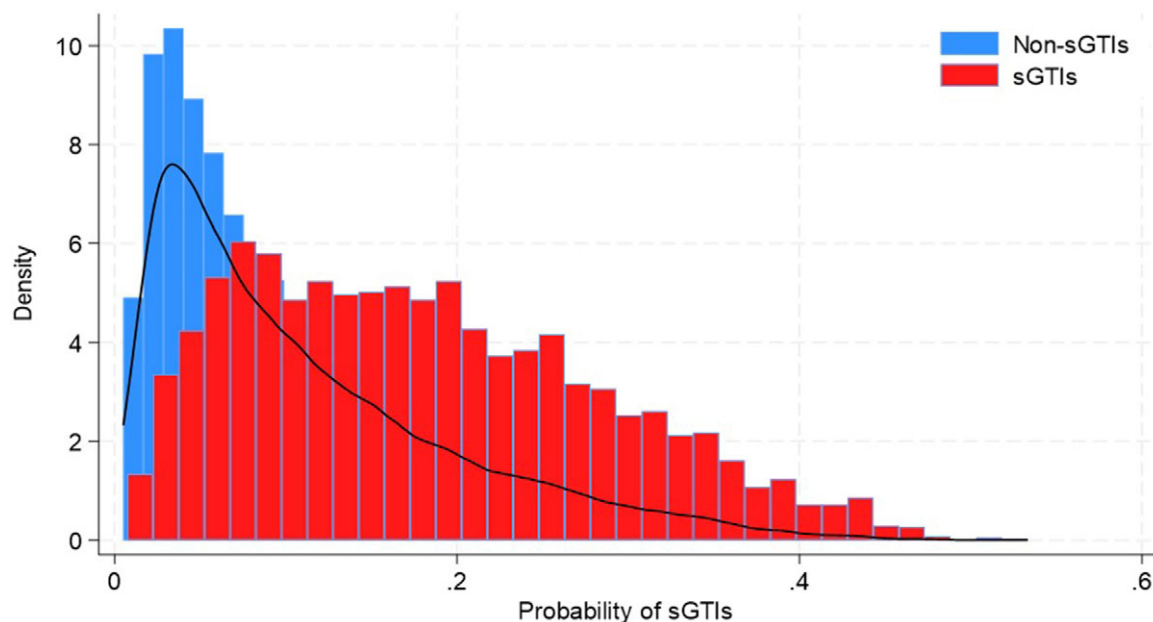


Figure 3. Predicted probability of self-reported genital tract infections among men and women aged 15–49 years, GDHS 2003–2014.

11 potential predictors in their predictive model for STIs among persons living with HIV [17]. Notably, our study also considered similar predictors, including age group and sex, which were among the six predictors included by Thivalapill et al. Drawing parallels between our findings and those of Thivalapill et al. highlights age group and sex consistencies across different populations and contexts, enhancing the external validity of our results. The additional variables integrated into the final model have been independently identified as associated with STIs by scholars in different research settings, underscoring their robustness and relevance in understanding and addressing STI transmission dynamics.

This prediction tool is a remarkable instrument for screening and detecting sGTIs at the community level in Ghana to promote early management. The model AUC was highly acceptable indicating over 70% ability to discriminate true sGTIs at the community level. The resulting individual risk score derived from the model may allow healthcare providers and other stakeholders to identify individuals with the highest predicted risk for early intervention. A self-reporting model can be applied in a community survey by trained healthcare workers at the community level to identify and refer cases requiring medical attention. It can be developed into an algorithm that public health nurses and community health workers can apply during home visits to improve linkage to diagnosis and treatment, and as a complement to current approaches in managing sexually transmitted infections and genital tract infections, even at the lowest level of the health care system. Secondary prevention of GTIs through early diagnosis and treatment is both a health and a development issue. The predictive model has the potential to contribute towards Sustainable Development Goal 3, target 3.3 which seeks to combat communicable diseases and end epidemics such as HIV/AIDS (which is facilitated by GTIs) by 2030.

Limitations and strengths of the study

The key limitation is the cross-sectional nature of the study. The design does not allow for establishing causality. In this regard, the findings of this study should be interpreted with caution. Even

though the AUC and calibration bootstrapping were acceptable, there is the need for external validity of the tool using current similar data in Ghana given that the latest version of GDHS was in 2014 at the time this study was conducted. Again, the outcome considered was self-reported which may not be the true reflection. This is because self-reported morbidity alone cannot serve as an indicator to measure the burden of any disease at the community level [38]. The self-reported measurements were not validated from records or clinical examination to confirm. Participants may or may not report the condition, leading to social desirability bias and under or over-estimations. Again, recall bias may occur because of the period specified for recall.

The study has some merits in that it provides useful information that can be explored for validation in field situations, and this is the subject of our next paper. In a follow-up paper, we will apply the predictive model to the next GDHS data to test the model's reliability. Again, we hope to conduct a nationally stratified survey on genital tract infection prevalence in Ghana by adopting clinical examination to test the model fidelity in field situations. If validated, it presents a useful opportunity to improve the diagnosis and treatment of GTIs at the community level. Healthcare facilities at the community level are not equipped with point-of-care diagnostics for field use based on the level of competence at this level of healthcare. Where point-of-care diagnostic is available, it is often used in independent surveys through donor funding. The predictive model can enhance linkage to higher-level facilities where they can be managed.

Conclusion

Generally, the model performance was very good and acceptable. With the absence of clinical measurement, this prediction model can be used to identify individuals aged 15–49 years with sGTIs at the community level in Ghana. Potential indicators including poverty, urban place of residence, male sex, and lower education were highly associated with sGTIs. By using the indicators, a risk score was derived for individual at the community level,

predicting the risk of sGTIs. Healthcare workers and other stakeholders can use this tool for screening and early detection at the community level to complement current approaches in managing sexually transmitted and genital tract infections, even at the lowest level of the healthcare system. We, therefore, propose field testing and external validation as the next step before adoption.

Supplementary material. The supplementary material for this article can be found at <http://doi.org/10.1017/S0950268824001444>.

Acknowledgments. The authors are grateful to DHS for providing the data for this research work.

Author contribution. JT and MYN conceptualized the study. JT sought approval for access to GDHS data. JT undertook the statistical analysis. MYN, JT, SMS, EAU, and AEY drafted the initial manuscript and provided intellectual content revisions. All authors read and approved the final review manuscript.

Data availability. The data that support the findings of this study are available from DHS but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of DHS from https://dhsprogram.com/data/dataset_admin/login_main.cfm

Funding statement. No funding support.

Competing interest. No competing interest.

Ethical considerations and consent to participate. The GDHS protocol was reviewed and approved by the Ghana Health Service Ethical Review Committee and the ICF Institutional Review Board examined. The ICF IRB guarantees that the survey follows all U.S. regulations. Regulations for the protection of human subjects issued by the Department of Health and Human Services (45 CFR 46). Individual women's written consent was obtained during the data collection process for all participants. Privacy and confidentiality were strictly adhered to.

Consent for publication. Not applicable.

References

- [1] Choe H-S, Lee S-J, Yang SS, et al. (2018) Summary of the UAA-AAUS guidelines for urinary tract infections. *International Journal of Urology Official Journal, Japanese Urology Association* **25**, 175–185.
- [2] Flores-Mireles AL, Walker JN, Caparon M, et al. (2015) Urinary tract infections: Epidemiology, mechanisms of infection and treatment options. *Nature Reviews Microbiology* **13**, 269–284.
- [3] Hooton TM (2012) Uncomplicated urinary tract infection. *New England Journal of Medicine* **366**, 1028–1037.
- [4] Liu J, Zeng M, Yang L, et al. (2022) Prevalence of reproductive tract infections among women preparing to conceive in Chongqing, China: trends and risk factors. *Reproductive Health* **19**, 197.
- [5] Tomas ME, Getman D, Donskey CJ, et al. (2015) Overdiagnosis of urinary tract infection and underdiagnosis of sexually transmitted infection in adult women presenting to an emergency department. *Journal of Clinical Microbiology* **53**, 2686–2692.
- [6] WHO (2023) Sexually Transmitted Infections (STIs). [https://www.who.int/news-room/fact-sheets/detail/sexually-transmitted-infections-\(stis\)](https://www.who.int/news-room/fact-sheets/detail/sexually-transmitted-infections-(stis)) (accessed 8 August 2023).
- [7] Dadzie LK, Agbaglo E, Okyere J, et al. (2022) Self-reported sexually transmitted infections among adolescent girls and young women in sub-Saharan Africa. *International Health* **14**, 545–553.
- [8] Seidu A-A, Ahinkorah BO, Dadzie LK, et al. (2020) A multi-country cross-sectional study of self-reported sexually transmitted infections among sexually active men in Sub-Saharan Africa. *BMC Public Health* **20**, 1884.
- [9] Ratnaprabha G, Thimmaiah S, Johnson AR, et al. (2015) Prevalence and awareness of reproductive tract infections among women in select underprivileged areas of Bangalore city. *International Journal of Medical Science and Public Health* **4**, 1691–1696.
- [10] Surya B, Shivsakthimani R, Muthathal S, et al. (2021) A cross-sectional study on health-seeking behavior in relation to reproductive tract infection among ever-married rural women in Kancheepuram district, Tamil Nadu. *Journal of Family Medicine and Primary Care* **10**, 3424–3428.
- [11] Gamberini C, Juliana NCA, de Brouwer L, et al. The association between adverse pregnancy outcomes and non-viral genital pathogens among women living in sub-Saharan Africa: a systematic review. *Frontiers in Reproductive Health*, **5**. <https://www.frontiersin.org/articles/10.3389/frph.2023.1107931> (2023, accessed 13 August 2023).
- [12] Moragianni D, Dryllis G, Andromidas P, et al. (2019) Genital tract infection and associated factors affect the reproductive outcome in fertile females and females undergoing in vitro fertilization. *Biomedical Reports* **10**, 231–237.
- [13] Tang Y, Kurths J, Lin W, et al. (2020) Introduction to focus issue: When machine learning meets complex systems: Networks, chaos, and non-linear dynamics. *Chaos Interdisciplinary Journal of Nonlinear Science* **30**, 063151.
- [14] Kavzoglu T and Teke A (2022) Predictive performances of ensemble machine learning algorithms in landslide susceptibility mapping using random forest, extreme gradient boosting (XGBoost) and natural gradient boosting (NGBoost). *Arabian Journal of Science and Engineering* **47**, 7367–7385.
- [15] Gunn HJ, Rezvan PH, Fernández MI, et al. (2023) How to apply variable selection machine learning algorithms with multiply imputed data: A missing discussion. *Psychological Methods* **28**, 452–471.
- [16] Comulada WS, Rotheram-Borus MJ, Arnold EM, et al. (2023) Using machine learning to identify predictors of sexually transmitted infections over time among young people living with or at risk for HIV who participated in ATN protocols 147, 148, and 149. *Sexually Transmitted Diseases* **50**, 739–745.
- [17] Thivalapill N, Jasumback CL, Perry SH, et al. (2020) Predicting sexually transmitted infections among HIV+ adolescents and young adults: A novel risk score to augment syndromic management in Eswatini. *JAIDS Journal of Acquired Immune Deficiency Syndrome* **85**, 543.
- [18] GSS, GHS & ICF International (2015) *Ghana Demographic and Health Survey 2014*. Rockville: GSS, GHS, and ICF International.
- [19] Fisher AA and Way AA (1988) The demographic and health surveys program: An overview. *International Family Planning Perspectives* **14**, 15–19.
- [20] Baker FB (2001) *The Basics of Item Response Theory*, 2nd Edn. For full text: <http://ericae>, <https://eric.ed.gov/?id=ED458219> (accessed 3 April 2024).
- [21] Yang FM and Kao ST (2014) Item response theory for measurement validity. *Shanghai Archives of Psychiatry* **26**, 171–177.
- [22] DerSimonian R and Laird N (1986) Meta-analysis in clinical trials. *Controlled Clinical Trials* **7**, 177–188.
- [23] Bender R, Friede T, Koch A, et al. (2018) Methods for evidence synthesis in the case of very few studies. *Research Synthesis Methods* **9**, 382–392.
- [24] Chintalapudi N, Angeloni U, Battineni G, et al. (2022) LASSO regression modeling on prediction of medical terms among seafarers' health documents using tidy text mining. *Bioengineering* **9**, 124.
- [25] Sharma A (2017) Cross Validation in Machine Learning. *GeeksforGeeks*. <https://www.geeksforgeeks.org/cross-validation-machine-learning/> (accessed 24 July 2023).
- [26] Steyerberg EW, Vickers AJ, Cook NR, et al. (2010) Assessing the performance of prediction models: A framework for some traditional and novel measures. *Epidemiology (Cambridge, Mass.)* **21**, 128–138.
- [27] Grün B and Miljkovic T (2023) The automated bias-corrected and accelerated bootstrap confidence intervals for risk measures. *North American Actuarial Journal* **27**, 731–750.
- [28] Fenlon C, O'Grady L, Doherty ML, et al. (2018) A discussion of calibration techniques for evaluating binary and categorical predictive models. *Preventative Veterinary Medicine* **149**, 107–114.
- [29] Anguzu G, Flynn A, Musaaazi J, et al. (2019) Relationship between socio-economic status and risk of sexually transmitted infections in Uganda:

- Multilevel analysis of a nationally representative survey. *International Journal of STD & AIDS* **30**, 284–291.
- [30] **Curtis TJ, Field N, Clifton S**, et al. (2018) Household structure and its association with sexual risk behaviours and sexual health outcomes: Evidence from a British probability sample survey. *British Medical Journal Open* **8**, e024255.
- [31] **Seidu A-A, Agbaglo E, Dadzie LK**, et al. (2021) Self-reported sexually transmitted infections among sexually active men in Ghana. *BMC Public Health* **21**, 993.
- [32] **Ademas A, Adane M, Sisay T**, et al. (2020) Does menstrual hygiene management and water, sanitation, and hygiene predict reproductive tract infections among reproductive women in urban areas in Ethiopia? *PLoS One* **15**, e0237696.
- [33] **Lim RBT, Wong ML, Cook AR**, et al. (2015) Determinants of chlamydia, gonorrhoea, and coinfection in heterosexual adolescents attending the National Public Sexually Transmitted Infection Clinic in Singapore. *Sexually Transmitted Diseases* **42**, 450.
- [34] **Birhane BM, Simegn A, Bayih WA**, et al. (2021) Self-reported syndromes of sexually transmitted infections and its associated factors among reproductive (15–49 years) age women in Ethiopia. *Heliyon* **7**(7).
- [35] **Dadzie LK, Agbaglo E, Okyere J**, et al. (2022) Self-reported sexually transmitted infections among adolescent girls and young women in sub-Saharan Africa. *International Health* **14**, 545–553.
- [36] **Shrestha R, Karki P and Copenhaver M** (2016) Early sexual debut: A risk factor for STIs/HIV acquisition among a nationally representative sample of adults in Nepal. *Journal of Community Health* **41**, 70–77.
- [37] **Seidu A-A, Ahinkorah BO, Dadzie LK**, et al. (2020) A multi-country cross-sectional study of self-reported sexually transmitted infections among sexually active men in Sub-Saharan Africa. *BMC Public Health* **20**, 1–11.
- [38] **Balamurugan SS and Bendigeri N** (2012) Community-based study of reproductive tract infections among women of the reproductive age group in the Urban Health Training Centre Area in Hubli, Karnataka. *Indian Journal of Community Medicine Official Publishers, Indian Association of Preventative and Social Medicine* **37**, 34–38.