



A Logistic Regression Model to Predict Risk Factors Associated with HIV Prevalence: A Comprehensive Review

Danung Friday Ezekiel, Kazem E. Lasisi, Muhammad Sanda Adamu, Abuzairu Ahmad, Termen Nanfwang Yunana

Department of Mathematical Sciences,
Abubakar Tafawa Balewa University, Bauchi

ABSTRACT

The Human Immunodeficiency Virus (HIV) is a virus that attacks immune cells known as CD4 cells, which are a type of T cell. These are white blood cells that circulate throughout the body, detecting defects and abnormalities in cells as well as infections. When HIV targets and infiltrates these cells, it reduces the body's ability to fight off other diseases. This raise the likelihood and severity of opportunistic infections and cancer. Since there has been no effective cure for the global HIV/AIDS epidemic for more than 40 years, current control strategies have centered on reducing (or eliminating) new infections through risk reduction. As a result, the modified logistic regression model is used to predict HIV risk factors. The model will be used to assess the strength of the association between each of the hypothesized risk factors, as well as to determine the efficiency of the modified model, using the statistical tool SPSS version 25.0

ARTICLE INFO

Article History

Received: July, 2021

Received in revised form: July, 2021

Accepted: August, 2021

Published online: September, 2022

KEYWORDS

HIV/AIDS, Developed, Logistic Regression,
Risk Factors, Prevalence

INTRODUCTION

HIV is a virus that attacks and alters the immune system, making other infections and diseases more dangerous and devastating. Without treatment, the infection could progress to a more advanced stage of the disease known as AIDS. People living with HIV in countries with good access to healthcare, on the other hand, very rarely develop AIDS once they start treatment. The life expectancy of a person who carries the HIV virus is now approaching that of a person who tests negative for the virus, as long as they continue to take a combination of medications known as antiretroviral therapy (ART). The Human Immunodeficiency Virus (HIV) is a virus that attacks immune cells known as CD4 cells, which are a type of T cell. These are white blood cells that circulate throughout the body, detecting cell defects and anomalies as well as infections. When HIV targets and infiltrates these cells, the body's ability to fight off other diseases suffers. This raises the likelihood and severity of opportunistic infections and

cancer. However, a person can have HIV for a long time without showing symptoms.

According to a 2016 Kaiser Permanente study, between 1996 and 2016, the difference in life expectancy between HIV positive and HIV negative people shrank from 44 years to 12 years. According to the World Health Organization (WHO), a person living with HIV can regain a high quality of life with treatment, and as of mid-2017, 20.9 million people worldwide were receiving ART. While HIV is a life-changing illness, it is possible to live a long and fulfilling life while infected. HIV is a virus that attacks immune cells known as CD4 cells, which are a type of T cell. These are white blood cells that circulate throughout the body, detecting cell defects and anomalies as well as infections. HIV reduces the body's ability to fight other diseases when it targets and infiltrates these cells. This raises the likelihood and severity of opportunistic infections and cancer. However, a person can have HIV for a long time without showing symptoms.

Statement of the Problem

According to the World Health Organization (WHO), 75 million people have been infected with HIV and approximately 32 million have died as a result of the virus. At the end of 2018, 37.9 million [32.7-44.0 million] people were living with HIV worldwide. Tuot et al. (2020) use logistic regression to investigate risk factors for HIV infections among female entertainment workers, including marital status, age, type of entertainment establishment, working duration, and education level. Only females are taken into account in his study. Using a modified logistic regression, the researchers will base their findings on gender, age, marital status, level of education, HIV family history, number of sexual partners per week, sexually transmitted infection, and sexual contact with foreigners, use of drugs, alcohol consumption, smoking, and blood transfusion history. We also plan to use the coefficient of determination (R^2) and the ROC curve to compare the modified model to the logistic model to determine which model best fits the data.

Aim and Objectives

The goal of this study is to create a modified logistic regression model that can identify risk factors for HIV prevalence.

The following objectives help to achieve the aforementioned goal:

- (i) To ascertain the role of risk factors in the prevalence of HIV/AIDS
- (ii) Model parameters should be tested
- (iii) To assess the model's applicability

LITERATURE REVIEW

This section provides a quick introduction to the computational techniques used in the field of logistic regression models as mentioned in the literature in order to properly understand how they operate, their merits and shortcomings, their limitations, and how they accomplish their goals.

Logistic Regression

Logistic regression describes and estimates the relationship between a dependent variable Y , which has only two possible values based on the occurrence or non-occurrence of an event, and independent variables influencing that phenomenon. Its popularity stems from the lack of the need to satisfy many requirements for linear regression and general linear models, such as the linearity of the relationship between a dependent and an independent variable, as well as the normality and homoscedasticity of the distribution of independent variables, or the requirement to use variables given using the metric system of measures. Borucka et al, (2020) If we compute the probability of success (assume Y is a dichotomous variable with values of 1 for the occurrence of the event of interest), (success) and 0 for the opposite case (failure) the logistic regression model is described by the following equation:

$$P(Y = 1|x_1, x_2, \dots, x_n) = \frac{e^{\beta_0 + \sum_{i=1}^n \beta_i x_i}}{1 + e^{\beta_0 + \sum_{i=1}^n \beta_i x_i}} \quad (1)$$

Where β represent logistic regression coefficients and x_1, x_2, \dots, x_n represent independent variables that can be manipulated both quantitatively and qualitatively.

OVERVIEW OF FUNDAMENTAL CONCEPT AND COMPUTATIONAL INTELLIGENCE

The subsections that follow provide an overview of some fundamental concepts that are frequently used in relation to this proposed research work.

Linear Regression

Logistic regression is essentially a supervised classification algorithm; the target variable (or output), y , can only takes discrete values for a given set of features (or inputs), X . Contrary to popular belief, logistic regression is a regression model. The model creates a regression model to predict the likelihood that a given data entry belongs to the category labeled "1." Logistic regression, like linear regression, assumes that the data follows a linear function.

$$g(z) = \frac{1}{1+e^{-z}} \quad (2)$$

Only when a decision threshold is introduced into the equation does logistic regression transform into a classification technique. The threshold value is an important aspect of logistic regression and is determined by the classification problem itself.

Binomial

A binomial logistic regression predicts the likelihood that an observation will fall into one of two categories of dichotomous dependent variables, which can be either continuous or categorical. Binomial logistic regression is similar to linear regression in many ways, with the exception of the dependent variable's measurement type (i.e., linear regression uses a continuous dependent variable rather than a dichotomous one).

Multinomial

The logistic regression algorithm is extended to solve multiclass possible outcome problems given one or more independent variables in multinomial logistic regression. The probabilities of a categorically dependent variable with two or more possible outcome classes are predicted using this model. When the dependent categorical variable has two outcome classes, such as when students can either "pass" or "fail" an exam, the logistic regression model is used.

Ordinal

Ordinal logistic regression is a statistical method for modeling the relationship between an ordinal response variable and one or more explanatory variables. An ordinal variable is a categorical variable with a distinct ordering of category levels. The explanatory variables can be categorical or continuous.

REVIEW OF RELATED EMPIRICAL STUDIES

Since the discovery of the Human Immunodeficiency Virus (HIV) in 1983 as the causative organism of Acquired Immunodeficiency Syndrome (AIDS), the infection has reached epidemic proportions around the world. HIV/AIDS is a unique crisis in that it is both an emergency and a long-term

development issue. According to Tumer et al. (2006), HIV/AIDS is one of the most complicated health issues of the twenty-first century. Despite increased funding, political commitment, and progress in expanding access to HIV treatment, the AIDS epidemic outpaces all responses. Infections with the human immunodeficiency virus (HIV) continue to be a major social problem around the world.

According to WHO estimates, 36.7 million people were living with HIV infection and acquired immunodeficiency syndrome (AIDS) at the end of 2015, with 2.1 million new infections and 1.1 million deaths due to HIV-related causes (Osman et al., 2017). According to genetic evidence, HIV originated in west-central Africa in the early twentieth century. The Centers for Disease Control and Prevention (CDC) first identified AIDS in 1981, and its cause—HIV infection—was identified early in the decade. Since its discovery, AIDS has killed nearly 30 million people and infected approximately 34 million more. AIDS is considered a pandemic, a disease outbreak that is widespread and actively spreading. HIV/AIDS has had a significant impact on society, both as an illness and as a result of many misconceptions about it, such as the belief that it can be transmitted through casual non-sexual contact (Peeters et al. 2013). Women are a vulnerable demographic that is more susceptible to sexually transmitted infections (STIs), particularly the Human Immunodeficiency Virus (HIV). This is because of their genital anatomy, which exposes them to more infections during sexual intercourse, and their low sociocultural and economic status.

Several studies have found that age, multiple sexual partners, poverty, literacy, and occupation are all associated with HIV infection. In most Sub-Saharan African countries, resources have had a significant impact on HIV transmission. Noncompliance with antiretroviral therapy, which results in treatment failure in patients, has been identified as an additional risk factor for virus spread or even death in patients. (2017) Konate et al. Tuot et al. (2020) used multiple logistic regressions to investigate factors related to HIV infection. The adjusted HIV prevalence in this study of 3149 FEWs with a



mean age of 26.2 years (SD 5.7) was 3.2% (95% CI: 1.76-5.75).

In the multiple regression model, the odds of HIV infection were significantly higher among FEWs aged 31 to 35 (AOR 2.72, 95% CI 1.35-8.25) and 36 or older (AOR 3.62, 95% CI 1.17-7.79); FEWs who were not married but lived with a sexual partner (AOR 3.00, 95% CI 1.17-7.79); FEWs with at least ten years of formal education (AOR 0.32, 95% CI. Since the introduction of highly active antiretroviral therapy (ART) in the late 1990s, a large percentage of people living with HIV have avoided death and lived longer in good health. Despite this, PLHA are at a higher risk of mental disorders, particularly depression, due to social stigma, social dysfunction, long-term physical discomfort and illness, antiretroviral therapy side effects, and neurobiological changes. Depression appears to be more common in HIV-positive people than in HIV-negative people or the general population, according to evidence. Even depression symptoms that do not meet all of the diagnostic criteria for a depressive disorder have been identified as a significant factor associated with poorer health outcomes among people living with HIV, including impaired immunological response and mortality. As a result, screening for depression or depressive symptoms is a primary concern in identifying significant risk factors for health outcomes among HIV/AIDS patients (Wang et al., 2018). Mobili et al. (2010) used both principal component analysis (PCA) and partial least squares discriminant analysis (PLS-DA) to analyze and interpret Roman spectra collected from microorganisms of various species and recorded in the 2000 to 200 cm⁻¹ spectral range. They used PLS-DA in LOOCV of a classification model to develop a classification rule; they claimed that the results obtained demonstrated an acceptable classification among the strains under study, implying that it could be useful for the classification and discrimination of similar samples.

Gunda et al. (2018) investigated the prevalence and risk factors for mortality among HIV patients receiving ART in Northwestern rural Tanzania. Demographic information, the date of HIV diagnosis, the WHO stage, opportunistic infections, CD4, hemoglobin, the ART regimen,

and the time and outcome of treatment as dead or alive were collected and analyzed using STATA version 11. The findings revealed that 740 patients were studied in total. The median age was 35 (27-42) years, with 465 (62.8%) women predominating. 261 (35.3%) of the participants were in WHO stages 3 and 4. The majority of participants, 258 (34.9%), had a CD4 count of 200 cells/l at the start. Male gender (16.0% versus 9.0%), being divorced (OR = 2.7), WHO stages 3 and 4 (OR = 2.3), CD4 cells/l. (OR = 3.4), and severe anemia (OR = 6.6) were all independently associated with death. Wang et al. (2018) investigate the prevalence of actual uptake and willingness to use pre-exposure prophylaxis to prevent HIV acquisition among men who have sex with men in Hong Kong, China, in their article "Prevalence of actual uptake and willingness to use pre-exposure prophylaxis to prevent HIV acquisition among men who have sex with men in Hong Kong, China," the willingness to use daily oral PrEP under two cost scenarios, and potential issues related to PrEP use among men who. Only 1% had ever used PrEP, according to the findings. After being briefed on some PrEP facts, the prevalence of willingness to use daily oral PrEP was 7.7% if it could be purchased at private hospitals or clinics for HK\$8,000 (US\$1,032) per month (market rate) and 45.2% if it was provided free by Hong Kong public hospitals or clinics (free PrEP). HIV positivity among ANC attendees reflects the diversity of the HIV epidemic in South India. Contextual factors must be considered in addition to individual factors for effective HIV interventions. It is critical for effective intervention to identify district- and individual-level factors influencing ANC positivity. Thamattoor et al. (2015) identify individual and district-level factors associated with ANC-HIV positivity using multilevel logistic regression. From 2004 to 2007, the average ANC-HIV prevalence in the 24 integrated biological and behavioral assessment (IBBA) districts ranged from 0.25 to 3.25%. HIV positivity was significantly higher among ANC women aged 25 years [adjusted odds ratio (AOR): 1.49% confidence interval (95% CI): 1.27 to 1.76] compared to those aged 25 years; illiterate (AOR: 1.62; 95% CI: 1.03 to 2.45) compared to literate; employed in agriculture

(AOR: 1.34; 95% CI: 1.11 to 1.62) or with occupations like driver/helper/indus.

In their study to determine the prevalence and factors influencing modern contraceptive use among HIV-positive women in northern Tanzania, Tewabe et al. (2020) used multivariate logistic regression analysis to determine independent predictors of modern contraceptive use. According to the findings, 672 HIV-positive women were enrolled in total. Their average age was 36.4 years old (7.7). Fifty-four percent (362) were currently using modern contraceptives, with male condoms (76%; 275) being the most commonly used method, followed by Depo-Provera (28%; 101). A total of 33% (121) of users reported using dual contraception. Women with a primary education [AOR = 7.54, 95% CI: 1.51 - 12.48, P=0.014]; a post-secondary education [AOR = 6.23, 95% CI: 1.14 - 14.07, P=0.035]; are not currently on ARVs [AOR = 11.29, 95% CI: 2, 60 - 19.94, P=0.001]; have ever discussed contraceptive use with partner [AOR = 3.68, 95%

METHODOLOGY

The Proposed System

The goal of this study is to create a modified logistic regression model that can identify risk factors associated with HIV prevalence.

Logistic Regression

When the dependent variable is dichotomous, logistic regression is the appropriate regression analysis to perform (binary). The logistic regression is a predictive analysis, as are all regression analyses. Logistic regression is a statistical method for describing and explaining the relationship between one dependent binary variable and one or more nominal, ordinal, or interval independent variables. Logistic regression is a type of statistical regression analysis that predicts the outcome of a categorical dependent variable based on a set of predictors or independent variables. Logistic regression is a mathematically modeled method for defining the relationship between independent variables such as $X_1, X_2, X_3, \dots, X_n$, and Y , are binary dependent variable coded as 0 or 1 for two possible categories.

The Logistic Regression assigns each predictor a coefficient that measures its independent contribution variation in the dependent variable. If the response is "Yes," the dependent variable Y is 1, otherwise it is 0.

Predicted probabilities are expressed in model form as a natural logarithm (ln) of the odd ratio:

$$\ln[(P(Y))/(1-P(Y))]= \beta_0+ \beta_1X_1+ \beta_2X_2+ \dots+ \beta_nX_n. (3)$$

$$\text{And, } (P(Y))/(1-P(Y)) = e^{(\beta_0+ \beta_1X_1+ \beta_2X_2+ \dots+ \beta_nX_n)} \quad (4)$$

$$P(Y) = \beta_0+ \beta_1X_1+ \beta_2X_2+ \dots+ \beta_nX_n - P(Y) \beta_0+ \beta_1X_1+ \beta_2X_2+ \dots+ \beta_nX_n \quad (5)$$

$$P(Y) = \frac{e^{\beta_0+ \beta_1X_1+ \beta_2X_2+ \dots+ \beta_nX_n}}{1+ e^{\beta_0+ \beta_1X_1+ \beta_2X_2+ \dots+ \beta_nX_n}} \quad (6)$$

The modified model

$$P(Y)= \left\{ \frac{e^{\beta_0+ \beta_1X_1+ \beta_2X_2+ \dots+ \beta_nX_n}}{1+ e^{\beta_0+ \beta_1X_1+ \beta_2X_2+ \dots+ \beta_nX_n}} \right\} + \alpha \quad (7)$$

Where α represent STI History

Y is the dichotomous outcome; $X_1, X_2, X_3, \dots, X_k$ are the predictor variables, $\beta_0, \beta_1, \beta_2, \dots, \beta_k$ are the regression (model) coefficients and β_0 is the intercept

Assumptions on Logistic Regression

- (i) The predictors should not have a high correlation (multicollinearity)
- (ii) Independent variables are (X_n) from one another
- (iii) The binary dependent variable Y is statistically insignificant

Evaluation Parameters and Performance Metrics

The effectiveness of the proposed system will be evaluated after it is implemented. Using the statistical tool SPSS version 25.0, the model will be used to assess the strength of the association between each of the hypothesized risk factors, as well as to determine the efficiency of the modified model.

Predicted probabilities are expressed in model form as a natural logarithm (ln) of the odd ratio:

$$\ln \left[\frac{P(Y)}{1-P(Y)} \right] = \beta_0+ \beta_1X_1+ \beta_2X_2+ \dots+ \beta_kX_k \quad (8)$$

$$\text{and, } \frac{P(Y)}{1-P(Y)} = e^{\beta_0+ \beta_1X_1+ \beta_2X_2+ \dots+ \beta_kX_k} \quad (9)$$

$$P(Y) = \beta_0+ \beta_1X_1+ \beta_2X_2+ \dots+ \beta_kX_k - P(Y) \beta_0+ \beta_1X_1+ \beta_2X_2+ \dots+ \beta_kX_k \quad (10)$$

$$P(Y) = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k}} \quad (11)$$

Where, $\ln\left[\frac{P(Y)}{1-P(Y)}\right]$ is the log (odds) of the outcome

Y is the dichotomous outcome;
 $X_1, X_2, X_3, \dots, X_k$ are the predictor variables,
 $\beta_0, \beta_1, \beta_2, \dots, \beta_k$ are the regression (model) coefficients and

β_0 is the intercept.

P should meet following criteria;

- (i) It must be positive since ($P \geq 0$)
- (ii) It must always be less than equal to 1 (since $P \leq 1$)

IMPLEMENTATION OF THE SYSTEM

SPSS version 25.0 was used as a statistical tool to assess the strength of the association between each of the hypothesized risk factors, as well as the efficiency of the modified model. A Pentium® Core i7 processor, 8 GB of RAM, and a 1 TB hard drive are used in the system.

METHOD OF DATA COLLECTION

Tuot et al journal's article was the primary source of data for this study (2020). Several other journals and websites were visited during data collection. As a result, to assess the risk factors for HIV infection, the study employs modified logistic regression. Gender, age, marital status, level of education, HIV family history, number of sexual partners per week, history of sexually transmitted infection, sexual contact with foreigners, use of drugs, alcohol consumption, smoking, and blood transfusion history are among the risk factors considered.

CONCLUSION

The study produced a modified logistics regression model analysis. For future work, we propose developing an ensemble learning model for HIV prediction that combines logistic regression, an artificial neural network, and a support vector machine. We hope that these models will assist many people in making future predictions or determining who will be HIV patients, or that this will provide health educators with information to disseminate as part of their general public sensitization on HIV management.

Gender, level of education, HIV family history, number of sexual partners per week, use of drugs, sexual contact with foreigners, history of sexually transmitted infection, and blood transfusion history are all risk factors for HIV. Non-HIV patients must therefore examine their lifestyles in order to avoid contracting the virus.

REFERENCES

- Curtis, J. R., Harrold, L. R., Asgari, M. M., Deodhar, A., Salman, C., Gelfand, J. M., ... & Herrinton, L. J. (2016). Diagnostic prevalence of ankylosing spondylitis using computerized health care data, 1996 to 2009: underrecognition in a US health care setting. *The Permanente Journal*, 20(4).
- Tuot, S., Teo, A. K. J., Chhoun, P., Mun, P., Prem, K., & Yi, S. (2020). Risk factors of HIV infection among female entertainment workers in Cambodia: Findings of a national survey. *PLoS one*, 15(12), e0244357.
- Borucka, A. (2020). Logistic regression in modeling and assessment of transport services. *Open Engineering*, 10(1), 26-34.
- Hotez, P. J., Molyneux, D. H., Fenwick, A., Ottesen, E., Ehrlich Sachs, S., & Sachs, J. D. (2006). Incorporating a rapid-impact package for neglected tropical diseases with programs for HIV/AIDS, tuberculosis, and malaria: a comprehensive pro-poor health policy and strategy for the developing world. *PLoS medicine*, 3(5), e102.
- Osman, S. M., Osman, M. M., Osman, M. M., Magzoub, M., & Ali, S. E. E. (2017). Clinical Presentations of HIV/AIDS among Sudanese Patients in Khartoum, Sudan. *Cough*, 14, 14-0.
- Peeters, M., Jung, M., & Ayouba, A. (2013). The origin and molecular epidemiology of HIV. *Expert review of anti-infective therapy*, 11(9), 885-896.
- Konaté, D., Dahourou, H., Traoré, W., Ouedraogo, C., Bambara-Kankouan, A., Somda, A., ... & Sangaré, L. (2017). Risk factors associated with HIV prevalence in pregnant women in



- Burkina Faso, from 2006 to 2014. *African Journal of Clinical and Experimental Microbiology*, 18(2), 92-101.
- Wang, T., Fu, H., Kaminga, A. C., Li, Z., Guo, G., Chen, L., & Li, Q. (2018). Prevalence of depression or depressive symptoms among people living with HIV/AIDS in China: a systematic review and meta-analysis. *BMC psychiatry*, 18(1), 1-14.
- Mobili, P., Londero, A., De Antoni, G., Gomez-Zavaglia, A., Araujo-Andrade, C., Avila-Donoso, H., ... & Frausto-Reyes, C. (2010). Multivariate analysis of Raman spectra applied to microbiology: Discrimination of microorganisms at the species level. *Revista mexicana de fisica*, 56(5), 378-385.
- Gunda, D. W., Maganga, S. C., Nkandala, I., Kilonzo, S. B., Mpondo, B. C., Shao, E. R., & Kalluvya, S. E. (2018). Prevalence and risk factors of active TB among adult HIV patients receiving ART in northwestern Tanzania: a retrospective cohort study. *Canadian Journal of Infectious Diseases and Medical Microbiology*, 2018.
- Thamattoor, U., Thomas, T., Banandur, P., Duchesne, T., Abdous, B., Washington, R., & Alary, M. (2015). Multilevel analysis of the predictors of HIV prevalence among pregnant women enrolled in annual HIV sentinel surveillance in four states in Southern India. *PloS one*, 10(7), e0131629.
- Tewabe, T., Ayalew, T., Abdanur, A., Jenbere, D., Ayehu, M., Talema, G., & Asmare, E. (2020). Contraceptive use and associated factors among sexually active reproductive age HIV positive women attending ART clinic at Felege Hiwot Referral Hospital, Northwest Ethiopia: A cross-sectional study. *Heliyon*, 6(12), e05653.