An Application of CATANOVA and Logistic Regression on the Most Prevalent Sexually Transmitted Infection (A Case Study of the University of Nigeria Teaching Hospital)

Nnaemeka Martin Eze[1]
Department of Statistics, University of Nigeria, Nsukka, Nigeria

Oluchukwu Chukwuemeka Asogwa
Department of Mathematics, Computer Science, Statistics and Informatics, Alex Ekwueme Federal University Ndufu-Alike Ikwo, Nigeria

Samson Offorma Ugwu, Chinonso Michael Eze
Felix Obi Ohanuba, Tobias Ejiofor Ugah
Department of Statistics, University of Nigeria, Nsukka, Nigeria

ABSTRACT

This research focused on the application of CATANOVA and logistic regression on the most prevalent Sexually Transmitted Infection (STI) reported in the University of Nigeria Teaching Hospital from 2010-2020. A population of 20,704 patients was recorded to have contracted eight(8) selected STIs. Prevalence analysis was computed to determine the most prevalent STI. Two-way CATANOVA cross-classification was computed to ascertain the age group and gender that suffer more from the most prevalent STI. Three-way CATANOVA was computed to ascertain the association among drug prescription, age, and gender of the Gonorrhea patients. A logistic regression model was fitted to predict infertility as an effect of the most prevalent STI. The prevalence analysis showed Gonorrhea infection as the most prevalent STI at 33.08%. A population of 6,850 patients recorded to have contracted Gonorrhea infection from 2010-2020 was employed for the analysis. Two-way CATANOVA cross-classification showed that gender, age, and interaction effects were statistically significant at a 5% significance level. Male (3,752; 54.8%) suffers Gonorrhea infection more than female (3,098;45.2%) and aged 30-39 years (1,946; 28.4%) suffers it more than any other age interval. The interaction effect shows that the rate of contracting Gonorrhea infection by gender differs from one age interval to another. Three-way CATANOVA results showed that drugs prescribed for the treatment of Gonorrhea infection depend on gender and age. Logistic regression results showed that an increase in age, body mass index, blood pressure, blood sugar, bacteria quantity, and Gonorrhea history were associated with an increased likelihood of the Gonorrhea patient being infertile.

Keywords: *Chi-square test, Prediction, Prevalence*

[1] **Address correspondence to Nnaemeka Martin Eze:** nnaeneka.eze@unn.edu.ng

1. Introduction

This study focuses on the occurrence of different kinds of Sexually Transmitted Infections (STIs) in our societies. Scientists have proved that several infections have their origin and some can be cured while some cannot be cured. The U.S. Department of Health and Human Services reported that there are several ways in which one can contract these infections and this can be through sexual practices (Scatterwhite et al., 2013). Sexually Transmitted Infections (STIs) also known as Sexually Transmitted Diseases (STDs) are harmful micro-organisms that are very hard to control their growth in the body of their host. These infections are easily contracted through sex. Most STIs initially do not show symptoms. According to medical experts, infections can be called diseases only when they show symptoms and this is the reason STDs are known as STIs. Medical experts had said that the infections can easily be spread when there is no presence of symptoms of these infections. Some of the symptoms of STIs are vaginal discharge, penile discharge, ulcers on or around the genitals, and pelvic pain. Some STIs may cause infertility in both males and females and also poor development of a baby if contracted before or during pregnancy. Different bacteria, viruses, fungi, and parasites pathogenic are the major causes of STIs. Some of the bacterial STIs are chlamydia infections, gonorrhea or gonococci infection, cancroids, granuloma inguinal, and syphilis. Some of the viral STIs are genital herpes, HIV/AIDS, Viral hepatitis (Hepatitis B virus), and genital warts. Some of the fungal STIs are candidiasis and Parasitic STIs include crab louse, scabies, and Trichomoniasis (Scatterwhite et al., 2013). Despite the contamination of some STIs through sex, one can contact them through blood and tissues, breastfeeding or during child delivery.

The contamination of STIs from one, and another or from surrounding objects can be prevented (Center for Diseases Prevention and Control, 2013). Azmi et al., (2008) presented their prevalence analysis from child-bearing-age women and the result showed that the prevalence of C. trachomatis infection was 0.6% and 0.5%, among symptomatic and asymptomatic women respectively, N. gonorrhoeae was 0.9% and 2.2%, T. pallidum 0.0% and 0.0%, and Tr. vaginalis was 0.7% and 0.5%. It was noted from the result that there was no significant difference in the prevalence rate between symptomatic and asymptomatic women. Kesah et al., (2013) stated that improvement in hand washing, clean toilets, abstaining from sex, condom usage, rational employment of examination methods, medical diagnostics testing for both men and women, attitude change, and prevention education should be consistently highlighted. Otaru and Ogbonda (2020) studied the application of categorical data-nested design of knowledge and control practices of Hepatitis B Virus (HBV) infection using the two-way CATANOVA technique. They considered frequency data from university students in three universities involving response rate of student's knowledge and control practices of HBV infection using a scale of good, fair, and poor. It was noted from the result that there was no significant difference in the knowledge and control practices of HBV infection of the students in the three considered universities at a 5% level of significance. Deyhoul et al., (2017) studied infertility rate risk factors and the result showed that infertility in men and women could be caused by sexually transmitted infections and hormonal disorders. Some lifestyle factors can also cause infertility such as obesity, nutrition, smoking/alcohol consumption, mobile phone use, sexual violence, and anxiety.

It has been known that Sexually Transmitted Infections (STIs) have sporadically increased over the years and of course have caused more harm than good in our societies. These infections could lead to various dangerous ailments such as infertility, pelvic inflammatory

disease in women, ectopic pregnancy, and serious effects on pregnancy which might lead to miscarriage, failure of development of a new baby, blindness, congenital defects, and so on. This study aims to know the most prevalent Sexually Transmitted Infection (STI) among the reported cases of STIs in the University of Nigeria Teaching Hospital; the gender and ages that always suffered from the most prevalent STI; examine if the prescribed drugs depend on patient's gender and age; and examine the reproductive status of the patients suffering the most prevalent STI, that is, to know if the carrier of the infection is fertile or infertile.

This study focuses only on eight (8) major sexually transmitted infections (Chlamydia, Gonorrhea, Syphilis, Herpes, Hepatitis B, Trichomoniasis, Human Immuno Deficiency Virus (HIV), Human Papilloma Virus (HPV)) contracted by both males and females which has attained sexual age as reported at University of Nigeria Teaching Hospital (UNTH) from 2010 to 2020. The significance of this study tends to educate Nigerians and the world at large about the existence of sexually transmitted infections in our societies and their risk factors. It will also notify people about the most prevalent STI, the gender and age interval that is more likely to be at risk of it, and more precisely, educate them on how to take precautionary measures.

2. Materials and Methods

2.1 Data and sampling design

The data used in this study were secondary data collected from eight (8) types of Sexually Transmitted Infections (STIs) reported in the Department of Micro Biology, University of Nigeria Teaching Hospital (UNTH). To determine the most prevalent STI, a population of 20,704 patients that reported to have contracted eight (8) selected STIs (Chlamydia (4,855), Gonorrhea (6,850), Syphilis (1,680), Trichomoniasis (1,770), Herpes (483), Hepatitis-B (602), Human Papilloma Virus (619) and Human Immune-deficiency Virus (3,845)) in the years 2010 through 2020 were collected and the prevalence method of analysis was used to ascertain the most prevalent STI among them. Furthermore, the record also showed that there were 6,850 reported cases of the most prevalent STI, and the data were presented using a randomized complete block design in which a K-dimensional vector $\left[n_{ijk}\right]$ of nominal responses are observed in frequencies in the ij^{th} plot (see Table 1). These most prevalent STI data were analyzed using categorical analysis of variance (CATANOVA) and logistic regression.

2.2 Ethical approval

The ethical issues in this study were addressed by making sure that anonymity and confidentiality are highly maintained when the need arises either from the data collection or any sources of information, and the consent of patients was respected. Therefore, all procedures performed in this research that involved patients and healthcare workers were in accordance with the ethical standards of the University of Nigeria Teaching Hospital (UNTH).

2.3 Models

2.3.1 Prevalence Rate: Prevalence is an epidemiology characteristic that is easily measured using survey data or medical records. To establish prevalence, researchers randomly select a sample (smaller group) from the entire population they want to describe. Using random selection methods increases the chances that the characteristics of the sample will be representative of (similar to) the characteristics of the population. For a representative sample, prevalence is the number of people in the sample with the characteristics of interest divided by the total number of people in the sample.

$$\left(\text{i. e., Prevalence formula} = \frac{\text{number of people in the sample with the characteristics of interest}}{\text{total number of people in the sample}}\right).$$

2.3.2 CATANOVA: The categorical analysis of variance (CATANOVA) is a technique designed to help the researcher identify the variation between treatments of interest**.** This CATANOVA is used to solve the problem in the analysis of variance when the observations are nominal without any underlying metric and it was also formulated to solve the erroneous analysis of nominal data by using the chi-square test (Onukogu, 1985; Otaru and Ogbonda, 2020). In addition, there are several methods for analyzing categorical data in which some of these methods use data transformation before proceeding to analyze the data. The transformation method to be used may depend on the classification of categorical data (Fienberg, 1973; Florian, 2008; Onukogu, 2014; Singh, 2004). In this research, two-way and three-way CATANOVAs are adopted and there is no loss in generality using the method for unequal levels of factors that do not differ significantly.

Table 1 shows the data layout for two-way cross classification or a randomized complete block design in which a K-dimensional vector $[n_{ijk}]$ of nominal responses are observed in frequencies in the ij^{th} plot. In this Table 1, the main factor A ranging from 1 to I and main factor B ranging from 1 to J have from 1 to K quanta responses per unit (D'Ambra et al., 2005; Anderson and Landis, 1980, 1982; Light and Margolin, 1971; Margolin and Light, 1974). Table 2 depicted the CATANOVA table that contains the source of variation, degrees of freedom (df), the sum of squares (SS) which is the trace of its variance-covariance matrix, test ratio from chi-square calculated, a critical value from chi-square tabulated and hypotheses for the study.

Furthermore, this study assumed that the data follows:

- Multi-nominal distribution

  $$P(\{n_{ijk}\};\{\pi_{ijk}\}) = \binom{n_{ij}}{n_{ij1},\ldots,n_{ijK}}\prod_{k=1}^{K}(\pi_{ijk})^{n_{ijk}}$$

  $n_{ijk} = 0, 1, \cdots, n_{ij}$ and $\pi_{ijk} = \frac{n_{ijk}}{n_{ij}};\ 0 \le \pi_{ijk} \le 1$
- Independence: The levels and blocks each act independently. That is, n_{ijk} and $n_{i'j'k}$ are statistically independent $\forall i \neq i'$ and $\forall j \neq j'$.
- Constant variance: $\text{var}(n_{ijk}) = n\pi_{ijk}(1-\pi_{ijk})$. The variance is not constant because it depends on i, j and k.
  $\pi_{ijk} > 0,\ \sum_{k=1}^{K}\pi_{ijk} = 1,\ \sum_{k}^{K} n_{ijk}$ is held fixed (i.e., grand total over k for j)

Table 1: The data layout for two-way CATANOVA cross-classification or randomized complete block design.

	B(j)												
	b1				b2				. . .	bJ			
A(i)	1	2	….	K	1	2	….	K	. . .	1	2	….	K
1	n_{111}	n_{112}	….	n_{11K}	n_{121}	n_{122}	….	n_{12K}	….	n_{1J1}	n_{1J2}	….	n_{1JK}
2	n_{211}	n_{212}	….	n_{21K}	n_{221}	n_{222}	….	n_{22K}	….	n_{2J1}	n_{2J2}	….	n_{2JK}
.	.	.		.	.	.		.		.	.		.
.	.	.	….	.	.	.	….	.	….	.	.	….	.
.	.	.		.	.	.		.		.	.		.
I	n_{i11}	n_{i12}	….	n_{i1K}	n_{i21}	n_{i22}	….	n_{i2K}	….	n_{iJ1}	n_{iJ2}	….	n_{iJK}

Table 2: Summary of two-way CATANOVA cross-classification of nominal data.

Source	df	SS	Test Ratio	Critical Value	Hypothesis
Row(Ai)	I-1	RSS	χ^2_{RT}	$\chi^2_{(I-1)(K-1)}$	$HO_R: \pi_{ijk} = \pi_{jk} \forall_i$
Column(Bj)	J-1	CSS	χ^2_{CT}	$\chi^2_{(J-1)(K-1)}$	$HO_C: \pi_{ijk} = \pi_{ik} \forall_j$
Interaction(AB)	(I-1)(J-1)	NSS	χ^2_{NT}	$\chi^2_{(I-1)(J-1)(K-1)}$	$HO_{RC}: \pi_{ijk} = \pi_k \forall_{ij}$
Weight Units	n-IJ	WUSS	-		-
Total	n-1	TSS	-		-

Computation of Sum of Squares

Total Sum of Square (TSS) = $n - \frac{\sum_k n^2_{..k}}{n}$; where $n_{..k} = \sum_{ij} n_{ijk}$ (1)

Within Unit Sum of Square (WUSS) = $n - \sum_{ij} \frac{\sum_k n^2_{ijk}}{n_{ij}}$ (2)

Between Row Sum of Square (BRSS) = $n - \sum_i \frac{\sum_k n^2_{i.k}}{n_i}$; where $n_{i.k} = \sum_j n_{ijk}$ (3)

Between Column Sum of Square (BCSS) = $n - \sum_j \frac{\sum_k n^2_{.jk}}{n_j}$; where $n_{.jk} = \sum_i n_{ijk}$ (4)

Row Sum of Square (RSS) = TSS – BRSS (5)

Column Sum of Square (CSS) = TSS – BCSS (6)

Interaction Sum of Square (NSS) = BCSS + BRSS – TSS – WUSS (7)

Two-way CATANOVA cross classification model

$$E(\hat{\pi}_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_{ij} \quad (8)$$

where $\hat{\pi}_{ijk}$ is the probability that k^{th} observation occurs in the i^{th} level of factor A and j^{th} level of factor B, i.e., $\hat{\pi}_{ijk} = P_{ijk} = \frac{n_{ijk}}{n_{ij}}$, ($n_{ijk}$ is the k^{th} observation in the ij^{th} cell, n_{ij} is the sum of k^{th} observation in the ij^{th} cells, i.e., $n_{ij} = \sum_k n_{ijk}$), μ is a constant for k^{th} observation, α_i ($i = 1, 2, \ldots, I$) is the effect of the i^{th} level of factor A, β_j ($j = 1, 2, \ldots, J$) is the effect of the j^{th} level of factor B, γ_{ij}(($i = 1, 2, \ldots, I$) and ($j = 1, 2, \ldots, J$)) is the interaction between the i^{th} level of factor A and j^{th} level of factor B. In nominal data, the sum of squares is the trace of its variance-covariance matrix and the parameter π_{ijk} may be considered fixed or random with probability density $h(\pi_{ijk})$ ranging from 0 to 1 depending on whether I and J are random or fixed (Anderson, 1958; Onukogu, 1985; Onukogu, 2014; Scheffe, 1959).

Hypotheses

$H_{0R}: \pi_{ijk} = \pi_{jk}$, i. e., $\alpha_i = 0\ \forall_i$ (There is no row effect)
$H_{1R}: \pi_{ijk} \neq \pi_{jk}$, i. e., $\alpha_i \neq 0\ for\ at\ least\ one\ (i)$ (There is row effect)

$H_{0C}: \pi_{ijk} = \pi_{ik}$, i. e., $\beta_j = 0\ \forall_j$ (There is no column effect)
$H_{1C}: \pi_{ijk} \neq \pi_{ik}$, i. e., $\beta_j \neq 0\ for\ at\ least\ one\ (j)$ (There is column effect)

$H_{0RC}: \pi_{ijk} = \pi_k$, i. e., $\gamma_{ij} = 0\ \forall_{ij}$ (There is no interaction effect)
$H_{1RC}: \pi_{ijk} \neq \pi_k$, i. e., $\gamma_{ij} \neq 0\ for\ at\ least\ one\ pair\ (ij)$ (There is an interaction effect)

Test Statistic

$$\chi^2_{RT} = \frac{(K-1)(n-1)RSS}{TSS} \sim \chi^2_{(I-1)(K-1)};\ \alpha$$

$$\chi^2_{CT} = \frac{(K-1)(n-1)CSS}{TSS} \sim \chi^2_{(J-1)(K-1)};\ \alpha$$

$$\chi^2_{NT} = \frac{(K-1)(n-1)NSS}{TSS} \sim \chi^2_{(I-1)(J-1)(K-1)};\ \alpha$$

Decision rule

Reject H_{0R} if $\chi^2_{RT} \geq \chi^2_{(I-1)(K-1)}$,

Reject H_{0C} if $\chi^2_{CT} \geq \chi^2_{(J-1)(K-1)}$, and

Reject H_{0RC} if $\chi^2_{NT} \geq \chi^2_{(I-1)(J-1)(K-1)}$, at specified level of significance (5%). Fail to reject if otherwise.

Table 3: The data layout for the 3-way contingency table.

	Y_1				Y_2					
	Z_1	Z_2	n_{+j+}	π_{+j+}	Z_1	Z_2	n_{+j+}	π_{+j+}	n_{i++}	π_{i++}
X_1	$n_{111}(\hat{f}_{111})$	$n_{112}(\hat{f}_{112})$	n_{11+}	π_{11+}	$n_{121}(\hat{f}_{121})$	$n_{122}(\hat{f}_{122})$	n_{12+}	π_{12+}	n_{1++}	π_{1++}
X_2	$n_{211}(\hat{f}_{211})$	$n_{212}(\hat{f}_{212})$	n_{21+}	π_{21+}	$n_{221}(\hat{f}_{221})$	$n_{222}(\hat{f}_{222})$	n_{22+}	π_{22+}	n_{2++}	π_{2++}
n_{++k}	n_{+11}	n_{+12}	n_{+1+}	-	n_{+21}	n_{+22}	n_{+2+}	-	n	-
π_{++k}	π_{+11}	π_{+12}		π_{+1+}	π_{+21}	π_{+22}		π_{+2+}		$\sum_{ijk} \pi_{ijk} = 1$

where; $X_i (i = 1, 2, \cdots, \mathrm{I})$, $Y_j (j = 1, 2, \cdots, \mathrm{J})$, $Z_k (k = 1, 2, \cdots, \mathrm{K})$, n_{ijk} is the observed frequency in ijk cell, $n = \sum_i \sum_j \sum_k n_{ijk}$ is the total observation, $n_{i++} = \sum_j \sum_k n_{ijk}$ is the marginal row total, $n_{+j+} = \sum_i \sum_k n_{ijk}$ is the marginal column total, and $n_{++k} = \sum_i \sum_j n_{ijk}$ is the marginal k^{th} observation total, $\hat{f}_{ijk} = n\left(\frac{n_{i++}}{n}\right)\left(\frac{n_{+j+}}{n}\right)\left(\frac{n_{++k}}{n}\right)$ is the estimated expected frequency in ijk cell, π_{ijk} is the probability value in ijk cell, $\pi_{i++} = \left(\frac{n_{i++}}{n}\right)$ is the row marginal probability, $\pi_{+j+} = \left(\frac{n_{+j+}}{n}\right)$ is the column marginal probability, $\pi_{++k} = \left(\frac{n_{++k}}{n}\right)$ is the k^{th} marginal probability, $\pi_{+jk} = \left(\pi_{+j+} \cap \pi_{++k}\right) = \left(\frac{n_{+jk}}{n}\right)$ is the intersection of column marginal probability and k^{th} marginal probability, and $\sum_{ijk} \pi_{ijk} = \sum_i \pi_{i++} = \sum_j \pi_{+j+} = 1$.
Note: $\cap$ is an intersection symbol.

Hypothesis for conditional independency test in the 3-way contingency table

$H_0: \pi_{ijk} = \pi_{i++} \times \left(\pi_{+j+} \cap \pi_{++k}\right)$ (X variable is independent of Y and Z variables)

$H_1: \pi_{ijk} \neq \pi_{i++} \times \left(\pi_{+j+} \cap \pi_{++k}\right)$ (X variable depends on Y and Z variables)

Test Statistic

$$\chi^2 = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{c} \frac{\left(n_{ijk} - \hat{f}_{ijk}\right)^2}{\hat{f}_{ijk}} \sim \chi^2_{ijk-(i+j+k)+2}$$

Decision Rule: Reject H_0 if $\chi^2{}_{cal} \geq \chi^2{}_{tab}$. Fail to reject if otherwise.

2.3.3 Logistic Regression: This is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analysis, logistic regression is a predictive analysis used to describe data and to explain the relationship between one dependent binary response variable, which takes values 1 and 0, and one or more nominal, ordinal, interval, or ratio level independent variable(s). The logistic regression gives each predictor a coefficient that measures its independent contribution to variation in the dependent variable. The dependent variable Y takes the value 1 if the response is "yes'' and takes a value 0 if the response is "no''. Logistic regression calculates the probability of success over the probability of failure. The results of the analysis are in the form of an odds ratio (Boateng and Abaye, 2019).

The model form for predicted probabilities is expressed as a natural logarithm (ln) of the odds ratio:

$$\mathrm{Ln}(ODDS) = \ln\left(\frac{\mathrm{P(Y)}}{1-\mathrm{P(Y)}}\right) = \beta_0 + \beta_1 \mathrm{X}_1 + \beta_2 \mathrm{X}_2 + \cdots + \beta_\mathrm{m} \mathrm{X}_\mathrm{m} \quad (9)$$

$$\frac{\mathrm{P(Y)}}{1-\mathrm{P(Y)}} = e^{\beta_0+\beta_1 \mathrm{X}_1+\beta_2 \mathrm{X}_2+\cdots+\beta_\mathrm{m} \mathrm{X}_\mathrm{m}} \quad (10)$$

$$\mathrm{P(Y)} = e^{\beta_0+\beta_1 \mathrm{X}_1+\beta_2 \mathrm{X}_2+\cdots+\beta_\mathrm{m} \mathrm{X}_\mathrm{m}} - \mathrm{P(Y)} e^{\beta_0+\beta_1 \mathrm{X}_1+\beta_2 \mathrm{X}_2+\cdots+\beta_\mathrm{m} \mathrm{X}_\mathrm{m}} \quad (11)$$

$$= \frac{e^{\beta_0+\beta_1 \mathrm{X}_1+\beta_2 \mathrm{X}_2+\cdots+\beta_\mathrm{m} \mathrm{X}_\mathrm{m}}}{1+e^{\beta_0+\beta_1 \mathrm{X}_1+\beta_2 \mathrm{X}_2+\cdots+\beta_\mathrm{m} \mathrm{X}_\mathrm{m}}} \quad (12)$$

where; $\frac{P(Y)}{1-P(Y)}$ is the odds ratio, $\ln\left(\frac{\mathrm{P(Y)}}{1-\mathrm{P(Y)}}\right)$ is the log odds or "logit'' of the outcomes, Y is the dichotomous outcome, P(Y = 1) is the probability of an event, $\times_i$ ($\mathrm{i} = 1,2, \ldots, \mathrm{m}$) are the predictors, β_i($\mathrm{i} = 1, 2, \ldots, \mathrm{m}$) are unknown regression parameters to be estimated and β_0 is the intercept (i.e., constant).

2.3.3.1 Goodness of Fit Test. It is also known as the Hosmer-Lemeshow test which represents a chi-square test used for testing the adequacy of the model for fitting the data. The null hypothesis is that the model is adequate to fit the data and we will only reject this null hypothesis if the p-value is less than 0.05 (Abdulqader, 2017). It is given as

$$H = \sum_{i=1}^{g} \frac{(O_i - E_i)^2}{E_i} \tag{13}$$

where O_i and E_i denote the observed and expected frequencies , respectively.

Table 4: Values of the logistic regression model when the independent variable is dichotomous.

Outcome variable (Y)	Independent variable (X)	
	X = 1	X = 0
Y = 1	$P(1) = \frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}$	$P(0) = \frac{e^{\beta_0}}{1+e^{\beta_0}}$
Y = 0	$1 - P(1) = \frac{1}{1+e^{\beta_0+\beta_1}}$	$1 - P(0) = \frac{1}{1+e^{\beta_0}}$

The odds ratio is then computed as:

$$Odds\ ratio\ (OR) = \frac{\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}} \Big/ \frac{1}{1+e^{\beta_0+\beta_1}}}{\frac{e^{\beta_0}}{1+e^{\beta_0}} \Big/ \frac{1}{1+e^{\beta_0}}} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{(\beta_0+\beta_1)-\beta_0} = e^{\beta_1} \tag{14}$$

Hence, for logistic regression with a dichotomous independent variable coded 1 and 0, the relationship between the odds ratio and the regression coefficient is $Odds\ ratio(OR) = e^{\beta_1}$.

3. Results and Discussions

Figure 1 is a pie chart representation of the yearly percentage of reported cases of Sexually Transmitted Infections (STIs) as depicted in Table 5. From this figure, the years 2019(12%) and 2017(12%) had the highest reported cases of eight (8) types of STIs and were followed by the years 2016(11%), 2012(11%), 2013(10%), 2011(10%), 2018(8%), 2015(8%), 2010(8%), 2020(5%) and 2014(5%). Figure 1 also shows that there is a difference in yearly reports of STIs. This may be due to a lack of knowledge about the harmfulness of STIs in society.

Table 5: Reported cases of selected STIs from 2010-2020.

YEARS	CHL.	GON.	SYPH.	TRICO.	HERPES	HEPA.B	HPV	HIV	TOTAL
2010	201	603	137	159	85	61	18	414	1678
2011	167	941	226	173	11	28	71	455	2072
2012	437	987	148	158	35	73	15	441	2294
2013	698	434	76	178	43	24	88	397	1938
2014	116	392	64	147	33	35	78	243	1108
2015	576	503	115	131	15	84	31	260	1715
2016	705	741	123	188	35	95	23	272	2182
2017	461	982	326	118	58	159	54	262	2420
2018	516	316	178	174	16	19	98	337	1654
2019	826	751	148	181	89	17	119	418	2549
2020	152	200	139	163	63	7	24	346	1094
Total	4855	6850	1680	1770	483	602	619	3845	**20,704**

CHL.= Chlamydia, GON.= Gonorrhea, SYPH.= Syphilis, Herpes, HPV = Human Papilloma Virus, TRICO.=Trichomoniasis, HEPA B.= Hepatitis B Virus, HIV= Human Imuno Deficiency Virus

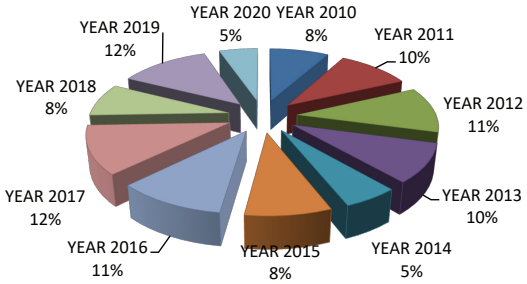

Figure 1: Yearly percentage of reported cases of Sexually Transmitted Infections.

Table 6: Prevalence rate of the eight (8) selected sexually transmitted infections (2010-2020)

Sexually Transmitted Infection (STI)	Prevalence Rate	Percentage of Prevalence Rate
Chlamydia	0.2344	23.44
Gonorrhea	0.3308	33.08
Syphilis	0.0812	8.12
Triconomiasis	0.0854	8.54
Herpes	0.0233	2.33
Hepatitis B Virus	0.0290	2.90
Human Papilloma Virus	0.0298	2.98
Human Immuno Deficiency Virus	0.1856	18.56

$$\text{Prevalence formula} = \frac{number\ of\ each\ STI}{Total\ number\ of\ STI} \times 100 \qquad (15)$$

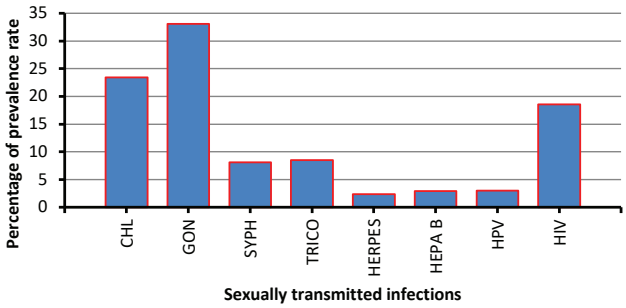

Figure 2: Bar chart for the prevalence rate of reported cases of sexually transmitted infections.

Figure 2 is a bar chart representation of the results of the prevalence rate percentage as depicted in Table 6. As can be seen from this Figure 2, Gonorrhea infection with a 33.08% rate appears to be the most prevalent among the eight selected sexually transmitted infections reported in the University of Nigeria Teaching Hospital from 2010-2020 when compared with Chlamydia, Syphilis, Trichomoniasis, Herpes, Human Papilloma Virus (HPV), Hepatitis B, Human Immuno-deficiency Virus (HIV) with 23.44%, 8.12%, 8.54%, 2.33%, 2.90%, 2.98%, and 18.56% respectively. From the data in Table 7, the percentage (54.8%) of males that suffered from Gonorrhea infection is more than the percentage (45.2%) of females. This shows that the male suffers from Gonorrhea infection more than female. Also, note that 28.4% of Gonorrhea patients are at the age interval of 30-39 years, 23.4% are at the age of 20-29 years, 18.5% are at the age of 40-49 years, 15% are the age of 50 years and above while14.7% are at the age of fewer than 20 years. These show that the age interval of 30-39 years suffers Gonorrhea infection more than any other age interval.

Table 7: Two-way contingency table depicting the response of gender and ages of gonorrhea patients reported in UNTH from 2010-2020.

Gender (i)	Age (j)															Total $n_{i.k}$		Total $n_{i..}$
	< 20 (j_1)			20-29 (j_2)			30-39 (j_3)			40-49 (j_4)			50+ (j_5)					
	History response			History response			History response			History response			History response					
	YES	NO	Total $n_{i1.}$	YES	NO	Total $n_{i2.}$	YES	NO	Total $n_{i3.}$	YES	NO	Total $n_{i4.}$	YES	NO	Total $n_{i5.}$	YES	NO	
Male	128	383	511	395	437	832	463	545	1008	388	415	803	347	251	598	1721	2031	**3752**
Female	154	340	494	272	498	770	385	553	938	283	181	464	234	198	432	1328	1770	**3098**
Total $n_{.jk}$	282	723	**1005**	667	935	**1602**	848	1098	**1946**	671	596	**1267**	581	449	**1030**	**3049**	**3801**	**6850**

Table 8: Table for the significance of gender, age, and interaction between gender and age effects.

Source	DF	Sum of Squares	Test Ratio	Critical Value	Decision
Gender (Row)	1	3.06	6.19	3.84	significant, (reject H_{0R})
Age (Column)	4	104.63	211.78	9.49	significant, (reject H_{0C})
Gender*Age	4	23.15	46.86	9.49	significant, (reject H_{0RC})
Within unit	6840	3252.88	-	-	
Total	6849	3383.72	-	-	

From the results in Table 8, we noticed a statistically significant difference in gender $\left(\chi^2_{RT(Cal)} = 6.19 > \chi^2_{RT(tab)} = 3.84\right)$, and a statistically significant difference in age $\left(\chi^2_{CT(Cal)} = 211.79 > \chi^2_{CT(tab)} = 9.49\right)$. The significant difference in gender means that a particular gender suffers more from Gonorrhea infection than another gender. The significant difference in age means that a particular age group is the most likely age group that suffers from Gonorrhea infection. It was noticed also that there is a statistically significant difference in the interaction between gender and age at a 5% significance level $\left(\chi^2_{NT(Cal)} = 46.86 > \chi^2_{NT(tab)} = 9.49\right)$. The significant difference in the interaction between gender and age intervals of Gonorrhea patients means that the rate of contracting Gonorrhea infection by males and females differs from one age interval to another. Moreover, the data in Table 7 showed that 3801(55.5%) of Gonorrhea patients do not have a Gonorrhea infection history, while 3049(44.5%) have it. It shows that there is a spread of Gonorrhea infection between the genders. (see Appendix A for the computation of results in Table 8)

Table 9.1: Three-way contingency table depicting gender, ages, and drug prescription for gonorrhea infection (2010 - 2020).

GENDER	**Male**						**Female**						
AGE	**Response in Male Ages**					$\mathbf{n_{+j_1+}}$	**Response in Female Ages**						
DRUG	**< 20**	**20-29**	**30-39**	**40-49**	**50 +**		**< 20**	**20-29**	**30-39**	**40-49**	**50+**	$\mathbf{n_{+j_2+}}$	$\mathbf{n_{i++}}$
CEFTR.	110	188	206	171	118	793	129	196	229	80	63	697	1490
STREPT.	107	163	202	164	112	748	111	135	216	55	75	592	1340
DOXY.	109	123	178	115	116	641	61	102	110	103	96	472	1113
GENTA.	130	192	248	205	159	934	104	211	184	128	132	759	1693
OFLO.	55	166	174	148	93	636	89	126	199	98	66	578	1214
$\mathbf{n_{++k}}$	511	832	1008	803	598	**3752**	494	770	938	464	432	**3098**	**6850**

CEFT=Ceftriaxone, STREPT.= Streptomycin, DOXY.= Doxycycline, GENTA.=Gentamicin , OFLO.= Ofloxacin

Table 9.2: $n_{++k}-$ table computed from table 9.1 for three-way contingency table.

Age	**Total responses in ages**		$\mathbf{n_{++k}}$
	Male	**Female**	
< 20	511	494	1005
20-29	832	770	1602
30-39	1008	938	1946
40-49	803	464	1267
50+	598	432	1030

Hypotheses:

$H_0: \pi_{ijk} = \pi_{i++} \times (\pi_{+j+} \cap \pi_{++k})$ (The drugs used to treat Gonorrhea infection are independent of gender and age)

$H_1: \pi_{ijk} \neq \pi_{i++} \times (\pi_{+j+} \cap \pi_{++k})$ (The drugs used to treat Gonorrhea infection are dependent on gender and age)

Computed Test Statistic:

$\chi^2_{cal} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\frac{(n_{ijk}-f_{ijk})^2}{f_{ijk}} = 221.30$ (see appendix Table 1 in Appendix B)

$\chi^2_{tab} = \chi^2_{40} = 55.75$ (see Appendix B)

The result of the analysis for Tables 9.1 and 9.2 showed that the prescribed drugs for patients suffering from Gonorrhea infection depend on the age and gender of the patient, $\chi^2_{cal} = 221.30$ is greater than $\chi^2_{tab} = 55.75$, at a 5% significance level.

Table 10: Logistic regression analysis code sheet for dependent and independent variables data.

Variable	Description	Codes/ Values	Name	Data type
$\times_1$	Age	Years	Age	Numerical
$\times_2$	History of Gonorrhea	0 = No 1 = Yes	History	Nominal
$\times_3$	Body Mass Index	kg/m^2	BMI	Numerical
$\times_4$	Blood Pressure	mm Hg	BP	Numerical
$\times_5$	Blood Sugar	mg/dl	BS	Numerical
$\times_6$	Bacteria Quantity	(cfu/ml)*10^8	BQ	Numerical
Y	Reproductive Status (Dependent variable)	0 = fertile 1 = infertile	Reproductive Status	Nominal

The way a particular data is presented goes a long way in determining its analytical case. In order to prevent some problems usually encountered in the poor presentation of data, extra care is taken, in Table 10, to present the independent variables and their data type and values.

Table 11: Test statistics for test on multi-collinearity.

Model	Unstandardized Coefficients		Standardized Coefficients	t	P-value	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
(Constant)	-.995	.046		-21.478	.001		
AGE	.013	.001	.322	18.895	.001	.380	2.631
History of Gonorrhea	.220	.086	.172	2.543	.012	.774	1.292
BMI (Kg/m^2)	.013	.002	.121	11.029	.001	.918	1.089
BP (mmHg)	.002	.001	.060	5.420	.001	.914	1.094
BS (mg/dl)	.002	.001	.169	10.619	.001	.437	2.287
BQ (cfu/ml)*10^8	.141	.013	.114	10.596	.001	.958	1.043

Multi-collinearity occurs when independent variables in a model are correlated. In logistic regression, this kind of correlation is a problem because independent variables should be have weak or no relationship at all among themselves. If there is a presence of multicollinearity, logistic regression estimates will be unstable and have high standard errors. A researcher can use the tolerance method or Variance Inflation Factor (VIF) method to check

presence of multi-collinearity. The high value of tolerance is an indication that there is no multi-collinearity in the model while the low value of tolerance is known to affect adversely the results associated with the model. The minimum tolerance value should be < 0.25. Variance Inflation Factor (VIF) is the reciprocal of tolerance. It identifies the correlation between independent variables and the strength of that correlation. The minimum value of VIF is 1 and has no upper limit. The value between 1 and 4 indicates that there is no correlation between this independent variable and any other independent variable and it suggests an absence of multi-collinearity. A VIF value between 5 and 9 indicated that there is a moderate correlation but it is not severe enough to cause a problem. A VIF value of more than 10 is said to be highly collinear and it indicates critical levels and causes a problem (Eze et al., 2021; Warner, 2013). From Table 11, the independent variables had no multi-collinearity since the tolerance values for the variables were greater than 0.25. Also, to confirm our claim the VIF values were between 1 and 4.

Omnibus Tests of Model Coefficients are used to assess the fitness of the overall logistic regression model. The overall model contains all the considered independent variables, unlike the null model which contains no independent variables. From Table 12, the Omnibus Tests of Model Coefficients tested the model fit to predict the reproductive status (i.e., fertility or infertility) of Gonorrhea patients. It tested the significance of the independent variables coded as age, history, blood sugar, bacteria quantity, body mass index, and blood pressure as predictors of the model with reproductive status as a dependent variable (fertile = 0 and infertile = 1). Also, the resuls show in Table 12, a chi-square value of 1678.063 with 6 degrees of freedom (df) and P-value less than 0.05 (i.e., $\chi^2_{(6)} = 1678.063$, P-value < 0.05). It means that the overall model is statistically significant, that is, the model as a whole fits significantly to predict the reproductive status of Gonorrhea patients better than a model with no predictors at a 5% significance level.

Table 12: Omnibus Tests of Model Coefficients

		Chi-square	df	P-value
Step 1	Step	1678.063	6	.000
	Block	1678.063	6	.000
	Model	1678.063	6	.000

The Cox & Snell R^2 and Nagelkerke R^2 seen in Table 13 are similar to R^2 which is in linear regression that gives us an idea of how much variance in the dependent variable is explained by the independent variables. The R^2 ranges from 0 to 1, with 1 being a perfect fit. These Cox & Snell R^2 and Nagelkerke R^2 values are sometimes called pseudo R^2 and have lower values than R^2 in linear regression (Laerd Statistics, 2018; Cox and Snell, 1989; Nagelkerke, 1991). The Cox & Snell R^2, both corrected and uncorrected, was discussed earlier by Maddala (1983) and Cragg and Uhler (1970). From the results in Table 13, we noticed that Cox & Snell R^2 is 0.215(21.7%) and Nagelkerke R^2 is 0.334(33.4%); this is to say that R^2 ranges between 21.7% to 33.4%. It is preferable to report the Nagelkerke R^2 because it is a modification of Cox & Snell R^2 that cannot achieve a value of 1but Nagelkerke R^2 can reach a maximum of 1 (Laerd Statistics, 2018). It can be seen from Nagelkerke's R^2 result that 33.4% of the variance in the outcome variable is affected by predictor variable and it can be said that there is evidence to say that the logistic model is adequate or a good fit for the data.

Table 13: Model summary statistics

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	5520.271	.217	.334

The Hosmer and Lemeshow Test in Table 14 tests the null hypothesis that predictions made by the logistic model fit perfectly with observed group memberships. The statistical test makes use of a chi-square statistic computed to compare the observed frequencies with those expected under the linear model. A nonsignificant chi-square statistic indicates that the model fits well with the data. This Hosmer and Lemeshow Test has several problems in which one of which is that it relies on a test of significance. The implication of this is that with large sample sizes, the test may be significant even when the fit is good, and with small sample sizes, it may not be significant even with a poor fit (Hosmer and Lemeshow, 2000; Wuensch, 2021). The result from the Hosmer and Lemeshow Test in Table 14 showed a chi-square value of 102.127 with 8 degrees of freedom (df) and a P-value greater than 0.05 (i.e., $\chi^2_{(8)} = 102.127$, P-value > 0.05) and this means that the model adequately fits the data perfectly well. Hence, there is no difference between the observed frequencies and the predicted model at a 5% significance level.

Table 14: Hosmer and Lemeshow test

Step	Chi-square	df	P-value
1	102.127	8	.0901

Table 15 shows the classification table for the reproductive status (fertility or infertility) of Gonorrhea patients. The logistic regression model estimates the probability of an event occurring using the values of the independent variables on a certain cut-off point, usually 0.5. If the estimated probability of the event occurring is greater than or equal to 0.5, the event is classified as occurring but if the probability is less than 0.5, the event is classified as not occurring. The classification table compares the actual and predicted groups to asses how many would be correctly classified. This method of classification of individuals into one of the outcome groups (YES or NO) is a way to assess the model's reliability for prediction. Therefore, it becomes necessary to have a method to examine the effectiveness of the predicted classification against the actual classification. In Table 15, the Gonorrhea patients with predicted probabilities of fertility greater than or equal to 0.5 are classified into the fertile group while those with predicted probabilities of infertility greater than or equal to 0.5 are classified into the infertile group. The model correctly classified 1121(74.8%) Gonorrhea patients into the infertile group; this is known as the sensitivity of prediction, that is, the percentage of occurrences correctly predicted. The model also correctly classified 5085(95%) Gonorrhea patients into the fertile group and this is known as a specificity of prediction, that is, the percentage of nonoccurrences correctly predicted. The overall correct prediction was 6206 out of 6850 Gonorrhea patients with an overall success rate of 90.6%. The model predicted the total number of infertility as 1387 against the actual observation of 1499 Gonorrhea patients. It predicted 266(19.2%) infertile Gonorrhea patients that were wrongly classified into the fertile group at the time of actual observation recording and this is known as a false positive prediction. Also, the model predicted the total number of fertility as 5463 against the actual observation of 5351 Gonorrhea patients. It predicted 378(7.1%) fertile Gonorrhea patients that were wrongly classified into the infertile group and this is known as a false negative of prediction.

Table 15: Classification Table[a] for reproductive status (fertility or infertility) of a Gonorrhea patient

Observed			Predicted		
			Reproductive Status		Percentage
			Fertile	Infertile	Correct
Step 1	Reproductive Status	Fertile	5085	266	95.0
		Infertile	378	1121	74.8
	Overall Percentage				90.6

a. The cut value used is 0.500

From Table 16, it was noticed that column 2 shows results for logistic regression coefficients, column 4 shows the Wald Chi-Square statistic that tests the unique contribution of each predictor to the model, and column 6 shows probability values (P-values). The unique contribution of each predictor is significant if P-value is less than the 5% level of significance. Since the P-values in this Table 16 are less than 0.05 (i.e., P-value < 0.05), we conclude that the variables coded as age, history, blood sugar, bacteria quantity, body mass index, and blood pressure and used as the predictors of the model are statistically significant at a 5% significance level. Column 7 in Table 16 shows the result for the odds ratio related to variables coded age, history, blood sugar, bacteria quantity, body mass index, and blood pressure, and column 8 shows a 95% confidence interval for the odds ratio. An odds ratio is used to predict the probability of an event occurring based on a one-unit change in a predictor when all other predictors are kept constant. The odds ratio (OR) can be less than 1 (< 1), greater than 1 (>1), or equal to 1 (= 1). There is no change in odds if the odds ratio is 1. The odd decreases for every unit change in the predictor variable if it is less than 1. The odd increases for every unit change in the predictor variable if it is greater than 1. Thus, the higher the odds ratio is above 1, the more likely a patient is to be infertile. The result from Table 16 shows that for every unit increase in the variables coded as age, blood sugar, bacteria quantity, body mass index, and blood pressure of a Gonorrhea patient, the odds ratio of being infertile are 1.086, 1.104, 1.013, 1.014, and 2.314 when other predictors are constant respectively. The odds ratio of a Gonorrhea patient with a Gonorrhea history is 3.718 times more likely to be infertile than a Gonorrhea patient without a Gonorrhea history when variables coded as age, blood sugar, bacteria quantity body mass index, and blood pressure are held constant. We noticed that the odds ratio for the constant is less than 0.001, that is, the odds ratio for the model without the variables coded as age, history, blood sugar, bacteria quantity, body mass index, and blood pressure as predictors is less than 0.001.

Table 16: The logistic regression model table to predict the reproductive status (fertility or infertility) of Gonorrhea patients

	B	**S.E.**	**Wald**	**Df**	**P-value**	**Exp(B)**	**95% C.I.for EXP(B)**	
							Lower	**Upper**
AGE	.082	.005	270.821	1	.000	1.086	1.075	1.096
HISTORY(1)	1.320	.475	7.645	1	.006	3.718	1.466	9.431
BMI (Kg/m^2)	.099	.010	104.038	1	.000	1.104	1.083	1.125
BP (mmHg)	.013	.002	29.562	1	.000	1.013	1.008	1.018
BS (mg/dl)	.014	.001	99.015	1	.000	1.014	1.011	1.016
BQ (cfu/ml)*10^8	.839	.099	71.873	1	.000	2.314	1.906	2.809
Constant	-9.921	.398	622.020	1	.000	.000		

Obtained Model: The model form for predicted probabilities is expressed as a natural logarithm (ln) of the odds ratio:

$$\ln\left(\frac{\mathrm{P(Y)}}{1-\mathrm{P(Y)}}\right) = -9.921 + 0.082(\mathrm{Age}) + 1.320(\mathrm{History}) + 0.099(\mathrm{BMI}) + 0.013(\mathrm{BP}) \quad + 0.014(\mathrm{BS}) + 0.839(\mathrm{BQ}) \tag{16}$$

3.1 Predictions

The odds ratio prediction that is formed from the model in *Equation* (16) is given as

$$\frac{\mathrm{P(Y)}}{1-\mathrm{P(Y)}} = \mathrm{e}^{-9.921+0.082(\mathrm{Age})+1.320(\mathrm{History})+0.099(\mathrm{BMI})+0.013(\mathrm{BP})+0.014(\mathrm{BS})+0.839(\mathrm{BQ})} \tag{17}$$

The conversion of the odds ratio in *Equation* (17) to general probability form for the prediction of Gonorrhea patients that are infertile is given as

$$P(Y) = \frac{e^{-9.921+0.082(\mathrm{Age})+1.320(\mathrm{History})+0.099(\mathrm{BMI})+0.013(\mathrm{BP})+0.014(\mathrm{BS})+0.839(\mathrm{BQ})}}{1+e^{-9.921+0.082(\mathrm{Age})+1.320(\mathrm{History})+0.099(\mathrm{BMI})+0.013(\mathrm{BP})+0.014(\mathrm{BS})+0.839(\mathrm{BQ})}} \tag{18}$$

The results in column 7 of Table 17 were obtained using *Equation* (18)

Table 17: Probability computations for classification of reproductive status of a gonorrhea patient.

Age	History of Gonorrhea	Body Mass Index (BMI)	Blood Pressure (BP)	Blood Sugar (BS)	Bacteria Quantity (BQ)	Probability (Y)	Reproductive status of a Gonorrhea patient
52	1	21.333	130	279	0.103	0.97***	Infertile
25	0	26.9	120	114	0.302	0.14**	Fertile
34	1	26.439	100	168	0.206	0.65***	Infertile
64	1	20.08	130	116	0.502	0.91***	Infertile
40	0	22.676	120	127	0.168	0.28**	Fertile
49	1	26.8	139	132	0.123	0.86***	Infertile
16	0	24.9	120	104	0.033	0.04**	Fertile
38	1	21.4	135	164	0.092	0.68**	Infertile

***P(Y) greater than 0.5 = Infertile; **P(Y) less than 0.5 = Fertile
*P(Y) equal to 0.5 = Equal chances of being Infertile or Fertile

4. Conclusions

In this study, we used data on the eight (8) types of sexually transmitted infections (STIs) recorded from 2010 through 2020 in the Department of Micro Biology, University of Nigeria Teaching Hospital to obtain the most prevalent sexually transmitted infection. Firstly, the prevalence analysis method was used to determine the most prevalent sexually transmitted infection among eight (8) selected infections (Chlamydia, Gonorrhea, Syphilis, Trichomoniasis, Hepatitis B, Herpes, Human papilloma Virus (HPV), and Human Immuno-deficiency Virus (HIV)). The results showed that Gonorrhea is the most prevalent STI at 33.08%. Secondly, two-way CATANOVA cross-classification was used to ascertain the gender and age of those who always suffer from Gonorrhea infection and the results showed that gender, age, and its interaction effect were statistically significant at a 5% level. This implies that a particular gender and age interval always suffer from Gonorrhea infection. The data showed that the percentage of males is more than the percentage of females that suffer from Gonorrhea infection. The percentage of 30-39 years old suffer Gonorrhea infection more than any other age interval. The data also showed that 55.5% out of 6850 Gonorrhea patients, do not have a Gonorrhea infection history.

The results also showed that there is a spread of Gonorrhea infection between the genders. The significance of the interaction effect showed that the rate of contracting Gonorrhea infection by gender differs from one age group to another. The three-way CATANOVA result showed that the drug prescription for the treatments of Gonorrhea infection depends on gender and age at a 5% significance level. A logistic regression was performed to ascertain the effects of the variables coded as age, history, blood sugar, bacteria quantity, body mass index, and blood pressure on the likelihood that a Gonorrhea patient is infertile. The logistic regression model was statistically significant, $\chi^2_{(6)} = 1678.063$, P-value < 0.001. The model explained 33.4% (Nagelkerke R^2) of the variance in reproductive status (fertility or infertility) of Gonorrhea patients and correctly classified 90.6% of cases into the fertile and infertile groups. A Gonorrhea patient with a Gonorrhea history is 3.718 times more likely to be infertile than a Gonorrhea patient without a Gonorrhea history. An increase in age, body mass index, blood pressure, blood sugar, and bacteria quantity of a Gonorrhea patient were associated with an increased likelihood of being infertile.

The findings of this study also showed drugs used in treating Gonorrhea infection depend on the patient's gender and age which means that some drugs are not for the treatment of a Gonorrhea patient because of the patient's gender or age. Gonorrhea patients are advised not to lie about their age as it helps the physicians who prescribe these medicines and the female gender should know their pregnancy status to avoid health complications. Moreover, we noticed a spread of Gonorrhea infection between the genders since the number of Gonorrhea patients that do not have an infection history is more than those with an infection history.

The previous studies showed that these infections, especially Gonorrhea, caused infertility if poorly treated or left untreated over a long period. In this study, we use the fitted logistic regression model to make some predictions on the fertility of Gonorrhea patients. Our findings (see Table 17) showed that a Gonorrhea patient with a certain age, gonorrhea history, body mass index, blood pressure, and bacterial quantities can be infertile.

This study is an eye-opener to different types of sexually transmitted infections for Nigerians. The findings in this study showed that significant steps are to be used to create awareness and motivate adults about the need for regular health check-ups for proper termination or cure of these infections. More precisely, the concerned authorities need to make efforts to educate people on STIs and this may be through mass media, social media, schools, and any other means of communication. The authorities should also provide appropriate healthcare facilities in both urban and rural areas with government intervention for the benefit of the poor ones. These measures against STIs, especially Gonorrhea infection, with their risk reduce STIs drastically in Nigeria.

Acknowledgements

The authors would like to acknowledge the department of Micro Biology, University of Nigeria Teaching Hospital (UNTH) who made their data available for free for this research.

Data availability statement

We used secondary data from the Department of Micro Biology, University of Nigeria Teaching Hospital. The data had been presented in this research work and any other information needed on the data used in this work will be made available.

Conflict of Interest

We (authors) declare that there are no conflicts of interest.

Literature Cited

ANDERSON, T. W. (1958). *An introduction to multivariate analysis*. John Wiley: New York

ANDERSON, R. J., LANDIS, J. R. (1980). CATANOVA for multidimensional contingency tables: Nominal-scale response. *Commun. Statist. Theor. Meth.* 9:1191–1206.

ANDERSON, R. J., LANDIS, J. R. (1982). CATANOVA for multidimensional contingency tables: Ordinal-scale response. *Commun. Statist. Theor. Meth.* 11:257–270.

AZMI M.,MUATAZ A.,ALI M.A. and MOHAMMAD S.E (2008). Prevalence of Sexually Transmitted Infections Among Sexually Active Jordanian Females. *Sex Transm. Dis.*, 35 (6):607-710

BOATENG, E.Y. and ABAYE, D. A. (2019). A Review of the Logistic Regression Model with Emphasis on Medical Research. *Journal of Data Analysis and Information Processing*, 7: 190-207

CENTERS FOR DISEASES CONTROL AND PREVENTION (2013). Incidence, prevalence, and cost of living in the United States.

COX, D. R and SNELL, E. J. (1989). *Analysis of Binary Data*. Second Edition. Chpman & Hall.

CRAGG, J. G. and UHLER, R. S. (1970). The demand for automobiles. *The Canadian Journal of Economics* 3: 386-406.

D'AMBRA, L., BEH, J. E. and AMENTA, P. (2005). Analysis of contingency tables: Catanova for two-way contingency tables with ordinal variables using orthogonal polynomials. *Communications in Statistics—Theory and Methods*, 34: 1755–1769 DOI:10.1081/STA-200066325

DEYHOUL N., MOHAMADDOOST T. and HOSSINI M. (2017). Infertility related risk factors:A systematic review. *International Journal of women health and reproduction science*, 5(1):24-29.

EZE, N. M., ASOGWA, O. C. and EZE, C. M. (2021). Principal component factor analysis of some development factors in southern Nigeria and its extension to regression analysis. *Journal of Advances in Mathematics and Computer Science, 36*(3): 132-160. DOI:10.9734/JAMCS/2021/v36i330351

FIENBERG, S. E. (1973). Analysis of incomplete multiway contingency tables. Biometrics, 28:177-202. DOI: https://doi.org/10.2307/2528967

FISHER, L. D (1998). Self-designing clinical trials. Statistics in Medicine; 17:1551-1562

FLORIAN, T. J.(2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *J Mem Lang, 59*(4):434-446.

HOSMER, D. and LEMESHOW, S. (2000). *Applied Logistic Regression (Second Edition)*. New York: John Wiley & Sons, Inc.

KESAH F.N.C., VINCENT K.P and AUGUSTINE A.(2013). Prevalence and etiology of Sexually Transmitted Infections I gynecologic unit of a developing country. *Annals of tropical medicine and public health, 6*(5):526.

LAERD STATISTICS (2018). Binomial Logistic Regression using SPSS Statistics. Accessed 14 August, 2021. Available: https://statistics.laerd.com/spss-tutorials/binomial-logistic-regression-using-spss-statistics.php

LIGHT, R. J., MARGOLIN, B. H. (1971). An analysis of variance for categorical data. *J. Amer. Statist. Assoc.* 66:534–544.

MADDALA, G. S. (1983). *Limited Dependent and Qualitative Variables in Econometrics*. Cambridge University Press.

MARGOLIN, B. H., LIGHT, R. J. (1974). An analysis of variance for categorical data II. Small samples comparisons with chi-square and other competitors. *J. Amer. Statist. Assoc.* 69:755–544

NAGELKERKE, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika* 78: 691-692.

ONUKOGU, I. B. (1985). Reasoning by analogy from ANOVA to CATANOVA, *Biom J,* 27:839-849.

_____ (2014). Analysis of variance of categorical data-nested designs. *Journal of Statistics: Advances in Theory and Applications*, 12: 109-116

OTARU O. P. and OGBONDA N. P (2020). CATANOVA analysis of knowledge and control practices of hepatitis B virus infection amongst tertiary university students. Galician Medical Journal, 27(1).

SCATTERWHITE, C.L.,TORRONE, E., MEITES, E., DUNNE, E.F., MAHAJAN, R., OCFEMIA, C.SU, J., XU,F. and WEINSTOCK, H. (2013). Sexually transmitted infections among U.S. women and men: Prevalence and incidence estimates, 2008 *Sexually Transmitted Diseases*;40 (3):187-193.

SCHEFFE, H. (1959). *The Analysis of Variance*. Wiley: New York

SINGH, B. (2004). CATANOVA for analysis of nominal data from repeated measures design. *J Ind Soc Agril Statist, 58*(3):257-268

WARNER, R.M. (2013). *Applied Statistics (2nd. Edition)*. Thousand Oaks, CA: SAGE.

WUENSCH, K. L. (2021). Binary Logistic Regression with SPSS. Accessed 14 September, 2021. https://core.ecu.edu/wuenschk/MV/Multreg/Logistic-SPSS.PDF

Appendix A

Computation of Sum of Squares for Two-way contingency table depicting the response of gender and ages of gonorrhea patients reported in UNTH from 2010-2020 (see Table 7).

$$TSS = 6850 - \frac{3049^2 + 3801^2}{6850} = 3383.72$$

$$WUSS = 6850 - \frac{128^2 + 383^2}{511} + \frac{395^2 + 437^2}{832} + \cdots + \frac{234^2 + 198^2}{432} = 3252.88$$

$$BRSS = 6850 - \frac{1721^2 + 2031^2}{3752} + \frac{1328^2 + 1770^2}{3098} = 3380.66$$

$$BCSS = 6850 - \frac{282^2 + 723^2}{1005} + \frac{667^2 + 935^2}{1602} + \cdots + \frac{581^2 + 449^2}{1030} = 3279.09$$

$$RSS = 3.06$$
$$CSS = 104.63$$
$$NSS = 23.15$$

Chi-square calculated

$$\chi^2{}_{RT} = \frac{(2-1) \times (6850-1) \times 3.06}{3383.72} = 6.19$$

$$\chi^2{}_{CT} = \frac{(2-1) \times (6850-1) \times 104.63}{3383.72} = 211.78$$

$$\chi^2{}_{NT} = \frac{(2-1) \times (6850-1) \times 23.15}{3383.72} = 46.86$$

Chi-square tabulated

$$\chi^2{}_{RT} = \chi^2{}_{(2-1)(2-1)} = \chi^2{}_{(1)} (at\ 5\%\ from\ chi - square\ table = 3.841)$$
$$\chi^2{}_{CT} = \chi^2{}_{(5-1)(2-1)} = \chi^2{}_{(4)} (at\ 5\%\ from\ chi - square\ table = 9.49)$$
$$\chi^2{}_{NT} = \chi^2{}_{(2-1)(5-1)(2-1)} = \chi^2{}_{(4)} (at\ 5\%\ from\ chi - square\ table = 9.49)$$

Appendix B

Computation for Three-way contingency table depicting gender, ages, and drug prescription for gonorrhea infection through 2010 – 2020 (see Table 9.1 and 9.2).

Test Statistic:

$$\chi^2 = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{c} \frac{(n_{ijk} - f_{ijk})^2}{f_{ijk}} \sim \chi^2_{ijk-(i+j+k)+2}$$

Where i = number of drugs, j = number of age intervals in age, k = number of genders

$$f_{ijk} = n\left(\frac{n_{i++}}{n}\right)\left(\frac{n_{+j+}}{n}\right)\left(\frac{n_{++k}}{n}\right)$$

Appendix Table 1: Summary table for calculated observed and estimated expected frequencies.

Cell	Observed Frequencies (n_{ijk})	Estimated Expected Frequencies $(\hat{f}_{ijk})$	$\frac{(n_{ijk}-\hat{f}_{ijk})^2}{\hat{f}_{ijk}}$
f_{111}	110	119.74	0.79
f_{112}	188	190.87	0.04
f_{113}	206	231.85	2.88
f_{114}	171	150.95	2.66
f_{115}	118	122.72	0.18
f_{121}	129	98.87	9.18
f_{122}	196	157.60	9.36
f_{123}	229	191.44	7.37
f_{124}	80	124.64	15.99
f_{125}	63	101.33	14.50
f_{211}	107	107.68	0.01
f_{212}	163	171.65	0.44
f_{213}	202	208.51	0.20
f_{214}	164	135.76	5.88
f_{215}	112	110.36	0.02
f_{221}	111	88.91	5.49
f_{222}	135	141.73	0.32
f_{223}	216	172.17	11.16
f_{224}	55	112.09	29.08
f_{225}	75	91.13	2.85
f_{311}	109	89.44	4.28
f_{312}	123	142.57	2.69
f_{313}	178	173.19	0.13
f_{314}	115	112.76	0.04
f_{315}	116	91.67	6.46
f_{321}	61	73.85	2.24
f_{322}	102	117.72	2.10
f_{323}	110	143.00	7.62
f_{324}	103	93.10	1.05
f_{325}	96	75.69	5.45
f_{411}	130	136.05	0.27
f_{412}	192	216.87	2.85
f_{413}	248	263.44	0.90
f_{414}	205	171.52	6.54
f_{415}	159	139.44	2.74
f_{421}	104	112.34	0.62
f_{422}	211	179.07	5.69
f_{423}	184	217.52	5.17
f_{424}	128	141.62	1.31
f_{425}	132	115.13	2.47
f_{511}	55	97.56	18.57
f_{512}	166	155.51	0.71
f_{513}	174	188.90	1.18
f_{514}	148	122.99	5.08
f_{515}	93	99.99	0.49
f_{521}	89	80.55	0.89
f_{522}	126	128.40	0.05
f_{523}	199	155.98	11.87
f_{524}	98	101.55	0.12
f_{525}	66	82.56	3.32

$$\chi^2_{cal} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\frac{(n_{ijk}-f_{ijk})^2}{f_{ijk}} = 221.30$$

$\chi^2_{tab} = \chi^2_{ijk-(i+j+k)+2}$

Where i = 5, j = 2, k = 5(from the contingency Table 9.1)

$\chi^2_{tab} = \chi^2_{(5\times2\times5)-(5+2+5)+2} = \chi^2_{40} = 55.75$ (From the chi-square table, size 40 under 5% level of significance)