Journal of Asian Scientific Research
ISSN(e): 2223-1331
ISSN(p): 2226-5724
DOI: 10.55493/5003.v15i2.5485
Vol. 15, No. 2, 260-269.

URL: www.aessweb.com

Binary logistic regression to assess the factors affecting the infection of toxoplasmosis

Sofian A.A. Saad[1+]
Husham M.A. Attaalfadeel[2]
Mufda Jameel Alrawashdeh[3]

[1]*Department of Mathematics, College of Science, Qassim University, Buraydah 51452, Saudi Arabia.*
Email: so.saad@qu.edu.sa
[2]*Department of Mathematics, College of Science, Northern Border University, Arar 91431, Saudi Arabia.*
Email: husham.alhassan@nbu.edu.sa
[3]*Department of Management Information Systems, College of Business and Economics, Qassim University, Saudi Arabia.*
Email: M.alrawashdeh@qu.edu.sa

(+ Corresponding author)

Article History
Received: 10 April 2025
Revised: 16 June 2025
Accepted: 1 July 2025
Published: 22 July 2025

Keywords
Binary logistic regression techniques
Factors affecting *T. gondii*
Odds ratio (OR)
Sampling technique
Statistical model
Toxoplasmosis.

ABSTRACT

Logistic regression method (LR) is one of the most widely used modeling techniques in various fields of science, especially in clinical medicine where variables are often dichotomous. The factors that cause toxoplasmosis (T. gondii) are well explained by most clinical investigators. Therefore, the main objective of this article is to identify the most significant factors leading to toxoplasmosis infection. The binary logistic regression method has been used to interpret the study's findings. A clustered sampling technique with an informative questionnaire was used in the survey to collect a relevant sample of 508 individuals from the most affected areas, specifically the northern part of Saudi Arabia. SPSS as well as AMOS are the typical statistical analysis tools used to investigate the results. The binary logistic techniques showed that the factors (stillbirth, women's direct contact with soil, and keeping indoor cats) were the most significant factors influencing infection with toxoplasmosis, without neglecting some other invisible factors. Only 18.9% of the variation in the dependent variable (Toxoplasmosis infection) is attributed to the independent variables (which is moderate**,** with Nagelkerke's R square = 0.189). Early medical follow-up and health awareness campaigns should be adopted, especially in remote rural communities.

Contribution/ Originality: All previous studies in the same field have mainly focused on identifying the factors influencing toxoplasmosis gondii as a whole, while this study is considered to be the first in the region to identify the most important factors affecting infection with the toxoplasmosis virus.

1. INTRODUCTION

Toxoplasmosis is an illness caused by a parasite called Toxoplasma gondii, which is the most common parasite infecting most genera of warm-blooded animals. Infection usually occurs when people eat undercooked meat or through direct contact with cats and their feces. The parasite can also pass to a baby during pregnancy.

Most people infected with the parasite do not exhibit clear symptoms, and some rarely show flu-like symptoms. Serious disease most often affects infants and individuals with weakened immune systems. Toxoplasmosis during pregnancy can cause miscarriage and birth defects. Most infections do not require treatment; drug therapy is generally reserved for more serious cases, pregnant individuals, newborns, and people with weakened immune systems. The danger of toxoplasmosis lies in the complications associated with the disease, such as encephalitis, headaches, seizures, confusion, coma, and pneumonia, which can result in coughing, fever, and shortness of breath, in addition to eye infections that may cause blindness if medical intervention is not implemented in a timely manner. The disease is not only limited to the infected individual but can also seriously affect the life of a fetus in a pregnant woman. Moreover, outbreaks of toxoplasmosis can lead to negative social and economic effects.

50% The infection of T. gondii disease has a high possibility to reach up to 50%, which is completely has a significant difference in vary places in the world [1]. Aggressiveness and impulsiveness are symptoms of Toxoplasmosis. Furthermore, they increase the likelihood of suicide. Patients most affected by the disease include those living in unhygienic conditions, elderly individuals, and patients who are illiterate or less educated [2]. Different studies reveal that having food such as (uncooked vegetables and drinking the inferior amount of will activate factors that associated and influential with Toxoplasmosis infection [3]. From this angle, the importance of the study is derived, thus motivating us to achieve the following objectives.

1. Build a concrete holistic binary logistic LR.
2. Identify the most important factors and their probabilities of causing the disease.
3. Demonstrate the importance of applying binary LR in medical fields, where the selected data are categorical.

1.1. Variables of the Study

- Dependent variable:

1. Toxoplasmosis infection (Ig G ELISA).

- Independent variables:

1. Nationality.
2. Still birth.
3. Source of drinking water.
4. Meat Handling (MH).
5. Washing hands after (MH).
6. Direct contact with cats.
7. Keeping indoor cats.
8. Cleaning cat's area.
9. Eating soil (Mud).
10. Contact Soil.
11. Washing kitchen utensils.
12. Type of drinking milk.
13. Eating meat.

Table 1. Study's variables.

Variables	Codes	
Ig G ELISA (Dependent)	Infected (0)	Not infected (1)
Nationality	Saudi (1)	Non-Saudi (2)
Stillbirth	Alive fetus (0)	Dead fetus (1)
Source of drinking water	Home water (1)	Mineral water (2)
Meat handling (MH)	Yes	No
Washing hands after (MH)	Yes	No
Direct contact with cats	Yes	No
Keeping indoor cats	Not keeping (0)	Keeping (1)
Cleaning cat's area	Yes	No
Eating soil (Mud)	Yes	No
Contact soil	Not contacted (0)	Contacted (1)
Washing kitchen utensils	Yes	No
Type of drinking milk	Bottled	Fresh
Eating meat	Weekly	Daily

Table 1 presents the coded variables used in the study to assess the significance of the independent variables and their impact on the dependent variable, considering other (unseen) factors such as random variables. The tested assumptions should be as follows: The significance of the regression coefficients B1, B2, B3, ..., B13 for all independent variables is equal to zero (H0: B1, B2, B3, ..., B13 = 0). Therefore, various types of statistical tests should be conducted. (Subtitle: Results).

1.2. Sample Structure

Since toxoplasmosis is more prevalent in the northern part of Saudi Arabia, the sample structure is shaped based on this geographical area.

1.3. Sampling and Data Gathering

In order to have clear, reliable, and sufficient answers to the study's questions and obtain accurate results, a representative sample must be carefully determined and drawn from the original population. To achieve the most accurate results, the study should address the following questions:

1) Do all the tested variables included in the model have the same effect on the dependent variable?
2) What are the main variables (Factors) that cause toxoplasmosis?
3) What are the chances that each independent variable causes toxoplasmosis?
4) Do we have a significant relationship between the regressor and the dependent variable?

To adopt reliable answers and examine the study questions, sampling with full probability clustering techniques was applied to capture a sample of 508 patients.

2. METHOD

In many cases, and according to the nature of the data provided, researchers find themselves in a position to analyze categorical data to determine the relationships between variables relevant to that data. The best way to perform this analysis when the variables or data are qualitative is to present them as coded variables consisting of two or more values, depending on the nature of the study. The least squares method, in such cases, does not guarantee accurate results, and the outputs may be unreliable, leading to false parameter estimates. Therefore, we cannot depend on these results to assess the relationship between variables. However, treating response variables similarly can lead to heterogeneity problems, as the variance of random errors in the estimated models may not be equal.

Relevant techniques such as binary LR, multinomial LR, and discriminant function analysis can be used to derive accurate results for models or functions that involve qualitative variables. In this study, because the dependent variable is categorical and coded as 0 or 1 (i.e., 0=infected with toxoplasmosis and 1=not infected), we adopt the technique of binary LR. Binary LR is particularly suitable for analyzing data in widely used cross-sectional and case–control research designs [4, 5]. It is one of the most commonly used approaches for predictive analytics in clinical medicine, human sciences, social sciences, and engineering sciences. Understanding this method is important for those in the fields of medicine and the health community. The binary LR model is generally related to statistical generalized linear models with specific characteristics: the dependent variable has two levels (i.e., yes or no), and at least one independent variable is used to predict the values of the dependent variable [6]. Unlike the dependent variable, independent variables can be either in two or more categories, and they can even be continuous. The advantage of modeling data using logistic regression (LR) is that the estimated coefficients are easy to interpret and understand, and that exponentiation of the coefficient yields the odds ratio, which is directly interpretable for clinicians [7].

Although the coefficients are easy to interpret through LR, the variables to be included in the model must be carefully selected due to the nature of the method. The use of this powerful modeling technique requires careful attention in the specification of the model form, as well as in the calculation and interpretation, especially regarding the interpretation of the model's coefficients. An incorrect interpretation of the model coefficients can lead to misspecification, which may be discovered at an advanced stage later during the analysis [8, 9].

Estimation of the binary logistic regression model depends on the estimation of the model coefficients.

Conducting exploratory data analysis (EDA) is crucial before running and estimating the binary logistic regression (LR) model. EDA includes descriptive statistical analysis (tables and/or graphs) and serves multiple purposes. For example, it ensures that data are correctly measured and labeled, and it identifies any problems related to data distributions [10].

In the estimated binary LR model, the probabilities of the model predictor's outcomes need to be calculated to determine the possible chances for each value of the dependent variable (i.e., 0, 1). In this case, the odds and odds ratio, as informative statistical indicators, will also be calculated, which explain the relationships between the dependent variable and predictors.

$$Odds = e^{a+bx} \qquad (1)$$

Where "e" is the exponential term. And the,

$$Odds\ Ratio(OR) = \frac{\text{Odds for certain response outcome}}{\text{Odds for the other response outcome}} \qquad (2)$$

The logistic regression model of the binary response always takes the form as:

$$\pi(x) = \frac{e^{(\alpha+\beta x)}}{1+e^{(\alpha+\beta x)}} = \frac{1}{1+e^{-(\alpha+\beta x)}} \qquad (3)$$

The formula gives the probability of the dependent variable equaling to one of the response variables depending on one of the two values given to the dependent variable by the researcher (0 vs 1).

Finally, the estimated model will be given by the log odds which is called the logit as follow:

$$logit\ [\pi(x)] = log\frac{\pi(x)}{1-\pi(x)} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i \qquad (4)$$

2.1. Analytic Criteria

2.1.1. Sufficient Events Per Variable

For the LR, the response variables can take only two possible outcomes. The response variable, which is toxoplasmosis (Ig G ELISA), is (0, 1) (0 for not infected with toxoplasmosis, and 1 for infected with toxoplasmosis). Additionally, there should be little to no evidence of multicollinearity among the predictors or explanatory variables to ensure the independence of their relationships. Multicollinearity relates to two or more highly correlated independent variables that do not provide unique information in the estimated regression model, leading to incorrect interpretations. Therefore, either all or one of them should be considered, with the highest acceptable correlation value being less than 0.7 [11].

The relationship between independent variables should exhibit clear linearity to the log odds. The model should include all relevant variables, and it must be proportionally relevant to the given number of observations [5, 6, 12]. For LR, a useful rule of thumb suggests that the number of observations divided by the number of predictor variables should be at least 10, preferably greater. The fewer events per variable, the greater the likelihood that the estimates of the model coefficients are unreliable; the variance of the model coefficients and the confidence intervals will also be less accurate [7]. However, the current study considered a sample size of 508 observations against 13 predictors. Independence between observations should be considered, and they must not be related to each other or result from repeated measurements of the same individual types.

Another crucial assumption of binary logistic regression is related to the extreme values of observations, which can seriously affect the outcomes produced by the model. In such cases, these values must be avoided, or the dataset should be subjected to appropriate preprocessing [13].

3. RESULTS

3.1. Collinearity

Two highly correlated variables present a problem for any type of regression analysis [14]. If two highly correlated variables are included in the same model, their estimated contributions and those of all other variables in the model will lead to imprecise results, and there will be an inflated variance associated with these variables' coefficients, which can result in a loss of statistical significance. This issue is known as a collinearity problem. If two variables are found to be highly correlated, one of them should be excluded.

Table 2. Collinearity test of the coefficients.

Model		Unstandardized coefficients		Coefficients	t	Sig.	Collinearity statistics	
		B	Std. error	Beta			Tolerance	VIF
	(Constant)	0.014	0.030	-	0.457	0.648	-	-
	Stillbirth	0.223	0.061	0.154	3.640	0.000	0.997	1.003
	Keeping indoor cat	0.185	0.034	0.232	5.448	0.000	0.980	1.020
	Contact soil	0.091	0.035	0.112	2.632	0.009	0.980	1.020

Variance Inflation Factor (VIF) is a test used to check for collinearity. VIFs greater than 10 may warrant further examination. As shown in Table 2, none of the calculated VIFs are greater than 10; therefore, no collinearity problem exists.

3.2. Validation

Model validity refers to the stability and reasonableness of the LR coefficients, the plausibility and usability of the fitted LR function, and the ability to generalize inferences drawn from the analysis. Validation of the fitted model is conducted to confirm the accuracy of the inferences about the population regarding the postulated results drawn from the sample [15].

A good fit of the model with respect to the sample data used in modeling does not guarantee a well-fitting model in the future. Therefore, the performance of the model on new data extracted from the population should be evaluated. This statistical examination is called model validation [16].

The main idea behind validation is to split the data into two parts: the first part is for testing the validation, which accounts for nearly 50% of the entire data [17, 18]. The second part focuses on mode estimation, aiming to develop a well-fitted LR model to identify the main factors contributing to toxoplasmosis infection. To achieve this, various statistical tests are employed, including Cox and Snell R-square, Nagelkerke adjusted R-square, and the area under the receiver operating characteristic curve (ROC).

Table 3. Classification table.

Observed		Predicted		
		IgG ELISA	Percentage correct	
		Negative	Positive	
Ig G ELISA	Negative	429	5	98.8
	Positive	70	4	5.4
Overall percentage			85.2	

Table 3 shows that the overall accuracy of the classification of the binary LR model for the categories of being infected by toxoplasmosis (Positive) versus not infected (Negative) was 85.2% which means that the overall existing model can explain and predict the data as accurately as possible because the classification accuracy is greater than 50%.

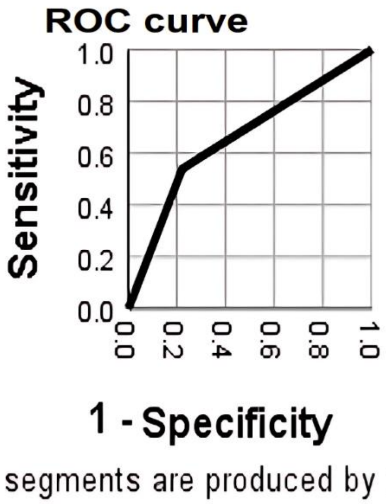

Figure 1. Area under the receiver operating characteristic curve (ROC).

Figure 1 presents the area under the receiver operating characteristic curve (AUC) is commonly used to measure the accuracy of diagnostic tests. The closer the receiver operating curve (ROC) is to the upper-left corner of the graph, the higher the accuracy of the test. This result occurs because the value of ROC ranges between (0 − 1), and the upper-left corner in the graph shows that the sensitivity = 1 and the false positive rate = 0

Some relevant studies have labeled the values of ROC as poor, moderate, good, or excellent. Most researchers consider AUC values lower than 0.6 as poor, but there is a large variation for AUC values higher than 0.7. In general, an AUC greater than 0.7 is considered acceptable [19]. In the current study, AUC = 0.71 illustrates that the diagnostic tests used in the study were more accurate.

Table 4. Hosmer & Lemeshow test.

Step	Chi-square	df	Sig.
1	5.943	8	0.654

Table 4 presents the values of the Hosmer-Lemeshow test, which is used to indicate a well-fitting model and reflect the model's performance. The null hypothesis of the test states that there is no significant difference between the observed and estimated data, while the alternative hypothesis suggests a perfect fit. Since we use a 95% significance level, the estimation terminated at iteration #6 because parameter estimates changed by less than .001. The table clearly shows that a p-value of 0.654 does not reject the null hypothesis. Therefore, the estimated model adequately fits the data.

3.3. Statistical Significance

3.3.1. Goodness-of-Fit Measures

When the final model is constructed, it should be examined in terms of the goodness of fit to describe how well the model fits the data. The R-squared value of logistic regression is usually low because of the nature of the technique, where the dependent variable is dichotomous. This R-squared is different from the R-squared calculated when dealing with linear regression. Hosmer Jr et al. [11] recommended performing the goodness of fit test instead of reporting the R square.

Table 5. Model summary.

Step	-2 Log likelihood	Cox & Snell R square	Nagelkerke R square
1	364.523[a]	0.107	0.189

Note: a. Estimation terminated at iteration # 6 because parameter estimates changed by less than 0.001.

The goodness of fit of binary logistic regression is an important issue to confirm the accuracy of the estimated probabilities. Usually, the power of goodness-of-fit tests increases as the sample size increases, which can influence the decision to reject the null hypothesis [20].

Table 5 presents some statistical measures used to assess the strength of the association between the dependent and independent variables included in the model (model performance). For instance, a value of 0.189 implies that the model moderately captures the relationship between the variables.

Table 6. Omnibus tests of model coefficients.

N	Chi-square	df	Sig.
Model	57.242	13	0.000

In terms of the null hypothesis that the estimated model is not significant, i.e., at least one of the coefficients is zero, Table 6 shows that there is no reason to accept the hypothesis since the p-value of 0.000 with a chi-square of 57.242 indicates the goodness of fit of the model.

3.4. Model Inference

The Wald ratio test is used for each of the independent variables and its associated *p*-value.

Table 7. Variables in the equation.

Variable	B	S.E	Wald	Sig.	Exp(B)	95.0% C.I. for EXP(B)	
						Upper	Lower
Nationality	-0.336	0.359	0.878	0.349	0.715	0.354	1.443
Stillbirth	1.299	0.442	8.637	0.003	3.665	1.541	8.714
Source of drinking water	0.184	0.276	0.442	0.506	1.202	0.699	2.064
Handling meat	0.546	0.946	0.334	0.563	1.727	0.271	11.020
Washing hands after meat handling	-0.929	1.023	0.825	0.364	0.395	0.053	2.933
Direct contact with cat	0.178	0.384	0.215	0.643	1.195	0.563	2.538
Keeping indoor cat	1.314	0.344	14.619	0.000	3.720	1.897	7.294
Cleaning cat area	-0.130	0.344	0.143	0.705	0.878	0.448	1.722
Eating soil (Mud)	0.472	0.304	2.411	0.121	1.603	0.884	2.907
Contact soil	1.139	0.455	6.275	0.012	3.124	1.281	7.617
Washing kitchen utensils	0.495	0.553	0.799	0.371	1.640	0.554	4.851
Type of drinking milk	-0.347	0.222	2.451	0.117	0.707	.457	1.091
Eating meat	-0.292	0.359	0.662	0.416	0.747	0.369	1.510
Constant	-2.444	0.971	6.329	0.012	0.087		

The dependent variable is a dichotomous (binary) variable, coded as 0 and 1. The predictive regression equation therefore, refers to Equation 4. Depending on the outcomes in Table 7, the estimated regression model can be designated as.

$$log\left[\frac{\pi(x)}{1-\pi(x)}\right] = -2.44 - 0.336x1 + 1.299x2 + 0.184x3 + 0.546x4 - 0.929x5 + 0.178x6 + 1.314x7 - 0.130\,x8 + 0.472\,x9 + 1.139x10 + 0.495x11 - 0.347x12 - 0.292x13$$

Where X1 refer to as nationality, X2 is still birth, X3 is source of drinking water, X4 is handling meat, X5 is washing hands after handling meat, X6 is a direct contact with cats, X7 keeping indoor cats, X8 cleaning cat's area, X9 eating soil (mud), X10 frequent contact with soil, X11 washing kitchen utensils, X12 type of drinking milk and X13 is an eating meat.

The essential revealed variables due to the analysis results were X2, X7 and X10, that is (still birth, direct contact with cats and frequent contact with soil). Therefore, the postulated final estimated binary regression model would be as follows:

$$log\left[\frac{\pi(x)}{1-\pi(x)}\right] = -2.444 + 1.299x2 + 1.314x7 + 1.139x10 \quad (5)$$

Based on the results, the three mentioned variables play a significant role in the incidence of toxoplasmosis infection, at least in the area where the study was conducted, considering other unseen variables.

The model can be used to predict the odds (How much more likely one outcome is over another outcome. in other words, the ratio of the probability that outcome # 0 will occur to the probability of outcome # 1) that is;

$$odds = \frac{p_1}{1-p_1} \quad (6)$$

$$= \frac{p_1}{p_2} \quad (7)$$

Where,

p_1 is a probability of outcome # 0, and $p_2 = 1 - p_1$ is a probability of outcome # 1.

Calculating the *logit* Equation 5, this will increase (or decrease) by *Bi* for a unit increase in predictor *Xi*.

The Exp(B) indicates the change in predicted odds of the outcome (Incident with toxoplasmosis) for a unit increase in the predictor.

For stillbirth, the odds of incident with toxoplasmosis increase by a factor of 3.665 for a dead fetus, making it more likely than an alive fetus.

For keeping an indoor cat, the odds of incident with toxoplasmosis increase by a factor of 3.72 for those keeping an indoor cat, making it more likely than not to keep an indoor cat.

Lastly, for direct contact with soil, the odds of incident with toxoplasmosis increase by a factor of 3.124 for contact with soil, thus making it more likely than for non-contact with soil.

An important step when dealing with binary logistic regression is to calculate the probability of membership in the target predictor group, which in this case is the incident with toxoplasmosis. This can be done by applying the following formula.

$$P_i = \frac{e^{y_i}}{1+e^{y_i}} \quad (8)$$

Where y_i is the predicted value postulated by using certain set of values of the predictors depending on the estimated model (5).

i.e. for a dead fetus $X_2 = 1$, keeping indoor cat $X_7 = 1$ and contacted with soil $X_{10} = 1$

$$y_i = -2.444 + 1.299(1) + 1.314(1) + 1.139(1) = 1.308$$

$$e^{y_i} = e^{1.308} = 3.698$$

$$P_1 = \frac{3.698}{1+3.698} = 0.787$$

So, for a woman with a dead fetus, keeping indoor cats with direct contact with soil increases the likelihood of infection by toxoplasmosis, with a probability of 0.787 or 78.7%.

The odds that a person birthing a dead fetus, keeping indoor cats, and having direct contact with soil will be infected with toxoplasmosis are;

$$\frac{p_1}{1-p_1} = \frac{0.787}{1-0.787} = 3.694$$

The same things can be done for the rest of categories, for example; for a dead fetus $X_2 = 1$ with non-keeping indoor cats $X_7 = 0$ and non-direct contact with soil $X_{10} = 0$.

$$y_i = -2.444 + 1.299(1) + 1.314(0) + 1.139(0) = -1.145$$

$$e^{y_i} = 0.318$$

$$P_1 = \frac{0.318}{1+0.318} = 0.241$$

However, the probability that a woman who gives birth to a stillborn fetus does not keep indoor cats and has no direct contact with soil is 0.241 or 24.1%.

The odds of a dead fetus, non-pregnant indoor cats with no direct contact with soil being infected with toxoplasmosis are.

$$\frac{p_1}{1-p_1} = \frac{0.241}{1-0.241} = 0.317$$

The odds indicate that women who have a dead fetus and keep indoor cats, while having direct contact with soil, are more likely to become infected.

4. CONCLUSION

Binary logistic regression is an essential statistical tool when dealing with dichotomous outcomes. Logistic regression was highly effective in analyzing the data, as indicated by the positive results, which align with the methodological assumptions. The estimated model showed that three factors stillbirth, direct contact with soil, and keeping indoor cats significantly affect toxoplasmosis infection, while other unseen factors may also have probabilistic effects. The findings suggest that different regressors influence the response variables. The research highlights the importance of binary logistic regression as a suitable statistical method in fields where most data are categorical. The results emphasize the significance of applying binary logistic regression in analysis, especially when the data are categorical, as in clinical medicine.

Funding: This research is supported by Deanship of Scientific Research at Northern Border University, Saudi Arabia (Grant number: NBU-FFR-2025-1153-03).
Institutional Review Board Statement:
Transparency: The authors state that the manuscript is honest, truthful, and transparent, that no key aspects of the investigation have been omitted, and that any differences from the study as planned have been clarified. This study followed all writing ethics.
Data Availability Statement:
Competing Interests: The authors declare that they have no competing interests.
Authors' Contributions: Concept development, literature review, findings interpretation, Sofian A.A. Saad (SAS); data collection, data analysis, conceptual input, Husham M.A. Attaalfadeel (HMA); research design, English language review, reference management, Mufda Jameel Alrawashdeh (MJA). All authors reviewed and approved the final version of the manuscript.

REFERENCES

[1] J. Flegr, J. Prandota, M. Sovičková, and Z. H. Israili, "Toxoplasmosis–a global threat. Correlation of latent toxoplasmosis with specific disease burden in a set of 88 countries," *PloS One*, vol. 9, no. 3, p. e90203, 2014. https://doi.org/10.1371/journal.pone.0090203

[2] P. B. Mortensen *et al.*, "Toxoplasma gondii as a risk factor for early-onset schizophrenia: analysis of filter paper blood samples obtained at birth," *Biological Psychiatry*, vol. 61, no. 5, pp. 688-693, 2007.

[3] M. Kudakwashe and K. M. Yesuf, "Application of binary logistic regression in assessing risk factors affecting the prevalence of toxoplasmosis," *American Journal of Applied Mathematics and Statistics*, vol. 2, no. 6, pp. 357-363, 2014.

[4] A. L. Njunda, J. C. Assob, D. S. Nsagha, H. L. Kamga, P. F. Nde, and V. C. Yugah, "Seroprevalence of Toxoplasma gondii infection among pregnant women in Cameroon," *Journal of Public Health in Africa*, vol. 2, no. 2, p. e24, 2011.

[5] J. Concato, A. R. Feinstein, and T. R. Holford, "The risk of determining risk with multivariable models," *Annals of Internal Medicine*, vol. 118, no. 3, pp. 201-210, 1993.

[6] J. Concato and A. R. Feinstein, "Monte Carlo methods in clinical research: applications in multivariable analysis," *Journal of investigative medicine: the official publication of the American Federation for Clinical Research*, vol. 45, no. 6, pp. 394-400, 1997.

[7] F. E. Harrell Jr, K. L. Lee, D. B. Matchar, and T. A. Reichert, "Regression models for prognostic prediction: advantages, problems, and suggested solutions," *Cancer Treatment Reports*, vol. 69, no. 10, pp. 1071-1077, 1985.

[8] J. Tolles and W. J. Meurer, "Logistic regression: relating patient characteristics to outcomes," *Jama*, vol. 316, no. 5, pp. 533-534, 2016.

[9] H. White, "Maximum likelihood estimation of misspecified models," *Econometrica: Journal of the Econometric Society*, vol. 50, no. 1, pp. 1-25, 1982.

[10] M. D. Begg and S. Lagakos, "On the consequences of model misspecification in logistic regression," *Environmental Health Perspectives*, vol. 87, pp. 69-75, 1990.

[11] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression*. Hoboken, NJ: John Wiley Sons, 2013.

[12] K. P. Vatcheva, M. Lee, J. B. McCormick, and M. H. Rahbar, "Multicollinearity in regression analyses conducted in epidemiologic studies," *Epidemiology (Sunnyvale, Calif.)*, vol. 6, no. 2, p. 227, 2016.

[13] P. Peduzzi, J. Concato, E. Kemper, T. R. Holford, and A. R. Feinstein, "A simulation study of the number of events per variable in logistic regression analysis," *Journal of Clinical Epidemiology*, vol. 49, no. 12, pp. 1373-1379, 1996.

[14] S. K. Sarkar, H. Midi, and S. Rana, "Detection of outliers and influential observations in binary logistic regression: An empirical study," *Journal of Applied Sciences*, vol. 11, no. 1, pp. 26-35, 2011.

[15] A. R. Feinstein, *Multivariable analysis: An introduction*. New Haven, CT: Yale University Press, 1996.

[16] C. M. Mallory and T. Valerio, *Statistics for evidence-based practice in nursing*. Burlington, MA: Jones Bartlett Learning, 2020.

[17] R. R. Rhinehart, *Nonlinear regression modeling for engineering applications: Modeling, model validation, and enabling design of experiments*. Hoboken, NJ: John Wiley, Sons, 2016.

[18] R. B. Haynes, *Clinical epidemiology: How to do clinical practice research*. Philadelphia, PA: Lippincott Williams Wilkins, 2012.

[19] M. Kim, C. Mallory, and T. Valerio, *Statistics for evidence-based practice in nursing*. Burlington, MA: Jones Bartlett Learning, 2020.

[20] A. A. De Hond, E. W. Steyerberg, and B. Van Calster, "Interpreting area under the receiver operating characteristic curve," *The Lancet Digital Health*, vol. 4, no. 12, pp. e853-e855, 2022.