

Binary Regression Analysis Using Three Link Function in Modeling Pulmonary Tuberculosis

Tugga H. A.^{1,3}, Bamigbala O. A.^{2,3}, Ahmad S. S.³

^{1,2}Federal University Wukari, Taraba State, Nigeria

³Federal University Dutsin-Ma, Katsina State, Nigeria

hussainatugga@fuwukari.edu.ng

Submitted:	Revised:	Accepted:	Published:
Sep 16, 2026	Oct 15, 2026	Oct 26, 2026	Oct 31, 2026

Abstract

Pulmonary tuberculosis (PTB) remains a substantial public health challenge, particularly in sub-Saharan Africa. Understanding demographic patterns in PTB occurrence and selecting appropriate statistical models can inform epidemiological analysis and intervention planning. This study examined associations between selected demographic and temporal characteristics and PTB status, comparing binary regression models with logit, probit, and complementary log-log link functions. A retrospective analysis was conducted using secondary data comprising 4,167 patient records from Haruna Tsokwa General Hospital, Takum, Nigeria. PTB status was specified as the binary response variable, while age, sex, and year of diagnosis were included as explanatory variables. Model parameters were estimated using maximum likelihood. Goodness-of-fit was assessed using deviance, Pearson chi-square, and Hosmer–Lemeshow tests, while model selection was informed by the Akaike information criterion (AIC) and Bayesian information criterion (BIC). Age and sex were statistically significant predictors across all three models, with male sex and older age associated with a higher modeled probability of PTB. The consistency of these findings across link functions supports the relevance of these demographic characteristics within the analyzed dataset. The logit model produced the lowest AIC and BIC values and was therefore preferred among the three candidate models. This preference concerns relative model fit within the study context and does not establish predictive performance in other populations. The findings provide hospital-based

evidence that may inform the planning of targeted tuberculosis interventions and demonstrate the value of comparing alternative link functions when analyzing binary health outcomes. Further research should evaluate the preferred model using independent datasets and examine additional clinical and contextual variables. The study concludes that logit regression offers the most suitable specification among the models assessed for these hospital records.

Keywords: Pulmonary Tuberculosis; Binary Regression; Logit; Probit; Complementary Log-Log

INTRODUCTION

Tuberculosis (TB), caused by *Mycobacterium tuberculosis*, remains one of the leading infectious causes of morbidity and mortality globally, particularly in low- and middle-income countries. Although it primarily affects the lungs pulmonary tuberculosis, it can also involve other organs such as the kidneys, brain, and spine. Despite significant advances in diagnosis, treatment, and global control strategies led by organizations such as the World Health Organization, TB continues to pose a substantial public health burden, especially in sub-Saharan Africa (WHO, 2023). In 2016, Africa accounts for a considerable proportion of global TB cases, with Nigeria ranking among the high-burden countries. Reports indicate that Nigeria contributes significantly to the global TB incidence, with persistent challenges such as underreporting, delayed diagnosis, and limited access to healthcare services exacerbating disease control efforts (Ogunsina, et.al, 2022). These epidemiological concerns underscore the need for robust analytical approaches to better understand the determinants of PTB and support evidence-based interventions.

Numerous research problems often call for modeling binary outcomes such as the presence or absence of disease is critical for identifying risk factors and predicting disease occurrence. Traditional methods such as ordinary least squares regression are unsuitable for such analyses due to violations of underlying assumptions. Consequently, binary regression models, particularly those within the generalized linear model (GLM) framework, have gained prominence (Tukur & Usman 2016).

While numerous studies have applied binary regression techniques to infectious disease data, limited attention has been given to comparative assessments of different link functions in modeling PTB. Such comparisons are essential, as the choice of link function can influence model fit, inference, and predictive performance. Therefore, this study aims

to model pulmonary tuberculosis using logit, probit, and complementary log-log regression models, identify significant socio-demographic predictors, and determine the most appropriate model for PTB data. The analysis is based on retrospective data obtained from a secondary healthcare facility in Nigeria, incorporating variables such as age, sex, and year of diagnosis.

Previous studies have demonstrated the usefulness of binary regression models in analyzing infectious diseases. Shahzad et al. (2022) applied stepwise logistic regression to identify determinants of hepatitis B and C co-infection among HIV patients. Hussain et al. (2021) utilized a logit-loglinear modeling approach to study tuberculosis outcomes in Southeast Asia. Tua Sinaga and Darnius (2023) employed binary logistic regression to identify factors influencing tuberculosis incidence, highlighting age and comorbidities as significant predictors. Similar applications have been reported in studies of hepatitis B and other infectious diseases (James et al., 2023). Collectively, these studies underscore the relevance of binary regression models while revealing a gap in comparative analyses of alternative link functions for PTB modeling.

METHODS

This section discusses the binary regression model with various link functions. Binary Regression Model provides the basis of multifaceted models for ordinal, nominal and count model, it explores how regressors' affects the chance of occurrence of an event. (Kamal and Pervaiz, 2011).

Binary Logistic Regression:

Binary Logistic Regression model is a statistical technique that is used for exploring or describing the relationship between a binary dependent variable and a set of explanatory (which might be two or more) variables. Logistic regression assumes a linear relationship between the logit of the response and the predictors. Binary logistic regression models are special case of GLMs with binomial random component and a logit link Function. Logistic regression models are nonlinear regression model and also called logit models. (Alan, 2007).

The binary logistic regression model with p predictor is expressed as

For $P = p(Y = 1 / X) = \pi(x_i)$

$$\log it[\pi(x)] = \log \left[\frac{\pi(x_i)}{1 - \pi(x_i)} \right] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p \quad (1)$$

Where $\pi(x)$: is the probability of having the disease, $(1 - \pi(x))$: is the probability of not having the disease, X_i : is the explanatory variable or the predictors (age, sex and year), α : Gives the odds of having the disease at the baseline i.e. when $x=0$, β_i : Show how the odds differ across the predictors. The probability of success below is an alternative formula for the logit function

$$\pi(x) = \frac{e^{(\alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p)}}{1 + e^{(\alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p)}} \quad (2)$$

Probit Model

The probit model is express as $\pi(x) = \phi(\alpha + \beta x)$ (3)

Equation (3) can also be express as:

$$\phi^{-1}[\pi(x)] = \alpha + \beta x \quad (4)$$

Where, $\phi^{-1}(\cdot)$. Is the probit link function

The probit model is another form of nonlinear model use to explore the relationship between a binary response and multiple explanatory variables. (Ruspriyanty and Sofro 2018). The probit model uses the cumulative distribution function (CDF) to explain the behavior of a dichotomous response variable

Complementary Log-Log Model

The Complementary log-log Regression Model is expressed as;

$$\log \{-\log(1 - p)\} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p \quad (5)$$

Another binary regression model called complementary log-log regression. This Model follows the Gompertz Distribution. The left-hand side of equation (5) is called Complementary log log Link (Kamal & Pervaiz, 2011).

Goodness of Fit Test

To test the adequacy of the model for fitting the data the Hosmer-Lemeshow test which represents a χ^2 (Chi-square) test is utilized in this study. The null hypothesis is that the model is adequate to fit the data. Null hypothesis will only be rejected if there are sufficiently strong grounds to do so (if the p-value is less than 0.05). Pearson's chi-square statistic and the likelihood ratio statistics (G2) which are based on the comparison of the fitted and observed counts were as well estimated to further assess the quality of the model fit. A significant chi-square in goodness-of-fit tests indicates poor model fit. (Kudakwashe & Yesuf 2014).

The Odds Ratio Interpretation

The odds and odds ratio give a crucial interpretation of logistic regression. Taking the Exponentiation of both sides of equation (2) gives the odds of success and also provides an interpretation for β : the odds are multiple by e^β for every one unit increase in x. When $\beta=0$, $e^\beta = 1$ (Alan Agresti, 2007).

$$\frac{\pi(x)}{1-\pi(x)} = \exp(\alpha + \beta x) = e^\alpha (e^\beta)^x \quad (6)$$

RESULTS AND DISCUSSION

Response Information

Variable Value Count		
Ptb	1	913 (Event)
	0	3254
Total		4167

Table 1: Goodness of fit tests for Binary Logistic Regression Using Various Link Functions

Link Function	Test	Chi-Square	P-Value
Logit	Deviance	3589.44	1.000
	Pearson	4259.45	0.131

	Hosmer-Lemeshow	6.19	0.518
Probit	Deviance	3591.18	1.000
	Pearson	4340.42	0.023
	Hosmer-Lemeshow	2.85	0.943
Complementary Log-Log	Deviance	3589.78	1.000
	Pearson	4220.31	0.243
	Hosmer-Lemeshow	10.12	0.182

The goodness of fit test for the logit function in Table 1 reveal that the Deviance Chi-Square and the Pearson Chi-Square has indicates a good fit, as it suggests the model explains the data well. The Hosmer-Lemeshow results from the Table reveal no significant difference between observed and predicted values. The goodness of fit for the probit link function shows Similar results as the Logit function. While the Pearson test indicates some potential issue with the fit. The complementary log log goodness of fit test as well shows that the Deviance and Hosmer-Lemeshow suggest a good fit, however the Pearson test indicates acceptable but potentially weaker fit compared to the Logit model.

Table 2: Parameter Estimate for Binary Logistic Regression Using Logit Functions

Term	Coefficient	Standard Error	Z-Value	P-Value	Odd Ratio exp(Coefficient)
Constant	-4.006	0.293	-13.68	0.000	0.0182
YEAR 2018 (Ref)					1.000
2019	-0.575	0.127	-4.52	0.000	0.5627
2020	-0.705	0.140	-5.02	0.000*	0.4941
2021	-0.552	0.167	3.31	0.001	0.5758
2022	-0.308	0.142	-2.17	0.030	0.7349
2023	0.119	0.134	0.89	0.373	1.1264
SEX Female (Ref)					1.000
Male	0.4524	0.0841	5.38	0.000*	1.5721
AGE 0-14 (Ref)					1.000
15-29 years	2.017	0.292	6.92	0.000	7.5157
30-49 years	3.455	0.289	11.97	0.000*	31.6583
50 years and above	3.834	0.296	12.93	0.000*	46.2472

* Significant at 5% Level

Parameter estimation for the logit model presented in Table 2 shows the coefficient value of -4.006, suggesting a low baseline probability of TB and also the odds of TB when all predictors are at their reference levels is lower than 1 (0.0182), highlighting a baseline decrease in the odds. The years Compared to 2018 in the Table reveals that each

subsequent year generally shows a decrease in the odds of TB. For example, the odds for 2019 are reduced by about 44% (OR = 0.5627). 2023 shows no significant change ($p = 0.373$). Similarly, Males have higher odds (OR = 1.5721) of TB compared to females. Using group (0-14) as the reference the odds for 15-29, 30-49 is 7.5157 and 31.6583, older age groups exhibit significantly higher odds, indicating strong age dependence.

Table 3: Parameter Estimate for Binary Logistic Regression Using Probit Functions

Term	Coefficient	Standard Error	Z-Value	P-Value
Constant	-2.115	0.124	-17.06	0.000
YEAR 2018 (Ref)				
2019	-0.3225	0.0732	-4.41	0.000*
2020	-0.3781	0.0794	-4.76	0.000*
2021	0.3381	0.0990	3.41	0.001
2022	-0.1689	0.0821	-2.06	0.040
2023	0.1080	0.0783	1.38	0.168
SEX Female (Ref)				
Male	0.2527	0.0483	5.23	0.000*
AGE 0-14 (Ref)				
15-29 years	0.938	0.121	7.73	0.000
30-49 years	1.763	0.121	14.62	0.000*
50 years and above	1.999	0.128	15.64	0.000*

* Significant at 5% Level

From Table 3 above present the Parameter Estimate for Binary Logistic Regression Using Probit Functions shows a Coefficient value of -2.115 which suggests a higher baseline probability than the Logit functions. The years compared to 2018 in the Table reveals that each subsequent year shows some decrease in of ptb probability. The result shows similar trends consistent with the logit model, ptb prevalence is higher in males compared to females. **AGE** in the Table above portray a significant increase in odds for older age groups compared to the younger age.

Table 4: Parameter Estimate for Binary Logistic Regression Using Complementary Log-Log Functionss

Term	Coefficient	Standard Error	Z-Value	P-Value	Odd Ratio exp(Coefficient)
Constant	-3.980	0.286	-13.93	0.000*	0.0187
YEAR 2018 (Ref)					1.000
2019	-0.485	0.106	-4.58	0.000*	0.6157
2020	-0.609	0.118	-5.16	0.000*	0.5439
2021	0.409	0.124	3.30	0.001	1.5053
2022	-0.264	0.116	s-2.27	0.023	0.7679

Term	Coefficient	Standard Error	Z-Value	P-Value	Odd Ratio exp(Coefficient)
2023	0.052	0.104	0.50	0.620	1.0534
SEX Female (Ref)					1.000
Male	0.3794	0.0692	5.48	0.000*	0.6073
AGE 0-14 (Ref)					1.000
15-29 years	1.939	0.287	6.76	0.000	6.9518
30-49 years	3.198	0.282	11.33	0.000*	24.4835
50 years and above	3.471	0.286	12.13	0.000*	32.1689

* Significant at 5% Level

The Complementary Log-Log Functions Parameter Estimate presented in Table 4 above shows that the Coefficient has a value of -3.980 which indicate a low baseline and significant probability of tuberculosis. The **YEAR** gives a decreasing trend, with the odds reducing for 2019 and 2020 with a statistically significant probability value, and an increase in odds for 2021 (OR = 1.5053) although the probability value is not significant. The **SEX** reveals that males show higher odds and significant probability of ptb compared to females (OR = 1.6073). The **AGE** from the Table exhibits similar trend compared to other models, older individuals have much higher odds and acceptable probability value of ptb

Model Selection

In this study to know the best model fit between the various links functions (logit, probit and complementary log log or Gombit model). The Akaike information criterion and the Bayesian information criteria were used in selecting the optimal model. The optimal model is the model having a minimal value of the AIC and BIC.

Table 6: Model Selection for Binary Logistic Regression for Various Link Functions

Link Function	AIC	BIC
Logit	3609.44	3672.79
Probit	3611.18	3674.53
Complementary Log-Log	3609.78	3673.13

* Best fit Model

Based on AIC and BIC the best model which described the data well is the logit link function since it has the smallest as AIC and BIC value

CONCLUSION

This study employed three binary regression models logit, probit, and complementary log-log to analyze pulmonary tuberculosis data and evaluate their relative performance in modeling disease occurrence. The findings reveal that age and sex are statistically significant predictors of PTB across all models, with higher likelihood observed among males and older age groups. These results are consistent with established epidemiological patterns of tuberculosis transmission and progression for instance the work of (Ogunsina, et.al, 2022). Model comparison based on Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) indicates that the logit model provides the best fit to the data, demonstrating superior performance relative to the probit and complementary log-log models. This suggests that the logit link function is more suitable for capturing the underlying structure of PTB data in this context. Overall, the study highlights the effectiveness of binary regression models in epidemiological analysis and emphasizes the importance of appropriate model selection. The results provide valuable insights for public health planning, particularly in identifying high-risk groups and informing targeted tuberculosis control strategies.

Recommendations

Based on the findings of this study, the following recommendations are proposed:

1. Public Health Interventions: Health authorities should focus tuberculosis prevention and control programs on high-risk groups, particularly males and older age groups, who were found to have significantly higher odds of PTB.
2. Early Screening Programs: Regular TB screening programs should be strengthened, especially for adults and elderly populations, in order to ensure early detection and timely treatment.
3. Health Awareness Campaigns: Public health education programs should be intensified to increase awareness about tuberculosis transmission, symptoms, and the importance of early diagnosis and treatment.
4. Use of Statistical Modeling in Epidemiology: Researchers and health planners should adopt appropriate statistical models such as the logit model when analyzing binary health outcomes, as it demonstrated superior performance in this study.

5. Further Research: Future studies should incorporate additional predictors such as socioeconomic status, living conditions, comorbidities (e.g., HIV status), and environmental factors to provide a more comprehensive understanding of PTB determinants.
6. Expansion of Study Area: Similar studies should be conducted using data from multiple hospitals or regions to improve generalizability and provide broader insights into tuberculosis patterns in Nigeria.

REFERENCES

- Agresti, A. (2002). *Categorical data analysis* (2nd ed.). John Wiley & Sons. <https://doi.org/10.1002/0471249688>
- Agresti, A. (2007). *An introduction to categorical data analysis* (2nd ed.). John Wiley & Sons. <https://doi.org/10.1002/0470114754>
- Güriş, S., Çağlayan, E., & Ün, T. (2011). Estimating of probability of home-ownership in rural and urban areas: Logit, probit and gompit model. *European Journal of Social Sciences*, 21(3), 405–411. <https://hdl.handle.net/11424/256653>
- Husain, M. S. B., Noor, N. M., & Abdullah, F. Z. S. (2021). Application of logit-loglinear model for tuberculosis disease. *International Journal of Mathematics and Computer Science*, 16(3), 1031–1043. <https://future-in-tech.net/16.3/R-Husain-Noor-Abdullah.pdf>
- James, T. O., Onwuka, G. I., & Bello, U. (2023). Binary logistic regression analysis on risk factors associated with hepatitis B disease in Gusau Local Government Zamfara State, Nigeria. *International Journal of Science for Global Sustainability*, 9(3), 79–84. <https://doi.org/10.57233/ijsgs.v9i3.531>
- Kamal, A., & Pervaiz, M. K. (2011). Factors affecting the family size in Pakistan: Cloglog regression model analysis. *Journal of Statistics*, 18, 29–53. <https://doi.org/10.65463/jstat.2011.115>
- Kudakwashe, M., & Yesuf, K. M. (2014). Application of binary logistic regression in assessing risk factors affecting the prevalence of toxoplasmosis. *American Journal of Applied Mathematics and Statistics*, 2(6), 357–363. <https://doi.org/10.12691/ajams-2-6-1>
- Ogunsina, M. A. (2022). Tuberculosis and associated risk factors: A 5-year review in a tertiary hospital, Kaduna, northwest Nigeria. *The Nigerian Health Journal*, 22(1), 66–78. <https://doi.org/10.60787/tnhj.v22i1.544>
- Ruspriyanty, D. I., & Sofro, A. (2018). Analysis of hypertension disease using logistic and probit regression. *Journal of Physics: Conference Series*, 1108(1), Article 012054. <https://doi.org/10.1088/1742-6596/1108/1/012054>
- Shahzad, F., Abid, F., Obaid, A. J., Rai, B. K., Ashraf, M., & Abdulbaqi, A. S. (2021). Forward stepwise logistic regression approach for determinants of hepatitis B & C among HIV/AIDS patients. *International Journal of Nonlinear Analysis and Applications*, 12(Special Issue), 1367–1396. <https://doi.org/10.22075/ijnaa.2022.5717>

- Sinaga, R. H. T., & Darnius, O. (2023). Binary logistic regression analysis using stepwise method on tuberculosis events. *Journal of Mathematics Education and Application*, 2(1), 1–10. <https://doi.org/10.30596/jmea.v2i1.12164>
- Tukur, K., & Usman, A. U. (2016). Binary logistic regression analysis on admitting students using JAMB score. *International Journal of Current Research*, 8(1), 25235–25239. <https://www.journalcra.com/sites/default/files/issue-pdf/12487.pdf>
- World Health Organization. (2023). *Tuberculosis in the WHO African Region: 2023 progress update*. WHO Regional Office for Africa. https://www.afro.who.int/sites/default/files/2023-09/Tuberculosis%20in%20the%20African%20Region_2023%20report.pdf