Brazilian Journal of Biometrics, **44**, 1–20 (2026)
doi:10.28951/bjb.v44i1.905

BRAZILIAN JOURNAL OF BIOMETRICS
ISSN:2764-5290

ARTICLE

Statistical count models for predicting the number of treatment dropouts in Pediatric Tuberculosis patients

Shalini Kumari,[1] Subhajit Das,[2] Alok Kumar,[1] Jai Krishna Mishra,[3] and Mukesh Kumar*,[4]

[1]Department of Statistics, Banaras Hindu University, Varanasi, India
[2]Department of Mathematics and Statistics, IISER, Kolkata, India
[3]Department of TB and Respiratory Diseases, I.M.S., Banaras Hindu University, Varanasi, India
[4]Department of Statistics, MMV, Banaras Hindu University, Varanasi-221005, India
*Corresponding author. Email: mukesh.mmv@bhu.ac.in

(Received: June 9, 2025; Revised: July 28, 2025; Accepted: July 29, 2025; Published: March 24, 2026)
Section Editor: Gajendra Kumar Vishwakarma

Abstract

Tuberculosis (TB) is a bacterial infectious disease that remains a challenging and unexplored public health concern in pediatric patients. This study aims to fit a suitable count-based regression model and identify the key factors associated with treatment dropout in paediatric TB patients. The retrospective data of 2086 pediatric TB patients was obtained from the Nikshay database at the Department of TB and Respiratory Diseases of Sir Sunderlal Hospital, Banaras Hindu University, Varanasi, Uttar Pradesh, India. The analysis utilized various count data models, including Poisson, Negative Binomial, Zero-Inflated, Hurdle and Zero- truncated models. Model performance was assessed using Akaike Information Criterion, Bayesian Information Criterion and log likelihood values which confirmed the models' effectiveness. The findings suggest that some key factors, such as "Missed followup" and "Contact tracing not done", as well as patients' failing to access the nutritional and financial support provided by the National Tuberculosis Elimination Programme, are significant factors contributing to patients' non-adherence to TB treatment. These insights can inform public health strategies aimed at increasing support services, enhancing TB control programs, and contributing to broader TB elimination efforts.

Keywords: Treatment dropouts; Poisson model; Negative binomial; Zero-inflated model; Hurdle model; Zero truncated model.

1. Introduction

TB is an infectious disease that continues to impact millions of individuals in developing countries like India. According to the India TB Report 2024, in 2022, the TB mortality rate in India was

23 per 1,00,000 population (CTD, 2024). Also, as per the WHO Global TB Report 2024, India is at the forefront of the global TB epidemic (WHO, 2024). India's determined effort to eradicate TB by 2025 has encountered obstacles. These efforts have been impacted due to the COVID-19 pandemic and the factors associated with negative impacts on treatment outcomes. However, TB is a curable disease, with effective and free drugs provided from the National Tuberculosis Elimination Program (NTEP), a centrally-sponsored program implemented within the aegis of the National Health Mission (NHM), under the Ministry of Health and Family Welfare (MoHFW), Government of India. It is a core component of a comprehensive approach to combat TB and address the inequities in TB outcomes across India.

TB in children and adolescents aged 1-18 years is defined as pediatric tuberculosis (PTB). According to the NTEP Pediatric TB Management Guideline 2022, an estimated 3.42 lakh children and adolescents in India are diagnosed with TB cases annually and reported to NTEP, comprising approximately 10% of the total TB caseload. Furthermore, India accounts for nearly one-third of the global pediatric TB burden. Following the National Strategic Plan (2017-25), the pediatric population faces a higher risk of acquiring and developing TB, resulting in considerable illness and death, (Central TB Division & Family Welfare, 2017). The pediatric population is particularly vulnerable to developing severe, life-threatening forms of TB. The nonspecific symptoms, paucibacillary nature, and limited access to healthcare services collectively contribute to the significant burden of the disease. A comprehensive understanding of this patient category in the population is crucial.

A successful treatment is defined as a patient who is cured and has completed the full course of treatment. Any other result is considered an unfavourable outcome. TB treatment dropout refers to a patient who, after their initial registration, stops taking anti-TB medication for two or more consecutive months. Treatment dropouts pose a significant obstacle and a critical challenge to effective treatment management, the TB prevention and cure program. Failure to complete the prescribed treatment regimen is a major contributor to treatment dropout and the occurrence of drug resistance (Basa & Venkatesh, 2015). Thus, the objective of the study is to estimate the number of PTB patients by examining some clinical and programmatic factors that may influence treatment outcomes in PTB patients in certain districts of India. Identifying the key factors associated with patient dropout from TB treatment may assist in mitigating the number of dropouts. So, this relationship can be modeled by applying different count regression models.

The data presents the number of TB dropout patients in various district, which forms an observed count. We introduced Poisson (PO) and negative binomial (NB) models, which may yield optimal results. We also discussed the challenges of excessive zeros in the sample and explored zero-inflated and hurdle models, which reflect districts without dropout occurrences. Additionally, we investigated the performance of zero-truncated models, where districts with zero counts were excluded from the sampling plan. The optimal model was selected using the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and log-likelihood as the evaluative metrics. To the best of our knowledge, previous studies on TB disease have mainly focused on modeling the number of TB cases and identifying the factors influencing disease transmission. While modeling treatment dropouts in PTB patients remains an unexplored area and had received limited attention. However, our study examines treatment adherence by modeling the number of patient dropouts and identifying potential risk factors. Furthermore, more advanced count regression models beyond Poisson and NB distributions have been explored to assess the model fit in different scenarios.

The study tries to integrate public health data and statistical modeling techniques to analyze the research question. Identifying the potential factors contributing to PTB patients' premature leaving of TB treatment is crucial for drawing the attention of TB control programs and healthcare providers. Promoting regular compliance to treatment regimens may contribute to reduced treatment dropouts. This patient-centric approach, especially in paediatrics, requires particular attention and support to achieve successful treatment outcomes. Adopting and implementing patient care

strategies on a larger scale can strengthen the healthcare system. This study may promote health equity by directing resources to the areas and populations that require the most support. The study's significance includes statistical and methodological advancements, public health implications, policy evaluation, equity-focused perspectives, and contributions towards the NTEP's objective of TB elimination.

2. Literature

The available literature provides support for the application of count regression models. A study reviewed TB in children, its diagnostic challenges, and factors contributing to the under-reporting of cases, emphasizing the importance of understanding the burden of the disease and its changing landscape in recent years. Despite the advancements in understanding and managing TB, some gaps remain in understanding childhood TB (Seddon *et al.,* 2015). A study on India's significant contribution to the global TB burden has focused minimally on PTB. It represents a vulnerable population where the delayed diagnosis can worsen morbidity (Balakrishnan & Varadharajan, 2025). An epidemiological study on the TB burden in Brazil used a multivariable logistic regression model with a stepwise backward elimination approach to identify age-related factors associated with adverse TB treatment outcomes. The findings suggest that treatment outcomes vary by age in Brazil and may inform the implementation of public health policies and the management of clinical practices to improve treatment results (Barreto-Duarte *et al.,* 2021).

A study was conducted in Indonesia to find TB cases. To address the rise in TB incidence, discrete count modeling using Poisson regression was performed, with the NB regression approach used to overcome overdispersion. The findings identified the determinants affecting the number of TB cases and determined the appropriate model based on the AIC value (Yotenka & Banapon, 2020). The study explores the application of count data models such as Poisson regression on a German demographic dataset related to fertility, divorce, and mobility. When the Poisson model's underlying assumptions were not satisfied, the NB models and proposed a generalized event count model was applied. This approach allowed a wider range of count data models to be considered, while also accounting for over- and under-dispersion in the data (Winkelmann & Zimmermann, 1994). The study examined the factors influencing unintentional injuries among children from 54 months to sixth grade, utilizing data from the National Institute of Child Health and Human Development study of early child care (Karazsia & Van Dulmen, 2008). To analyze the data, the study offered a practical demonstration of regression models such as ordinary least squares (OLS), Poisson, NB, and zero-inflated Poisson (ZIP) models. The findings suggest that a ZIP model provided the best fit for the data, while the NB model exhibited overdispersion. Similarly, (Kibria, 2006) examined run-off-road crash data, collected on arterial roads in the southern region of Florida, and compared the predictive performance of Poisson, NB, ZIP, and zero-inflated negative binomial (ZINB) models. The results indicate that the NB and ZINB models demonstrate superior predictive performance when the data exhibits over-dispersion or excess zeros.

A study suggests the demand for medical care using a microeconomic cross-section data set by applying count data regression models such as Poisson, quasi-Poisson, NB, hurdle, and ZINB. These models account for over-dispersion and excessive zeros in the data. The data analysis reveals that both the hurdle model and the zero-inflated models yield comparable qualitative findings and model fit, although the hurdle model is slightly preferred due to its clear interpretation (Zeileis *et al.,* 2008). The paper uses count models, such as Poisson, NB, COM-Poisson, and Generalized Poisson distributions, to analyze overdispersed data, obtained from the state of Alagoas, Brazil, and also to determine the association between TB incidence and the Human Development Index (HDI). The study determined that the COM-Poisson distribution regression model exhibited the lowest AIC and BIC values, indicating it as the most appropriate model to describe the overdispersion and the relationship between TB notifications and HDI (Azevedo *et al.,* 2023). The study examines the ap-

plication of Poisson regression and NB regression as more suitable statistical models, compared to OLS regression, to analyze discrete count data that occur infrequently, like the number of pregnancies among adolescent females. The authors evaluate the advantages, disadvantages, and unique considerations of these three modeling approaches, utilizing empirical data from the National Longitudinal Survey of Adolescent Health (Hutchinson & Holtman, 2005). The research that uses zero truncated Poisson (ZTP) and zero truncated Generalized Poisson regression models (ZTGP) to find the factors influencing the number of Children Ever Born (CEB) among women of reproductive age in Andhra Pradesh. The analysis used data from the National Family Health Survey (NFHS) conducted between 2019-2021, and the ZTP regression model was identified as the most effective model in determining the key factors impacting CEB (Muniswamy & Lavanya, 2025).

3. Methodology

3.1 Source of data and variables

The data used in this research is retrospective data of 2086 drug-susceptible (DS) PTB patients extracted from the Nikshay database at the Department of TB and Respiratory Diseases of Sir Sunderlal Hospital, Banaras Hindu University, Varanasi, Uttar Pradesh, India from January 2017 to December 2023 where the patients were registered for TB treatment. The study cohort of PTB aged 1-18 years includes patients who drop out of treatment and are from 69 distinct districts across. Here, we modeled the number of patients leaving the treatment incomplete (dropouts) from each district. The dataset consists of a total of eleven variables, ten independent and one dependent variable. The outcome of interest is the count of PTB patients who dropout of TB treatment in a district. The descriptions of 10 independent variables are presented in Table 1.

Table 1. Variables and their descriptions considered in the analysis

S.No	Independent variables	Descriptions
1	HIV negative	Proportion of PTB patients who were HIV-negative.
2	Diabetic negative	Proportion of PTB patients who were diabetes-negative.
3	Not microbiologically confirmed	Proportion of PTB patients who were not confirmed through microbiological diagnostic methods.
4	New type of case	Proportion of PTB patients who were diagnosed with TB and initiated treatment for the first time.
5	Bank details not added	Proportion of PTB patients whose bank details were not available.
6	Contact tracing not done	Proportion of PTB patients for whom identification and screening of close contacts were not conducted.
7	Not provided TB support	Proportion of PTB patients who failed to access TB support provided under the NTEP.
8	Missed follow-up	Proportion of PTB patients who failed to attend scheduled follow-up visits for treatment monitoring.
9	Patients < 5 years of age	Proportion of PTB patients who were under five years of age.
10	Not delayed treatment	Proportion of PTB patients who initiated anti-TB treatment without delay after diagnosis in the respective district.

3.2 Checking multicollinearity

Multicollinearity is a serious issue that can significantly affect the utility of a regression model. The near-linear dependence among the predictor variables can adversely impact the precision of the estimated regression coefficients. So, the diagnosis and mitigation of multicollinearity is a crucial aspect of regression modeling. The variance inflation factors (VIFs) constitute an important diagnostic tool for assessing the presence and severity of multicollinearity (Shrestha, 2020). It is given

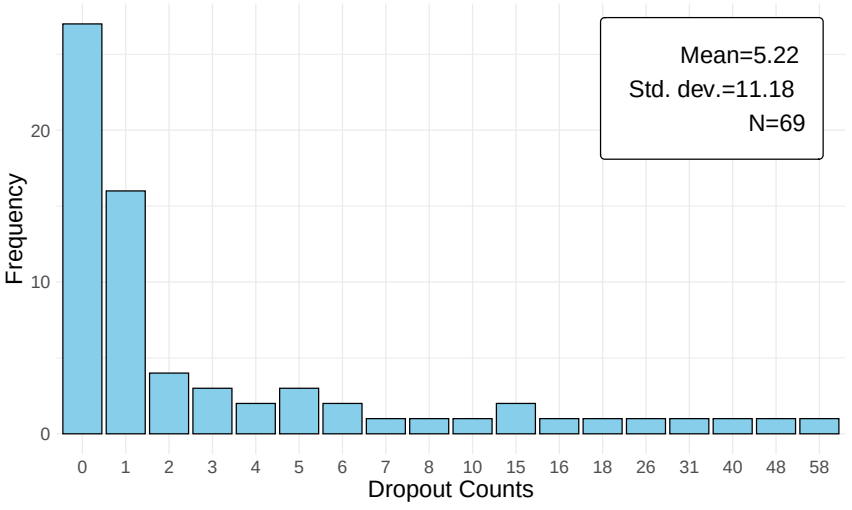

Figure 1. Frequency plot of the dropout counts

by:

$$VIF_j = \frac{1}{1 - R_j^2}$$

where Rj^2 is the coefficient of determination from regressing the j^th predictor variable on the other predictor variables. If a predictor variable x_j is nearly linearly dependent on some of the other predictors, then Rj^2 will be close to unity, resulting in a large VIF value. Values of the VIF greater than 10 are generally regarded as signifying the presence of severe multicollinearity. Which can adversely impact the reliability of the regression models.

3.3 Regression models

The dependent variable, which represents the number of TB patients who discontinue their treatment, is a count variable. The standard statistical modeling approach for analyzing count-based outcome variables, which are restricted to non-negative integer values, is the Poisson model. However, when the Poisson model exhibits distributional overdispersion, the NB model represents a more appropriate alternative. Additionally, in these situations, the Poisson and NB models may be inadequate, as they fail to properly account for an excessive number of zero-count observations, thereby violating the fundamental assumptions of these statistical models. To address this issue, zero-augmented models, such as zero-inflated and hurdle models, are introduced to account for abnormal zero-count situations. Furthermore, zero-truncated models are employed to investigate the model behaviour in the absence of zero-count observations.

3.3.1 Standard count models

Poisson Model (PO): If the interest is to examine the relationship between observed counts and independent variables, Poisson regression is the most appropriate model (Haight, 1967). It utilizes the probability mass function (pmf) of the Poisson distribution as:

$$Pr(Y_i = y_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}, \quad y_i = 0, 1, 2, \ldots$$

The dependent variable, Y_i, is a random variable representing the observed count variable y_i. The parameter $\mu_i > 0$ is the mean parameter of the number of occurrences of the given event, with mean and variance $E(Y_i) = \mu_i = V(Y_i)$. In the context of regression analysis, we assume that there is a function 'g' that relates the mean μ_i to a linear predictor:

$$g(\mu_i) = \boldsymbol{X}_i'\boldsymbol{\beta} \tag{1}$$

Where $\boldsymbol{X_i} = (1, x_1, ...x_k)^T$ is a (k+1) vector of independent variables and $\boldsymbol{\beta} = (\beta_0, \beta_1, ...\beta_k)$ is also (k+1) vector is a set of unknown model parameters. The link function 'g' has been selected as the logarithmic function in this case, and the parameters are estimated using the maximum likelihood estimation (MLE) approach. Then for n random sample $(\boldsymbol{X_1}, y_1), (\boldsymbol{X_2}, y_2)...(\boldsymbol{X_n}, y_n)$ the log-likelihood of the Poisson model is:

$$\mathscr{L}_P(\mathbf{Y}, \boldsymbol{\beta}) = \sum_{i=1}^{n} \left[y_i \ln \mu_i - \mu_i - \ln(y_i!)\right] \tag{2}$$

where $\boldsymbol{Y} = (y_1, \ldots, y_n)^T$, $\mu_i = e^{\boldsymbol{X_i'}\boldsymbol{\beta}}$. The MLE of the parameters can be obtained through the use of iteratively weighted least squares (IWLS) (Myers *et al.,* 2012). If the estimated parameters are $\hat{\boldsymbol{\beta}}$, then the fitted Poisson regression model is $\hat{y_i} = e^{\boldsymbol{X_i'}\hat{\boldsymbol{\beta}}}$.

Negative Binomial Model (NB): To account for Poisson overdispersion, an NB model is used. It is a Poisson-gamma mixture model with Y_i as a random variable representing the frequency of failures before r^{th} success in a successive draws of n Bernoulli trials (Hilbe, 2014). Its pmf is given by:

$$P(Y_i = y_i) = \frac{\Gamma(y_i + \frac{1}{\theta})}{y!\Gamma(\frac{1}{\theta})} \left(\frac{1}{\theta\mu_i + 1}\right)^{\frac{1}{\theta}} \left(\frac{\theta\mu_i}{\theta\mu_i + 1}\right)^{y_i}, \quad y_i = 0, 1, 2, \ldots$$

The parameter $\mu_i > 0$ is the mean parameter of the number of occurrences of the given event, with mean $E(Y_i) = \mu_i$ and variance $V(Y_i) = \mu_i + \mu_i^2\theta$. Here, θ is the shape parameter corresponding to extra dispersion in the model. Similar to Poisson regression, a log link function 'g' defines the linkage between the μ_i and the linear predictor given as:

$$\mu_i = e^{\boldsymbol{X_i'}\boldsymbol{\beta}} \tag{3}$$

Where $\boldsymbol{X_i} = (1, x_1, ..., x_k)^T$ is a (k+1) vector of independent variable and $\boldsymbol{\beta} = (\beta_0, \beta_1, ...\beta_k)$ is also (k+1) vector of unknown model parameters. Again we use MLE to estimate the parameters. Then for n random sample $(\boldsymbol{X_1}, y_1), (\boldsymbol{X_2}, y_2)...(\boldsymbol{X_n}, y_n)$ the log-likelihood function for the NB model is:

$$\mathscr{L}_{NB}(\boldsymbol{Y}, \boldsymbol{\beta}, \theta) = \sum_{i=1}^{n} \left\{ \ln\left(\frac{\Gamma(y_i + \frac{1}{\theta})}{\Gamma(\frac{1}{\theta})\Gamma(y_i + 1)}\right) + y_i \ln(\theta\mu_i) - \left(\frac{1}{\theta} + y_i\right) \ln(1 + \theta\mu_i) \right\} \tag{4}$$

Where μ_i is given as in equation 3. Estimation of the parameters of the log-likelihood function for the NB model can be obtained through the use of IWLS (Cameron & Trivedi, 2013; Myers *et al.,* 2012).

3.3.2 Zero inflated models

Zero-inflated models are primarily used when the excessive zeros arise from a mixture of true zeros and stochastic zeros. If we consider treatment dropout as the target outcome, the population of interest would consist of individuals with no risk of dropout as true zeros, as well as those with a high risk of dropout as stochastic zeros. The zero-inflated model can be used to address not only

the overdispersion arising from excessive zeros but also the unobserved variability within the PTB patient (Kibria, 2006). The model simultaneously estimates the zero-inflated and count models in order to optimise the likelihood function. This indicates that the zero and count components are interrelated, such that the parameters of one component affect the fit of the other component. As a result, modifying the predictors for one part of the model could impact the approximations of the counterpart.

Zero Inflated Poisson model (ZIP): The ZIP model constructs two separate regression equations: one that models the probability of the event happening, and another that models the count of the event, conditional on its occurrence, i.e. first part generates zero and the other part generates counts through the Poisson model. The ZIP model is defined as:

$$P(Y_i = y_i) = \begin{cases} \phi_i + (1-\phi_i)e^{-\mu_i}, & \text{if } y_i = 0 \\ (1-\phi_i)e^{-\mu_i}\frac{\mu_i^{y_i}}{y_i!}, & \text{if } y_i > 0 \end{cases}$$

Where the dependent variable, Y_i, is a random variable representing the count. The parameter $\mu_i > 0$ is the mean parameter of the number of occurrences of the given event modeled as $\mu_i = e^{X_i'\boldsymbol{\beta}}$, with mean and variance $E(Y_i) = (1-\phi_i)\mu_i$ and $V(Y_i) = (1-\phi_i)\mu_i(1+\phi_i\mu_i)$. The parameter ϕ_i: $0 < \phi_i < 1$ is the probability of zero count depending on the covariates $\boldsymbol{X_i}$ through a *probit* link function. It is modeled as:

$$probit(\phi_i) = \Phi^{-1}(\phi_i) = \boldsymbol{X_i'\alpha} \tag{5}$$

Here, the cumulative distribution function of the standard normal distribution is represented by Φ, and $\boldsymbol{\alpha}$ is the unknown model parameters to be estimated. In the zero-inflated modeling approach, the true zeros are modeled using *probit* regression, while the stochastic zeros are addressed through the zero-inflated count component. In the current study, the stochastic zeros represent districts that have a high probability of treatment dropout but do not have actual dropouts owing to random occurrences. The regression coefficients $\boldsymbol{\beta}$ represent the log-transformed rate ratios for the count outcomes of the Poisson model, while the coefficients $\boldsymbol{\alpha}$ represent the *probit*-transformed odds ratios for the probability of structural zeros in the probit model. The parameters in both sets are determined by maximizing the log-likelihood function (Lambert, 1992).

Zero inflated negative binomial model (ZINB): Similar to the ZIP model, an extension called the the ZINB model is a combination of a dual-state system: an NB model and a binary process. The ZINB also model the over-dispersed data and allows for extra variation relative to the standard NB model, enabling it to handle over-dispersion more effectively (Mwalili *et al.,* 2008; Kibria, 2006). The ZINB model is defined as:

$$P(Y_i = y_i) = \begin{cases} \phi_i + (1-\phi_i)\left(\frac{1}{\theta\mu_i+1}\right)^{\frac{1}{\theta}}, & \text{if } y_i = 0 \\ (1-\phi_i)\frac{\Gamma(y_i+\frac{1}{\theta})}{y_i!\Gamma(\frac{1}{\theta})}\left(\frac{1}{\theta\mu_i+1}\right)^{\frac{1}{\theta}}\left(\frac{\theta\mu_i}{\theta\mu_i+1}\right)^{y_i}, & \text{if } y_i > 0 \end{cases}$$

Where the dependent variable, Y_i, is a random variable representing the observed count. The parameter $\mu_i > 0$ is the mean parameter of the number of times a phenomenon is observed is modeled as $\mu_i = e^{X_i'\boldsymbol{\beta}}$ with mean $E(Y_i) = (1-\phi_i)\mu_i$ and variance $V(Y_i) = \mu_i(1-\phi_i)[1+(\phi_i+\theta)\mu_i]$. θ is the shape parameter corresponding to extra dispersion in the model, and ϕ_i is the probability of zero count depending on the covariates $\boldsymbol{X_i}$ through a *probit* link function as given in equation 5. The parameters $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$ and θ are estimated through the MLE (Myers *et al.,* 2012).

3.3.3 Hurdle model

Hurdle models address excess zero values and overdispersion in count data. This approach also employs a two-part framework: the first component uses a binary probability approach to determine if the count variable is zero, known as the hurdle component. The second component models positive counts via a truncated-at-zero distribution once the hurdle has been crossed, known as the count process. The two-stage decision-making process reflected in hurdle regression models corresponds to common human behaviours and thus offers an interpretable framework (Mullahy, 1986; Cameron & Trivedi, 2013). The model's two parts can feature distinct sets of predictor variables. Altering the predictors for one component does not impact the other. A Poisson and NB model are specified for the count portions of the model.

Hurdle Poisson model (HP): The Logit-Poisson model integrates a *logit* component that models the two-category outcome: zero or non-zero. The truncated Poisson model models the positive count values conditional on the outcome being non-zero. The HP model is expressed as:

$$P(Y_i = y_i) = \begin{cases} \phi_i, & \text{if } y_i = 0 \\ (1-\phi_i) \cdot \frac{e^{-\mu_i}\mu_i^{y_i}/y_i!}{1-e^{-\mu_i}}, & \text{if } y_i > 0 \end{cases}$$

Where the outcome variable, Y_i, is a random variable representing the observed count. The parameter $\mu_i > 0$ is the mean parameter of the number of occurrences of the given event modeled as $\mu_i = e^{X_i'\beta}$. The parameter ϕ: $0 < \phi_i < 1$ is the risk of zero count depending on the covariates $\boldsymbol{X_i}$ through a *logit* link function, modeled as:

$$logit(\phi) = \log_e\left(\frac{\phi}{1-\phi}\right) = \boldsymbol{X_i'\alpha} \tag{6}$$

The parameters $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$ are estimated through the MLE (Myers *et al.,* 2012).

Hurdle negative binomial model (HNB): The HP model addresses the issue of over-dispersion arising from an excess of zeros, but it does not account for unobserved heterogeneity. In such situations, the HNB model may be more appropriate. Similarly, this model uses a *logit* component to describe all zero observations, while the positive count data is represented through the use of a zero truncated NB model. The HNB model is given as:

$$P(Y_i = y_i) = \begin{cases} \phi_i, & \text{if } y_i = 0 \\ (1-\phi_i) \cdot \frac{\Gamma(y_i+\frac{1}{\theta})}{y_i!\,\Gamma(\frac{1}{\theta})} \left(\frac{\theta\mu_i}{1+\theta\mu_i}\right)^{y_i} \left(\frac{1}{1+\theta\mu_i}\right)^{\frac{1}{\theta}} \Big/ \left(1-\left(\frac{1}{1+\theta\mu_i}\right)^{\frac{1}{\theta}}\right), & \text{if } y_i > 0 \end{cases}$$

The HNB model shares the same set of parameters as the ZINB model; the only difference is in the data generating process for zero and positive observations (Mullahy, 1986; Cameron & Trivedi, 2013).

3.3.4 Zero truncated model

Our dataset included many observations with zero counts (districts with no treatment dropouts). We modeled these using zero-inflated and hurdle models. Now, we aim to analyze the data without the zero counts, so we have excluded all zero count observations. However, excluding the zero counts means we cannot use regular Poisson or NB models, as these models incorporate zero counts. Therefore, we must adjust the pmf and log-likelihoods of the standard Poisson or NB model to exclude zeros (Muniswamy & Lavanya, 2025).

Zero Truncated Poisson model (ZTP): For the Poisson model, the pmf was defined under Poisson model. The pmf for the ZTP model (Cameron & Trivedi, 2013; Cruyff & Van Der Heijden, 2008) is defined by conditioning on $y > 0$ is expressed as :

$$P(Y_i = y_i \mid Y_i > 0) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!(1-e^{-\mu_i})}, \quad y_i = 1, 2, 3, \ldots$$

In this case, the corresponding log-likelihood function would be reformulated to exclude the zero count observations and would be of the form.

$$\mathscr{L}(\boldsymbol{Y}, \boldsymbol{\beta} | \boldsymbol{Y} > 0) = \mathscr{L}_P(\boldsymbol{Y}, \boldsymbol{\beta}) - \sum_{i=1}^{n} \ln(1 - e^{-\mu_i})$$

where $\mathscr{L}_P(\boldsymbol{Y}, \boldsymbol{\beta})$ is given in equation 2.

Zero Truncated negative binomial model (ZTNB): Similarly, for the NB model, the pmf was defined under the NB model. The pmf for the ZTNB model (Cameron & Trivedi, 2013; Cruyff & Van Der Heijden, 2008) is defined by conditioning on $y > 0$ is expressed as :

$$P(Y_i = y_i \mid Y_i > 0) = \frac{\frac{\Gamma(y_i + \frac{1}{\theta})}{y_i!\,\Gamma(\frac{1}{\theta})} \left(\frac{1}{\theta\mu_i + 1}\right)^{1/\theta} \left(\frac{\theta\mu_i}{\theta\mu_i + 1}\right)^{y_i}}{1 - \left(\frac{1}{\theta\mu_i + 1}\right)^{1/\theta}}, \quad y_i = 1, 2, 3, \ldots$$

The corresponding log-likelihood function for the ZTNB model is of the form:

$$\mathscr{L}(\boldsymbol{Y}, \boldsymbol{\beta}, \theta | \boldsymbol{Y} > 0) = \mathscr{L}_{NB}(\boldsymbol{Y}, \boldsymbol{\beta}, \theta) - \sum_{i=1}^{n} \ln(1 - (1 + \theta\mu_i)^{\frac{1}{\theta}}) \tag{7}$$

where $\mathscr{L}_{NB}(\boldsymbol{Y}, \boldsymbol{\beta}, \theta)$ is given in equation 4. In both cases, μ_i is the mean parameter and is related to the predictors through a log link function $\mu_i = e^{X_i'\boldsymbol{\beta}}$.

3.4 Identifying suitable models

To determine the best-fitting model for treatment dropouts among PTB patients, different goodness-of-fit measures are used. These measures assess how well a statistical model aligns with the observed data. The predictive performance of the model approaches is evaluated using metrics like log-likelihood, AIC, and BIC values. The values of these model fit indices have been reported in the results section.

The value of the maximized log likelihood function, $\mathscr{L}$, represents the optimal alignment between the model parameters and the observed data. The higher the maximized log likelihood, the greater the agreement between the statistical model and the observed data. The AIC is a data-driven model selection tool that evaluates the goodness-of-fit of a statistical model while penalizing for model complexity. The AIC value is calculated for each model, and the model with the lowest AIC is selected as the optimal balance between complexity and goodness-of-fit (Akaike *et al.,* 1973; Chakrabarti & Ghosh, 2011). It is expressed as:

$$\text{AIC} = 2k - 2\ln(\mathscr{L})$$

Here, the number of parameters estimated in the model is denoted by k, and $\ln(\mathscr{L})$ represents the maximum value of the log-likelihood function.

BIC is another criterion for model selection that balances the trade-off between the model's goodness-of-fit and complexity, imposing a penalty for models with a greater number of parameters. It differs from AIC only in the first term, which is the sample size n. The models that minimize the BIC value are selected (Schwarz, 1978). It is expressed as:

$$\text{BIC} = \ln(n) - 2\ln(\mathscr{L})$$

where n is the total count of observations in the dataset.

Additionally, a probability plot that visually compares the observed count data with the predicted count data of a selected model to evaluate the model's goodness of fit is plotted for each model. All model estimates and plot constructions were generated using the open-source R statistical software.

4. Result

The summary statistics of the dropout counts include a mean of 5.21, and a variance of 125.11. Additionally, the data shows a maximum of 58 dropouts and a minimum of 0 dropouts across 69 districts as shown in Figure 1 . The relation between the number of TB treatment dropout cases and the factors that impact treatment outcomes is examined through the regression analyses. The regression models include all ten independent variables as shown in Table 1, to preserve the comparative findings in the multivariate analysis. A multicollinearity assessment of all independent variables was conducted to ensure the modeling assumptions were met, as presented in Table 2. The results indicated no significant multicollinearity concerns, as all VIFs values were less than 10. Therefore, all variables were retained in the models for analysis.

4.1 Findings from the Poisson model

In this case, the Poisson model is used, as the treatment dropout counts observed in the PTB patient population are presumed to adhere to a Poisson distribution. The model results are presented in Table 3. This approach enables the estimation of the model coefficients, its associated standard errors (SE), and the p-values to statistically test the null hypothesis (H_0): that each predictor variable's coefficient is equal to zero (the predictor has no effect on the dropout counts) against the alternative hypothesis (H_1): that the coefficient is non-zero (the predictor have some effect) for each predictor variable. A two-sided Z-test is used to evaluate these hypotheses, and the p-values are reported at the 5% level of significance. Except for "New Type of case" and "Diabetic negative", all the remaining variables were found to be associated with treatment dropouts.

The key assumption of the Poisson model is equidispersion, assuming equal mean and variance for the outcome variable. However, in the present case, this assumption is violated as the variance is over five times the mean. When the equidispersion condition is violated, over- or underdispersion is assessed using the dispersion ratio (DR), defined as the residual deviance divided by the residual degrees of freedom. If the DR > 1, the data exhibits overdispersion when variance is greater than mean; if DR < 1, the data is underdispersed when variance is less than mean. It is the most common phenomenon in the Poisson model that biases parameter estimates and fitted values. Here, DR is 8.26 greater than one, indicating overdispersion in the model. Therefore, the NB model may be more suitable for achieving a superior model fit.

Table 2. Assessment of multicollinearity among independent variables using VIFs

S.No	Independent variables	VIFs	S.No	Independent variables	VIFs
1	HIV negative	1.93	6	Contact tracing not done	3.46
2	Diabetic negative	2.18	7	Not provided TB support	1.45
3	Not microbiologically confirmed	1.48	8	Missed follow-up	1.37
4	New type of case	1.62	9	Patients < 5 years of age	1.31
5	Bank details not added	2.90	10	Not delayed treatment	1.86

4.2 Findings from the NB model

The NB model can effectively model overdispersion in count data. The results of the NB model presented in Table 3 indicate that variables such as "Bank details not added", "Contact tracing not done", "Not provided TB support", and "Missed follow up" have a significant effect on the model. The model also revealed that of the four significant variables, the estimates for two are only positive. This confirms that in estimating the coefficients using the NB model, a one-unit increase in "Not provided TB support" or "Missed follow-up" is associated with $e^{6.0} \approx 426$ or $e^{7.6} \approx 2059$ fold increase in expected counts of dropouts respectively, and have a positive impact on the model. However, "Bank details not added" and "Contact tracing not done" have negative coefficients, which suggests that an increase of one unit in those variables is associated with a factor of $e^{-2.3}$ and $e^{-3.0}$ times decrease in the expected count of dropouts i.e the expected count decreases by about 90.4% and 95.1% respectively, indicating that these variables have a negative impact on the model.

Table 3. Estimates of regression coefficients for the independent variables in the Poisson and negative binomial models

	PO model			NB model		
Variables	**Estimate**	**SE**	**P-value**	**Estimate**	**SE**	**P-value**
Intercept	-2.773	2.699	0.007**	-6.196	2.726	0.023*
HIV negative	-0.717	2.175	0.030*	-0.410	0.856	0.632
Diabetic negative	-0.298	0.829	0.407	-0.693	0.921	0.452
Not microbiologically confirmed	-0.630	2.014	0.044*	-0.663	0.843	0.431
New type of case	-0.508	0.596	0.551	-1.100	2.241	0.623
Bank details not added	-2.026	5.513	0.000***	-2.342	0.919	0.011*
Contact tracing not done	-2.234	5.768	0.000***	-3.024	1.156	0.009**
Not provided TB support	3.769	4.579	0.000***	6.055	2.097	0.004**
Missed follow-up	5.240	11.753	0.000***	7.631	1.240	0.000***
Patients < 5 years of age	-5.526	4.304	0.000***	-3.222	2.752	0.242
Not delayed treatment	0.727	2.390	0.017*	0.562	0.786	0.475

*p < 0.05; **p < 0.01; ***p < 0.001

4.3 Findings from the zero-inflated models

The data may exhibit excessive zero counts, which can contribute to overdispersion in the count data. This suggests that the data has a substantial number of PTB patients who experienced no treatment dropouts, such a pattern that cannot be adequately addressed using a standard NB model. In

such cases, the ZIP and ZINB models can effectively account for both excess zeros and overdispersion. These models have a dual-component framework, where the first part utilizes a *probit* link function to model a binary process that determines whether a treatment dropout occurs or not, and the second part considers Poisson or NB models to model the positive count of treatment dropouts, if any. The results of the ZIP and ZINB regression analyses, presented in Table 4, focus solely on the estimates of coefficients for the positive count component, as none of the variables were found to be significant for the *probit* part. The result reveals that for the ZIP model, factors such as "Bank details not added", "Contact tracing not done", "Not provided TB support", and "Patients < 5 years of age" were significant factors associated with the treatment dropouts. Similarly, the ZINB model indicates that "Bank details not added" and "Not provided TB support" were significantly linked to increased treatment dropouts.

Table 4. Estimates of regression coefficients for independent variables in the count component of the ZIP and ZINB models

	ZIP model			ZINB model		
Variables	**Estimate**	**SE**	**P-value**	**Estimate**	**SE**	**P-value**
Intercept	1.916	1.241	0.123	-0.324	3.850	0.933
HIV negative	-0.720	0.383	0.060	-0.218	1.018	0.831
Diabetic negative	-0.392	0.408	0.337	-0.642	1.067	0.547
Not microbiologically confirmed	-0.505	0.369	0.172	-0.204	1.064	0.848
New type of case	-0.428	0.930	0.645	-0.845	2.875	0.769
Bank details not added	-2.454	0.378	0.000***	-2.423	1.202	0.044*
Contact tracing not done	-2.190	0.408	0.000***	-2.407	1.382	0.082
Not provided TB support	3.418	0.871	0.000***	4.954	2.492	0.047*
Missed follow-up	0.763	0.712	0.284	1.295	2.905	0.656
Patients < 5 years of age	-6.910	1.380	0.000***	-3.390	2.953	0.251
Not delayed treatment	0.840	0.352	0.017*	0.753	1.030	0.465

*p < 0.05; **p < 0.01; ***p < 0.001

4.4 Findings from the hurdle models

Two-part hurdle models, similar to zero-inflated models, are used to analyze count data with excess zeros and overdispersion, but with a slightly different interpretation. The framework of the current study indicates that all PTB patients have an equal risk of dropout, and the zeros originate from the same cohort. In this analytical approach, the zero-valued observations are modeled through *logit*, while the positive count data is fitted using either a truncated Poisson distribution for equidispersed data or a truncated NB model for overdispersed data. The results of the positive count components of the HP and HNB models are presented in Table 5. The hurdle model's binary component (*logit*) did not identify any significant variables. Further, the analysis found a strong consistency between the ZIP and HP models, in addition to the ZINB and HNB, in identifying the factors associated with higher rates of PTB treatment dropout. For the ZIP and HP models, the results indicate that variables such as "Bank details not added", "Contact tracing not done", "Not provided TB support", and "Patients < 5 years old" were significantly associated with increased treatment dropout rates. Similarly, the ZINB and HNB models revealed that "Bank details not added" and "Not provided TB support" were significantly linked to higher dropout rates. This concordance in the assessment of statistical significance across these various modeling approaches suggests

a robust determination of key factors for non-completion of TB treatment.

Table 5. Estimates of regression coefficients for independent variables in the count component of the HP and HNB models

	HP model			HNB model		
Variables	**Estimate**	**SE**	**P-value**	**Estimate**	**SE**	**P-value**
Intercept	1.985	1.336	0.137	-9.013	9.430	0.339
HIV negative	-0.755	0.428	0.078	2.519	3.751	0.502
Diabetic negative	-0.319	0.450	0.478	-0.422	3.154	0.894
Not microbiologically confirmed	-0.492	0.403	0.222	1.323	2.263	0.559
New type of case	-0.828	1.003	0.409	-5.195	5.906	0.379
Bank details not added	-2.781	0.413	0.000***	-9.076	4.390	0.039*
Contact tracing not done	-2.331	0.440	0.000***	-2.665	3.368	0.429
Not provided TB support	3.848	0.966	0.000***	13.796	5.833	0.018*
Missed follow-up	0.848	0.732	0.247	3.917	6.856	0.568
Patients < 5 years of age	-8.245	1.512	0.000***	-19.058	13.373	0.154
Not delayed treatment	0.964	0.396	0.015*	2.635	2.931	0.369

*p < 0.05; **p < 0.01; ***p < 0.001

4.5 Findings from the zero truncated models

The models discussed in the preceding tables addressed scenarios characterized by an excess of zero counts. The issue pertains to the statistical modeling of count-based data, which includes the potential exclusion of zero-valued observations, particularly emphasizing districts that have experienced at least one treatment dropout. To model such count data, a zero-truncated model is used. The study utilizes the ZTP model and its extension, the ZTNB model, to address overdispersed count data, and the findings are displayed in the Table 6. From Table 6, the consistent findings across the ZTP/HP/ZIP models indicate the robustness of the identified significant variables. Specifically, the variables such as “Bank details not added", “Contact tracing not done", “Not provided TB support", “Number of patients < 5 years", and “Not delayed treatment" were found to be significant for treatment dropouts, despite the differing approaches these models used in handling zero values. So, the effect of these variables remains strong. Also, from Table 6, we can see that in the ZTNB model, the analysis did not identify any variables that achieved statistical significance. This may be due to the model’s adjustment for overdispersion, which often increases SE and reduces statistical power. Additionally, the combined effects of zero-truncation and overdispersion modeling could have contributed to wider confidence intervals, thus making it difficult to determine the significance of variables.

Table 6. Estimates of regression coefficients for independent variables in truncated Poisson and negative binomial models

	ZTP model			ZTNB model		
Variables	**Estimate**	**SE**	**P-value**	**Estimate**	**SE**	**P-value**
Intercept	1.916	1.241	0.133	0.204	5.951	0.973
HIV negative	-0.720	0.383	0.070	-0.104	1.704	0.952
Diabetic negative	-0.392	0.408	0.344	-0.467	1.760	0.793
Not microbiologically confirmed	-0.505	0.369	0.181	-0.206	1.773	0.908
New type of case	-0.428	0.930	0.648	-0.281	3.811	0.942
Bank details not added	-2.454	0.378	0.000***	-1.632	2.076	0.438
Contact tracing not done	-2.190	0.408	0.000***	-1.545	1.834	0.406
Not provided TB support	3.418	0.871	0.000***	2.949	3.866	0.452
Missed follow-up	0.763	0.713	0.292	0.433	4.211	0.919
No. of patients < 5 years	-6.910	1.380	0.000***	-2.837	3.506	0.425
Not delayed treatment	0.840	0.352	0.023*	0.525	1.585	0.743

*p < 0.05; **p < 0.01; ***p < 0.001

4.6 Models comparisons

To determine the most suitable model for the count based data, different goodness-of-fit measures have been evaluated. We compared Poisson, NB, ZIP, ZINB, HP, and HNB models, which were trained on all 69 discrete data, including the zero counts. In contrast, the ZTP and ZTNB models were trained exclusively on only 42 districts, the reduced dataset containing only positive counts. Therefore, it would be inappropriate to compare these zero-truncated models directly with the other models. As a result, we present the comparison of the zero-truncated models separately. The model characteristics are given in Table 7. The HNB model demonstrated the lowest AIC value (251.23), followed by the NB model. The remaining comparative measures, such as the maximum log-likelihood (-102.61) and minimum BIC values (302.62), demonstrate the HNB model's clear advantages over other count models.

Table 7. Comparison of models using goodness-of-fit measures

Models	Log-likelihood	AIC	BIC
PO	-306.25	634.51	659.08
NB	-126.15	276.30	303.11
ZIP	-276.81	597.61	646.76
ZINB	-122.39	290.77	342.16
HP	-272.50	588.99	638.15
HNB	**-102.61**	**251.23**	**302.62**
ZTP	-276.81	575.61	594.73
ZTNB	**-114.75**	**253.97**	**274.35**

Figure 2 (A) depicts the graphical illustration of the difference between observed and predicted probabilities of treatment dropouts at each count for the count regression models used in the study. Figure 2 (A) indicates that the NB and HNB models outperformed the Poisson model in predicting the number of dropouts. The differences between observed and predicted dropout probabilities for

the HNB are close to the reference zero line, suggesting a better fit compared to the other count models.

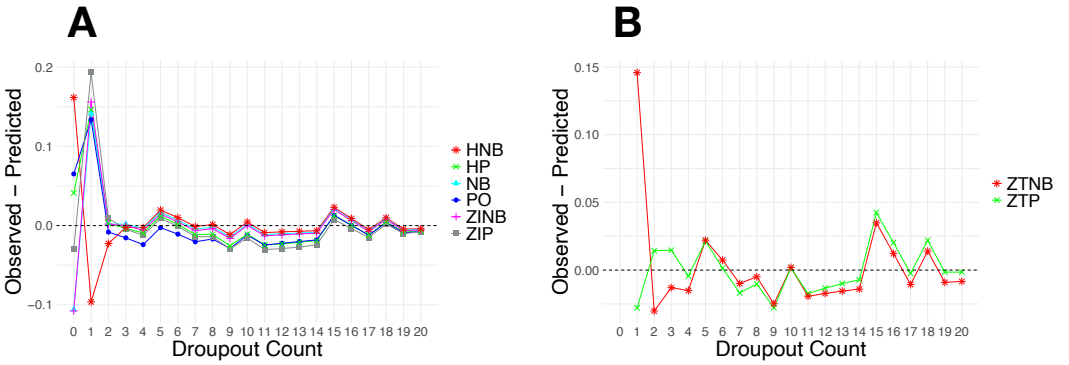

Figure 2. Graphical comparison of observed and predicted probabilities to compare different models for dropout counts (A) Models trained on all 69 data points, including zero counts; (B) ZTP and ZTNB models trained on the positive count data.

For the zero-truncated model, the key difference is that the log-likelihood of the model is calculated while truncating at zero. According to the AIC values in Table 7, the ZTNB model is strongly preferred over the ZTP model, as the ZTNB model has the minimum AIC (253.97), BIC values (274.35) and maximum log-likelihood (-114.75) values. Figure 2 (B) shows the plot of the difference between observed and predicted probabilities of treatment dropouts at each count for the ZTP and ZTNB models. The findings indicate that the ZTNB outperformed the ZTP model, more precisely in predicting the number of dropouts. This is evidenced by the ZTNB model's lines depicting the difference between observed and predicted dropout probabilities being closer to the reference zero line, suggesting a superior model fit. The x-axis of the plots depicted in Figures 2 (A) and 2 (B) is limited to 20 counts. Beyond this count, the model fitting remains largely consistent.

5. Discussion

Count regression models are appropriate when the objective is to model an outcome variable characterised by count data with highly skewed distributions and heteroscedasticity. Our study found that when using the Poisson model to analyze treatment dropouts, there was excess variability in the data distributions; as a result, the NB model was utilized to fit better and predict the count data. These findings are corroborated by prior studies that have also determined the NB model to have a superior fit compared to the Poisson model when modeling the number of TB cases in young children (Fajri *et al.,* 2024; Yotenka & Banapon, 2020). A study supporting the use of count models in the application of TB data found the COM-Poisson distribution regression model as the best fit for the data, with the lowest values for the AIC and BIC criteria (Azevedo *et al.,* 2023). The study (MacCallum *et al.,* 2002) encourages researchers to consider statistical methodologies according to the characteristics of their data.

The data concerning the number of treatment dropouts exhibits an excess of zero values, which leads to overdispersion. Consequently, there is a need to apply advanced statistical regression techniques to analyze count data characterized by overdispersion, an excessive zeros, and the absence of zeros in order to conduct a comprehensive analysis and achieve accurate predictions. Accordingly, we have fitted zero-inflated and zero-hurdle models, which can evaluate the variability associated with the excess zero observations (Greene, 1994), and have also examined the fit of zero-truncated models on the data having no zero observations (Muniswamy & Lavanya, 2025). The study used

ZIP, ZINB, HP, and HNB regression models to examine the factors contributing to over-dispersion and to evaluate the predictive capabilities of these approaches in estimating the number of treatment dropouts among PTB patients. The result showed that the HNB model is the most robust and accurate in estimating the risk factors responsible for TB treatment dropouts. This model indicates that factors such as a lack of patients' bank details to receive financial support and a lack of nutritional support were found to contribute significantly to an increased likelihood of treatment dropouts. The current findings hold significant indications for interventions intended to enhance TB treatment outcomes, as they illuminate potential factors that may explain the increased risk of treatment dropouts.

A relevant study investigated the use of various count regression models to model the average number of spontaneous abortions experienced by women in Punjab and several northern Indian states, using standard Poisson, NB, zero-inflated and hurdle regression approaches. The findings, considering important variables related to the frequency of spontaneous abortions, indicated that the ZINB model demonstrated the optimal fit to the data. (Verma *et al.,* 2020). In our case, HNB was the best fit for predicting treatment dropouts in PTB patients. This suggests that certain regression models may be more effective than others in accurately representing the observed data, and the associated factors can vary depending on the specific model used (Hall & Zhang, 2004). The treatment dropout data exhibit overdispersion, which is not solely attributable to an excess of zero values but also due to unobserved heterogeneity among PTB patients across various districts of India. The Poisson model has demonstrated the greatest predictive error when estimating the frequency of dropout occurrences, a consequence of the presence of overdispersion in the data. Conversely, the HNB models account for various sources of excessive variability and have resulted in more accurate predictions. These findings are supported by prior research, which has also found strong agreement between ZHNB and ZINB models in the analysis of vaccine adverse event data, representing only "at-risk" zeros (Rose *et al.,* 2006). Additionally, a related study that compared the performance of various statistical models in estimating the lymph node status of patients with breast cancers, similarly determined the ZHNB and ZINB models as the best fit (Dwivedi *et al.,* 2010), analogous to the data used in the current study.

In a nutshell, the HNB model is the most suitable choice to predict PTB treatment dropout, considering two key factors: "Bank details not added" and "contact tracing not done". Furthermore, the modeling process revealed that additional variables including "Not provided TB support", "Missed follow up", "No Patients < 5 years of age ", and "Not delayed treatment", may also contribute to an elevated risk of TB treatment dropout. Patients without accessible bank account details may miss out on potential government-provided cash transfer benefits, which are provided to reduce financial barriers to treatment. Additionally, the lack of nutritional support to the patients provided to boost their immunity and those patients whose contact tracing is not done with a lost to follow-up mechanism are found to be significant factors for dropouts, as they have limited engagement with healthcare providers, potentially leading to poor treatment adherence. Furthermore, the two most important factors are critically ill children under 5 years of age, and those patients who did not experience treatment delays were more prone to becoming dropouts. These factors highlight important considerations for strengthening TB control programs. The study that investigated the factors influencing antenatal care visit frequency utilized various count data models and found the hurdle model to be an appropriate choice for modeling antenatal service uptake among expectant women in Ethiopia Bekalo & Kebede, 2021. This aligns with the findings from our study, which also identified the HNB as the best model for analyzing treatment dropout patterns. This may be attributed to the capability of the hurdle model to distinctly handle the zero and positive counts; the hurdle model avoids ambiguity. When the zeros are fundamentally structural and cannot occur in the count-generating process, the hurdle model appropriately assigns them to the binary model. This often results in lower deviance, better AIC, BIC, and log-likelihood as well as improved pre-

diction accuracy, thereby making the hurdle model a preferred choice.

The count regression models and zero-augmented models have limitations in their underlying assumptions, such as the requirement for the count data to have at least some zero or excess zero values, and the significance attached to these zero values. Further, we have explored the model fit when the outcome variable does not contain any zero values, such as in a district where dropouts are structurally impossible. To address this, we utilized the ZTP and ZTNB models, which can account for the "missing" zeros. The ZTNB model's superior performance, as shown by its lower AIC, BIC values as well as its higher log-likelihood, indicates a more optimal fit compared to the standard Poisson and NB models. A similar kind of study on the factors influencing the number of children ever born to the reproductive-aged female residing in Andhra Pradesh found ZTP as a better statistical fit model, indicated by higher log-likelihood and lower AIC and BIC values (Muniswamy & Lavanya, 2025). However, in our case, the ZTNB model did not find any significant variables influencing the model. This underscores the importance of using the zero-truncated model to correct the issue and avoid bias in estimating the relationships between the predictors and the outcome, which could have been caused by the regular Poisson or NB models that assume the presence of zeros. The absence of TB treatment dropouts in certain districts emphasizes the pivotal role of the NTEP in facilitating successful TB treatment outcomes. The program has demonstrated flexibility and responsiveness in incorporating global and Indian evidence-based practices into treating and caring for individuals affected by TB. The country has also achieved significant advancements in TB management by enhancing access to treatment and delivering high-quality care for TB patients across all age groups (Central TB Division & Family Welfare, 2017).

The sample size of this study may constrain the statistical strength of the evaluation. Additionally, the presence of binary independent variables (zeros and ones) could potentially introduce multicollinearity concerns in the data. Furthermore, some socio-demographic factors, such as residential area, socioeconomic status, parental education, family support systems, etc., are factors missing beyond those examined in this study that may constitute potential factors for TB treatment dropout among PTB patients. Therefore, further investigation is warranted to enhance the proposed model.

6. Conclusions

This study analysed count data of treatment dropout and determined its associated factors among PTB patients in various districts of India. The regression analysis revealed that the among the models evaluated, the HNB model demonstrated the best fit for data with excess zeros and overdispersion. Conversely, when the data has no zero observations, the ZTNB model was more appropriate than ZTP model. Addressing the identified risk factors can enhance the quality of care delivered to TB patients, particularly among the high-risk population, which requires focused attention and periodic monitoring by healthcare providers. This approach also enables healthcare workers to track patients more effectively, ensuring they remain engaged in treatment. The findings highlight the benifits of count regression models in analysing complex datasets involving multiple variables, offering valuable insights for public health planning. These models can support evidence-based decision-making to optimize resource distribution and strengthen TB control and prevention strategies. We can use mixed-effects or interaction effects models to analyze count data for future research. The statistical approach used in this study could also be applied to other types of count data, such as hospital visits, clinical epidemiology, biometrics, and environmental data.

Acknowledgments

Authors are very thankful to Editor-in-Chief and Co-Editors, and learned reviewers for their suggestions to improve the quality of the manuscript. Shalini Kumari acknowledges research fel-

lowship (UGC-Ref.No.: 200510034092) from UGC, New Delhi. Authors also acknowledge to Department of TB and Respiratory Diseases, Institute of Medical Sciences (IMS), Banaras Hindu University, Varanasi and NTEP, Central TB Division (CTD), Ministry of Health and Family Welfare, Government of India.

Funding

This research did not receive any funding.

Ethics Approval

Ethical approval for this study was granted by the Institutional Ethics Committee of the Institute of Medical Sciences, Banaras Hindu University, Varanasi, India, vide reference number (IMS/IEC/2024/7501).

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

Conceptualization: KUMARI, S., KUMAR, M. **Data curation**: KUMARI, S., MISHRA, J.K., KUMAR, M. **Formal analysis**: KUMARI, S., DAS, S., KUMAR, M. **Investigation**: KUMARI, S., KUMAR, A., KUMAR, M. **Methodology**: KUMARI, S., KUMAR, A., MISHRA, J.K., KUMAR, M. **Software**: KUMARI, S., DAS, S., KUMAR, M. **Resources**: MISHRA, J.K., KUMAR, A., KUMAR, M. **Supervision**: KUMAR, A., KUMAR, M. **Validation**: KUMARI, S., DAS, S., KUMAR, A., MISHRA, J.K., KUMAR, M. **Visualization**: KUMARI, S., DAS, S., KUMAR, A., MISHRA, J.K., KUMAR, M. **Writing-original draft**: KUMARI, S., DAS, S., KUMAR, A., MISHRA, J.K., KUMAR, M. **Writing-review and editing**: KUMARI, S., DAS, S., KUMAR, A., MISHRA, J.K., KUMAR, M.

References

1. Akaike, H., Petrov, B. N. & Csaki, F. *Second international symposium on information theory* 1973.
2. Azevedo, A. M., Silva, Í. J., Nery, M. C., Rocha, H. P. & Santana, R. A. Counting models for overdispersed data: A review with application to tuberculosis data. *Brazilian Journal of Biometrics* **41,** 274–286 (2023).
3. Balakrishnan, M. & Varadharajan, R. Spatial patterns and multilevel analysis of factors associated with paediatric tuberculosis in India. *Indian Journal of Tuberculosis* **72,** S12–S17 (2025).
4. Barreto-Duarte, B. *et al.* Tuberculosis burden and determinants of treatment outcomes according to age in Brazil: a nationwide study of 896,314 cases reported between 2010 and 2019. *Frontiers in Medicine* **8,** 706689 (2021).
5. Basa, S. & Venkatesh, S. Study on default and its factors associated among Tuberculosis patients treated under DOTS in Mayurbhanj District, Odisha. *Journal of Health Research and Reviews (In Developing Countries)* **2,** 25–28 (2015).
6. Bekalo, D. B. & Kebede, D. T. Zero-inflated models for count data: an application to number of antenatal care service visits. *Annals of data science* **8,** 683–708 (2021).
7. Cameron, A. C. & Trivedi, P. K. *Regression analysis of count data* **53** (Cambridge university press, 2013).

8. Central TB Division, M. o. H. & Family Welfare, G. o. I. *National Strategic Plan for Tuberculosis Elimination 2017–2025* https://tbcindia.mohfw.gov.in/wp-content/uploads/2023/05/National-Strategic-Plan-2017-25.pdf. Accessed: 2025-05-01. 2017.
9. Chakrabarti, A. & Ghosh, J. K. AIC, BIC and recent advances in model selection. *Philosophy of statistics,* 583–605 (2011).
10. Cruyff, M. J. & Van Der Heijden, P. G. Point and interval estimation of the population size using a zero-truncated negative binomial regression model. *Biometrical Journal: Journal of Mathematical Methods in Biosciences* **50,** 1035–1050 (2008).
11. CTD, C. T. D. *India TB Report 2024* https://tbcindia.mohfw.gov.in/2024/10/11/india-tb-report-2024/. Accessed: 2025-05-01. 2024.
12. Dwivedi, A. K., Dwivedi, S. N., Deo, S., Shukla, R. & Kopras, E. Statistical models for predicting number of involved nodes in breast cancer patients. *Health* **2,** 641 (2010).
13. Fajri, A., Rahmi, N., Maharani, P. A. & Amal, M. I. Modeling tuberculosis in children under five using poisson and negative binomial regression. *Desimal: Jurnal Matematika* **7,** 311–320 (2024).
14. Greene, W. H. Accounting for excess zeros and sample selection in Poisson and negative binomial regression models (1994).
15. Haight, F. A. Handbook of the Poisson distribution. *(No Title)* (1967).
16. Hall, D. B. & Zhang, Z. Marginal models for zero inflated clustered data. *Statistical modelling* **4,** 161–180 (2004).
17. Hilbe, J. M. *Modeling count data* (Cambridge University Press, 2014).
18. Hutchinson, M. K. & Holtman, M. C. Analysis of count data using poisson regression. *Research in nursing & health* **28,** 408–418 (2005).
19. Karazsia, B. T. & Van Dulmen, M. H. Regression models for count data: illustrations using longitudinal predictors of childhood injury. *Journal of pediatric psychology* **33,** 1076–1084 (2008).
20. Kibria, B. G. Applications of some discrete regression models for count data. *Pakistan Journal of Statistics and Operation Research,* 1–16 (2006).
21. Lambert, D. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* **34,** 1–14 (1992).
22. MacCallum, R. C., Zhang, S., Preacher, K. J. & Rucker, D. D. On the practice of dichotomization of quantitative variables. *Psychological methods* **7,** 19 (2002).
23. Mullahy, J. Specification and testing of some modified count data models. *Journal of econometrics* **33,** 341–365 (1986).
24. Muniswamy, B & Lavanya, M. ZERO TRUNCATED POISSON REGRESSION MODEL FOR REPRODUCTIVE PATTERNS ON COUNT DATA. *Reliability: Theory & Applications* **20,** 1070–1088 (2025).
25. Mwalili, S. M., Lesaffre, E. & Declerck, D. The zero-inflated negative binomial regression model with correction for misclassification: an example in caries research. *Statistical methods in medical research* **17,** 123–139 (2008).
26. Myers, R. H., Montgomery, D. C., Vining, G. G. & Robinson, T. J. *Generalized linear models: with applications in engineering and the sciences* (John Wiley & Sons, 2012).
27. Rose, C. E., Martin, S. W., Wannemuehler, K. A. & Plikaytis, B. D. On the use of zero-inflated and hurdle models for modeling vaccine adverse event count data. *Journal of biopharmaceutical statistics* **16,** 463–481 (2006).

28. Schwarz, G. Estimating the dimension of a model. *The annals of statistics,* 461–464 (1978).

29. Seddon, J. A. *et al.* Counting children with tuberculosis: why numbers matter. *The International Journal of Tuberculosis and Lung Disease* **19,** S9–S16 (2015).

30. Shrestha, N. Detecting multicollinearity in regression analysis. *American journal of applied mathematics and statistics* **8,** 39–42 (2020).

31. Verma, P., Swain, P. K., Singh, K. K. & Khetan, M. Count data regression modeling: an application to spontaneous abortion. *Reproductive Health* **17,** 1–9 (2020).

32. WHO, W. H. O. *Global tuberculosis report 2024* https : / / www . who . int / teams / global - programme-on-tuberculosis-and-lung-health/tb-reports/global-tuberculosis-report-2024/. Accessed: 2025-05-01. 2024.

33. Winkelmann, R. & Zimmermann, K. F. Count data models for demographic data. *Mathematical Population Studies* **4,** 205–221 (1994).

34. Yotenka, R. & Banapon, A. *Modelling the Number of Tuberculosis (TB) Cases in Indonesia using Poisson Regression and Negative Binomial Regression* in *The 2nd International Seminar on Science and Technology (ISSTEC 2019)* (2020), 36–42.

35. Zeileis, A., Kleiber, C. & Jackman, S. Regression models for count data in R. *Journal of statistical software* **27,** 1–25 (2008).