STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **x**, Month 202x, pp 0–12.
Published online in International Academic Press (www.IAPress.org)

Modeling Pulmonary Tuberculosis Case Based on HIV and AIDS Cases in Indonesia Using Negative Binomial Regression Least Square Spline

Arip Ramadan[1], Nur Chamidah[2,3,*], I Nyoman Budiantara[4], Naufal Ramadhan Al Akhwal Siregar[5], Dursun Aydin[6,7]

[1]*Information System Study Program, School of Industrial and System Engineering, Telkom University, Surabaya Campus, Jl. Ketintang No.156, Surabaya 60231, East Java, Indonesia*
[2]*Department of Mathematics, Faculty of Science and Technology, Airlangga University, Surabaya 60115, Indonesia*
[3]*Research Group of Statistical Modeling in Life Science, Faculty of Science and Technology, Airlangga University, Surabaya 60115, Indonesia*
[4]*Statistics Department, Faculty of Science and Data Analytics, Institut Teknologi Sepuluh Nopember, Surabaya 60111, Indonesia*
[5]*Mathematics Master Study Program, Department of Mathematics, Faculty of Science and Technology, Airlangga University, Surabaya 60115, Indonesia*
[6]*Department of Statistics, Faculty of Science, Muğla Sıtkı Koçman University, Muğla 48000, Turkey*
[7]*Research Scholar at Department of Mathematics, University of Wisconsin, Oshkosh Algoma Blvd, Oshkosh, WI 54901, USA*

Abstract The rising incidence of pulmonary tuberculosis (TB) among HIV/AIDS patients in Indonesia poses a significant challenge to public health. However, statistically modeling this co-epidemic is frequently impeded by overdispersion in epidemiological count data, which is a phenomenon where the variance exceeds the mean. Standard parametric models often fail to account for this variability, resulting in biased estimates and reduced predictive validity. To overcome these limitations, this study proposes a Nonparametric Negative Binomial Regression model utilizing a Least Square Spline (NNBR-LSS) estimator as a robust alternative to address these limitations. Applied to the 2023 Indonesian Health Profile data, the model investigates the functional dependence of TB incidence on HIV/AIDS prevalence. The empirical results reveal a strong positive correlation, supported by a Pseudo R-square of 64.8%, confirming that HIV/AIDS case numbers are a critical predictor of TB distribution. Furthermore, the NNBR-LSS model demonstrated superior performance over standard parametric regression, as indicated by a significant reduction in the deviance statistic from 40.921 to 19.525. These findings establish the NNBR-LSS as a powerful methodological tool for capturing nonlinear patterns in overdispersed data, offering more precise insights for strategic public health interventions. It recommended the findings for epidemiological intervention planning and aligning efforts with Sustainable Development Goals (SDGs) 3.3, which aims to end the TB epidemic by 2030.

Keywords Pulmonary Tuberculosis, HIV AIDS, Negative Binomial Regression, Least Square Spline, Epidemiological Modeling

DOI: 10.19139/soic-2310-5070-3152

1. Introduction

Pulmonary tuberculosis (TB) remains one of the most persistent global health challenges, disproportionately affecting developing nations with limited sanitation and healthcare access. According to the Indonesian Ministry of Health and the National Guidelines for Tuberculosis Control, pulmonary TB specifically infects lung parenchyma, distinct from extra-pulmonary manifestations [1, 2]. While latent TB infection can remain dormant without symptoms [3, 4], immunosuppression can trigger progression to active disease, manifesting in chronic symptoms and high mortality. This progression is critically exacerbated by HIV co-infection. Individuals living with

*Correspondence to: Nur Chamidah (Email: nur-c@fst.unair.ac.id). Department of Mathematics, Faculty of Science and Technology, Airlangga University, Surabaya 60115, Indonesia

ISSN 2310-5070 (online) ISSN 2311-004X (print)

HIV/AIDS face an 18-fold increased risk of developing active TB due to the depletion of CD4 T-helper cells, which compromises the immune system's defense mechanisms [5, 7]. As reported in the 2024 WHO TB Reports, Indonesia ranks second globally in TB burden, contributing 10% of total cases, underscoring the urgency of integrated management strategies for this co-epidemic [6, 8].

Despite the clinical understanding of TB-HIV co-infection, statistical modeling of this relationship faces significant methodological challenges. Previous studies, such as the case study by Rewata et al. [9] in Bandung, highlighted the difficulty in identifying clear patterns between TB and HIV using standard descriptive approaches. A fundamental research problem in epidemiological modeling is that count data, such as TB case numbers, frequently exhibit *overdispersion*, where the variance exceeds the mean [15]. This phenomenon violates the equidispersion assumption of standard Poisson regression, leading to biased parameter estimates and underestimated standard errors.

To address overdispersion, Negative Binomial Regression (NBR) has been widely adopted. Kusuma et al. [11] demonstrated that NBR outperformed Poisson regression in modeling TB cases in NTB, yielding a significantly lower deviance and AIC. However, parametric NBR assumes a rigid functional form (e.g., linear or log-linear) between the response and predictors. As noted by Darma et al. [10], real-world health data often follow complex, nonlinear patterns that parametric models fail to capture. In such cases, nonparametric approaches offer superior flexibility. Nonparametric regression allows the data to determine the shape of the regression curve without pre-specified assumptions.

Recent advancements have integrated NBR with nonparametric estimators to simultaneously address overdispersion and nonlinearity. Tohari et al. [12, 13, 14] successfully applied Nonparametric Negative Binomial Regression (NNBR) using local linear estimators to model HIV/AIDS dynamics in East Java. Furthermore, the use of spline estimators, particularly the Least Square Spline (LSS), has gained attention due to its ability to accommodate local data variations and optimize smoothness through the selection of knot points [22]. This approach enables the model to capture abrupt local variations and structural changes (i.e., critical thresholds at which HIV prevalence significantly alters TB incidence) while maintaining an explicit and interpretable mathematical formulation. Ramadan et al. [23, 24] further extended this approach by employing the NNBR-LSS estimator to model HIV cases across 30 provinces. However, the specific application of NNBR-LSS to model the functional dependence of pulmonary TB incidence on HIV/AIDS prevalence remains underexplored.

The performance of the NNBR-LSS model relies crucially on identifying optimal knot points, specific locations in the predictor space where the data's behavioral trend shifts. In this study, knot locations are systematically determined through a grid search procedure across the data distribution, rather than relying on arbitrary placement. To select the most optimal knot locations and combinations, we employ the Maximum Likelihood Cross-Validation (MLCV) criterion. The MLCV is calculated by evaluating the cross-validated log-likelihood of the Negative Binomial distribution while penalizing for model complexity (the number of knots). Consequently, the particular knot combinations chosen for the final model are strictly those that yield the minimum MLCV value, ensuring the optimal balance between goodness-of-fit and curve smoothness to prevent overfitting.

Therefore, the primary objective of this study is to model the number of pulmonary TB cases based on HIV/AIDS cases in Indonesia using the Nonparametric Negative Binomial Regression with Least Square Spline (NNBR-LSS) estimator, optimized via the MLCV criterion. This study specifically aims to overcome the limitations of parametric models in handling overdispersed and nonlinear data and provide a more accurate empirical basis for epidemiological interventions. By utilizing the 2023 Indonesian Health Profile data, this research offers a novel contribution to biostatistical modeling, facilitating more targeted public health strategies for the TB-HIV co-epidemic. It recommended the findings for epidemiological intervention planning and aligning efforts with Sustainable Development Goals (SDGs) 3.3, which aims to end the TB epidemic by 2030.

2. Research Methods

2.1. Data Sources

The data analyzed in this research comes from the Indonesian Ministry of Health's 2023 health profile. It encompasses the number of confirmed pulmonary TB cases, as well as HIV and AIDS cases, observed across 38 Indonesian provinces.

2.2. Research Variables

The research variables used in this study include the number of pulmonary tuberculosis cases (y), the number of HIV cases (x_1), and the number of AIDS cases (x_2). The response variable, y, represents the number of bacteriologically confirmed pulmonary tuberculosis cases and is classified as discrete data. The predictor variables include x_1, which indicates the number of confirmed HIV-positive cases, and x_2, representing the number of confirmed AIDS-positive cases, both of which are also classified as discrete data. These variables were chosen to analyze the relationship between HIV/AIDS cases and the incidence of pulmonary tuberculosis, highlighting the epidemiological connection and addressing overdispersion in count data.

2.3. Research Steps

The steps taken in this study for modeling pulmonary tuberculosis using the LSS estimator are as follows:

1. Inputting paired data (x_{1i}, x_{2i}, y_i); $i = 1, 2, \ldots, n$.
2. Describe the pulmonary tuberculosis variable using descriptive statistics.
3. Test the overdispersion indication of the pulmonary tuberculosis response variable using descriptive statistics.
4. Determine smoothing parameters, including the number of knot points and the location of knot points with quantiles based on MLCV criteria.
5. Estimate the regression function based on the NNBR-LSS estimator using smoothing parameters obtained from step 4.
6. Plot the estimated result $\hat{y}$ and observed values y.
7. Calculate deviance and pseudo-R^2 values to test model fit.

Deviance formula can be presented in equation

$$D = 2\sum_{i=1}^{n} \left\{ \ell(y_i; y_i) - \ell(\mu_i; y_i) \right\}, \tag{1}$$

with, $l(y_i; y_i)$ is log-likelihood function where μ is the value of mean of itself. While $l(\mu_i; y_i)$ is log-likelihood function for model estimated [9]. Deviance for NNBR-LSS can be presented as follows: . Deviance for NNBR-LSS can be presented as follows:

$$D_{NB} = 2\sum_{i=1}^{n} \left\{ y_i \ln\left(\frac{y_i}{\mu_i}\right) - \left(\frac{1}{\alpha} + y_i\right) \ln\left(\frac{1+\alpha y_i}{1+\alpha\mu_i}\right) \right\} \tag{2}$$

After that, the pseudo-R^2 formula is presented in the equation

$$R_p^2 = 1 - \frac{L_F}{L_{\text{null}}} \tag{3}$$

where L_F is the log-likelihood of the full model and L_{null} is the log-likelihood of the intercept-only model [9].

3. Results and Discussion

An overview of pulmonary TB and HIV/AIDS cases in Indonesia, including two factors thought to influence them, can be found in Table 1

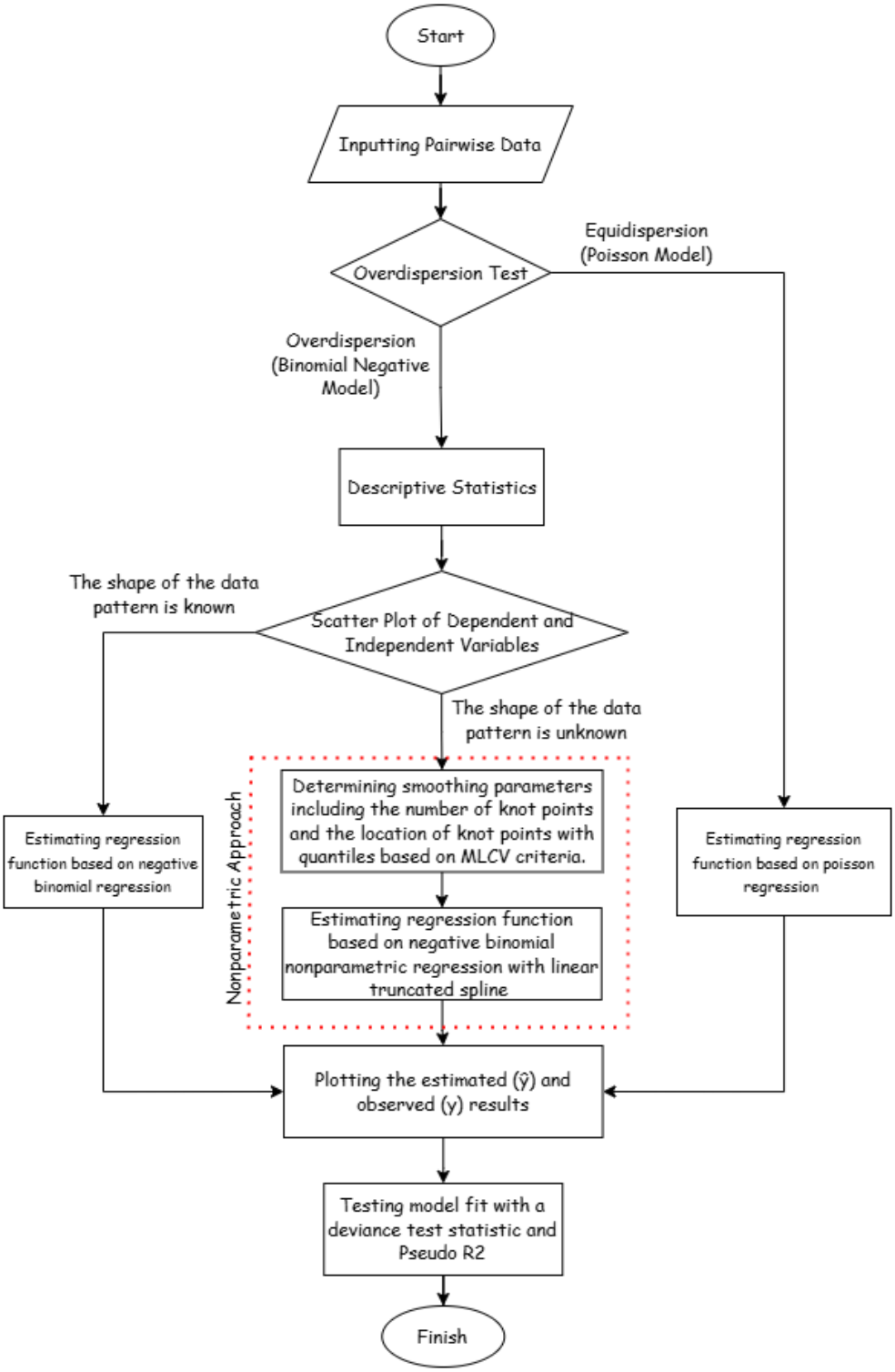

Figure 1. Research Steps of Analysis.

Table 1. Characteristics of the Number of Pulmonary Tuberculosis (TB) Cases and the Number of HIV and AIDS Cases in Indonesia

Variable	Mean	Variance	Minimum	Maximum
Y	10,618.82	264,369,586.6	383	83,252
X_1	1,507.87	4,571,405	75	9,500
X_2	431.84	437,673.4	16	2,575

Based on Table 1, it can be shown that the characteristics of the variable number of cases of pulmonary tuberculosis (Y) indicate that the average number of cases of pulmonary tuberculosis in Indonesia is 10,618.82 with a variance value of 264,369,586.6. The lowest number of cases of pulmonary tuberculosis is in Papua Province with 383 cases, and the highest is in West Java Province with 83,252 cases. The difference between the two regions is 82,869, indicating that the number of pulmonary tuberculosis (TB) cases is influenced by the different conditions of each region.

Furthermore, the characteristics of the variable number of HIV cases (X_1) show that the average number of HIV cases in Indonesia is 1,507.87 with a variance value of 4,571,405. The lowest number of HIV cases is in West Sulawesi Province and the highest is in East Java Province. The difference between the two regions is 9,425. Then, the characteristics of the variable number of AIDS cases (X_2) show that the average number of AIDS cases in Indonesia is 431.84 with a variance value of 437,673.4. The lowest number of AIDS cases is in South Papua Province and the highest is in West Java Province. The difference between the two regions is 2,559. Based on the Indonesian health profile data in 2023, the distribution of the number of pulmonary TB, HIV, and AIDS cases in Indonesia can be presented in Figure 2.

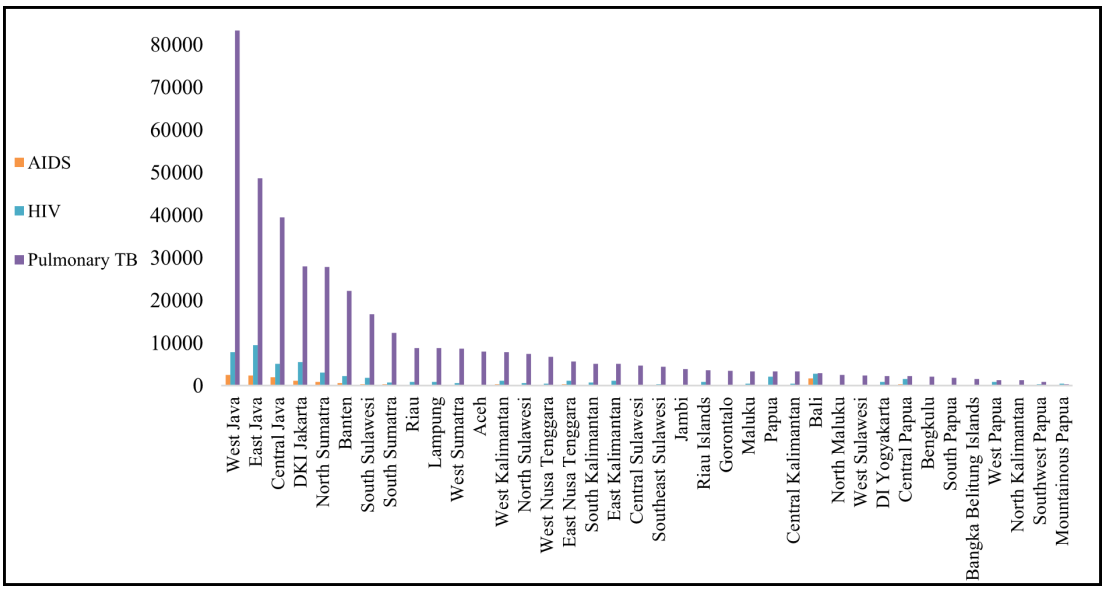

Figure 2. The Distribution of The Number of Pulmonary TB, HIV, and AIDS In Indonesia.

The first step taken before analyzing the data is by plotting the number of pulmonary tuberculosis patients with each predictor variable if other predictor variables are considered constant. The scatter plot results are presented in Figure 3. From the scatter plot in Figure 3 (a), it can be seen that the distribution of the dots shows a pattern that may illustrate a relationship between the two variables. If most of the points show an upward trend (like an upward sloping line), this indicates a positive relationship. This suggests that the more HIV cases, the higher the number of pulmonary tuberculosis cases. Furthermore, based on Figure 3 (b) for the scatter plot of the relationship of the number of AIDS cases to pulmonary tuberculosis, the distribution of points is more concentrated at a certain

value, although the pattern of the relationship is similar to the first graph. If the pattern of dots also shows a positive relationship, it can be said that the increase in AIDS cases tends to go hand in hand with the increase in pulmonary tuberculosis cases. Based on Figure 3 (a) and (b), it can be seen from each plot of pulmonary tuberculosis cases with their respective predictor variables that the distribution is irregular and does not form a certain pattern. Therefore, the data on the number of pulmonary tuberculosis cases in Indonesia can be estimated by nonparametric regression.

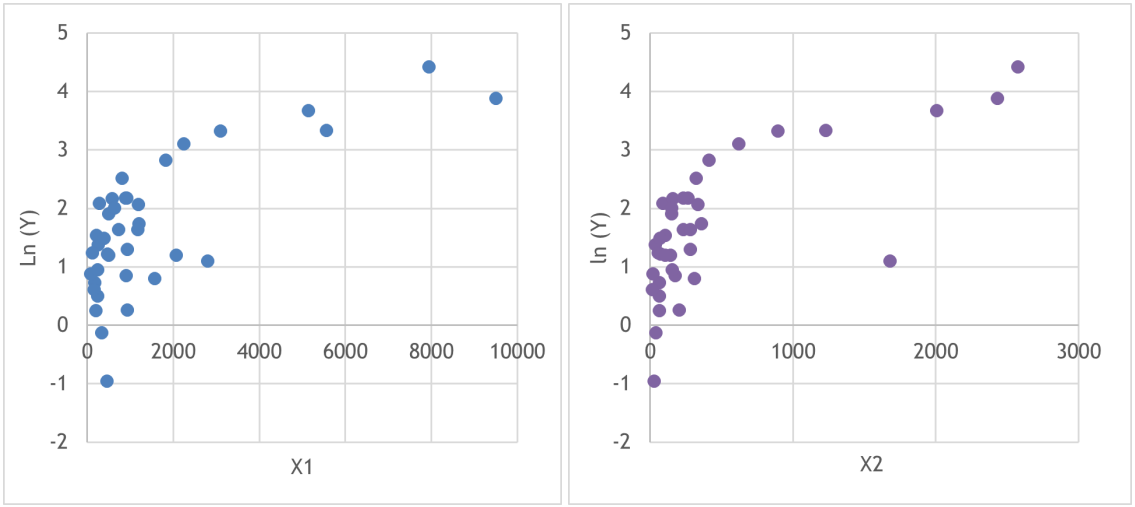

Figure 3. Scatter plot of the data: (a) The number of HIV cases (X_1), (b) The number of AIDS cases (X_2).

Modeling with the NBR method uses a response variable in the form of count data and negative binomial distribution. Before modeling the number of pulmonary TB cases, overdispersion testing of the Poisson regression model is first carried out. Detection of overdispersion cases can be done by examining the mean and variance of the response variable. Next, the results of the dispersion test on the pulmonary TB data are presented in Table 2.

Table 2. Dispersion Test for Pulmonary TB Data

Test	Value
Dispersion Ratio (α)	3398.76
Z-test Statistics	4.0616
P-Value	0.000
Overdispersion Detected?	Yes

Based on Table 2, the Poisson regression model shows that there is a case of overdispersion by observing that the ratio between the mean value and the variance value of the response y (the number of pulmonary TB patients) is greater than 1. This indicates the presence of overdispersion in the Poisson model.

The occurrence of overdispersion can also be detected using the Z-test (T_z) statistical test. The hypothesis used to test the negative binomial distribution for the response variable of the number of AIDS patients is as follows:
$H_0 : \alpha = 1$ (Equidispersion case)
$H_1 : \alpha > 1$ (Overdispersion case occurs).
The p-value obtained is 0.000. This value is compared with the significance level $\alpha = 5\%$, so the decision can be made to reject H_0. The conclusion is that overdispersion exists in the data on the number of pulmonary TB patients. These test results indicate that the data of pulmonary TB patients in Indonesia are more suitably analyzed and modeled using Negative Binomial Regression (NBR).

The number of pulmonary TB patients in Indonesia will be analyzed with both predictor variables to determine the optimum smoothing parameters (number of knots and location of knot points) based on the Maximum Likelihood Cross Validation (MLCV) criteria. Based on the program to determine smoothing parameters, the

process of estimating the number of pulmonary TB patients with both predictor variables obtained optimum smoothing parameters which has been presented in Tables 3–5.

Table 3. Selection of Number of Knots with 1 Knot Point Combination

Combination with 1 Knot				MLCV
Predictor 1	Predictor 2			
	295			-6109.711
	768			-12780.41
	1481.5			-9474.248
70.5	295	768		-7980.273
	295	1481.5		-9918.661
	768	1481.5		-11931
	295	768	1481.5	-9925.619
	295			-17759.87
	768			-12236.6
	1481.5			-13508.66
165	295	768		-13474.19
	295	1481.5		-9255.633
	768	1481.5		-21738.44
	295	768	1481.5	-10110.88
	295			-27393.93
	768			-16783.19
	1481.5			-13126.22
329.75	295	768		-26144.23
	295	1481.5		-21131.93
	768	1481.5		-15480.51
	295	768	1481.5	-15565.44

After determining the number of knots and their respective locations for the predictor variables in Table 3 , the next step involves exploring combinations of additional knot points to refine the model's smoothing parameters. Table 4 presents the selection of combinations with two knot points for each predictor variable. These combinations aim to further optimize the regression model, improving the fit by accounting for more nuanced changes in the data patterns. By adjusting the number and positioning of the knot points, the model becomes more flexible and better suited to capture the variations in the data, ensuring a more accurate representation of the relationship between the number of HIV/AIDS cases and pulmonary TB incidence in Indonesia.

Following the exploration of two-knot combinations in Table 4, Table 5 presents the results for combinations involving three knot points for each predictor variable. The inclusion of three knots allows for even greater flexibility in capturing the complexities of the data. By introducing an additional knot, the model can account for more intricate variations and non-linear relationships between the variables.

This step aims to further enhance the model's accuracy, ensuring that the smooth curves better reflect the underlying trends in the relationship between HIV/AIDS cases and pulmonary TB incidence. The comparison of the MLCV values across different combinations of knots will guide the selection of the optimal smoothing parameters for the regression model.

Determining the optimum smoothing parameters on the predictor variables involved is by looking at the maximum MLCV value.

Based on Table 7, it can be seen that from the combination of knots in the first predictor variable and the second predictor variable, the maximum MLCV value is -3627.227. The maximum MLCV value is obtained when the number of knots in the first predictor variable is 2 knots with knot locations at points 165 and 329.75, and the number of knots in the second predictor variable is 3 knots with locations at points 295, 768, and 1481.5. Based on the data in Table 3–5, we derived parameter estimates for both the parametric and least squares spline nonparametric

Table 4. Selection of Number of Knots with 2 Knot Point Combination

Combination with 2 Knot					MLCV
Predictor 1		Predictor 2			
		295			-22487.85
		768			-18301.18
		1481.5			-34732.42
70.5	165	295	768		-10766.09
		295	1481.5		-20551.84
		768	1481.5		-9884.305
		295	768	1481.5	-11133.34
		295			-33343.53
		768			-22758.37
		1481.5			-30540.01
70.5	329.75	295	768		-24502.27
		295	1481.5		-18385.44
		768	1481.5		-18016.54
		295	768	1481.5	-22054.70
		295			-45252.81
		768			-38034.79
		1481.5			-19301.83
165	**329.75**	295	768		-10631.85
		295	1481.5		-21563.35
		768	1481.5		-16615.81
		295	**768**	**1481.5**	**-3627.227**

Table 5. Selection of Number of Knots with 3 Knot Point Combination

Combination with 3 Knot						MLCV
Predictor 1			Predictor 2			
			295			-45149.09
			768			-21997.78
			1481.5			-25705.11
70.5	**165**	**329.75**	295	768		-12439.84
			295	1481.5		-19272.53
			768	**1481.5**		**-10454.29**
			295	768	1481.5	-27480.35

Table 6. Summary of optimal knot selection across different numbers of knot points

Combination Scenario	**Optimal Knots for Predictor 1**	**Optimal Knots for Predictor 2**	**Optimal MLCV**
1 Knot Point Combination	70.5	295	-6109.711
2 Knot Point Combination	165; 329.75	295; 768; 1481.5	**-3627.227***
3 Knot Point Combination	70.5; -10454.29	768; 1481.5	-10454.29

regression models. These results can be found in Table 7. The estimation equation of the NNBR model based on

Table 7. Parameter estimation results for NBR and NNBR-LSS model

NBR Estimation		NNBR-LSS Estimation	
Parameters	Value	Parameters	Value
$\hat{\beta}_0$	8.241517	$\hat{\beta}_0$	8.847461
$\hat{\beta}_1$	0.0004195	$\hat{\beta}_1$	0.0009138
$\hat{\beta}_2$	0.0002646	$\hat{\beta}_{1,1}$	-0.000359
$\hat{\alpha}$	2.123528	$\hat{\beta}_{1,2}$	-0.0001899
		$\hat{\beta}_2$	-0.0039403
		$\hat{\beta}_{2,1}$	0.0061628
		$\hat{\beta}_{2,2}$	-0.0020813
		$\hat{\beta}_{2,3}$	0.0000437
		$\hat{\alpha}$	0.4804481

the truncated spline estimator on the data of the number of pulmonary TB cases in Indonesia is as follows:

$$\hat{y} = \exp\Big(8.847641 + 0.0009138\,x_1 - 0.000359(x_1 - 165) - 0.0001899(x_1 - 329.75) - 0.0039403\,x_2 + 0.0061628(x_2 - 295) - 0.0020813(x_2 - 768) + 0.0000437(x_2 - 1481.5)\Big) \tag{4}$$

If the model of the number of HIV cases (x_1) is known against the number of pulmonary TB cases by holding the other predictor variables constant, we obtain the following equation:

$$\hat{s}(x_1) = 8.847641 + 0.0009138\,x_1 - 0.000359(x_1 - 165)_+ - 0.0001899(x_1 - 329.75)_+ \tag{5}$$

where,

$$\hat{s}(x_1) = \begin{cases} 8.847641 + 0.0009138\,x_1, & x_1 \leq 165 \\ 8.906876 + 0.0005548\,x_1, & 165 < x_1 \leq 329.75 \\ 8.9694955 + 0.0003649\,x_1, & x_1 > 329.75 \end{cases} \tag{6}$$

If the model of the number of AIDS cases (x_2) is known against the number of pulmonary TB cases by holding the other predictor variables constant, we obtain the following equation:

$$\hat{s}(x_2) = 8.847641 - 0.0039403\,x_2 + 0.0061628\,(x_2 - 295)_+ - 0.0020813\,(x_2 - 768)_+ + 0.0000437\,(x_2 - 1481.5)_+ \tag{7}$$

$$\hat{s}(x_2) = \begin{cases} 8.847641 - 0.0039403\,x_2, & x_2 \leq 295 \\ 7.029615 + 0.0022225\,x_2, & 295 < x_2 \leq 768 \\ 8.6280534 + 0.0001412\,x_2, & 768 < x_2 \leq 1481.5 \\ 8.5633118 + 0.00005782\,x_2, & x_2 > 1481.5 \end{cases} \tag{8}$$

Based on the results of the NNBR-LSS estimator according to Equation (1), referring to the polynomial cuts in Equation (3) and Equation (5), if initially it is known that the average number of HIV cases is 100 people and AIDS cases is 500 people, then the regression model is as follows:

$$\hat{y} = \exp\big(8.847641 + 0.0009138\,x_1 + 7.029615 + 0.0022225\,x_2\big) \tag{9}$$

In a certain region, if the number of HIV patients is less than 165, every increase of 100 HIV patients will result in a 1.096 times increase in pulmonary TB cases, assuming other variables remain constant. This is calculated using the exponential function $exp(0.0009138 \times 100)$. In another region, where the number of AIDS patients ranges between 295 and 768, every increase of 100 AIDS patients will lead to a 1.25 times increase in pulmonary TB cases, assuming other variables remain constant. This is derived from the exponential function $\exp(0.0022225 \times 100)$.

Building upon the findings detailed in Table 7 and as informed by Equation (1), we further refined our analysis to obtain parameter estimates for both the parametric regression model and the least squares spline nonparametric regression model. These estimates were then rigorously compared against the observed, real-world pulmonary TB case data from various provinces within Indonesia. Table 8 provides a comprehensive presentation of these comparative results. Furthermore, a visual representation of the relationship between our model-derived estimations and the actual observed number of pulmonary TB patients in Indonesia is illustrated in Figure 4, offering a clear depiction of the model's performance.

Table 8. Estimation Results of NBR and NNBR-LSS Models for the Pulmonary TB Cases in Provinces of Indonesia

No	Provinces	Actual Data	NBR Estimation	NNBR-LSS Estimation
1	Aceh	8090	4239.789	2511.241
2	North Sumatra	27886	12554.66	14544.24
3	West Sumatra	8700	4722.945	4716.528
4	Riau	8855	5397.708	7801.699
5	Jambi	3957	4128.59	2570.718
6	South Sumatra	12415	5382.52	7938.316
7	Bengkulu	2069	4081.025	3679.838
8	Lampung	8853	5295.055	7645.89
9	Bangka Belitung Islands	1644	4159.532	2770.888
10	Riau Islands	3656	5463.989	7902.202
11	DKI Jakarta	28025	27746.94	25929.8
12	West Java	83252	91714.45	65842.62
13	Central Java	39503	34360.63	31837.1
14	DI Yogyakarta	2335	5187.528	7416.419
15	East Java	48701	130073	83192.19
16	Banten	22242	8938.811	11250.54
17	Bali	2998	16107.03	18315.12
18	West Nusa Tenggara	6753	4614.614	3963.279
19	East Nusa Tenggara	5711	6060.422	8518.204
20	West Kalimantan	7872	5970.275	8420.629
21	Central Kalimantan	3315	4590.848	3854.971
22	South Kalimantan	5165	5067.474	6800.985
23	East Kalimantan	5127	5818.045	8163.311
24	North Kalimantan	1286	4109.388	3367.16
25	North Sulawesi	7415	4773.609	5267.508
26	Central Sulawesi	4673	4200.641	3256.36
27	South Sulawesi	16828	7311.585	9631.381
28	Southeast Sulawesi	4428	4332.191	2858.232
29	Gorontalo	3444	4011.763	4483.723
30	West Sulawesi	2420	3903.995	5272.26
31	Maluku	3370	4426.258	3405.492
32	North Maluku	2577	4318.154	3088.83
33	Papua	3330	6860.918	8440.272
34	West Papua	1298	5292.296	7570.759
35	South Papua	1837	3985.06	3772.807
36	Central Papua	2219	6557.57	8827.228
37	Mountainous Papua	383	4332.677	3192.277
38	Southwest Papua	883	4226.249	2510.677

Based on Figure 4, it can be observed that the distance between the actual observation data and the estimation results is not large; at some points, the estimation results correspond closely to the actual data. This indicates a good match between the observed data and the estimated values for the number of pulmonary TB cases in Indonesia. The model fit criteria for both the parametric and NNBR-LSS approaches, based on the estimation output, are presented in Table 9.

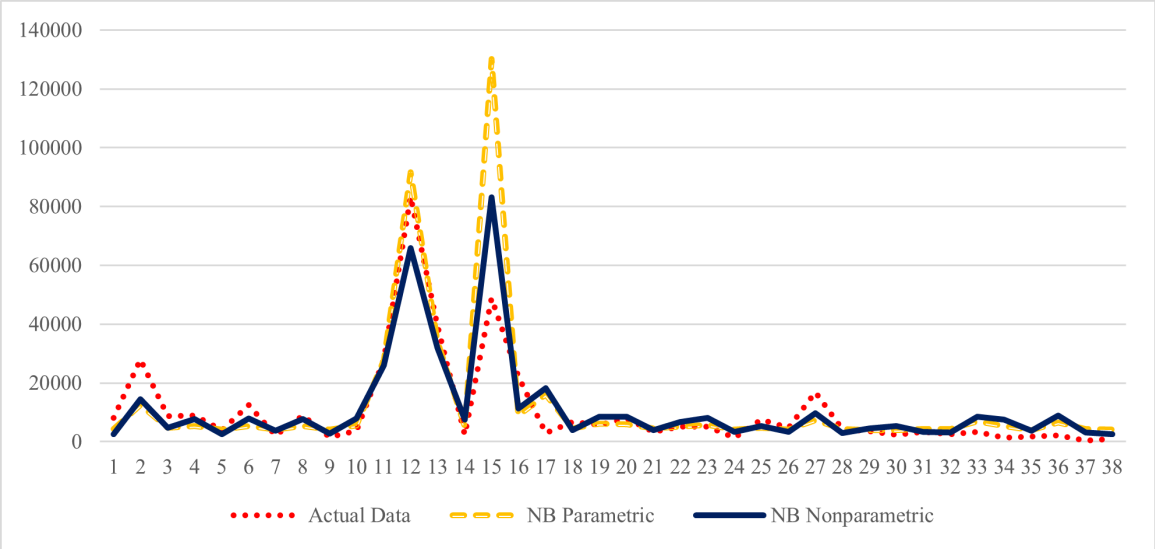

Figure 4. Plot of Observations and Estimation Results

Table 9. Comparison of Goodness of Fit Criteria Between Poisson, NBR, and NNBR-LSS

Goodness of Fit (GoF) Criteria	Poisson Model	Negative Binomial	
		Parametric	LSS
Deviance	131.331	40.920	19.525
Pseudo R-Square	0.769	0.639	0.648

Based on Table 9, it can be seen, if when compared to the Poisson model, the deviance value of 131.331 and the Pseudo R-square value are higher than other models. This indicates that there is a case of overdispersion in the Poisson model so that if it is used to estimate it will cause biased results. This is supported by chi-squared testing with p-value of $0.000 < \alpha$ (5%). It can be concluded that poisson model is not suitable for use as a model. So that, the negative binomial model is more suitable for use as a model.

Furthermore, from the result on Table 9, NNBR-LSS provides better results than the parametric regression approach. This is evidenced by deviance test statistical value in the NNBR-LSS model is 19.525 and the p-value with α of 5%. Because the p-value $> \alpha$, so it is concluded that the nonparametric binomial regression model is appropriate. The deviance test statistic value of NBR with a nonparametric regression approach is 19.525, this value is smaller than the deviance test statistic value with a parametric regression approach which is 40.920. So it can be concluded that NNBR-LSS provides better results than the parametric regression approach.

4. Limitations of the Study

While the Nonparametric Negative Binomial Regression with Least Square Spline (NNBR-LSS) approach successfully addresses overdispersion and captures the complex nonlinear relationship between HIV/AIDS and pulmonary TB, several methodological and data-related limitations must be acknowledged.

First, the analysis is limited by the omission of key covariates. Tuberculosis is a complex, multifactorial disease heavily influenced by socio-demographic and environmental determinants. Important variables such as poverty rates, population density, nutritional status, sanitation quality, and healthcare access metrics (e.g., diagnostic coverage and distance to healthcare facilities) were not included in the model. The exclusion of these covariates may result in omitted variable bias, preventing a fully comprehensive understanding of TB transmission dynamics.

Second, the dataset utilizes ecological data aggregated at the provincial level, resulting in a relatively small sample size ($n = 38$ provinces). A sample size of this magnitude can constrain the statistical power of the model and limit the generalizability of the findings to more granular administrative levels, such as municipalities or regencies. Furthermore, while the MLCV criterion mitigates the risk of overfitting in nonparametric modeling, small sample sizes remain inherently more sensitive to extreme local data variations.

Finally, regarding model robustness, although the knot locations were systematically selected by evaluating the global minimum MLCV criterion across a grid, the study lacks a dedicated sensitivity analysis. Future research should implement rigorous sensitivity testing—such as bootstrapping, alternative cross-validation algorithms, or comparisons with other penalized spline methods (e.g., P-splines)—to definitively confirm the stability of the chosen knot locations and the overall predictive robustness of the model.

5. Conclusion

This study analyzing a significant relationship between HIV/AIDS cases and the incidence of pulmonary TB. The NBR model was identified as an appropriate approach for handling overdispersion in the data, characterized by a variance greater than the mean, as opposed to the Poisson model. The model estimation results reveal that an increase in HIV/AIDS cases contributes to a rise in pulmonary TB cases. With a pseudo R-square value of 64.8%, the model explains a substantial portion of the variability in the data, indicating a strong epidemiological linkage between these diseases. Furthermore, the model evaluation using the deviance test highlights that the NNBR-LSS outperforms the parametric regression approach. A deviance value of 19.525 for the nonparametric model is notably smaller than the 40.92087 deviance value for the parametric approach, underscoring the advantages of the nonparametric method in this analysis. Overall, these findings emphasize the importance of using appropriate statistical methodologies, such as the NNBR-LSS to better understand the complex epidemiological interactions between HIV/AIDS and pulmonary TB. This serves as a foundation for evidence-based public health policy formulation.

REFERENCES

1. W. H. Organization, *Evidence and research gaps identified during development of policy guidelines for tuberculosis*, World Health Organization, 2024.
2. M. Wang, *Infectious Disease and Deep Mycosis*, in Textbook of Pathologic Anatomy: For Medical Students, Springer, 2024, pp. 495–532.
3. M. H. Arsyad, I. Syafina, H. Hapsah, and H. Hervina, *Knowing and Understanding the Tuberculosis (Tb) Disease of the Lung (Literature Review)*, International Journal of Natural Science Studies and Development (IJOSS), vol. 1, no. 2, pp. 56–85, 2024.
4. H. M. Swalehe and E. I. Obeagu, *Tuberculosis: Current diagnosis and management*, Elite Journal of Public Health, vol. 2, no. 1, pp. 23–33, 2024.
5. W. H. Organization, *Global technical consultation report on proposed terminology for pathogens that transmit through the air*, World Health Organization, 2024.
6. W. H. Organization, *Health expenditure by type of service and provider*, 2024.
7. J. Harper *et al.*, *Progress Note 2024: Curing HIV; Not in My Lifetime or Just Around the Corner?*, Pathogens and Immunity, vol. 8, no. 2, p. 115, 2024.
8. K. Bi, Y. Zhao, S. Lu, K. Xu, and Y. Zhang, *Comment on WHO Global Tuberculosis Report 2024: Progress, Challenges and Innovations for Achieving End TB Strategy by 2035*, Lippincott Williams & Wilkins (LWW), 2023.
9. P. Cummings, *Negative Binomial Regression*, in Analysis of Incidence Rates, Chapman and Hall/CRC, 2019, pp. 271–292.
10. I. Darma, M. Ratna, and I. N. Budiantara, *Pemodelan Angka Kasus Tuberkulosis di Surabaya menggunakan Pendekatan Regresi Nonparametrik Spline Truncated*, Jurnal Sains dan Seni ITS, vol. 8, no. 2, pp. 2337–3530, 2019.
11. W. Kusuma, C. F. Utomo, S. Tervia, and R. N. S. Setiawan, *Pemodelan Kasus Tuberkulosis (Tb) di Nusa Tenggara Barat Menggunakan Model Regresi Binomial Negatif*, Jurnal Serunai Matematika, vol. 14, no. 2, pp. 142–147, 2022.

12. A. Tohari and N. Chamidah, *Modeling of HIV and AIDS in Indonesia Using Bivariate Negative Binomial Regression*, in IOP Conference Series: Materials Science and Engineering, IOP Publishing, 2019, p. 052079.
13. A. Tohari, N. Chamidah, and Fatmawati, *Estimating model of the number of HIV and AIDS cases in East Java using Bi-response negative binomial regression based on local linear estimator*, AIP Conference Proceedings, vol. 2264, no. 2, pp. 215–219, 2020, doi: 10.1063/5.0023451.
14. A. Tohari, N. Chamidah, and F. Fatmawati, *Modelling of HIV and AIDS cases in Indonesia using bi-response negative binomial regression approach based on local linear estimator*, Annals of Biology, vol. 36, no. 2, pp. 215–219, 2020.
15. R. Yotenka and A. Banapon, *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), Atlantis Press, 2020, pp. 36–42.
16. A. J. Syafiqoh, R. Mahardika, S. Amaria, E. Winaryati, and M. Al Haris, *Pemodelan Regresi Binomial Negatif untuk Mengevaluasi Faktor-faktor yang Mempengaruhi Kasus Tuberkulosis di Provinsi Jawa Barat*, Jurnal MSA (Matematika dan Statistika serta Aplikasinya), vol. 12, no. 1, pp. 15–23, 2024.
17. E. H. Payne, M. Gebregziabher, J. W. Hardin, V. Ramakrishnan, and L. E. Egede, *An empirical approach to determine a threshold for assessing overdispersion in Poisson and negative binomial models for count data*, Communications in Statistics—Simulation and Computation, vol. 47, no. 6, pp. 1722–1738, 2018.
18. W. Gardner, E. P. Mulvey, and E. C. Shaw, *Regression analyses of counts and rates: Poisson, overdispersed Poisson, and negative binomial models*, Psychological Bulletin, vol. 118, no. 3, p. 392, 1995.
19. N. Chamidah and B. Lestari, *Estimation of covariance matrix using multi-response local polynomial estimator for designing children growth charts: A theoretical discussion*, in Journal of Physics: Conference Series, IOP Publishing, 2019, p. 12072, doi: 10.1088/1742-6596/1397/1/012072.
20. N. Chamidah, B. Lestari, T. Saifudin, R. Rulaningtyas, P. Wardhani, and I. N. Budiantara, *Estimating the number of malaria parasites on blood smears microscopic images using penalized spline nonparametric Poisson regression*, Communications in Mathematical Biology and Neuroscience, vol. 2024, p. Article-ID, 2024.
21. E. Tjahjono, M. F. F. Mardianto, and N. Chamidah, *Prediction of electricity consumption using Fourier series estimator in bi-response nonparametric regression model*, Far East Journal of Mathematical Sciences, vol. 103, pp. 1251–1263, 2018.
22. A. Islamiyati, A. Kalondeng, N. Sunusi, M. Zakir, and A. K. Amir, *Biresponse nonparametric regression model in principal component analysis with truncated spline estimator*, Journal of King Saud University—Science, vol. 34, no. 3, p. 101892, 2022.
23. A. Ramadan, N. Chamidah, and I. N. Budiantara, *Modelling the number of HIV cases in Indonesia using negative binomial regression based on least square spline estimator*, Communications in Mathematical Biology and Neuroscience, vol. 2024, p. Article-ID, 2024.
24. A. Ramadan, N. Chamidah, I. N. Budiantara, B. Lestari, and D. Aydin, *Method for Modelling the Number of HIV and AIDS Cases Using Least Square Spline Biresponse Nonparametric Negative Binomial Regression*, MethodsX, p. 103336, 2025.