



# Negative Binomial Modeling of District-Level Diarrhea Incidence in West Java Province

Nurul Nabilah<sup>1</sup> and A'yunin Sofro<sup>2\*</sup>

<sup>1</sup>*Department of Mathematics, Faculty of Mathematics and Natural Sciences, State University of Surabaya, Surabaya, Indonesia*

<sup>2</sup>*Department of Actuarial Science, Faculty of Mathematics and Natural Sciences, State University of Surabaya, Surabaya, Indonesia*

## Abstract

This study investigated district-level diarrhea incidence in West Java Province during 2022–2024 and identified an appropriate count-data regression model for longitudinal epidemiological data. The response variable was the annual number of diarrhea cases, while explanatory variables included sanitation access, inpatient health centers, community health workers, nutritional indicators, drinking water facilities, and households implementing clean and healthy living behavior (PHBS). A population offset term was incorporated to account for differences in population exposure across districts/cities. Poisson Regression was initially fitted as a baseline model; however, severe overdispersion was detected. Therefore, Negative Binomial Regression and Generalized Linear Mixed Models (GLMM) were estimated and compared using Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), log-likelihood values, and simulation-based residual diagnostics. The results showed that Negative Binomial Regression provided the best overall fit, yielding substantially lower AIC and BIC values than competing models. Residual diagnostics indicated no evidence of remaining overdispersion, zero inflation, or serious outliers. Sanitation access was significantly associated with lower diarrhea incidence rates, with a one-standard-deviation increase corresponding to an estimated 15.7% reduction in incidence. These findings highlight the importance of accounting for overdispersion and suggest that Negative Binomial Regression is an appropriate framework for modeling regional diarrhea incidence data.

**Keywords:** Diarrhea incidence; Negative Binomial Regression; Overdispersion; Epidemiology; Count data

Copyright © 2026 by Authors, Published by CAUCHY Group. This is an open access article under the CC BY-SA License (<https://creativecommons.org/licenses/by-sa/4.0>)

## 1. Introduction

Diarrhea remains a major public health challenge worldwide, particularly in developing countries, where it contributes substantially to morbidity and mortality among children under five years of age. Its occurrence is closely associated with environmental sanitation, access to clean water, nutritional status, hygiene behavior, and healthcare availability [1, 2]. Beyond its direct health impact, diarrhea also reflects broader inequalities in public health infrastructure and socioeconomic conditions across regions.

---

\*Corresponding author. E-mail: [ayuninsofro@unesa.ac.id](mailto:ayuninsofro@unesa.ac.id)

In epidemiological research, disease incidence is commonly represented as count data and modeled using Poisson Regression, which assumes a Poisson-distributed response variable with equality between the mean and variance (equidispersion) [3]. Although widely applied in public health studies [4, 5], this assumption is frequently violated because epidemiological data often exhibit overdispersion and clustered dependence arising from regional heterogeneity or repeated observations [6, 7]. Ignoring these characteristics may lead to biased standard errors and unreliable statistical inference [8, 9]. Consequently, alternative approaches such as Negative Binomial Regression and mixed-effects count models have been developed to accommodate extra variability and hierarchical dependence structures [10–15].

Previous studies have shown that hierarchical count-data models improve disease-risk estimation by accounting for regional variability and temporal dependence, while recent developments in mixed-effects and Bayesian approaches have expanded their application in epidemiology and disease mapping [16–19]. However, comprehensive evaluations incorporating population offsets, overdispersion diagnostics, mixed-effects assessment, and residual diagnostics for longitudinal diarrhea surveillance data in Indonesia remain limited.

West Java Province, the most populous province in Indonesia, exhibits substantial disparities in sanitation conditions, healthcare facilities, population density, and nutritional status across districts and cities. These differences are likely to influence diarrhea incidence and generate heterogeneous count-data structures. Therefore, this study compares Poisson Regression, Negative Binomial Regression, and Generalized Linear Mixed Models (GLMM) to identify the most appropriate framework for modeling diarrhea incidence in West Java Province. Model performance is evaluated using information criteria and residual diagnostics, with the aim of providing a robust statistical framework for regional disease-surveillance analysis and public health planning.

Unlike many previous epidemiological studies that apply a single count-regression framework, this study provides a comprehensive methodological evaluation of longitudinal disease-incidence modeling by integrating population-offset adjustment, formal overdispersion assessment, mixed-effects count modeling, information-criterion-based model comparison, and simulation-based residual diagnostics within a unified analytical framework. The study therefore contributes not only to the empirical understanding of diarrhea incidence patterns in West Java Province but also to the methodological validation of alternative count-data regression approaches for overdispersed longitudinal epidemiological datasets. By systematically evaluating whether additional model complexity, such as random effects, provides meaningful improvements beyond standard count-regression models, this study offers practical guidance for selecting appropriate statistical methods in regional disease-surveillance research.

From a methodological perspective, this study evaluates whether the additional complexity introduced by mixed-effects count models provides meaningful improvement over standard Negative Binomial regression for relatively short longitudinal epidemiological datasets. By systematically assessing population-offset adjustment, overdispersion handling, random-effect structures, information-criterion-based model selection, and simulation-based residual diagnostics within a unified framework, the study contributes to the statistical validation of count-data modeling strategies commonly used in disease-surveillance research.

## **2. Methods**

This section describes the overall analytical framework used to model district-level panel count data in West Java Province. The methodology includes data preparation, variable definition, descriptive analysis, multicollinearity assessment, count-data regression modeling, model comparison, and residual diagnostic evaluation. These procedures were implemented to identify the most appropriate statistical framework for analyzing overdispersed epidemiological count data.

## **2.1. Type and Design of Research**

This study employed an observational longitudinal panel-data design using count-data regression methods to analyze diarrhea incidence in West Java Province. The study design followed a longitudinal panel-data framework in which districts/cities were repeatedly observed during the 2022–2024 period.

The dataset possesses both temporal and regional dimensions, where repeated annual observations are nested within districts/cities. Such data structures may produce heterogeneity and correlation among observations from the same region over time [20, 21]. Therefore, count-data regression approaches capable of accommodating overdispersion and hierarchical characteristics were considered in the analysis.

The analytical procedure began with descriptive statistical analysis and multicollinearity testing, followed by estimation of several count regression models. Poisson Regression was initially employed as the baseline model because the response variable represents count data [3]. However, since epidemiological count data frequently exhibit overdispersion, the analysis was extended using Negative Binomial Regression, which introduces an additional dispersion parameter to relax the equidispersion assumption [6–8].

To evaluate the possibility of regional heterogeneity, Generalized Linear Mixed Models (GLMM) with Poisson and Negative Binomial distributions were also considered by incorporating district/city random effects [13, 15, 22]. The final model selection was conducted using model comparison criteria and residual diagnostic procedures.

The analytical stages in this study consisted of:

- Descriptive statistical analysis,
- Multicollinearity testing using Variance Inflation Factor (VIF),
- Estimation of Poisson Regression,
- Overdispersion testing,
- Estimation of Negative Binomial Regression,
- Estimation of Poisson and Negative Binomial GLMM,
- Model comparison using AIC, BIC, and log-likelihood,
- Residual diagnostic evaluation.

## **2.2. Research Framework**

The analytical framework consisted of descriptive statistical analysis, multicollinearity assessment using the Variance Inflation Factor (VIF), and estimation of several count-data regression models. Poisson Regression was initially fitted as the baseline model, followed by overdispersion diagnostics using the Pearson dispersion ratio. When overdispersion was detected, Negative Binomial Regression was employed to accommodate extra-Poisson variability. In addition, Poisson and Negative Binomial Generalized Linear Mixed Models (GLMM) were estimated to evaluate potential regional heterogeneity. Model performance was subsequently compared using Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), log-likelihood values, and residual diagnostic procedures to identify the most appropriate model for the data [3, 6, 10, 15, 23, 24].

## **2.3. Data Sources and Types**

This study used secondary panel data obtained from the Open Data Jabar Portal managed by the Government of West Java Province (<https://opendata.jabarprov.go.id/>). Data retrieval was conducted between January and February 2026. The analysis combined eight publicly available district/city-level datasets covering the 2022–2024 period, including: (1) reported diarrhea cases, (2) inpatient health centers, (3) community health workers, (4) households with proper sanitation access, (5) undernourished toddlers, (6) stunted toddlers, (7) drinking water facilities, and (8)

households implementing clean and healthy living behavior (PHBS).

Because the datasets were derived from routine administrative surveillance records, potential reporting bias, under-reporting, and differences in data collection practices across districts/cities cannot be completely ruled out. Nevertheless, the Open Data Jabar Portal is the official public data repository maintained by the Government of West Java Province and is widely used for regional planning and policy evaluation. The use of standardized administrative reporting procedures improves data consistency across districts and years, although some measurement error may remain.

Table 1 summarizes the datasets used in this study. All datasets were obtained from the Open Data Jabar Portal and correspond to district/city-level observations during 2022–2024.

**Table 1:** Datasets Used in the Study

| Variable | Dataset Description                      | Years     |
|----------|------------------------------------------|-----------|
| $Y$      | Reported diarrhea cases by district/city | 2022–2024 |
| $X_1$    | Households with proper sanitation access | 2022–2024 |
| $X_2$    | Inpatient health centers                 | 2022–2024 |
| $X_3$    | Community health workers                 | 2022–2024 |
| $X_4$    | Undernourished toddlers                  | 2022–2024 |
| $X_5$    | Stunted toddlers                         | 2022–2024 |
| $X_6$    | Drinking water facilities                | 2022–2024 |
| $X_7$    | Households implementing PHBS             | 2022–2024 |

All datasets were publicly available through the Open Data Jabar Portal maintained by the Government of West Java Province. Detailed dataset URLs are provided in Appendix and the Data Availability section to improve reproducibility and facilitate future data verification. All datasets were accessed during January–February 2026.

The initial dataset consisted of 81 observations representing 27 districts/cities observed annually during the 2022–2024 period. Data integration was performed by matching district/city identifiers and observation years across datasets. No observations were excluded during the data preparation stage. Missing-value assessment indicated that all study variables contained complete information, with no missing observations detected. Consequently, the final analytical dataset remained a balanced panel comprising 81 observations (27 districts/cities observed over three years).

The response variable was the annual number of reported diarrhea cases, while population size was incorporated as an offset term to estimate incidence rates rather than raw case counts [3]. Because the data contained repeated observations across districts/cities and years, mixed-effects count regression models were also evaluated to assess potential regional heterogeneity [14, 25].

## 2.4. Operational Definition of Variables

The response variable in this study is denoted by  $Y$ , representing the annual number of diarrhea cases recorded in each district/city. Because the response variable consists of non-negative integer counts, count-data regression approaches were considered appropriate for the analysis [3].

Seven explanatory variables were included in the analysis. Sanitation access ( $X_1$ ) was measured as the number of households with proper sanitation facilities. Healthcare availability was represented by the number of inpatient health centers ( $X_2$ ) and community health workers ( $X_3$ ). Nutritional conditions were represented by the number of undernourished toddlers ( $X_4$ ) and stunted toddlers ( $X_5$ ). Environmental and behavioral conditions were represented by the number of drinking water facilities ( $X_6$ ) and the number of households implementing clean and healthy living behavior (PHBS) ( $X_7$ ). All explanatory variables were measured at the district/city level and recorded as annual count indicators [1, 2].

In addition, dummy variables for observation years were incorporated to capture temporal

differences across the study period. The year 2022 was treated as the baseline category, while indicator variables for 2023 and 2024 were included in the regression models.

For notation consistency, uppercase symbols ( $X_j$ ) denote variables in descriptive analysis, whereas lowercase symbols ( $x_{ij}$ ) denote observed values in the regression models.

**Table 2:** Operational Definition of Variables

| Variable | Description                                        | Unit       | Scale | Expected Association | Transformation |
|----------|----------------------------------------------------|------------|-------|----------------------|----------------|
| $Y$      | Reported diarrhea cases per district/city per year | Cases      | Ratio | –                    | None           |
| $X_1$    | Households with proper sanitation access           | Households | Ratio | Negative             | Standardized   |
| $X_2$    | Inpatient health centers                           | Facilities | Ratio | Negative             | Standardized   |
| $X_3$    | Community health workers                           | Persons    | Ratio | Negative             | Standardized   |
| $X_4$    | Undernourished toddlers                            | Children   | Ratio | Positive             | Standardized   |
| $X_5$    | Stunted toddlers                                   | Children   | Ratio | Positive             | Standardized   |
| $X_6$    | Drinking water facilities                          | Facilities | Ratio | Negative             | Standardized   |
| $X_7$    | Households implementing PHBS                       | Households | Ratio | Negative             | Standardized   |

All explanatory variables were obtained as district/city-level count indicators from the Open Data Jabar Portal. Prior to model estimation, continuous explanatory variables were standardized using z-score transformation to improve numerical stability, reduce scale imbalance among predictors, and facilitate comparison of regression coefficients. Consequently, the estimated coefficients and incidence rate ratios (IRRs) represent the effect associated with a one-standard-deviation increase in the corresponding predictor rather than a one-unit increase in its original measurement scale.

Continuous explanatory variables were standardized prior to model estimation to improve numerical stability and facilitate coefficient interpretation. Consequently, the estimated coefficients represent the effect of a one-standard-deviation increase in each predictor. Because the models incorporated a population offset term, the estimated effects should be interpreted as associations with diarrhea incidence rates adjusted for population exposure differences across districts/cities rather than changes in raw case counts.

Accordingly, the Incidence Rate Ratios (IRR) reported in this study should be interpreted as the multiplicative effect associated with a one-standard-deviation increase in the corresponding explanatory variable rather than a one-unit increase in the original measurement scale.

The standardization procedure was performed using the z-score transformation:

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j},$$

where  $x_{ij}$  denotes the observed value of predictor  $j$  for observation  $i$ ,  $\bar{x}_j$  represents the mean of predictor  $j$ , and  $s_j$  denotes its standard deviation.

## 2.5. Descriptive Statistical Analysis

Descriptive statistical analysis was conducted to summarize the characteristics of the response and explanatory variables prior to model estimation. The descriptive measures included minimum values, maximum values, means, and standard deviations.

This preliminary analysis was intended to identify variability, dispersion patterns, and potential heterogeneity across districts/cities and observation years. Such exploration is important in epidemiological studies because disease count data often exhibit substantial variation across regions and periods [20, 21].

## 2.6. Multicollinearity Test

Before model estimation, multicollinearity testing was conducted among explanatory variables using the Variance Inflation Factor (VIF). Multicollinearity occurs when predictor variables are highly correlated, potentially resulting in unstable coefficient estimates and misleading statistical inference [26].

The VIF for the  $i$ -th explanatory variable is defined as

$$\text{VIF}_i = \frac{1}{1 - R_i^2},$$

where  $R_i^2$  is the coefficient of determination obtained by regressing the  $i$ -th predictor variable on all remaining explanatory variables. A predictor variable is generally considered free from serious multicollinearity if its VIF value is below 10 [27, 28]. Therefore, variables satisfying this criterion were retained in the regression analysis.

## 2.7. Poisson Regression Model

Poisson Regression was initially employed as the baseline model because the response variable represents count data. Let  $Y_i$  denote the number of diarrhea cases for the  $i$ -th observation. The response variable is assumed to follow a Poisson distribution:

$$Y_i \sim \text{Poisson}(\mu_i),$$

The Poisson distribution assumes equidispersion, meaning that the mean and variance are equal:

$$E(Y_i) = \text{Var}(Y_i) = \mu_i.$$

The expected count is linked to explanatory variables through the logarithmic link function:

$$\log(\mu_i) = \beta_0 + \sum_{j=1}^p \beta_j x_{ij} + \log(\text{population}_i),$$

where  $\log(\text{population}_i)$  represents the offset term used to account for population exposure differences across districts/cities.

The inclusion of the population offset allows the regression model to estimate incidence rates rather than raw disease counts, thereby reducing potential bias caused by substantial differences in population size across districts/cities.

Parameter estimation was performed using Maximum Likelihood Estimation (MLE) [4]. Because Poisson Regression assumes equidispersion, an overdispersion test was subsequently conducted by evaluating the dispersion ratio:

$$\text{Dispersion Ratio} = \frac{\chi^2}{df},$$

where  $\chi^2$  denotes the Pearson chi-square statistic and  $df$  represents the residual degrees of freedom. A dispersion ratio substantially greater than one indicates overdispersion [6, 7].

## 2.8. Overdispersion Test

One of the primary assumptions of Poisson Regression is equidispersion, where the mean and variance of the response variable are equal. However, epidemiological count data frequently exhibit overdispersion, a condition in which the variance exceeds the mean [6, 7].

Overdispersion may arise due to unobserved heterogeneity, excess variability across observational units, omitted variables, or clustering effects within the data. When overdispersion occurs, the Poisson Regression model tends to underestimate the standard errors of the regression

coefficients, potentially leading to misleading statistical inference and inflated significance levels [8, 9].

To evaluate the presence of overdispersion, the dispersion ratio was calculated using the Pearson chi-square statistic divided by the residual degrees of freedom. Overdispersion was assessed using the Pearson dispersion ratio described in Equation (6). Where  $\chi^2$  denotes the Pearson chi-square statistic and  $df$  represents the residual degrees of freedom [29–31].

A dispersion ratio close to one indicates that the equidispersion assumption is satisfied. In contrast, a dispersion ratio substantially greater than one indicates overdispersion, suggesting that the Poisson Regression model may not be appropriate for the data [6, 7].

When overdispersion was detected, the analysis was extended using the Negative Binomial Regression model, which introduces an additional dispersion parameter to accommodate extra-Poisson variability [10, 11].

## 2.9. Negative Binomial Regression Model

Because the Poisson Regression model exhibited overdispersion, Negative Binomial Regression was employed as an alternative approach. The Negative Binomial model extends the Poisson model by introducing an additional dispersion parameter that allows the variance to exceed the mean [8, 12].

The Negative Binomial model is specified as

$$Y_i \sim NB(\mu_i, \alpha),$$

where  $\mu_i$  denotes the expected count and  $\alpha > 0$  is the dispersion parameter controlling extra-Poisson variability. The Negative Binomial distribution can be derived as a Poisson-Gamma mixture distribution, allowing the variance to exceed the mean.

Under this model, the variance becomes

$$\text{Var}(Y_i) = \mu_i + \alpha\mu_i^2,$$

Parameter estimation for the Negative Binomial Regression model was conducted using Maximum Likelihood Estimation (MLE). The additional dispersion parameter allows the model to accommodate extra-Poisson variability commonly observed in epidemiological count data.

The systematic component of the model is expressed as

$$\log(\mu_i) = \beta_0 + \sum_{j=1}^p \beta_j x_{ij} + \log(\text{population}_i).$$

Regression coefficients were interpreted using the Incidence Rate Ratio (IRR), obtained from

$$IRR_j = \exp(\beta_j).$$

The corresponding 95% confidence intervals for the incidence rate ratios were obtained from

$$CI_{95\%} = (\exp(\beta_j - 1.96SE(\beta_j)), \exp(\beta_j + 1.96SE(\beta_j))),$$

where  $SE(\beta_j)$  denotes the standard error of the estimated regression coefficient. Confidence intervals provide additional information regarding the precision and statistical uncertainty associated with the estimated incidence rate ratios.

The IRR represents the multiplicative effect of a one-standard-deviation increase in the predictor variable on the expected diarrhea incidence rate because all continuous explanatory variables were standardized prior to model estimation. Standardization was performed to improve numerical convergence, reduce scale imbalance among predictors, and facilitate comparison of regression effect sizes across explanatory variables [26].

## 2.10. Generalized Linear Mixed Models (GLMM)

To evaluate potential regional heterogeneity, Generalized Linear Mixed Models (GLMM) with Poisson and Negative Binomial distributions were also estimated. The GLMM framework extends generalized linear models by incorporating random effects that capture unobserved variability across clusters or regions [13, 15].

Let  $Y_{ij}$  denote the number of diarrhea cases in district/city  $i$  during year  $j$ . The mixed-effects model is specified as

$$\log(\mu_{ij}) = \mathbf{x}_{ij}^T \boldsymbol{\beta} + u_i + \log(\text{population}_{ij}),$$

where  $\mathbf{x}_{ij}$  denotes the vector of explanatory variables,  $\boldsymbol{\beta}$  represents fixed-effect parameters, and  $u_i$  denotes the random intercept associated with the  $i$ -th district/city.

The random effects are assumed to follow a normal distribution:

$$u_i \sim N(0, \sigma_u^2).$$

Only random-intercept models were considered because the dataset contained a relatively limited number of repeated observations per district/city during the study period. Random-slope models were not considered because the relatively short observation period (three years) could lead to unstable variance estimation and overparameterization in mixed-effects count models [15, 22]. Parameter estimation was conducted using likelihood-based numerical approximation methods such as the Laplace approximation [32, 33].

The inclusion of random effects was evaluated using variance estimates, model fit criteria, and singularity diagnostics. For Poisson GLMM estimated with the `lme4` package, singularity was assessed using the `isSingular()` function. Random-effect variances approaching zero or singular model fits indicate that the mixed-effects specification provides little additional explanatory value beyond the fixed-effects model. Nevertheless, GLMMs were evaluated because the longitudinal data structure may contain regional heterogeneity arising from repeated observations within districts/cities.

## 2.11. Best Model Selection

Model selection was performed using the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and log-likelihood values.

The AIC is defined as

$$AIC = -2 \log L(\theta) + 2k,$$

while the BIC is defined as

$$BIC = -2 \log L(\theta) + k \log(n),$$

where  $L(\theta)$  denotes the maximized likelihood function,  $k$  is the number of estimated parameters, and  $n$  represents the number of observations.

In addition to information criteria, Nagelkerke's pseudo- $R^2$  was calculated to provide an overall measure of explanatory power for the selected count regression model. Although pseudo- $R^2$  values in generalized linear models do not have the same interpretation as the coefficient of determination in ordinary least squares regression, they provide a useful indication of the proportion of variation explained by the fitted model. Models with lower AIC and BIC values are considered preferable because they achieve a better balance between model fit and complexity [23, 24].

## 2.12. Residual Diagnostics and Model Evaluation

Residual diagnostic analysis was conducted to evaluate the adequacy of the final selected model. Simulation-based residual diagnostics were employed to assess dispersion, zero inflation, and outlier behavior in the count regression models.

The dispersion test was used to determine whether overdispersion remained after model estimation. In addition, zero-inflation and outlier diagnostics were evaluated to detect potential violations of model assumptions [7].

A model was considered adequate if:

- the dispersion test was not statistically significant,
- no substantial zero inflation was detected,
- no serious outlier patterns were observed.

These diagnostic procedures ensured that the selected model provided reliable parameter estimates and appropriate statistical inference for diarrhea incidence data in West Java Province.

Although the residual-versus-predicted plot indicated significant quantile deviations, the overall diagnostic results suggest that the Negative Binomial model provides an adequate representation of the data. The observed quantile pattern may indicate minor nonlinear structures or unmeasured regional characteristics not fully captured by the available predictors.

All statistical analyses were conducted using the R statistical software environment. Poisson and Negative Binomial Regression models were estimated using generalized linear modeling procedures, while mixed-effects models were estimated using the `lme4` and `glmmTMB` packages. Residual diagnostic analyses were performed using simulation-based methods implemented in the `DHARMA` package.

### 3. Results and Discussion

This section presents the empirical findings of the study. The results begin with descriptive analysis and multicollinearity assessment, followed by estimation of count regression models, model comparison, evaluation of random effects, and residual diagnostics. The findings are subsequently discussed from both statistical and epidemiological perspectives.

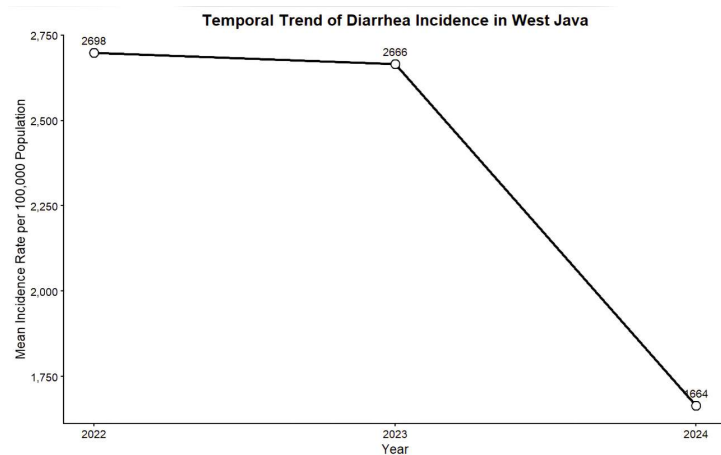
#### 3.1. Descriptive Analysis

This study analyzed district/city-level diarrhea cases in West Java Province during 2022–2024 using panel count data. The response variable was the number of diarrhea cases, while seven environmental, health service, nutritional, and behavioral indicators were included as explanatory variables. Descriptive statistics were calculated to summarize the distribution and variability of the study variables (Table 3).

**Table 3:** Descriptive Statistics of Research Variables

| Variable | Min   | Max     | Mean      | Std. Dev  |
|----------|-------|---------|-----------|-----------|
| $Y$      | 4096  | 166464  | 41581.17  | 31221.55  |
| $X_1$    | 54951 | 1324263 | 490783.63 | 309148.40 |
| $X_2$    | 1     | 35      | 10.94     | 10.15     |
| $X_3$    | 31    | 310     | 118.41    | 58.70     |
| $X_4$    | 340   | 13458   | 3569.75   | 2725.00   |
| $X_5$    | 425   | 30266   | 5737.22   | 5930.83   |
| $X_6$    | 2     | 63653   | 1957.58   | 9187.42   |
| $X_7$    | 15413 | 749222  | 223596.14 | 180914.80 |

Table 3 shows substantial variation in diarrhea incidence and explanatory variables across districts/cities in West Java Province. The response variable exhibited high dispersion, with variance substantially exceeding the mean, indicating potential overdispersion. Considerable variability was also observed in sanitation, nutritional, behavioral, and healthcare indicators, reflecting regional disparities in public health conditions. These findings suggest substantial heterogeneity in the data and support the use of count-data regression models for further analysis.



**Fig. 1:** Temporal trend of reported diarrhea incidence rates in West Java Province during 2022–2024.

Fig. 1 illustrates the temporal pattern of diarrhea incidence rates across the study period. A general decline in incidence is observed in 2024 compared with previous years, supporting the negative coefficient estimated for the 2024 indicator variable in subsequent regression analyses.

### 3.2. Multicollinearity Test

Prior to model estimation, multicollinearity testing was conducted to assess the linear relationships among explanatory variables using the Variance Inflation Factor (VIF). The results are shown in Table 4.

**Table 4:** Variance Inflation Factor (VIF) Results

| Variable | VIF   |
|----------|-------|
| $X_1$    | 5.061 |
| $X_2$    | 1.211 |
| $X_3$    | 1.501 |
| $X_4$    | 6.622 |
| $X_5$    | 3.176 |
| $X_6$    | 1.218 |
| $X_7$    | 2.149 |

All predictor variables have VIF values below the commonly accepted threshold of 10, indicating that serious multicollinearity is not present in the dataset. Although sanitation access ( $X_1$ ) and undernourished toddlers ( $X_4$ ) show relatively higher VIF values, they remain within acceptable limits. These results indicate that the explanatory variables do not exhibit excessively strong linear relationships that could compromise parameter estimation stability. Therefore, all predictor variables were retained for subsequent modeling.

### 3.3. Poisson Regression Model

Poisson Regression was initially employed as the baseline model because the response variable represents count data. The model incorporated an offset term based on population size to account for differences in population exposure across districts/cities. Thus, the estimated regression coefficients represent effects on the incidence rate rather than the raw count of diarrhea cases. Parameter estimation was performed using Maximum Likelihood Estimation (MLE). The estimation results are presented in Table 5.

**Table 5:** Poisson Regression Parameter Estimates

| Variable  | Estimate | Std. Error | z value   | p-value |
|-----------|----------|------------|-----------|---------|
| Intercept | -3.571   | 0.001      | -3794.719 | < 0.001 |
| $X_1$     | -0.090   | 0.001      | -77.302   | < 0.001 |
| $X_2$     | -0.029   | 0.001      | -49.737   | < 0.001 |
| $X_3$     | 0.008    | 0.001      | 12.375    | < 0.001 |
| $X_4$     | 0.059    | 0.001      | 50.302    | < 0.001 |
| $X_5$     | -0.017   | 0.001      | -19.908   | < 0.001 |
| $X_6$     | 0.010    | 0.001      | 22.369    | < 0.001 |
| $X_7$     | 0.007    | 0.001      | 8.924     | < 0.001 |
| Year 2023 | -0.030   | 0.001      | -23.818   | < 0.001 |
| Year 2024 | -0.704   | 0.002      | -437.401  | < 0.001 |

The Poisson Regression results indicate that all explanatory variables appear statistically significant at the 5% significance level. However, because the model subsequently exhibited severe overdispersion, these significance tests should be interpreted with caution. Severe overdispersion can substantially underestimate coefficient standard errors, thereby inflating statistical significance and producing unreliable inference.

However, despite the statistical significance of the predictors, the Poisson Regression model showed severe overdispersion. The dispersion ratio was calculated as 1819.395, which is substantially greater than 1. In count data regression, a dispersion ratio far exceeding 1 indicates that the variance of the response variable is much larger than its mean, violating the equidispersion assumption of the Poisson distribution.

$$\text{Dispersion Ratio} = 1819.395$$

The presence of severe overdispersion implies that the Poisson Regression model underestimates the variability in the data, potentially leading to biased standard errors and misleading statistical inference. Therefore, a more flexible modeling approach capable of accommodating overdispersion is required.

### 3.4. Negative Binomial Regression Model

To address the overdispersion identified in the Poisson Regression model, Negative Binomial Regression was subsequently employed. The Negative Binomial model extends the Poisson model by introducing an additional dispersion parameter, allowing the variance to exceed the mean. The estimation results are presented in [Table 6](#).

**Table 6:** Negative Binomial Regression Parameter Estimates

| Variable  | Estimate | Std. Error | z value | p-value |
|-----------|----------|------------|---------|---------|
| Intercept | -3.619   | 0.048      | -75.652 | < 0.001 |
| $X_1$     | -0.170   | 0.061      | -2.776  | 0.006   |
| $X_2$     | -0.045   | 0.030      | -1.525  | 0.127   |
| $X_3$     | 0.009    | 0.034      | 0.265   | 0.791   |
| $X_4$     | 0.083    | 0.070      | 1.181   | 0.238   |
| $X_5$     | -0.008   | 0.049      | -0.173  | 0.862   |
| $X_6$     | 0.029    | 0.030      | 0.962   | 0.336   |
| $X_7$     | 0.011    | 0.040      | 0.282   | 0.778   |
| Year 2023 | 0.003    | 0.067      | 0.039   | 0.969   |
| Year 2024 | -0.499   | 0.070      | -7.156  | < 0.001 |

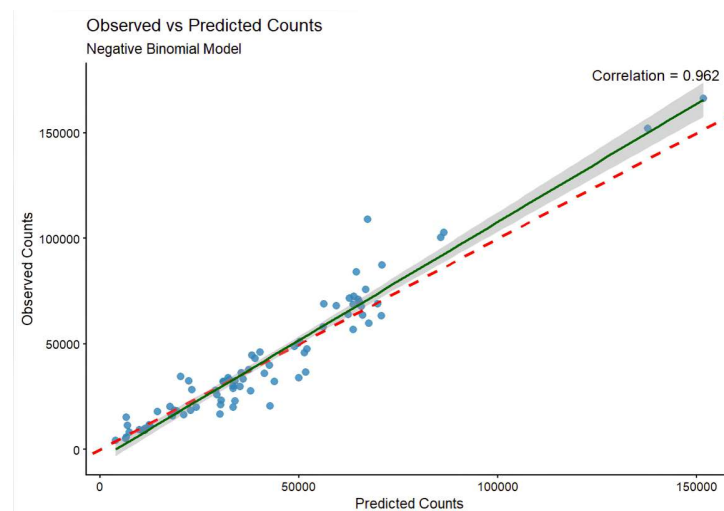
The Negative Binomial Regression results identified sanitation access ( $X_1$ ) and the year 2024 indicator as statistically significant predictors of diarrhea incidence. Sanitation access exhibited a negative association with diarrhea incidence, indicating that improved sanitation was associated

**Table 7:** Incidence Rate Ratios and 95% Confidence Intervals

| Variable  | IRR   | 95% CI         |
|-----------|-------|----------------|
| $X_1$     | 0.843 | (0.749, 0.951) |
| $X_2$     | 0.956 | (0.902, 1.014) |
| $X_3$     | 1.009 | (0.945, 1.079) |
| $X_4$     | 1.086 | (0.947, 1.248) |
| $X_5$     | 0.992 | (0.901, 1.093) |
| $X_6$     | 1.029 | (0.971, 1.097) |
| $X_7$     | 1.011 | (0.933, 1.097) |
| Year 2023 | 1.003 | (0.879, 1.144) |
| Year 2024 | 0.607 | (0.530, 0.695) |

with lower incidence rates. Similarly, the negative coefficient for the year 2024 variable suggests that diarrhea incidence was significantly lower in 2024 than in the baseline year after adjusting for the explanatory variables. In contrast, the remaining predictors were not statistically significant at the 5% level after accounting for overdispersion.

The estimated IRR for sanitation access was 0.843 (95% CI: 0.749–0.951), indicating that a one-standard-deviation increase in sanitation access was associated with an estimated 15.7% reduction in the expected diarrhea incidence rate. The estimated dispersion parameter ( $\alpha = 17.2136$ ) confirmed the presence of substantial extra-Poisson variability in the data. Furthermore, likelihood-based model comparison demonstrated that the Negative Binomial model provided a substantially better fit than the Poisson model, highlighting the importance of accommodating overdispersion to obtain reliable statistical inference.



**Fig. 2:** Observed versus predicted diarrhea case counts from the final Negative Binomial Regression model.

Fig. 2 compares observed and model-predicted diarrhea case counts. The close agreement between observed and predicted values indicates that the Negative Binomial model adequately captures the overall variation in the data while accounting for population exposure and overdispersion.

### 3.5. Model Adequacy and Goodness-of-Fit

The Negative Binomial Regression model achieved a Nagelkerke pseudo- $R^2$  value of 0.750, indicating substantial explanatory power for regional diarrhea incidence patterns. In addition, the estimated dispersion parameter ( $\alpha = 17.214$ ) confirmed the presence of substantial overdispersion and demonstrated the ability of the Negative Binomial specification to accommodate extra-Poisson variability, resulting in more reliable inference than the standard Poisson model. These findings

suggest that the model provides an adequate representation of the underlying epidemiological process and can be used for reliable statistical interpretation.

### **3.6. Interpretation of the Final Negative Binomial Model**

The final Negative Binomial Regression model identified sanitation access as the only predictor significantly associated with diarrhea incidence after adjusting for population exposure and overdispersion. The negative coefficient ( $\beta = -0.170$ ,  $p = 0.006$ ; IRR = 0.843, 95% CI: 0.749–0.951) indicates that improved sanitation access was associated with lower diarrhea incidence rates. The 2024 indicator was also significant ( $\beta = -0.499$ ,  $p < 0.001$ ; IRR = 0.607, 95% CI: 0.530–0.695), suggesting a lower incidence rate compared with 2022. In contrast, the remaining explanatory variables were not statistically significant after accounting for overdispersion. These results indicate that the apparent significance observed in the Poisson model was partly attributable to underestimated standard errors, while the Negative Binomial model provided more reliable inference. From a public health perspective, the findings highlight sanitation improvement as a key strategy for reducing diarrhea incidence.

Therefore, an IRR should be interpreted as the multiplicative change in diarrhea incidence rate associated with a one-standard-deviation increase in the corresponding predictor while holding all other variables constant.

The limited observation period (2022–2024) may also have reduced the ability to detect longer-term temporal patterns and delayed effects of public health interventions. Future studies incorporating longer surveillance periods may provide more robust assessment of temporal dynamics and district-level heterogeneity.

### **3.7. Discussion of Findings**

The findings of this study are consistent with previous epidemiological research emphasizing the importance of sanitation in reducing diarrheal disease transmission. Poor sanitation increases the likelihood of fecal contamination in household environments and water sources, thereby facilitating pathogen transmission through the fecal–oral route [1, 2].

The significant protective effect of sanitation access is consistent with previous studies highlighting sanitation as a key determinant of diarrheal disease prevention. In contrast, healthcare-related and nutritional variables were not statistically significant after overdispersion adjustment, suggesting that environmental factors may play a more direct role in shaping regional diarrhea incidence patterns than healthcare availability or aggregated nutritional indicators.

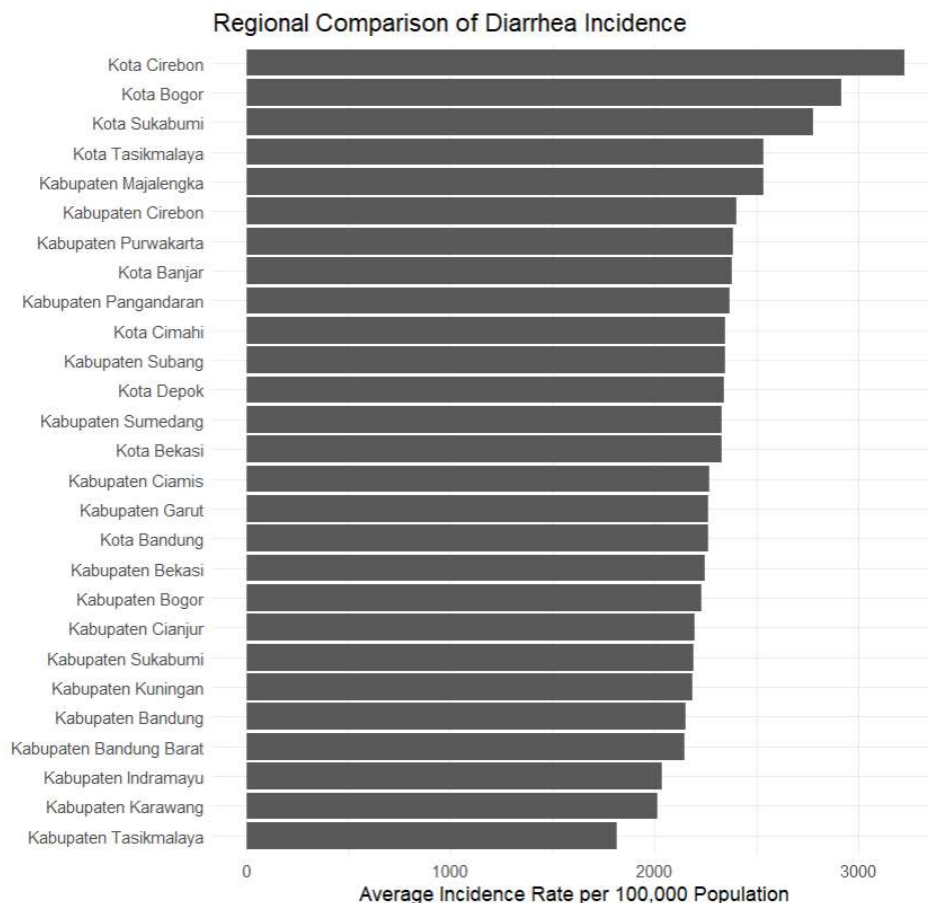
One possible explanation for the lack of statistical significance among healthcare-service, nutritional, drinking-water, and PHBS-related variables is that these indicators may reflect broader regional characteristics already partially captured by the population-adjusted incidence rates. In addition, the substantial overdispersion observed in the data suggests the presence of unmeasured contextual factors that may reduce the independent contribution of individual predictors once extra-Poisson variability is properly accommodated.

Fig. 3 demonstrates substantial variation in diarrhea incidence rates among districts and cities. The observed heterogeneity supports the need for count-data regression approaches capable of accommodating regional variability and overdispersion.

Overall, the results suggest that environmental health factors play a more direct role in shaping regional diarrhea incidence patterns than healthcare service availability alone.

### **3.8. Generalized Linear Mixed Model Results**

To evaluate whether district/city-level heterogeneity contributed substantially to diarrhea incidence variation, mixed-effects count regression models were estimated using both Poisson and Negative Binomial distributions. A random intercept was specified for districts/cities to account for repeated observations across the study period.



**Fig. 3:** Regional variation in diarrhea incidence rates across districts/cities in West Java Province during 2022–2024.

**Table 8:** Estimated Random-Intercept Variance in GLMM Models

| Model                  | Random-Intercept Variance |
|------------------------|---------------------------|
| Poisson GLMM           | 0.04266                   |
| Negative Binomial GLMM | $3.255 \times 10^{-10}$   |

Table 8 shows that the estimated district-level random-intercept variance in the Negative Binomial GLMM was essentially zero. This finding indicates that little unexplained heterogeneity remained after accounting for the fixed effects and population offset. Consequently, the mixed-effects specification provided negligible improvement over the standard Negative Binomial Regression model.

Although the Poisson GLMM slightly improved model fit compared with standard Poisson Regression, substantial overdispersion remained evident. In the Negative Binomial GLMM, the estimated random-intercept variance was essentially zero ( $3.255 \times 10^{-10}$ ), with a random-effect standard deviation of only  $1.804 \times 10^{-5}$ . These results indicate that district-level heterogeneity was largely explained by the observed predictors and population offset, and that the mixed-effects specification provided minimal improvement over the standard Negative Binomial Regression model.

These findings indicate that the incorporation of random effects was not necessary for the present dataset and that the standard Negative Binomial model was sufficient to capture the observed variation in diarrhea incidence.

### 3.9. Model Comparison

Model selection was conducted using the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and log-likelihood values. Lower AIC and BIC values indicate superior model performance through a better balance between goodness-of-fit and model complexity. The comparison results are presented in Table 9.

**Table 9:** Comparison of Count Regression Models

| Model                        | AIC      | BIC      | Log-Likelihood |
|------------------------------|----------|----------|----------------|
| Poisson Regression           | 147318   | 147344   | -73649         |
| Poisson GLMM                 | 146921   | 146951   | -73448         |
| Negative Binomial Regression | 1701.066 | 1727.405 | -839.533       |
| Negative Binomial GLMM       | 1703.066 | 1731.799 | -839.533       |

The Poisson Regression model exhibited extremely large AIC and BIC values compared with all alternative models. This result reflects the severe violation of the equidispersion assumption, as evidenced by the dispersion ratio of 1819.395. Consequently, the Poisson model failed to adequately accommodate the substantial extra-Poisson variability present in the diarrhea incidence data.

Both Negative Binomial models substantially improved model fit by incorporating an additional dispersion parameter. Among all competing models, the standard Negative Binomial Regression model produced the lowest AIC (1701.066) and BIC (1727.405), indicating the most favorable balance between model fit and model complexity.

Although the Negative Binomial Generalized Linear Mixed Model (NB-GLMM) was estimated to account for potential district-level heterogeneity, its AIC (1703.066) and BIC (1731.799) were slightly higher than those of the standard Negative Binomial model. Furthermore, the estimated district-level random intercept variance was extremely close to zero ( $3.255 \times 10^{-10}$ ), indicating that little residual variation remained attributable to differences between districts/cities after accounting for population size and explanatory variables.

The extremely large difference between the Poisson and Negative Binomial information criteria reflects the severe model misspecification caused by overdispersion in the Poisson model.

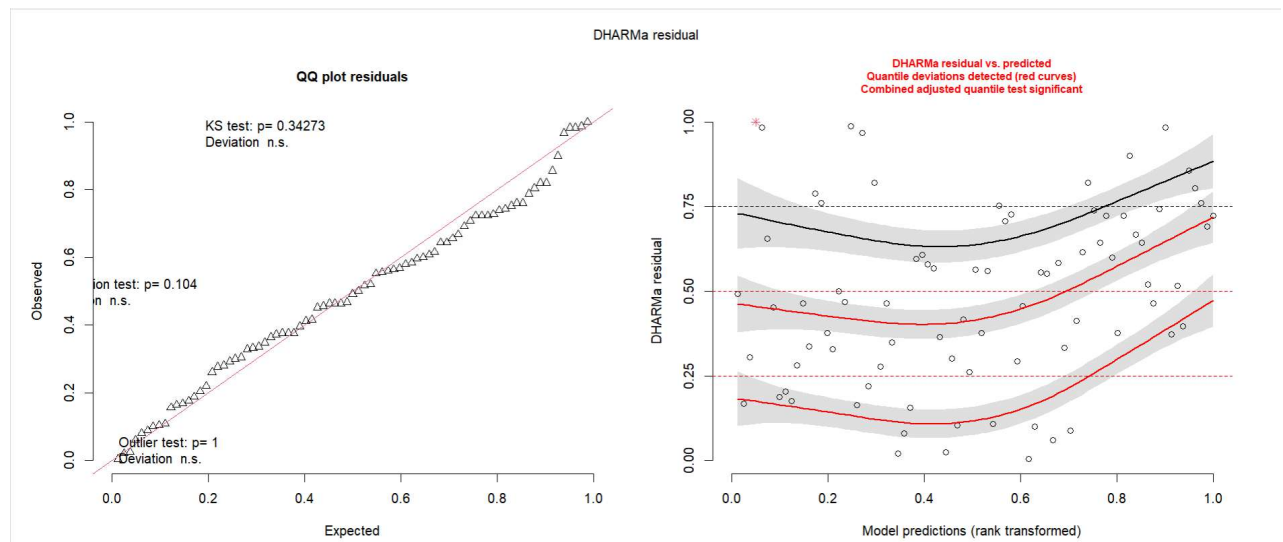
These findings suggest that the random-effect component contributed minimally to model performance. Therefore, the standard Negative Binomial Regression model was selected as the final model because it provided superior parsimony, computational stability, and overall model fit.

**Table 10:** Summary of Model Diagnostics

| Diagnostic Measure                       | Result   |
|------------------------------------------|----------|
| Poisson Dispersion Ratio                 | 1819.395 |
| NB Dispersion Parameter ( $\alpha$ )     | 17.214   |
| DHARMa Uniformity Test ( $p$ -value)     | 0.343    |
| DHARMa Dispersion Test ( $p$ -value)     | 0.104    |
| DHARMa Zero-Inflation Test ( $p$ -value) | 1.000    |
| DHARMa Outlier Test ( $p$ -value)        | 1.000    |
| Nagelkerke Pseudo- $R^2$                 | 0.750    |

The model comparison results further demonstrate the importance of evaluating overdispersion in epidemiological count-data studies. Failure to accommodate extra-Poisson variability can lead to substantial model misspecification and misleading statistical inference.

### 3.10. Residual Diagnostics and Model Performance



**Fig. 4:** Simulation-based residual diagnostics for the final Negative Binomial Regression model using DHARMA.

Figure 4 presents the simulation-based residual diagnostics obtained using the DHARMA package. The Kolmogorov–Smirnov uniformity test was not statistically significant ( $p = 0.343$ ), indicating consistency between the observed residual distribution and the expected uniform distribution. Likewise, the dispersion test was not significant ( $p = 0.104$ ), suggesting that the Negative Binomial model successfully accommodated the severe overdispersion observed in the Poisson model.

Residual diagnostics using the DHARMA package indicated satisfactory model performance. The dispersion test was non-significant (dispersion ratio = 0.606,  $p = 0.104$ ), confirming that the Negative Binomial model adequately accommodated overdispersion. Likewise, the zero-inflation test ( $p = 1.000$ ) and outlier test ( $p = 1.000$ ) showed no evidence of excess zeros or unusual observations.

Although the DHARMA residual-versus-predicted plot indicated significant quantile deviations, the dispersion, zero-inflation, and outlier tests were not statistically significant. These quantile deviations may suggest the presence of remaining nonlinear relationships or unmeasured regional heterogeneity not captured by the available predictors. Nevertheless, the overall diagnostic results indicate that the Negative Binomial model adequately accommodated the major distributional characteristics of the data.

Furthermore, the model achieved a Nagelkerke pseudo- $R^2$  value of 0.750, indicating substantial explanatory power. Collectively, these results support the adequacy and reliability of the selected Negative Binomial Regression model.

### 3.11. Methodological Contribution

This study contributes a practical framework for analyzing longitudinal epidemiological count data by integrating population-offset adjustment, overdispersion assessment, comparison of Poisson and Negative Binomial models, evaluation of random effects, and simulation-based residual diagnostics. The results demonstrated that Negative Binomial Regression provided the most appropriate and parsimonious approach for the analyzed diarrhea surveillance data, successfully accommodating substantial overdispersion while offering minimal benefit from additional mixed-effects specifications. Furthermore, the study provides empirical evidence that increased model complexity does not necessarily improve model performance when the observed heterogeneity can already be explained by the available predictors and population-adjustment mechanisms.

### 3.12. Implications for Public Health Policy

The findings highlight sanitation improvement as a key strategy for reducing diarrhea incidence in West Java Province. Investments in sanitation infrastructure, waste management, clean water access, and hygiene promotion may therefore have substantial public health benefits. In addition, the results emphasize the importance of evaluating overdispersion and regional heterogeneity when selecting count-data models, as failure to account for extra-Poisson variability may lead to misleading statistical inference. These findings may assist policymakers in prioritizing preventive interventions that target environmental determinants of diarrheal disease while supporting evidence-based public health planning.

### 3.13. Study Limitations

Several limitations should be considered. First, the analysis relied on secondary administrative data, which may be subject to reporting inaccuracies and incompleteness. Second, the use of aggregated district/city-level data limits interpretation to ecological associations rather than individual-level effects. Third, spatial and temporal dependencies were not explicitly modeled. Finally, several potentially relevant socioeconomic, environmental, and climate-related variables were unavailable, and the relatively short three-year observation period may not fully capture long-term disease dynamics. Future studies incorporating longer observation periods, additional contextual variables, and spatial-temporal modeling approaches may provide a more comprehensive understanding of diarrhea incidence patterns and their underlying determinants.

## 4. Conclusion

This study evaluated several count-data regression approaches for modeling district-level diarrhea incidence in West Java Province during 2022–2024. The initial Poisson Regression model exhibited severe overdispersion, indicating violation of the equidispersion assumption and demonstrating the need for alternative modeling approaches.

Among the competing models, Negative Binomial Regression provided the best overall performance based on AIC, BIC, log-likelihood values, and simulation-based residual diagnostics. Although mixed-effects models were considered to account for regional heterogeneity, the estimated random-effect variance approached zero, indicating that district-level random effects contributed negligibly after adjustment for explanatory variables and population exposure.

Sanitation access was the only predictor that showed a statistically significant association with diarrhea incidence after adjustment for population exposure and overdispersion. The results suggest that improvements in sanitation infrastructure may substantially reduce diarrhea burden across districts and cities in West Java Province.

Methodologically, the study demonstrates the importance of combining population-offset adjustment, overdispersion assessment, model comparison, and residual diagnostics when analyzing epidemiological count data. The findings indicate that Negative Binomial Regression offers a parsimonious and statistically reliable framework for analyzing overdispersed regional disease-incidence data.

Future studies may incorporate spatial dependence, temporal correlation, climatic variables, or Bayesian hierarchical approaches to further improve disease-surveillance modeling and understanding of regional transmission dynamics.

## Author Contributions

**Nurul Nabilah:** Conceptualization, Data Curation, Formal Analysis, Methodology, Software, Visualization, Writing–Original Draft. **A’yunin Sofro:** Supervision, Methodology Validation, Writing–review and Editing.

## Ethical Statement

This study used publicly available secondary data obtained from the Open Data Portal of West Java Province. No individual-level or personally identifiable information was involved in the analysis. Therefore, ethical approval was not required.

## Conflict of Interest

The authors declare that there is no conflict of interest regarding the publication of this study.

## Funding Statement

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

## Data Availability

The datasets analyzed during the current study are publicly available from the Open Data Jabar Portal managed by the Government of West Java Province and were accessed between January and February 2026: <https://opendata.jabarprov.go.id/>

The processed dataset and R analysis scripts are available from the corresponding author upon reasonable request.

## References

- [1] Karen L. Kotloff et al. “The Burden and Etiology of Diarrheal Illness in Developing Countries”. In: *The Lancet* 382.9888 (2013), pp. 209–222. DOI: [10.1016/S0140-6736\(13\)60844-2](https://doi.org/10.1016/S0140-6736(13)60844-2).
- [2] Frank Osei. “Spatial variation and hot-spots of district-level diarrhea incidences in Ghana: 2010–2014”. In: *BMC Public Health* 17 (2017). DOI: [10.1186/s12889-017-4541-z](https://doi.org/10.1186/s12889-017-4541-z).
- [3] A. C. Cameron and P. K. Trivedi. *Regression Analysis of Count Data*. 2nd ed. Cambridge University Press, 2013. DOI: [10.1017/CB09780511814365](https://doi.org/10.1017/CB09780511814365).
- [4] E. Oral. “Modified maximum likelihood estimation in Poisson regression”. In: *Biometrics & Biostatistics International Journal* 6.3 (2017). DOI: [10.15406/bbij.2017.06.00154](https://doi.org/10.15406/bbij.2017.06.00154).
- [5] A. M. Alomair, F. Ayyaz, S. Tariq, and M. Ahsan-ul-Haq. “Discrete Poisson Haq distribution with mathematical properties and count data modeling”. In: *Scientific Reports* 15 (2025), p. 23281. DOI: [10.1038/s41598-025-07223-y](https://doi.org/10.1038/s41598-025-07223-y).
- [6] J. Hinde and C. G. B. Demétrio. “Overdispersion: Models and estimation”. In: *Computational Statistics & Data Analysis* 27.2 (1998), pp. 151–170. DOI: [10.1016/S0167-9473\(98\)00007-3](https://doi.org/10.1016/S0167-9473(98)00007-3).
- [7] E. H. Payne, J. W. Hardin, L. E. Egede, V. Ramakrishnan, A. Selassie, and M. Gebregziabher. “Approaches for dealing with various sources of overdispersion in modeling count data: Scale adjustment versus modeling”. In: *Statistical Methods in Medical Research* 26.4 (2017), pp. 1802–1823. DOI: [10.1177/0962280215588569](https://doi.org/10.1177/0962280215588569).
- [8] A. M. Azevedo, Í. J. Silva, M. C. Nery, H. P. Rocha, and R. A. Santana. “Counting models for overdispersed data: A review with application to tuberculosis data”. In: *Brazilian Journal of Biometrics* 41.3 (2023), pp. 274–286. DOI: [10.28951/bjb.v41i3.635](https://doi.org/10.28951/bjb.v41i3.635).
- [9] Bijesh Yadav, Lakshmanan Jeyaseelan, Visalakshi Jeyaseelan, Jothilakshmi Durairaj, Sebastian George, K. G. Selvaraj, and Shrikant I. Bangdiwala. “Can Generalized Poisson model replace any other count data models? An evaluation”. In: *Clinical Epidemiology and Global Health* 11 (2021), p. 100774. DOI: [10.1016/j.cegh.2021.100774](https://doi.org/10.1016/j.cegh.2021.100774).

- [10] E. Espinoza, U. Saputri, F. H. Fadilah, and D. Devianto. “Modeling the count data of public health service visits with overdispersion problem by using negative binomial regression”. In: *Journal of Physics: Conference Series* 1940.1 (2021), p. 012021. DOI: [10.1088/1742-6596/1940/1/012021](https://doi.org/10.1088/1742-6596/1940/1/012021).
- [11] S. Q. Aina, Darnah, M. Fauziyah, and W. P. Nurmayanti. “Comparison of generalized Poisson regression and negative binomial regression models based on Akaike information criterion values”. In: *STATISTIKA: Journal of Theoretical Statistics and Its Applications* 25.1 (2025). DOI: [10.29313/statistika.v25i1.5402](https://doi.org/10.29313/statistika.v25i1.5402).
- [12] Antonio J. Sáez-Castillo and Antonio Conde-Sánchez. “A hyper-Poisson regression model for overdispersed and underdispersed count data”. In: *Computational Statistics & Data Analysis* 61 (2013), pp. 148–157. DOI: [10.1016/j.csda.2012.12.009](https://doi.org/10.1016/j.csda.2012.12.009).
- [13] N. E. Breslow and D. G. Clayton. “Approximate inference in generalized linear mixed models”. In: *Journal of the American Statistical Association* 88.421 (1993), pp. 9–25. DOI: [10.2307/2290687](https://doi.org/10.2307/2290687).
- [14] Youngjo Lee and John A. Nelder. “Hierarchical Generalized Linear Models”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.4 (Nov. 1996), pp. 619–656. DOI: [10.1111/j.2517-6161.1996.tb02105.x](https://doi.org/10.1111/j.2517-6161.1996.tb02105.x).
- [15] B. M. Bolker, M. E. Brooks, C. J. Clark, S. W. Geange, J. R. Poulsen, M. H. H. Stevens, and J. S. White. “Generalized linear mixed models: A practical guide for ecology and evolution”. In: *Trends in Ecology & Evolution* 24.3 (2009), pp. 127–135. DOI: [10.1016/j.tree.2008.10.008](https://doi.org/10.1016/j.tree.2008.10.008).
- [16] Bertho Tantular, I Gede Nyoman Mindra Jaya, and Zulhanif. “Multilevel Models Approach on Longitudinal Studies for Disease Data”. In: *Journal of Physics: Conference Series* 1025.1 (2018), p. 012108. DOI: [10.1088/1742-6596/1025/1/012108](https://doi.org/10.1088/1742-6596/1025/1/012108).
- [17] B. E. Yirdaw, L. K. Debusho, and A. Samuel. “Application of longitudinal multilevel zero inflated Poisson regression in modeling of infectious diseases among infants in Ethiopia”. In: *BMC Infectious Diseases* 24 (2024), p. 927. DOI: [10.1186/s12879-024-09820-0](https://doi.org/10.1186/s12879-024-09820-0).
- [18] Andrew B. Lawson. *Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology*. 3rd ed. Boca Raton: Chapman and Hall/CRC, 2018. DOI: [10.1201/9781351271769](https://doi.org/10.1201/9781351271769).
- [19] S. Xu, M. A. R. Ferreira, E. M. Porter, and C. T. Franck. “Bayesian model selection for generalized linear mixed models”. In: *Biometrics* 79.4 (2023), pp. 3266–3278. DOI: [10.1111/biom.13896](https://doi.org/10.1111/biom.13896).
- [20] Peter J. Diggle, Patrick Heagerty, Kung-Yee Liang, and Scott L. Zeger. *Analysis of Longitudinal Data*. 2nd ed. Oxford: Oxford University Press, 2002. DOI: [10.1093/acprof:oso/9780198524847.001.0001](https://doi.org/10.1093/acprof:oso/9780198524847.001.0001).
- [21] Geert Molenberghs and Geert Verbeke. *Models for Discrete Longitudinal Data*. 1st ed. Springer Series in Statistics. New York: Springer, 2005, p. 687. DOI: [10.1007/0-387-28980-1](https://doi.org/10.1007/0-387-28980-1).
- [22] H. Goldstein. *Multilevel Statistical Models*. John Wiley & Sons, 2011. DOI: [10.1002/9780470973394](https://doi.org/10.1002/9780470973394).
- [23] Clifford M. Hurvich and Chih-Ling Tsai. “Regression and time series model selection in small samples”. In: *Biometrika* 76.2 (1989), pp. 297–307. DOI: [10.1093/biomet/76.2.297](https://doi.org/10.1093/biomet/76.2.297).
- [24] N. M. Mangan, J. N. Kutz, S. L. Brunton, and J. L. Proctor. “Model selection for dynamical systems via sparse regression and information criteria”. In: *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 473.2204 (2017). DOI: [10.1098/rspa.2017.0009](https://doi.org/10.1098/rspa.2017.0009).

- [25] A. F. Zuur, E. N. Ieno, N. Walker, A. A. Saveliev, and G. M. Smith. *Mixed Effects Models and Extensions in Ecology with R*. Springer, 2009. DOI: [10.1007/978-0-387-87458-6](https://doi.org/10.1007/978-0-387-87458-6).
- [26] Jae H. Kim. “Multicollinearity and misleading statistical results”. In: *Korean Journal of Anesthesiology* 72.6 (2019), pp. 558–569. DOI: [10.4097/kja.19087](https://doi.org/10.4097/kja.19087).
- [27] Christopher G. Thompson, R. S. Kim, Ariel M. Aloe, and Betsy J. Becker. “Extracting the variance inflation factor and other multicollinearity diagnostics from typical regression results”. In: *Basic and Applied Social Psychology* 39.2 (2017), pp. 81–90. DOI: [10.1080/01973533.2016.1277529](https://doi.org/10.1080/01973533.2016.1277529).
- [28] Román Salmerón, Catalina Garcia, and Jose Pérez. “Overcoming the inconsistencies of the variance inflation factor: A redefined VIF and a test to detect statistical troubling multicollinearity”. In: *arXiv* (2020). DOI: [10.48550/arXiv.2005.02245](https://doi.org/10.48550/arXiv.2005.02245).
- [29] M. Edmondson, C. Luo, M. N. Islam, N. Sheils, J. Buresh, Z. Chen, J. Bian, and Y. Chen. “Distributed quasi-Poisson regression algorithm for modeling multi-site count outcomes in distributed data networks”. In: *Journal of Biomedical Informatics* 132 (2022), p. 104097. DOI: [10.1016/j.jbi.2022.104097](https://doi.org/10.1016/j.jbi.2022.104097).
- [30] Abobaker Mohamed Jaber, Mohamed Amraja Mohamed, Aissa Omar Assrhani, Hanadi Abdullah Amhimmid, and Kasem Abdinibi Farag. “Robust Modeling of Over Dispersed Count Data Using an Outlier-Weighted Poisson Regression Approach”. In: *Communications on Applied Nonlinear Analysis* 32.10s (2025), pp. 3191–3200. DOI: [10.52783/cana.v32.5959](https://doi.org/10.52783/cana.v32.5959).
- [31] A. R. Jamaludin, F. Yusof, and Suhartono. “Modelling asthma cases by count analysis approach: Poisson INGARCH and negative binomial INGARCH”. In: *Matematika* 36.2 (2020), pp. 173–184. DOI: [10.11113/matematika.v36.n2.1217](https://doi.org/10.11113/matematika.v36.n2.1217).
- [32] D. Handayani, K. Notodiputro, K. Sadik, and A. Kurnia. “A comparative study of approximation methods for maximum likelihood estimation in generalized linear mixed models (GLMM)”. In: *AIP Conference Proceedings*. Vol. 1827. 2017, p. 020033. DOI: [10.1063/1.4979449](https://doi.org/10.1063/1.4979449).
- [33] S. N. Wood. “Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73.1 (2011), pp. 3–36. DOI: [10.1111/j.1467-9868.2010.00749.x](https://doi.org/10.1111/j.1467-9868.2010.00749.x).

## Appendix

### Dataset Sources

The following datasets were used in this study:

1. Diarrhea cases: <https://opendata.jabarprov.go.id/id/dataset/jumlah-kasus-penyakit-diare-berdasarkan-kabupatenkota-di-jawa-barat>
2. Inpatient health centers: <https://opendata.jabarprov.go.id/id/dataset/jumlah-puskesmas-berdasarkan-jaringan-puskesmas-dan-kabupatenkota-di-jawa-barat>
3. Community health workers: <https://opendata.jabarprov.go.id/id/dataset/jumlah-tenaga-kesehatan-masyarakat-berdasarkan-kabupatenkota-di-jawa-barat>
4. Proper sanitation access: <https://opendata.jabarprov.go.id/id/dataset/jumlah-kepala-keluarga-dengan-akses-terhadap-fasilitas-sanitasi-yang-layak-jamban-sehat-berdasarkan-kabupatenkota-di-jawa-barat>
5. Undernourished toddlers: <https://opendata.jabarprov.go.id/id/dataset/jumlah-balita-bergizi-kurang-berdasarkan-kabupatenkota-di-jawa-barat>
6. Stunted toddlers: <https://opendata.jabarprov.go.id/id/dataset/jumlah-balita-stunting-berdasarkan-kabupatenkota-di-jawa-barat>
7. Drinking water facilities: <https://opendata.jabarprov.go.id/id/dataset/jumlah-sarana-air-minum-berdasarkan-kabupatenkota-di-jawa-barat>
8. PHBS households: <https://opendata.jabarprov.go.id/id/dataset/jumlah-rumah-tangga-dengan-perilaku-hidup-bersih-dan-sehat-phbs-berdasarkan-kabupatenkota-di-jawa-barat>