*Parameter: Journal of Statistics*
Available at https://bestjournal.untad.ac.id/index.php/parameter
Vol. 3 No. 2 2023, 54 - 60
DOI: *https://doi.org/10.22487/27765660.2023.v3.i2.16918*
eISSN: 2776-5660

# MODELING PNEUMONIA CASES IN TODDLERS IN INDONESIA USING GENERALIZED ADDITIVE MODEL FOR LOCATION, SCALE, AND SHAPE (GAMLSS) WITH LOESS SMOOTHING

**Sofya Syahar[1], Agustifa Zea Tazliqoh[2*], Harison[3]**

[1]Statistic Study Program, Tadulako University
[2]Management Study Program, Singaperbangsa Karawang University
[3]Statistic Study Program, Riau University

***e-mail***: agustifa.tazliqoh@fe.unsika.ac.id

**ABSTRACT**

*Pneumonia is an acute respiratory infectious disease that is the primary death cause due to infection in children worldwide, including Indonesia. Pneumonia case modelling is necessary to predict the incidence, especially pneumonia in toddlers. In this study, case modelling using the Generalized Additive Model for Location, Scale, and Shape (GAMLSS) method with LOESS smoothing to determine the model form and the factors influencing pneumonia cases in toddlers in Indonesia in 2021. The research results obtained indicate that with the Inverse Gaussian distribution, the model form for the location parameter is* $log(\hat{\mu}) = 9{,}719 + 0{,}013x_2 + 0{,}001x_3 + 0{,}031x_5$ *and for the scale parameter is* $log(\hat{\sigma}) = -10{,}897 + 0{,}125x_5$*. The resulting model is accurate and suitable for use because the model residuals follow a normal distribution. Along with factors that influence pneumonia cases in toddlers in Indonesia are the percentage of babies receiving exclusive breastfeeding* $(x_2)$*, population density* $(x_3)$*, and the percentage of toddlers receiving measles immunization* $(x_5)$*.*

**Keywords**: *GAMLSS, LOESS Smoothing, Pneumonia*.

**Cite:** Syahar, S., Tazliqoh, Z. A., & Harison (2023). *Modeling Pneumonia Cases in Toddlers in Indonesia Using Generalized Additive Model for Location, Scale, and Shape (GAMLSS) with Loess Smoothing*. *Parameter: Journal of Statistics*, *3*(2), 54-60, https://doi.org/10.22487/27765660.2023.v3.i2.16918

Copyright © 2023 Syahar et al. This is an open-access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

*Received: Dec 2023; Reviewed: Dec 2023; Published: Dec 2023

## INTRODUCTION

Pneumonia is an acute respiratory infection (ARI) that causes inflammation of the lung tissue (alveoli). Pneumonia is the primary cause of death due to transmission in children throughout the world (WHO, 2022). Pneumonia kills more children than any other infectious disease, claiming the lives of more than 800.000 children under the age of five or about 2.200 children every day. These include more than 153.000 newborns. Pneumonia affects more than 1.400 per 100.000 children worldwide, with the highest incidence in Central and West Africa (1.620 incidents per 100.000 children) and South Asia (2.500 incidents per 100.000 children) (Unicef, 2021).

According to WHO data in 2017, Indonesia is one of the highest countries with pneumonia burden, ranking seventh in the world. These was recorded because as many as 25.481 deaths among children under five were caused by acute respiratory infections (Edy, 2021). The Directorate General of Disease Prevention and Control reported that in 2021, the prevalence of pneumonia in children under five in Indonesia was 3,55 percent, with the percentage of pneumonia cases discovered in children under five being 31,41 percent, and the death rate due to pneumonia in children under five was 0,16 percent. It is also known that of the 2.310 deaths of children under five according to the primary causes in Indonesia, 217 deaths were caused by pneumonia cases, which is the second highest incidence after diarrhea cases (Indonesian Ministry of Health, 2022).

Regression analysis is a study that examines the relationship between a response variable and one or more predictor variables. The aim is to estimate and predict the population average or average response variable value based on the values of known predictor variables (Pertiwi, 2020). The Generalized Linear Model (GLM) is a development of the linear regression model that is capable of analyzing data with an exponential distribution, including the normal distribution. However, when the modeled response data does not meet the linearity assumption, in this case, smoothing will be carried out which is included in an additional model known as the Generalized Additive Model (GAM) (Fauziah, 2015). It is known that both models cannot model variance, skewness, and kurtosis explicitly in predictor variables but rather implicitly through dependence on $\mu$ (Wahyuni et al., 2021).

Based on this, a method called the Generalized Additive Model for Location, Scale, and Shape (GAMLSS) was developed. GAMLSS is known to be more flexible because it includes an extension of the exponential family of distributions that can handle overdispersed data, including continuous and discrete distributions of highly skewed or kurtosis response data. As a semiparametric model, GAMLSS also accommodates smoothing functions (Fauziah, 2015).

One of the smoothing functions is LOESS (Locally Estimated Scatterplot Smoothing). LOESS is a curve-smoothing approach from empirical data that provides a graphical summary of the relationship between a response variable and predictor variables. The LOESS procedure allows good flexibility because no assumptions regarding the parametric shape of the regression surface are required (Wahyuni et al., 2021).

Several previous studies, including Fauziah (2015) modeled data on the number of deaths among children under five due to pneumonia in all provinces in Indonesia (except West Java and Bengkulu Provinces) using the GAMLSS method with LOESS smoothing. The results obtained indicate that the Negative Binomial II distribution is the best distribution for modeling this data. Then, to analyze the factors that influence the incidence of pneumonia in toddlers in the city of Surabaya using nonparametric spline regression, it is known that the variables that significantly affect cases of pneumonia in toddlers are the variables of poor nutrition, measles immunization, percentage of low birth weight babies, health services toddlers, and the percentage of toddlers receiving vitamin A supplements (Nugroho, 2015). From the exposure above, this research will discuss the modeling of pneumonia cases in toddlers in Indonesia using the GAMLSS method with LOESS smoothing.

## MATERIALS AND METHODS

### Data Sources and Research Variables

The data used in this research is secondary data obtained from the Central Statistics Agency and the Ministry of Health of the Republic of Indonesia in 2021. The sample used as many as 34 provinces in Indonesia. The identification of research variables used in this research is as follows:

$Y$ : Percentage of Pneumonia Case Discovery in Toddlers
$X_1$ : Number of Low Birth Weight (LBW) Babies
$X_2$ : Percentage of Babies Receiving Exclusive Breastfeeding
$X_3$ : Population Density
$X_4$ : Percentage of Households that Have Access to Adequate Sanitation

$X_5$ : Percentage of Toddlers Receiving Measles Immunization
$X_6$ : Percentage of Coverage for Vitamin A Administration in Toddlers

**Analysis Method**

Data analysis in this study used the Generalized Additive Model for Location, Scale, and Shape (GAMLSS) with LOESS smoothing via the gamlss package provided by RStudio software. The stages of data analysis carried out are as follows:

1. Data exploration.
2. Determining the best distribution of the response variable based on a linear model of several continuous distributions, namely the Inverse Gaussian (IG), Log Normal (LOGNO), Gamma (GA), Generalized Inverse Gaussian (GIG), and Box-Cox Cole and Green (BCCG) distributions.
3. GAMLSS semiparametric modeling with LOESS smoothing:
   a. Selection of degrees of polynomials.
   b. Span selection.
   c. Model selection using step GAIC().
   d. Model diagnostic test. This test consists of a residual normal distribution test using the Shapiro-Wilk normality test.
   e. Test the significance of parameters in the model simultaneously using the likelihood ratio test and partially using the Wald test.

## RESULTS AND DISCUSSION

### Data Exploration

In this research, data exploration was carried out, namely to identify patterns of relationships between a response variable and predictor variables. This exploration will also determine what variables are modeled in parametric and nonparametric form. The following is a visualization of the relationship between a response variable and predictor variables using the scatterplotmatrix command in R.

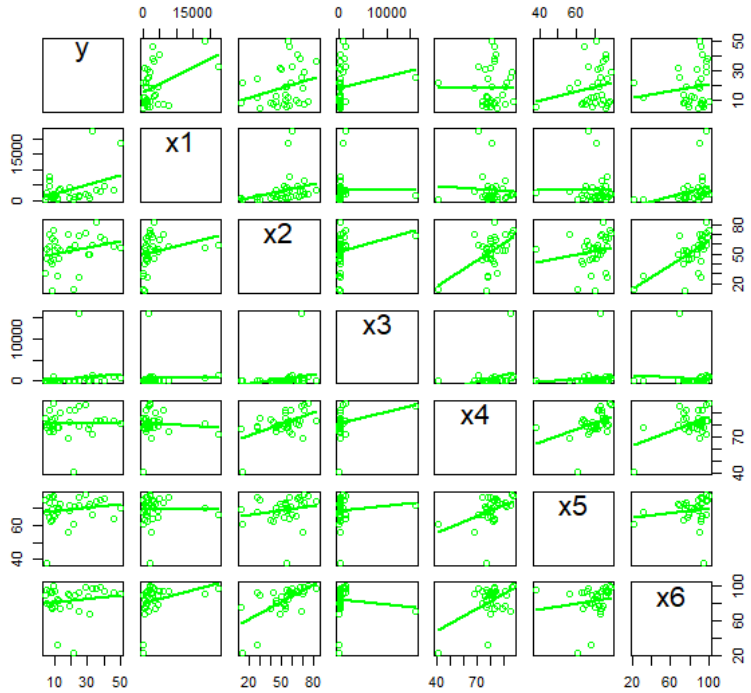


Figure 1. Scatterplot of the Relationship between Response Variables and Predictors

Figure 1 shows a scatterplot of the relationship between the variable percentage of pneumonia case discovery in toddlers ($y$) with the variables number of babies with low birth weight ($x_1$), percentage of babies receiving exclusive breastfeeding ($x_2$), population density ($x_3$), percentage of toddlers receiving measles immunization ($x_5$), and percentage coverage of vitamin A administration to toddlers ($x_6$) with a linear relationship pattern that follows a straight line and linearity that shows the right direction. This can be seen based on the green regression line in the first row of the scatterplot matrix

visualization and means that the linear relationship is positive, so variables $x_1$, $x_2$, $x_3$, $x_5$, and $x_6$ will be modeled parametrically.

Meanwhile, the relationship between the variable percentage of pneumonia case discovery in toddlers ($y$) and the variable percentage of households that have access to adequate sanitation ($x_4$) shows a relationship pattern that does not follow linearity or does not form a particular pattern, so this variable will be modeled nonparametrically. Because the variables in the data show parametric and non-parametric properties, semiparametric modeling needs to be carried out, which is then used by the GAMLSS semiparametric approach with LOESS smoothing. However, before doing GAMLSS semiparametric modeling, will first determine the distribution of the response variable based on a linear model.

**Determining the Best Distribution Based on Linear Models**

To determine the best distribution to be used in GAMLSS semiparametric modeling, the next step is to compare the AIC values for each distribution based on the linear model. The AIC value for each distribution is presented in Table 1.

Table 1. AIC Values for Linear Models in Several Distributions

| No | Distribution | $GD$ | $df$ | AIC |
|---|---|---|---|---|
| 1 | Inverse Gaussian (IG) | 242,6087 | 8 | 258,6087 |
| 2 | Log Normal (LOGNO) | 244,6661 | 8 | 260,6661 |
| 3 | Gamma (GA) | 245,0383 | 8 | 261,0383 |
| 4 | Generalized Invers Gaussian (GIG) | 244,1396 | 9 | 262,1396 |
| 5 | Box-Cox Cole and Green (BCCG) | 243,6647 | 9 | 261,6647 |

Based on Table 1, it can be seen that the Inverse Gaussian distribution has the smallest AIC value, of 258,6087 compared to other distributions. So, it can be said that the Inverse Gaussian distribution is the best distribution that will be used to analyze the response variable of the percentage of pneumonia case discovery in toddlers ($y$) and will be used in the next GAMLSS semiparametric modeling steps.

**GAMLSS Semiparametric Modeling with LOESS Smoothing**

At this stage, GAMLSS semiparametric modeling is carried out using an inverse Gaussian distribution and the predictor variable that will be estimated using LOESS smoothing is the variable percentage of households that have access to adequate sanitation ($x_4$). The modeling steps taken are as follows.

**Selection of Polynomial Degrees in LOESS Smoothing**

Selection of degree polynomial in LOESS smoothing provides two options, namely degree = 1 and degree = 2. To get the better one polynomials degree is determined from the smallest AIC value. A summary of the AIC values is presented in Table 2.

Table 2. AIC Values with The Best Degree of Polynomial

| Distribution | *Degree* | $GD$ | $df$ | AIC |
|---|---|---|---|---|
| Inverse Gaussian (IG) | 1 | 249,2256 | 16,90679 | 283,0392 |
| | 2 | 257,0224 | 16,96932 | 290,9611 |

Based on Table 2, it is known that the smallest AIC value is when selecting the degree of the polynomial which is better for the data at degree = 1 of 283,0392. Then, this value will be used in the next stage, namely selecting the span.

**Election Span in LOESS Smoothing**

According to Cleveland (1979) in Wahyuni (2021), a better span value is determined from 0,2 to 0,8. However, in this study, it was used span value from 0,6 to 0,8. This is because span values from 0,2 to 0,5 produce errors. The selection of the span value is determined from the smallest AIC value. The following is a summary of AIC values for selecting the best span.

Table 3. AIC Values with the Best Span

| Distribution | *Span* | $GD$ | $df$ | AIC |
|---|---|---|---|---|
| Inverse Gaussian (IG) | 0,6 | 249,4617 | 16,89562 | 283,2529 |
| | 0,7 | 245,4678 | 16,20112 | 277,8700 |
| | 0,8 | 248,3396 | 16,10264 | 280,5449 |

Based on Table 3, it is known that a span of 0,7 is the best span value that can represent the data distribution with the smallest AIC value, namely 277,8700. Next, after obtaining the best span value, model selection is carried out.

**Selection Using stepGAIC( )**

In this study, model selection was carried out using the stepGAIC ( ) function which aims to obtain a model with significant predictor variables. This function performs model selection in stages based on the GAIC value. The steps for selecting a model with the stepGAIC ( ) function are (a) Inputting the model to be tested and (b) Fitting the model with stepGAIC ( ) for each Inverse Gaussian distribution parameter in the form of location parameters and scale parameters. The results of the model fitting process with stepGAIC ( ) are presented in Table 4 and Table 5.

1. Model selection on scale parameters ($\sigma$)

   Model selection on the scale parameter ($\sigma$) is carried out using the model given by the location parameter, namely $y$~$x_1 + x_2 + x_3 + lo(\sim x_4, degree = 1, span = 0{,}7, family =$ "$symmetric$") $+ x_5 + x_6$. Then this model is input into the stepGAIC ( ) function to produce several models with AIC values. A summary of the AIC values of several models formed for the scale parameters is presented in Table 4.

Table 4. AIC Values of Several Models on the Parameter Scale

| No | Model | AIC |
|---|---|---|
| 1 | $y$~$x_1 + x_2 + x_3 + x_5 + x_6$ | 302,66 |
| 2 | $y$~$x_1 + x_2 + x_5 + x_6$ | 286,59 |
| 3 | $y$~$x_1 + x_2 + x_5$ | 268,76 |
| 4 | $y$~$x_2 + x_5$ | 265,75 |
| 5 | $y$~$x_5$ | 262,86 |

Based on Table 4, it can be seen that the predictor variables $x_1$, $x_2$, $x_3$, and $x_6$ are variables in the model that are gradually eliminated. From the first stage of elimination, the predictor variable that was eliminated was $x_3$, resulting in the second model, namely $y$~$x_1 + x_2 + x_5 + x_6$ with an AIC value of 286,59. This process gradually runs until the smallest AIC value with a significant predictor variable is produced. The best model for scale parameters with the smallest AIC value, namely 262,86, is model number 5 which consists of the predictor variable of the percentage of toddlers receiving measles immunization ($x_5$).

2. Model selection on the location ($\mu$) parameter

   Model selection on the location ($\mu$) parameter is done by using models given by the scale parameter, namely $y$~$x_5$. Then this model is input into the stepGAIC ( ) function to produce several models with AIC values. A summary of the AIC values of several models formed for the location parameter is presented in Table 5.

Table 5. AIC Values from Several Models on Location

| No | Model | AIC |
|---|---|---|
| 1 | $y$~$x_1 + x_2 + x_3 + lo(\sim x_4, degree = 1, span = 0{,}7, family =$ "$symmetric$") $+ x_5 + x_6$ | 262,86 |
| 2 | $y$~$x_2 + x_3 + lo(\sim x_4, degree = 1, span = 0{,}7, family =$ "$symmetric$" ) $+ x_5 + x_6$ | 259,33 |
| 3 | $y$~$x_2 + x_3 + lo(\sim x_4, degree = 1, span = 0{,}7, family =$ "$symmetric$") $+ x_5$ | 258,17 |

Based on Table 5, it can be seen that the predictor variables are $x_1$ and $x_6$ are the variables in the model that are gradually eliminated. From the first stage of elimination, the predictor variable that was eliminated was $x_1$, resulting in the second model, namely ~$x_2 + x_3 + lo(\sim x_4, degree = 1, span = 0{,}7, family =$ "$symmetric$" ) $+ x_5 + x_6$ with an AIC value of 259,33. This process gradually runs until the smallest AIC value with a significant predictor variable is produced. Model number 3, namely $y$~$x_2 + x_3 + lo(\sim x_4, degree = 1, span = 0{,}7, family =$ "$symmetric$") $+ x_5$ is a better model for location parameters with an AIC value of 258,17. The model consists of predictor variables, the percentage of babies receiving exclusive breastfeeding ($x_2$), population density ($x_3$), the percentage of toddlers receiving measles immunization ($x_5$), and the variable percentage of households having access to adequate sanitation ($x_4$) which are estimated using LOESS smoothing.

**Model Diagnostic Test**

Model diagnostic tests are carried out to determine the feasibility criteria and accuracy of the model produced in the previous stage by checking residual assumptions. This test consists of a residual normal distribution test using the Shapiro-Wilk normality test. The results are presented in Table 6.

Table 6. Shapiro-Wilk Normality Test

| **Shapiro-Wilk *Normality Test*** | ***p-value*** |
|---|---|
| 0,97991 | 0,77 |

Based on Table 6, the p-value obtained is 0,77 which is greater than the error rate of 0,05, so it fails to reject $H_0$, so it can be concluded that the residuals follow a normal distribution.

**Parameter Significance Test**

**Simultaneous Test (Likelihood Ratio Test)**

A summary of the values of the likelihood ratio test statistic is presented in Table 7.

Table 7. Likelihood Rasio Test

| $GD_0$ | $GD_1$ | ***Likelihood Rasio Test*** | ***p-value*** |
|---|---|---|---|
| 253,6645 | 223,035 | 30,6295 | 0,000158 |

Based on Table 7, the p-value obtained is 0,000158, which is smaller than the error rate of 0,05, so reject $H_0$, so it can be concluded that there is at least one predictor variable that has a significant effect on the response variable. Table 7 also shows the deviance value of the complex model (full model) namely $GD_1$ of 223,035 which is smaller than the deviance value of the simple model (reduction model) namely $GD_0$ of 253,6645, this shows that the complex model is the GAMLSS model with LOESS smoothing is better at fitting the data.

**Partial Test (Wald Test)**

The summary results of the partial significance test of parameters in the model for each parameter of the Inverse Gaussian distribution are in the form of coefficient values and link functions are presented in Table 8 and Table 9.

Table 8. Coefficient Location Parameter with Log as Link Function

| **Parameter** | ***Estimate*** | ***Std. Error*** | ***t-value*** | ***p-value*** |
|---|---|---|---|---|
| $\beta_{01}$ | 9,718503 | 0,190610 | 50,986 | $< 2 \times 10^{-16}$ |
| $\beta_{21}$ | 0,013271 | 0,002844 | 4,667 | $9{,}64 \times 10^{-5}$ |
| $\beta_{31}$ | 0,001024 | 0,000210 | 4,877 | $5{,}65 \times 10^{-5}$ |
| $\beta_{51}$ | 0,030611 | 0,002119 | 14,473 | $2{,}28 \times 10^{-13}$ |

Based on Table 8, it is known that the p-value of the variable parameters $x_2$, $x_3$, and $x_5$ is smaller than the error level of 0,05, so the rejection $H_0$ or parameters of these three variables partially have a significant effect on the response variable in the location parameter, so the model form for the location ($\mu$) parameter is obtained as follows:

$$log(\hat{\mu}) = 9{,}719 + 0{,}013x_2 + 0{,}001x_3 + 0{,}031x_5 \quad (1)$$

The model explains that for every percentage of babies receiving exclusive breastfeeding ($x_2$) increases by 1 percent, this causes the percentage of pneumonia cases discovered in toddlers to increase by *exp*(0,013) = 1,013 percent. Then, for every increase in population density ($x_3$) by 1 unit, the percentage of pneumonia cases discovered in toddlers increases by 100 percent, and every time the percentage of toddlers receiving measles immunization ($x_5$) rises by 1 percent, then the percentage of pneumonia cases discovered in toddlers increases by *exp*(0,031) = 1,031 percent.

Table 9. Scale Parameter Coefficients with Log as Link Function

| **Parameter** | ***Estimate*** | ***Std. Error*** | ***t-value*** | ***p-value*** |
|---|---|---|---|---|
| $\beta_{02}$ | -10,8969 | 2,12393 | -5,131 | $2{,}97 \times 10^{-5}$ |
| $\beta_{52}$ | 0,12524 | 0,03053 | 4,103 | 0,000405 |

Based on Table 9, it can be seen that the p-value of the variable parameter $x_5$ is smaller than the error rate of 0,05, so the reject $H_0$ or variable parameter $x_5$ has a significant effect on the response variable partially on the scale parameter, so the model form for the scale parameter ($\sigma$) is obtained as follows:

$$log(\hat{\sigma}) = \text{-}10{,}897 + 0{,}125x_5 \quad (2)$$

with estimated value $\hat{\sigma} = exp(\text{-}10{,}897 + 0{,}125x_5)$

## CONCLUSION

Based on the results and discussion previously explained, several conclusions were obtained that is the Inverse Gaussian distribution is the best distribution for modeling pneumonia cases in toddlers in Indonesia with the model form for parameter location ($\mu$) of $log(\hat{\mu}) = 9{,}719 + 0{,}013x_2 + 0{,}001x_3 + 0{,}031x_5$ and for the scale parameter ($\sigma$) of $log(\hat{\sigma}) = \text{-}10{,}897 + 0{,}125x_5$. The resulting model is accurate and suitable for use because the model residuals follow a normal distribution. The factors that influence pneumonia cases in toddlers in Indonesia are the percentage of babies receiving exclusive breastfeeding, population density, and the percentage of toddlers receiving measles immunization.

## REFERENCES

Edy, S. (2021). Save the Children : Imunisasi PCV Cegah Pneumonia the Silent Killer pada Anak. Retrieved from Website INAnews.co.id: https://www.inanews.co.id/2021/07/save-the-children-imunisasi-pcv-cegah-pneumonia-the-silent-killer-pada-anak/.

Fauziah, L. (2015). Aplikasi GAMLSS dengan Pemulusan Loess dan Algoritma Rigby-Stasinopoulos pada Data Cacahan. Skripsi. Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Jember, Jember.,

W., Nurfitra., Satriani., & Junaidi. (2022). Analisis Spasial Penyebaran Penyakit *Schistosomiasis* Menggunakan Indeks Moran Untuk Mendukung Eradikasi *Schistosomiasis* di Provinsi Sulawesi Tengah Berbasis *Web Dashborad*. *Jambura Journal Probability and Statistics*. *3*(2), 120-127.

[Kemenkes RI] Kementrian Kesehatan Republik Indonesia. (2021). Profil Kesehatan Indonesia 2021. Jakarta: Kementerian Kesehatan Republik Indonesia.

Nugroho, A. C. D. (2015). Analisis Faktor-faktor yang Mempengaruhi Pneumonia pada Balita di Kota Surabaya Menggunakan Regresi Nonparametrik Spline. Skripsi. Fakultas Matematika dan Ilmu Pengetahuan Alam, Institut Teknologi Sepuluh Nopember, Surabaya.

Pertiwi, F. (2020). Pengaruh Pendapatan Premi dan Hasil Investasi terhadap Cadangan Dana Tabarru pada Perusahaan Asuransi Syariah di Indonesia. Skripsi. Program Studi S1 Akuntansi, Sekolah Tinggi Ilmu Ekonomi Indonesia, Jakarta.

UNICEF. (2021). Pneumonia in Children Statistics. Retrieved from Website UNICEF: https://data.unicef.org/topic/child-health/pneumonia/.

Wahyuni, S. T., Utami, T. W., dan Darsyah, M. Y. (2021). Pemodelan Generalized Additive Model For Location, Scale, and Shape (Gamlss) dengan Pemulusan Locally Estimated Scatterplot Smoothing (Loess) pada Kasus Hiv/Aids Di Jawa Timur. *Jurnal Litbang Edusaintech*, *2*(1), 18-26.

[WHO] World Health Organization. (2022). Pneumonia in Children. Retrieved from Website World Health Organization: https://www.who.int/news-room/fact-sheets/detail/pneumonia.