JURNAL BIOMETRIKA DAN KEPENDUDUKAN
(Journal of Biometrics and Population)

UNDERREPORTING IN INFECTIOUS DISEASE CASES: A TIME SERIES REGRESSION-SIR APPROACH

***Emmanuel Alphonsus Akpan[1,2], Florence Gregory Ikono[3], Collins F. Udom[1], Uduakobong James Umondak[1], Onyekachukwu Henry Ikeh[1], Emem Godwin Nyong[1]**

[1]Department of Basic Sciences, Federal College of Medical Laboratory Science and Technology, 930105 Jos, Plateau, Nigeria
[2]Department of Mathematics, University of Jos, 930105 Jos, Plateau, Nigeria
[3]Department of Pharmacy, Federal Medical Centre, 569101 Yenagoa, Bayelsa, Nigeria
***Corresponding author**: Emmanuel Alphonsus Akpan; **Email:** eubong44@gmail.com

Published by *Fakultas Kesehatan Masyarakat Universitas Airlangga*

Keywords:
COVID-19,
Infectious diseases,
Quasi-Poisson distribution,
SIR model,
Time series regression

ABSTRACT

Failure to account for the underreporting of infectious disease cases distorts the understanding of infectious disease dynamics. Underreporting creates a false sense of security, allowing the disease to persist or resurge and undermining the effectiveness of public health interventions. This study aims to address underreporting and identify the underlying distribution that best describes the Coronavirus disease 2019 (COVID-19) cases in Nigeria. A Time Series Regression Susceptible-Infected-Recovered (TSIR) model, incorporating Poisson, Gaussian, and Quasi-Poisson distributions with various link functions, was applied to weekly cumulative COVID-19 case data. This dataset spans from February 28, 2020, to July 3, 2022, and includes 110 weekly records. It was sourced from the Nigerian Centre for Disease Control (NCDC) through publicly available weekly epidemiological reports. Microsoft Office Excel 2016 was utilized to collate the database, and the NCDC's online platform served as the primary data source. The data were divided into two sets: training data from February 28, 2020, to March 20, 2022, comprising 100 cases for modeling TSIR, and testing data from March 27, 2022, to July 3, 2022, encompassing 10 weekly cases for model performance evaluation. These research findings revealed that the reporting rate of COVID-19 data under study is about 35%, indicating underreporting. When accounting for underreporting, the transmission rate was reduced by approximately 0.15. The quasi-Poisson distribution with the log function was the best at describing the distribution of the incidence cases. The study established that the COVID-19 incidence cases in Nigeria are underreported and follow a quasi-Poisson distribution.

Kata Kunci:
COVID-19,
Penyakit menular,
Distribusi Quasi-Poisson,
Model SIR,
Time series regression

ABSTRAK

Kegagalan untuk memperhitungkan kurangnya pelaporan kasus penyakit menular mendistorsi pemahaman dinamika penyakit menular. Kurangnya pelaporan menciptakan rasa aman yang salah, yang memungkinkan penyakit tersebut bertahan atau muncul kembali dan melemahkan efektivitas intervensi kesehatan masyarakat. Studi ini bertujuan untuk mengatasi kurangnya pelaporan dan mengidentifikasi distribusi mendasar yang paling menggambarkan kasus penyakit Coronavirus 2019 (COVID-19) di Nigeria. Model Time Series Regression Susceptible-Infected-Recovered (TSIR), yang menggabungkan distribusi Poisson, Gaussian, dan Quasi-Poisson dengan berbagai link functions, diterapkan pada data kasus COVID-19 kumulatif mingguan. Dataset ini mencakup rentang 28 Februari 2020 hingga 3 Juli 2022, dan mencakup 110 catatan mingguan. Data ini bersumber dari Pusat Pengendalian Penyakit Nigeria atau Nigerian Centre for Disease Control (NCDC) melalui laporan epidemiologi mingguan yang tersedia untuk umum. Microsoft Office Excel 2016 digunakan untuk menyusun basis data, dan platform daring NCDC berfungsi sebagai sumber data utama. Data dibagi menjadi dua set: data pelatihan dari 28 Februari 2020 hingga 20 Maret 2022, yang terdiri dari 100 kasus untuk pemodelan TSIR, dan data pengujian dari 27 Maret 2022 hingga 3 Juli 2022, yang mencakup 10 kasus mingguan untuk evaluasi kinerja model. Temuan penelitian ini mengungkapkan bahwa tingkat pelaporan data COVID-19 yang diteliti adalah sekitar 35%, yang menunjukkan adanya underreporting. Ketika memperhitungkan underreporting, tingkat penularan berkurang sekitar 0,15. Distribusi kuasi-Poisson dengan fungsi logaritma adalah yang terbaik dalam menggambarkan distribusi kasus insiden. Penelitian ini menetapkan bahwa kasus insiden COVID-19 di Nigeria tidak dilaporkan dan mengikuti distribusi Quasi-Poisson.

Received 04 July 2024 ; **Reviewed** in 02 December 2024 ; **Accepted** in 11 June 2025 ; **p-ISSN** 2302–707X - **e-ISSN** 2540–8828 ; **DOI:** https://doi.org/10.20473/jbk.v14i1.2025.22-33 ; **Cite this as :** Akpan EA, Ikono FG, Udom CF, Umondak UJ, Ikeh OH, Nyong EG. Underreporting in Infectious Disease Cases: A Time Series Regression-Sir Approach. J Biometrika dan Kependud [Internet]. 2025;14(1):22–33. Available from: https://doi.org/10.20473/jbk.v14i1.2025.22-33

INTRODUCTION

The SIR (Susceptible-Infected-Recovered) model is one of the most widely used techniques for modelling infectious disease dynamics. This compartmental model employs ordinary differential equations to represent the time-based changes in the populations of susceptible (S), infected (I), and recovered (R) individuals. (1,2). The susceptible (S) compartment consists of individuals prone to contracting the disease at the initial stage of the outbreak and who are not yet exposed. The infected (I) compartment consists of individuals who are already infected with the infection and can pass it on to the susceptible individuals in one way or another. The recovered (R) compartment comprises individuals who have recovered, developed immunity, and are no longer vulnerable.

The Susceptible-Infected-Recovered (SIR) model is based on several assumptions: it considers a large, closed population with a short-lived outbreak, excludes natural births and deaths, assumes that individuals become infectious immediately upon infection (no latent period), grants lifetime immunity after recovery, and relies on a mass-action mixing of individuals within the population (3–6). The assumptions underlying the SIR model also apply to its variants, including the SEIR (Susceptible-Exposed-Infected-Recovered) model, the SIRD (Susceptible-Infected-Recovered-Deceased) model, the SIRV (Susceptible-Infected-Recovered-Vaccinated) model, and the SEIS (Susceptible-Exposed-Infected-Susceptible) model (3,4). Different studies have successfully applied the SIR model to study infectious diseases, as evident in the literature. Few such studies include (7–16).

Despite the successes recorded by the SIR model and its variants, the SIR model is still deficient in addressing issues bordering on the scarcity of data, quantifying the uncertainty in the model parameter, underreporting, and over-reporting (4,17). Therefore, adopting a model that can integrate observed data to estimate parameter values based on available information is essential. One model suited for this purpose is the Time Series Regression Susceptible-Infected-Recovered (TSIR) model.

The Time Series Regression Susceptible-Infected-Recovered (TSIR) model is a modified version of the SIR model that incorporates a time-varying parameter to enhance its ability to capture changes in disease transmission rates over time, based on historical data. TSIR combines the time series regression model and the SIR model, and it can perform better than either of the component models. However, the application of TSIR receives very little attention and lacks wider coverage from researchers globally in modeling infectious diseases. The studies of (1,18–22) provide successes in the application of the TSIR model in infectious disease modelling.

As outlined in (4), the TSIR model is based on two key assumptions: 1) the infectious period aligns with the data's sampling interval and remains consistent over a sufficiently long period, and 2) the sum of births and cases should be roughly equivalent owing to the significant contagiousness of pathogens before vaccines were widely available. This second assumption enables regression analysis between cumulative cases and cumulative births. The TSIR model offers a practical advantage over the SIR model due to its analytical tractability, making it particularly useful for examining infectious disease dynamics and incorporating variations in reporting rates.

The COVID-19 pandemic exemplifies an infectious disease that presents significant challenges to the global public health system (23). The post-COVID-19 era is characterized by the availability of data whose significance and usefulness cannot be undermined. For the TSIR model to work effectively, it needs reliable data on infection numbers over time, ideally broken down into consistent intervals. The COVID-19 used in this study is complete, well-structured, and consistent, all of which are essential for applying the TSIR model. With reliable weekly case counts from a trusted source, the assumption of data availability is fulfilled.

Therefore, modeling post-COVID-19 data is essential for diverse reasons; from extracting information that could aid public health planning, policy-making, vaccination strategies, and preparation for future public health emergencies to providing a foundation for ongoing research. However, understanding and controlling infectious diseases could be significantly hampered by underreporting. Underreporting comes into play when cases of an infection are not correctly recorded or not reported at all. The danger of underreporting

could be seen as the resultant effect of uncertainty, leading to significant deviations in predicting the true prevalence and spread of the infection.The fact that accurate reporting of infectious diseases is one essential way for effective disease control and prevention, this study seeks to apply the TSIR model to account for the under-reporting of COVID-19 cases in Nigeria and at the same time, identify the distribution that best describes the reported cases in which no prior studies have attempted. The choice of COVID-19 reporting data in this study was motivated by the availability of COVID-19 data and aimed to demonstrate that the TSIR model can effectively capture the complex dynamics of COVID-19.

METHODS

Research Design

This retrospective study utilized secondary surveillance data obtained from the Nigeria Centre for Disease Control (NCDC), covering the period from February 28, 2020, to July 3, 2022. The dataset comprised 110 weekly reported COVID-19 cases. For model development, the dataset was split into two subsets: the training dataset included 100 weekly observations spanning February 28, 2020, to March 20, 2022, which were used to calibrate the Time Series Susceptible-Infectious-Recovered (TSIR) model. The remaining 10 weekly observations, from March 27, 2022, to July 3, 2022, were used as the testing dataset for evaluating the model's predictive performance.

Demographic data, including annual birth rates and total population size, were obtained from the Nigerian Population Commission. These figures were interpolated to weekly values based on a COVID-19 infectious period of seven days. The analysis used a constant birth rate of 36.026 per 1,000 population and an estimated national population size of approximately 200 million.

One key limitation of this study is the exclusion of genomic surveillance data, particularly information on circulating SARS-CoV-2 variants. Since different variants may have varying levels of transmissibility, virulence, and immune evasion, their absence could influence the accuracy of the model's predictions. Furthermore, the dataset did not account for non-pharmaceutical interventions (NPIs) or vaccination coverage during the study period. The lack of these variables restricts the ability to assess the impact of control measures on disease dynamics and may contribute to unexplained variability in case counts.

Susceptible-Infectious-Recovered (SIR) Model

The mathematical representation of the SIR model is wrapped up in equations 2.1-2.3 as follows:

$$\frac{dS}{dt} = -\beta S(t)I(t), \quad 2.1$$

$$\frac{dI}{dt} = \beta S(t)I(t) - \gamma I(t), \quad 2.2$$

$$\frac{dR}{dt} = \gamma I(t), \quad 2.3$$

The initial conditions are $S(t) \geq 0, I(t) \geq 0,$ and $R(t)$ and where $S(t)$ represents the number of susceptible individuals, $I(t)$ represents the number of infected individuals, β denotes the transmission rate, and γ signifies the recovery rate [see also (8,13,20)].

Estimation of Basic Reproduction Number (R_0), Transmission Rate (β) and Recovery Rate (γ)

$$Attack\ rate\ (AR) = \frac{Total\ number\ of\ new\ cases}{Total\ population} \times 100\%. \quad 2.4$$

The attack rate (AR, which represents the percentage of the population that eventually becomes infected) is connected to the basic reproduction number by

Basic Reproduction Number,

$$R_0 = \frac{-\ln((1-AR)/S_0)}{(AR-(1-S_0))}, \quad 2.5$$

and S_0 is the proportio n of susceptible individuals (24,25).

The recovery rate, $\gamma = \frac{1}{T_i}$, 2.6

where T_i is the infection period.

The transmission rate, $\beta = \gamma R_0$, 2.7

[see (26,27) for further details.

Time Series Regression-SIR Model

The SIR model is modified as below to account for underreporting or overreporting,

$$S_t = S_{t-1} + B_{t-1} - I_t + a_t, \quad 2.8$$

$$E[I_{t+1}] = \beta_t {I^{\alpha}}_t S_t, \quad 2.9$$

where B_{t-1} represents the number of births at the previous time point, $S_t$denotes the number of susceptible hosts and I_{t+1} are the one-day-ahead forecasted number of reported cases, Equation 2.9, is the expected number of cases

in the next time step, indicating the multiplicative relationship between susceptible and infected individuals. Equation 2.8 describes the relationship for the susceptible, α allows for nonlinearities in the transmission rates. Similarly, the transmission rate, denoted by β_t, is determined by the TSIR model.

Assuming that the actual number of cases, I_t is related to the reported cases, C_t, by the equation

$$I_t = \rho_t C_t, \quad 2.10$$

where ρ_t represents the reporting rate at time, t. When $\rho_t < 1$, this indicates that the true number of cases is under-reported (20). We assume that ρ_t follows a probability function with an expected value $E(\rho_t) = \rho$. Putting equation 2.10 into 2.8 generates equation 2.11:

$$S_t = B_{t-1} + S_{t-1} - \rho_t C_t + a_t, \quad 2.11$$

Taking the expectation of equation 2.11, we have $E(S_t) = \bar{S}$ so that $S_t = \bar{S} + Z_t$ with $E(Z_t) = 0$. The deviations Z_t from their mean adhere to the same recursive relationship as S_t,

$$Z_t = B_{t-1} + Z_{t-1} - \rho_t C_t + a_t. \quad 2.12$$

Repeatedly iterating Equation 2.12 starting from the initial condition, Z_0 produces

$$Z_t = Z_0 + \sum_{i=1}^{t} B_{i-1} - \sum_{i=1}^{t} \rho_i C_i + \sum_{i=1}^{t} a. \quad 2.13$$

Equation 2.13 demonstrates that the susceptible group reflects the balance over time between new births entering and infected individuals exiting the compartment. Furthermore, without adjusting cases for the reporting rate, Z_t would not remain stable, as the gap between cumulative births and reported cases would increase indefinitely due to under-reporting.

to simplify notation, let

$$X_t = \sum_{i=1}^{t} C_i, \quad 2.14$$

$$Y_t = \sum_{i=1}^{t} B_{i-1}, \quad 2.15$$

$$A_t = \sum_{i=1}^{t} a_i, \quad 2.16$$

$$R_t = \sum_{i=1}^{t} (\rho_t - \rho)\, C_i, \quad 2.17$$

for $t = 1, \dots, n$. Then

$R_t = R_{t-1} + (\rho_t - \rho)\, C_i$ and $A_t = A_{t-1} + a_i$

are random walk processes, as their conditional mean depends on the preceding value, that is, $E(R_t|R_{t-1}) = R_{t-1}$ and $E(A_t|A_{t-1}) = A_{t-1}$, these processes may show extended periods of deviation from a mean of zero.

So far, equation 2.13 can be rearranged as

$$Y_t = -Z_0 + \rho X_t + R_t + Z_t - A_t. \quad 2.18$$

$A_t \approx 0$ for the case with minimal noise and a constant reporting rate $R_t \approx 0$ (22).Thus, equation 2.14 represents a straightforward linear regression relationship between cumulative births Y_t and cumulative cases X_t with a constant slope, ρ. The unobserved susceptible dynamics Z_t are then precisely derived as the residual of the regression.

Also, given equation 2.8, one can obtain equation 2.19 as follows:

$$log(I_{t+1}) = log(\beta_{t+1}) + \log((Z_t + \bar{S}) + \alpha log(I_t) \quad 2.19$$

The unknown parameters, β, α and $\bar{S}$ can be estimated by an equivalent generalized linear model with identity-link and log-link.

The normal and Poisson distributions are considered in this study. The relationship between the link functions for an ith case and the distributions are given as follows:

The identity Link:

$$\mu_i = X_i'\varphi, \quad 2.20$$

which is used when the error follows a normal distribution.

The Log-Link:

$$\log(\mu_i) = X_i'\varphi \quad 2.21$$

which is applied when the error follows a Poisson distribution. φ is a vector of unknown parameters [See (28)for more details].

Measurement of Prediction Performance

The performance of the TSIR model is assessed by analyzing the error distributions and their associated link functions using multiple evaluation metrics: Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Root Mean Square Error (RMSE), as defined in Equations 2.22–2.24. The distribution yielding the lowest error across these metrics is considered the most suitable.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|e_i - \hat{e}_i|, \quad 2.22$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|e_i - \hat{e}_i|}{e_i} \times 100\%, \quad 2.23$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(e_i - \hat{e}_i)^2}{n}}, \quad 2.24$$

where:

n is the number of observations in the training dataset.

e_i represents the observed case for the ith observation.

$\hat{e}_i$ represents the predicted case for the ith observation.

Elsewhere, the Lyapunov exponent will be used to quantify the impact of underreporting on the dynamics of COVID-19 reported cases [see (29–31)for details].

RESULTS

The SIR was estimated using *EpiDynamics*, TSIR by runtsir function of the TSIR package, and R_0 by R0 package. Figure 1 presents the trend of weekly COVID-19 cases.

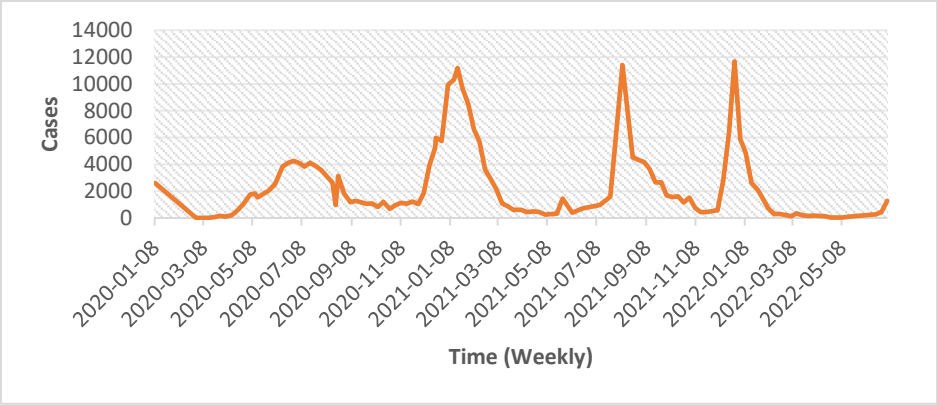

Figure 1: Plot of Weekly COVID-19 Cases

Building of SIR Model

The development of the SIR model for COVID-19 data depends on the parameter estimates shown in Table 1. The attack rate was estimated using equation (2.4) from the data, AR = 0.13, $R_0 = 1.08$ based on the attack rate obtained in equation (2.5), $\gamma = 0.143$, which is the inverse of the infection period taken to be 7 days, and $\beta = 0.15$, which is the product of γ *and* R_0. Figure 2 shows the development of S, I, and R over time. The drastic decline of the black line indicates that individuals who are not yet infected are rapidly becoming infected by week 1.05. The red line represents the number of infected individuals, reaching a maximum of 1.8 million people in week 1.05 before slowly declining until about week 10, a period that signifies nearly every individual has recovered. The green line depicts the number of individuals recovering. It increases steadily and levels off at week 10, marking the time when more individuals recover than are susceptible.

Table 1. SIR Parameters

Parameters	Values
R_0	1.08
AR	0.13
β	0.15
γ	0.143
$S(t)$	199999999
$I(t)$	1

Building Time Series Regression SIR (TSIR) Model

Analyzing the dynamics of COVID-19 incidence in Nigeria using the TSIR model, we applied a straightforward linear regression, treating cumulative births as the dependent variable and weekly cumulative COVID-19 cases as the independent variable. This approach was tested under Poisson, Gaussian, and Quasi-Poisson distributions, each with different link functions. The parameter estimates for the fitted TSIR model using the Poisson distribution with a log function, the Gaussian distribution with a log function, the quasi-Poisson with a log function, and the Gaussian distribution with an identity function are shown in Table 2. The parameter of interest, the reporting rate, is estimated to be 0.35 across the different distributions. The reporting rate less than 1 indicates possible underreporting of COVID-19 cases. $\bar{\beta}$ is the mean Transmission rate: 4.65e-08, 9.68e-08, 6.82e-08, and 3.85e-08 under Poisson distribution with log function, Gaussian distribution with log function, Quasi-Poisson distribution with log function, and Gaussian distribution with identity function, respectively, indicating that each infected individual infects less than one other individual on average. S_0 is the initial proportion of susceptible: 1.06e-01, 5.07e-02, 4.45e-02, and 4.45e-02under Poisson distribution with log function, Gaussian distribution with log function, Quasi-Poisson distribution with log

function and Gaussian distribution with identity function, respectively, indicating the average number of individuals in the population who are susceptible to infection. I_0 is the initial proportion of infected individuals: 3.52e-03, 3.52e-03, 3.52e-03, and 3.52e-03 under Poisson distribution with log function, Gaussian distribution with log function, Quasi-Poisson distribution with log function, and Gaussian distribution with identity function, respectively, indicating the fraction of the population initially infected. $\alpha = 1$ is the correction factor allowing for the mixing of the contact process, indicating that individuals have an equal chance of interacting with every other individual in the population.

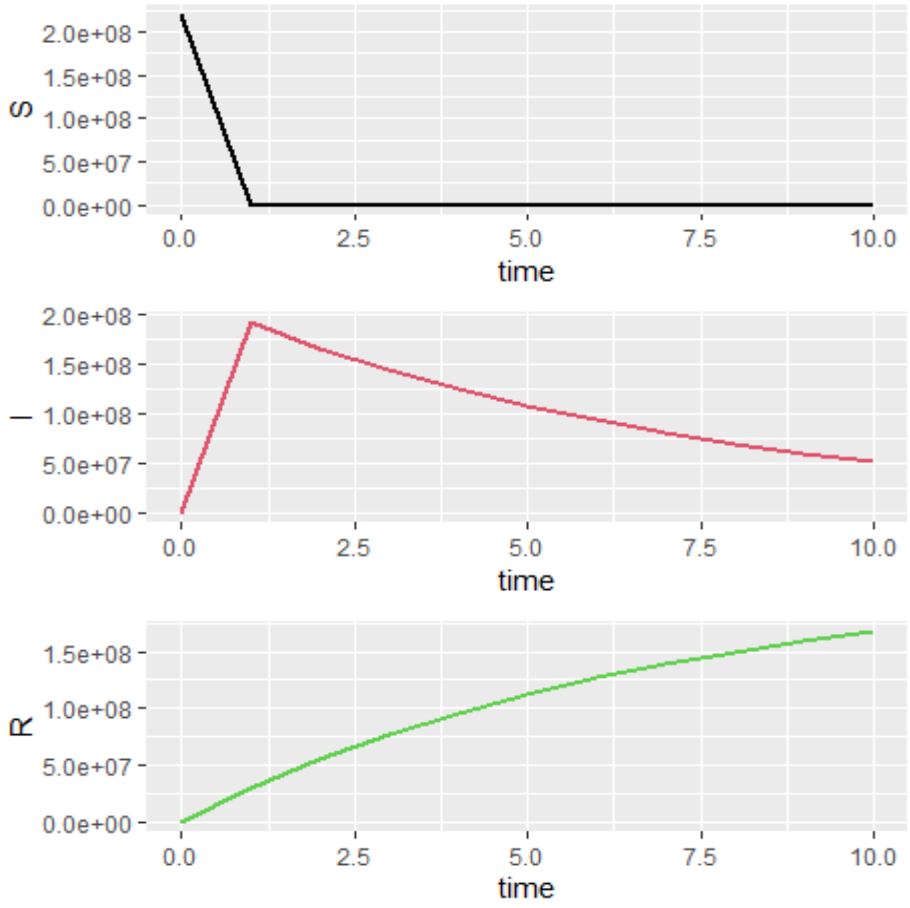

Figure 2: Plot of SIR Model Simulation

Table 2. Estimation of Parameters of Time Series Regression-SIR (TSIR) Model

Parameters	TSIR Model with Different Distributions			
	Poisson Distribution with Link = Log	**Gaussian Distribution with Link = Log**	**Quasi-Poisson Distribution with Link = Log**	**Gaussian Distribution with Link = Identity**
$\bar{\beta}$	4.65e-08	9.68e-08	6.82e-08	3.85e-08
$\bar{\rho}$	**0.35**	**0.35**	**0.35**	**0.35**
$\bar{S}$	2.15e+07	1.03e+07	9.05e+06	2.94e+07
S_0	1.06e-01	5.07e-02	4.45e-02	1.26e-01
I_0	3.52e-03	3.52e-03	3.52e-03	3.52-03
α	1.00	1.00	1.00	1.00

$\bar{\beta}$=mean Transmission rate, $\bar{\rho}$= mean reporting rate, $\bar{S}$ = mean susceptible, S_0 = initial proportion of susceptible, I_0 = initial proportion of infected and α = correction factor allowing for mixing of the contact process.

Gauging the Effect of Underreporting on Transmission Rate (β) of COVID-19

Measuring the effect of underreporting of COVID-19 cases on the transmission rate (β), we examine the value of β obtained by fitting the SIR model without accounting for underreporting, in comparison to those of TSIR with different distributions, where underreporting is properly accounted for. The findings from Table 3 reveal that when underreporting is considered, the transmission rate is significantly reduced by 0.1499999535, 0.1499999032, 0.1499999318, and 0.1499999615 under the Poisson distribution with a log function, the Gaussian distribution with a log function, the quasi-Poisson distribution with a log function, and Gaussian distribution with an identity function, respectively.

Table 3. Effect of Underreporting on Transmission Rate (β)

	Poisson Distribution with Link = Log		**Gaussian Distribution with Link = Log**		**Quasi-Poisson Distribution with Link = Log**		**Gaussian Distribution with Link = Identity**	
	SIR	**TSIR**	**SIR**	**TSIR**	**SIR**	**TSIR**	**SIR**	**TSIR**
β	0.15	4.65e-08	0.15	9.68e-08	0.15	6.82e-08	0.15	3.85e-08
Difference/bias	-0.1499999535		-0.1499999032		-0.1499999318		-0.1499999615	

The Performance of the TSIR Model Using Different Distributions

Determining the distribution that most accurately represents COVID-19 cases involves analyzing the error criteria for 10 weeks of simulated data and 10 weeks of forecasted cases. Table 4 shows that the TSIR model using the quasi-Poisson distribution with the log function outperforms the Poisson distribution with the log function and the Gaussian distribution with both the log and identity functions, as evidenced by the lowest MAPE, MAE, and RMSE values for the simulated cases. It is important to note that the high values of MAPE, MAE, and RMSE in this study indicate a significant discrepancy between the reported incidence cases (which are relatively low) and the cases adjusted for the 35% underreporting rate. Comparable results are observed for the predicted cases, as demonstrated in Table 5.

Figure 3 is a 10-week simulation using the quasi-Poisson distribution with a log function. The plot shows that the number of cases indicated by the brown line rises and falls in a wave-like pattern, likely due to seasonal changes, public behaviour, or shifting policies. Despite the fluctuations, the overall average, indicated by the blue line, remains fairly steady, suggesting no strong upward or downward trend in the long run. However, over time, the gray shaded area around the line, representing uncertainty, gets wider, especially after mid-2020. This reflects growing uncertainty in future projections due to model variability and structural noise. Despite the steady average trend, the model captures oscillating case patterns, which may result from seasonality, behavioural changes, or policy shifts. These features highlight the model's utility in forecasting and the importance of incorporating variability and uncertainty into public health planning.

Figure 4 presents predictions for 10 weeks ahead using the quasi-Poisson distribution with a log function. The plot shows two significant epidemic waves, one around mid-2020 and another around mid-2021. Both peaks are sharp and intense, with predicted cases spiking above 6 million, indicating very rapid and severe outbreaks. The narrow nature of these peaks highlights how cases surge quickly and then decline just as fast, pointing to short-lived but intense transmission periods. This pattern likely reflects sudden shifts, such as the virus spreading rapidly through a susceptible population or swift changes in public behaviour or policy. Between the waves, predicted case numbers drop close to zero, resembling the natural dynamics of an epidemic: surge, decline, pause. The use of a quasi-Poisson model addresses the irregular and bursty nature of epidemic data, where events seldom follow smooth patterns. These sudden waves could be due to herd immunity, policy interventions, environmental changes, or behavioural shifts that temporarily suppress transmission.

Table 4. Measurement of 10 Weeks Ahead Simulation Performance using Different Distributions

Simulation Performance Criteria	Poisson Distribution with Link = Log	Gaussian Distribution with Link = Log	Quasi-Poisson Distribution with Link = Log	Gaussian Distribution with Link = Identity
MAPE	274.094	274.172	74.676	274.115
MAE	259.441	259.528	73.956	259.478
RMSE	146.441	214.840	103.833	214.840

Table 5. Measurement of 10 Weeks Ahead Prediction Performance using Different Distributions

Prediction Performance Criteria	Poisson Distribution with Link = Log	Gaussian Distribution with Link = Log	Quasi-Poisson Distribution with Link = Log	Gaussian Distribution with Link = Identity
MAPE	127.428	250.854	60.143	251.203
MAE	162.992	268.217	62.59	271.183
RMSE	173.237	277.794	82.647	281.698

Measuring the Impact of Underreporting on COVID-17 Disease Dynamics

The effect of underreporting was measured using the Lyapunov exponent for both simulated and predicted cases, as shown in Table 6. Given that the TSIR model with a quasi-Poisson distribution and log function best captures the underlying dynamics of the reported data, the Lyapunov exponent for this model is considered. The Lyapunov exponent value is 0.014 for both simulation and prediction.

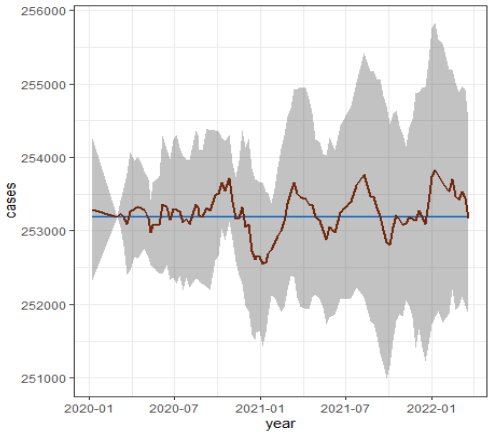

Figure 3. Plots of 10 Weeks Simulation based on fitted TSIR Model using Quasi Poisson with Link = Log

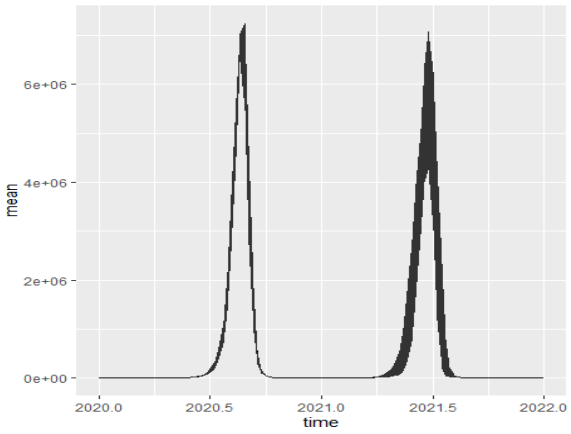

Figure 4. Plots of 10 Weeks ahead Prediction based on fitted TSIR Model using Quasi Poisson with Link = Log

Table 6. Impact of Underreporting on COVID-19 Disease Dynamics

Lyapunov Exponent	Data			
	Poisson Distribution with Link = Log	Gaussian Distribution with Link = Log	Quasi-Poisson Distribution with Link = Log	Gaussian Distribution with Link = Identity
Simulation	0.069	-0.00008	0.014	-0.0009
Prediction	0.032	0.0026	0.014	-0.0019

DISCUSSION

Going forward, we have simulated the traditional SIR model based on initial parameters as shown in Table 1, where the basic reproduction number, $R_0 = 1.08$. However, the SIR model is often useful at the outbreak stage of the infection, but it loses its potency with the availability of data and its inability to account for underreporting. According to the data, the estimated basic reproduction number, $R_0 > 1$, indicates the likelihood of the infection's spread (25). Sufficing the study's aim, the TSIR model was applied to the data, yielding a reporting rate of 0.35, below 1, indicating that the data used are underreported [see Table 2]. The transmission rate decreases when accounting for underreporting in the TSIR model compared to the SIR model due to underreporting. That is, from β = 0.15 to 4.65e-08, 9.68e-08, 6.82e-08, and 3.85e-08 under Poisson distribution with log function, Gaussian distribution with log function, quasi-Poisson distribution with log function, and Gaussian distribution with identity function, respectively. Contrary to the findings of (20), which suggests that failing to account for underreporting results in an underestimation of the transmission rate, β, and that directly applying an SIR model to raw underreported incidence leads to an underestimated contact rate. Our study shows that, due to significant underreporting, the transmission rate is substantially underestimated when compared to directly fitting the SIR model to the raw, underreported cases. This means that underreporting could introduce bias and result in misrepresenting the disease spread.

The underlying distribution that best describes the data is the quasi-Poisson distribution. In other words, there are elements of over-dispersion in the data that imply potential sources of unaccounted variation, further supporting the case that the data is significantly underreported. This study examined the temporal dynamics of predicted case counts using a quasi-Poisson regression model with a log link function, utilizing both 10-week-ahead simulations and forecasts. The model was specifically selected to address overdispersed count data, a frequent characteristic in epidemiological surveillance, while allowing for flexible, nonlinear growth patterns through the log link transformation. In the simulation, we observed relatively stable mean case counts over time, with consistent fluctuations around the central trend. These oscillations likely reflect underlying periodic dynamics such as seasonality, behavioural changes, or intervention effects. The widening confidence intervals over time indicate increasing prediction uncertainty, a common feature in extended forecasting. This aligns with the findings of (32,33), who emphasized the significance of probabilistic forecasts in conveying the uncertainty inherent in epidemic projections. In contrast, the 10-week-ahead prediction revealed two pronounced and sharply defined epidemic waves, peaked around mid-2020 and mid-2021. The predicted case counts exceeded six million at the peak of each wave, followed by rapid declines to near-zero levels. This pattern suggests intense but short-lived outbreaks, possibly driven by sudden increases in transmission followed by the depletion of the susceptible population or rapid implementation of control measures. Such behaviors have been modeled in previous studies of respiratory virus transmission, where nonlinearity and intervention timing produced wave-like dynamics with narrow peaks (34,35).

The impact of underreporting is well quantified by indicating the Lyapunov exponent of 0.014, indicating that underreporting leads to irregular and volatile dynamics of COVID-19 spread (30). The implication is that underreporting, if not accounted for, could bias the model parameters, distort the accuracy of predictions of future outbreaks, lead to a false sense of security, allow the disease to persist or resurge, and hinder the effectiveness of public health interventions (19).

The application of the TSIR model in this study shares some similarities with the studies of (1,18–22,36) However, it differs in identifying that the dynamics of COVID-19 in Nigeria can be robustly predicted using quasi-Poisson distributions.

The contribution of this study, in its originality, is underpinned by the possible combination of the time series regression model and the SIR model, with probabilistic frameworks incorporated to account for the uncertainty in COVID-19 data. This approach makes the TSIR model more robust and tractable than either the time series regression model or the SIR model alone in capturing the complex dynamics of infectious diseases, particularly COVID-19 infections.

CONCLUSIONS AND SUGGESTIONS

Conclusions

Overall, the dynamics of COVID-19 in Nigeria appear irregular and volatile, largely due to underreporting. These patterns can be more appropriately captured using a Quasi-Poisson Time Series Susceptible-Infectious-Recovered (TSIR) model, which accounts for overdispersion in the data. The inclusion of post-COVID-19 data is especially valuable, as it provides insight into the fluctuating and often unpredictable nature of transmission patterns. Neglecting the issue of underreporting can introduce significant bias in estimating the transmission rate and may obscure important signals, leading to unreliable predictions of outbreak behavior. Therefore, to accurately assess and interpret infectious disease dynamics, it is critical to ensure consistent, transparent, and comprehensive reporting of cases.

Suggestions

The weaknesses of this study include the failure to account for vaccination interventions and other forms of public health interventions, as well as the limitation of examining a single homogeneous strain of COVID-19, which overlooks the different variants of COVID-19. Therefore, it is recommended that further studies be conducted to address the limitations of this study.

ACKNOWLEDGMENT

The insightful comments and suggestions from both the reviewers and the editor, which have helped improve the manuscript, are greatly appreciated.

AUTHOR CONTRIBUTIONS

EAA supervised, conducted formal analysis and carried out validation, FGI wrote the original draft and processed data, CFU and UJU wrote the main manuscript text. OHI and EGN developed the methodology. All authors reviewed the manuscript.

REFERENCES

1. Singh RA, Lal R, Kotti RR. Time-discrete SIR model for COVID-19 in Fiji. Epidemiol Infect [Internet]. 2022 Apr 7 [cited 2024 Oct 31];150:e75. Available from: https://doi.org/10.1017/S0950268822000590
2. Lee TJ, Kakehashi M, Rao ASRS. Network models in epidemiology. Handb Stat [Internet]. 2021;44:235–56. Available from: https://doi.org/10.1016/bs.host.2020.09.002
3. Cooper I, Mondal A, Antonopoulos CG. A SIR model assumption for the spread of COVID-19 in different communities. Chaos, Solitons & Fractals [Internet]. 2020;139. Available from: https://doi.org/10.1016/j.chaos.2020.110057
4. Becker AD, Grenfell BT. tsiR: An R package for time-series Susceptible-Infected-Recovered models of epidemics. PLoS One [Internet]. 2017 Sep 28 [cited 2024 Oct 31];12(9):e0185528. Available from: https://doi.org/10.1371/journal.pone.0185528
5. Martignoni MM, Raulo A, Linkovski O, Kolodny O. SIR+ models: accounting for interaction-dependent disease susceptibility in the planning of public health interventions. Sci Rep [Internet]. 2024 Jun 5 [cited 2024 Nov 2];14(1):12908. Available from: https://doi.org/10.1038/s41598-024-63008-9
6. Chen J, Lewis B, Marathe A, Marathe M, Swarup S, Vullikanti AKS. Individual and collective behavior in public health epidemiology. In 2017 [cited 2024 Nov 2]. p. 329–65. Available from: https://doi.org/10.1016/bs.host.2017.08.011
7. Duan Y. Simulations of the COVID-19 epidemic in Nigeria using SIR model. J Phys Conf Ser [Internet]. 2021 Apr 1 [cited 2024 Oct 31];1893(1):012016. Available from: https://iopscience.iop.org/article/10.1088/1742-6596/1893/1/012016

8. Silva A da. Modeling COVID-19 in Cape Verde Islands - An application of SIR model. Comput Math Biophys [Internet]. 2021;9(1). Available from: https://doi.org/10.1515/cmb-2020-0114
9. Law KB, Peariasamy KM, Ibrahim HM, Abdullah NH. Modelling infectious diseases with herd immunity in a randomly mixed population. Sci Rep [Internet]. 2021;11(1). Available from: https://doi.org/10.1038/s41598-021-00013-2
10. Liu T, Bai Y, Du M, Gao Y, Liu Y. Susceptible-Infected-Removed Mathematical Model under Deep Learning in Hospital Infection Control of Novel Coronavirus Pneumonia. J Healthc Eng [Internet]. 2021;(1). Available from: https://doi.org/10.1155/2021/1535046
11. Bahari MF, Utami R, Rosyida A. SIR Model for the Spread of Tuberculosis in Kudus Regency. World Sci News [Internet]. 2022;163:128–38. Available from: https://worldscientificnews.com/sir-model-for-the-spread-of-tuberculosis-in-kudus-regency/
12. Gounane S, Barkouch Y, Atlas A, Bendahmane M, Karami F, Meskine D. An adaptive social distancing SIR model for COVID-19 disease spreading and forecasting. Epidemiologic Methods [Internet]. 2021 Mar 3 [cited 2024 Oct 31];10(s1). Available from: https://doi.org/10.1515/em-2020-0044
13. Espinosa P, Quirola-Amores P, Teran E. Application of a Susceptible, Infectious, and/or Recovered (SIR) Model to the COVID-19 Pandemic in Ecuador. Frontiers (Boulder) [Internet]. 2020;6. Available from: https://doi.org/10.3389/fams.2020.571544
14. Tolles J, Luong T. Modeling Epidemics With Compartmental Models. JAMA - J Am Med Assoc [Internet]. 2020;323(24):2515–6. Available from: https://doi.org/10.1001/jama.2020.8420
15. Akinwumi T, Adegboyegun B. Modelling an Infectious Disease Prediction and control using S-I-R Model. Math Theory [Internet]. 2015;5(4). Available from: https://www.iiste.org/Journals/index.php/MTM/article/view/21581
16. Rao F. Dynamics Analysis of a Stochastic SIR Epidemic Model. Abstr Appl Anal [Internet]. 2014;2014(1). Available from: https://doi.org/10.1155/2014/356013
17. Hong HG, Li Y. Estimation of time-varying reproduction numbers underlying epidemiological processes: A new statistical tool for the COVID-19 pandemic. PLoS One [Internet]. 2020;15(7). Available from: https://doi.org/10.1371/journal.pone.0236464
18. Trejo I, Hengartner NW. A modified Susceptible-Infected-Recovered model for observed under-reported incidence data. PLoS One [Internet]. 2022;17(2). Available from: https://doi.org/10.1371/journal.pone.0263047
19. Taimoor M, Ali S, Shah I, Muwanika FR. COVID-19 Pandemic Data Modeling in Pakistan Using Time-Series SIR. Comput Math Methods Med [Internet]. 2022;2022(1). Available from: https://doi.org/10.1155/2022/6001876
20. Katris C. A time series-based statistical approach for outbreak spread forecasting: Application of COVID-19 in Greece. Expert Syst Appl [Internet]. 2021;166. Available from: https://doi.org/10.1016/j.eswa.2020.114077
21. Ajbar A, Alqahtani RT, Boumaza M. Dynamics of a COVID-19 Model with a Nonlinear Incidence Rate, Quarantine, Media Effects, and Number of Hospital Beds. Symmetry (Basel) [Internet]. 2021;13(6). Available from: https://doi.org/10.3390/sym13060947
22. Shanbehzadeh M, Orooji A, Kazemi-Arpanahi H. Comparing of Data Mining Techniques for Predicting In-Hospital Mortality Among Patients with COVID-19. J Biostat Epidemiol [Internet]. 2021;7(2). Available from: https://jbe.tums.ac.ir/index.php/jbe/article/view/504
23. Alenezi MN, Al-Anzi FS, Alabdulrazzaq H, Alhusaini A, Al-Anzi AF. A study on the efficiency of the estimation models of COVID-19. Results Phys [Internet]. 2021;26. Available from:

https://doi.org/10.1016/j.rinp.2021.104370
24. Chen Y, Aldridge T, Ferraro C, Khaw FM. COVID-19 outbreak rates and infection attack rates associated with the workplace: a descriptive epidemiological study. BMJ Open [Internet]. 2022 Jul 18 [cited 2024 Nov 2];12(7):e055643. Available from: doi: https://doi.org/10.1136/bmjopen-2021-055643
25. Houben RMAC, Maanen K van, Kemp-Symonds JG, Waller AS, Oldruitenborgh-Oosterbaan MMS van, Heesterbeek H. Estimation of the basic reproduction number for Streptococcus equi spp. equi outbreaks by meta-analysis of strangles outbreak reports. Equine Vet J [Internet]. 2022;5(3):50–514. Available from: https://doi.org/10.1111/evj.13865
26. Alenezi MN, Al-Anzi FS, Alabdulrazzaq H, Alhusaini A, Al-Anzi AF. A study on the efficiency of the estimation models of COVID-19. Results Phys [Internet]. 2021 Jul [cited 2024 Oct 31];26:104370. Available from: https://doi.org/10.1016/j.rinp.2021.104370
27. A. Iboi E, Sharomi O, N. Ngonghala C, B. Gumel A. Mathematical modeling and analysis of COVID-19 pandemic in Nigeria. Mathematical Biosciences and Engineering [Internet]. 2020 [cited 2024 Oct 31];17(6):7193–221. Available from: https://doi.org/10.3934/mbe.2020369
28. Moffat I, Akpan E. A Probabilistic application of generalized linear model in discrete-time stochastic series. J Sci Res Rep [Internet]. 2018 Jun 5 [cited 2024 Oct 31];19(3):1–9. Available from: https://doi.org/10.9734/JSRR/2018/40909
29. Chen H, Haus B, Mercorelli P. Extension of SEIR compartmental models for constructive lyapunov control of COVID-19 and analysis in terms of practical stability. Mathematics [Internet]. 2021 Aug 27 [cited 2024 Oct 31];9(17):2076. Available from: https://doi.org/10.3390/math9172076
30. Matouk AE. Complex dynamics in susceptible-infected models for COVID-19 with multi-drug resistance. Chaos Solitons Fractals [Internet]. 2020 Nov [cited 2024 Oct 31];140:110257. Available from: https://doi.org/10.1016/j.chaos.2020.110257
31. Mahayana D, Anwari S, Sulistyo B, Rahman FN, Natanael DP. lyapunov stability analysis of COVID-19 SIR modeling. International Journal on Electrical Engineering and Informatics [Internet]. 2021 Mar 31 [cited 2024 Nov 3];13(1):73–86. Available from: DOI: 10.15676/ijeei.2021.13.1.4
32. Marshall M, Parker F, Gardner LM. When are predictions useful? A new method for evaluating epidemic forecasts. BMC Global and Public Health [Internet]. 2024 Oct 3[cited 2025 Jun 05] ;2(1):67. Available from: doi:10.1186/s44263-024-00098-7
33. Held L, Meyer S, Bracher J. Probabilistic forecasting in infectious disease epidemiology: the 13th Armitage lecture. Stat Med [Internet]. 2017 Sep 30 [cited 2025 Jun 05];36(22):3443–60. Available from: doi:10.1002/sim.7363
34. Kissler SM, Tedijanto C, Goldstein E, Grad YH, Lipsitch M. Projecting the transmission dynamics of SARS-CoV-2 through the postpandemic period. Science (1979) [Internet]. 2020 May 22 [cited 2025 Jun 05];368(6493):860–8. Available from: doi:10.1126/science.abb5793
35. Camacho A, Kucharski AJ, Funk S, Breman J, Piot P, Edmunds WJ. Potential for large outbreaks of Ebola virus disease. Epidemics [Internet]. 2014 Dec 9 [cited 2025 Jun 05];9:70–8. Available from: doi:10.1016/j.epidem.2014.09.003
36. 'Grenfell B, 'Bjørnstad O, 'Finkenstadt B. Dynamics of Measles Epidemics: Scaling Noise, Determinism, and Predictability with the TSIR Model. Ecol Monogr [Internet]. 2002 [cited 2024 Nov 3];72(2):185–202. Available from: https://doi.org/10.1890/0012-9615(2002)072[0185:DOMESN]2.0.CO;2