**FMDB** *Transactions on Sustainable Health Science Letters*

# Predictive Modelling of COVID-19 Diagnosis: Logistic Regression Analysis of SARI and RdRp Testing Parameters

**K. Anitha[1,*], S. Silvia Priscila[2], S. Belina V. J. Sara[3], Gnaneswari Gnanaguru[4], M. Sakthivanitha[5]**

[1]Department of Mathematics, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Chennai, Tamil Nadu, India.
[2]Department of Computer Science, Bharath Institute of Higher Education and Research, Chennai, Tamil Nadu, India.
[3]Department of Computer Applications, Faculty of Science and Humanities, SRM Institute of Science and Technology, Kattankulathur, Chennai, Tamil Nadu, India.
[4]Department of Computer Applications, CMR Institute of Technology, Bengaluru, Karnataka, India.
[5]Department of Information Technology, Vels Institute of Science Technology and Advance Studies, Chennai, Tamil Nadu, India.
k_anitha@ch.amrita.edu[1], silviaprisila.cbcs.cs@bharathuniv.ac.in[2], sbelinav@srmist.edu.in[3], gnaneswari@yahoo.com[4], sakthivanithamsc@gmail.com[5]

**Abstract:** The logistic regression analysis in the present study explores the relationship between COVID-19 test outcomes and some of the principal clinical indicators. The main point was made upon SARI and RdRp confirmatory testing. It also considered patients across five discrete categories, age-stratified according to WHO criteria, into four groups. The methodology used a two-stage analytical approach: first, Chi-Square tests to establish associations between sociodemographic variables and the COVID-19 test results, followed by logistic regression to develop predictive models. Statistical analysis revealed distinct patterns in both SARI and RdRp patient groups. In SARI patients, gender was not significantly related to COVID-19 test results. However, RdRp analysis revealed high correlations with age, and some demographic factors are statistically insignificant. The obtained results were a basis for creating the equations of the predictive models through logistic regression. This work is within the research scope on COVID-19 testing dynamics and characteristics of patients. This study identifies key indicators of a positive test result, helping doctors assess risk and manage patients. This research may aid clinical decision-making and resource allocation in COVID-19 testing facilities with predictive models.

**Keywords:** Logistic Regression; Predictive Modelling; Univariate and Bivariate Model; Real-Time Tracking; Viral Spread; QML Algorithms-E-QSVM; DL-Based Models; Real-Time Tracking.

**Received on:** 20/06/2024, **Revised on:** 05/09/2024, **Accepted on:** 28/10/2024, **Published on:** 03/12/2024

**Journal Homepage:** https://www.fmdbpub.com/user/journals/details/FTSHSL

**DOI:** https://doi.org/10.69888/FTSHSL.2024.000276

**Cite as:** K. Anitha, S. S. Priscila, S. B. V. J. Sara, G. Gnanaguru, and M. Sakthivanitha, "Predictive Modelling of COVID-19 Diagnosis: Logistic Regression Analysis of SARI and RdRp Testing Parameters," *FMDB Transactions on Sustainable Health Science Letters*., vol.2, no.4, pp. 221–230, 2024.



## 1. Introduction

[*]Corresponding author.

As of October 4, 2023, WHO has reported a total of 771,151,224 cases of COVID-19 so far, along with 6,960,783 deaths; India has reported 44,998,838 confirmed cases of COVID-19 and 532,032 deaths; analysis of data is the most crucial component in the knowledge and fight against the COVID-19 disease. These measures relate to real-time tracking of viral spread, tracing localized outbreaks, assessing interventions, drawing epidemiological information, demographic trend analysis, anticipating future trends, and maximizing resource utilization. Applications based on data from COVID-19 tests would involve various statistics and machine-learning techniques. Machine learning is important in the statistical analysis of COVID-19 test data. Published studies have applied different machine-learning techniques to improve diagnosis accuracy. For instance, QML algorithms E QSVM and QRF have been used on the data sets of COVID-19, with high accuracies ranging from 78% by E-QSVM to 75% by QRF [12]. Laboratory findings-based DL-based and ML-based models have also been developed to identify COVID-19, the former outperforming the latter.

The LSTM model reached the highest accuracy at 96.78% with an F1-score of 96.58% [13]. Such approaches depict the effectiveness of machine learning in the analysis of COVID-19 testing data towards its proper and efficient diagnosis. These are time series analysis, spatial analysis, correlation analysis, epidemiological models, predictive modelling, classification, and anomaly detection. Predictive modelling is one of the major aspects of disease data analysis as it holds utility in forecasting case trends and optimizing resource allocation in settings where care is being dispensed. It guides decisions related to vaccine distribution; it forms a basis for assessing risks; it evaluates how interventions would work; in the meantime, it monitors variant presence. This paper introduces applications of predictive modelling to data from the COVID-19 pandemic. A predictive model for the COVID-19 test data is developed using some machine learning techniques like XGBoost, LightGBM, and random forest [14]-[16].

In addition, a multi-criteria text mining model was designed along with an integration of a temporal predictive classification model for rural underserved areas of COVID-19 testing results. The model used a dataset of 6895 testing appointments and 14 features, and the most important factors for classification are patient history, age, testing reasons, and time [16]. The researchers integrated these models to predict COVID-19 severity and risk factors associated with severity from laboratory and clinical data, highlighting the importance of features such as high-sensitivity C-reactive protein (hs-CRP) [14]. These models provide effective tools for early screening, diagnosis, and retrospective studies of infectious diseases. Predictive modelling based on COVID-19 data involves the application of mathematical and statistical methods for analyzing past information as well as current data that is available regarding the COVID-19 pandemic. Its ultimate purpose is to create predictions about future trends and results, which may be helpful in healthcare decisions, public health sectors, and policymaking circles. Many research projects have utilized predictive modeling with data coming from COVID-19. For instance, it would take time series models to predict infections shortly by some place utilizing historical case information along with vaccination statistics.

It further allows researchers to tinker with their methods of machine learning algorithms for comparing the strengths and effectiveness of various public health interventions as measures of lockdown or contact tracing. In reaction to the COVID-19 outbreak in late 2019, tens of thousands of MDSSs have been designed and developed to aid disease identification and predict its intensity. As described in references [1]-[5], several predictive models are developed.

## 2. Background and Related works

Sharma et al. [6] developed an analytical machine learning-based predictive model that explored three critical dimensions of the COVID-19 pandemic: gender, global growth rate, and social distancing. The traditional classifiers, along with innovative ensemble techniques like bagging, feature-based ensemble, voting, and stacking, were proposed in the developed analytical model. Zhang et al. [7] proposed a hybrid model for COVID-19 prediction using the Auto-Regressive and LSTM techniques. The data came from Japan, Canada, Brazil, Argentina, Singapore, Italy, and the UK. The results for the hybrid model indicated the capability of producing accurate predictions while providing insights into the pattern of phases of virus transmission. Thus, the hybrid model holds the potential to enlighten and guide the policymaking process in public health. Deng et al. [8] bridged these knowledge gaps by determining the major laboratory and clinical markers of COVID-19. Then, it exploits machine learning models to differentiate patients suffering from COVID-19 from patients with CAP. These results will likely be useful for screening, diagnosis, and monitoring cases of COVID-19 based on large datasets of routine clinical data.

At present, several different types of predictive models have been developed for the analysis of COVID-19 data. Regression stands out among these models as a statistical technique closely related to predictive modelling. At its core, predictive modelling is constructing models to make predictions or estimates based on input data. Regression is a core technique in predictive modelling through which predictions regarding continuous target variables are made. In this paper, we propose using regression analysis as an integral part of our approach toward predictive modelling. SARI represents a condition known as severe acute respiratory illness, which could be SARS-CoV-2.

The definition of SARI helped significantly in prioritization during COVID-19 surveillance and testing, mainly in the early stages when testing capacity is very low. Trends in monitoring SARI helped the public health officer evaluate the potential healthcare burden caused by acute cases of severe COVID-19. This allowed better allocation of the available resources and planning related to healthcare. Studies that focus on the risk factors associated with SARI among confirmed cases of COVID-19 will also be able to present predictors of disease progression, helping high-risk groups targeted for more effective interventions. On a molecular level, the critical role of RNA-dependent RNA polymerase enzyme encoded by the SARS-CoV2 virus is replication of this virus, hence, forms the most vital target during the diagnosis process.

Detecting the RdRp gene and other virus indicators is used as a confirmed diagnostic tool for COVID-19 in real-time RT-PCR tests, which since recently have been the gold standard for identifying infections. This paper presents a modelling approach to predictive analytics by COVID-19 testing data that analyses the clinical presentation of SARI in conjunction with the RdRp detection status toward an improved understanding of the severe disease dynamics and more precise diagnosis. The coupling of clinical and molecular data affords a broad framework of studies into severe cases of COVID-19 and optimizing test protocols.

Regression analysis is a statistical method that studies the relationship between one or more independent or predictor variables and a dependent or response variable. This analytic method can be expressed as an equation containing regression coefficients. These numerical measures describe the strength and direction of the relationship between each independent and dependent variable. They state how variations in the independent variables cause variations in the dependent variable, keeping all other variables constant. Various regression models exist, such as simple and multiple linear regression, logistic regression, polynomial regression, and regularization methods like ridge and lasso regression. The type and nature of the data at hand determine the choice of a regression model.

**3. Materials and Methods**

In this experiment, we applied a logistic regression model. Logistic regression is helpful when the dependent variable is either binary or categorical and if one intends to describe the probability of an observation belonging to a certain class or category. It is a method of statistics that is quite flexible and applicable in numerous contexts for classification as well as prediction of probability. The logistic or sigmoid function is the mathematical basis of the logistic regression model (Figure 1). This function models the probability that a binary outcome variable, denoted as Y, assumes the value 1 (Yes), contingent on a given set of predictor variables $\mathfrak{X}_1, \mathfrak{X}_2, \ldots \mathfrak{X}_p$ Concerning the following equation.

$$P(\mathcal{Y} = 1) = \frac{1}{1+e^{-(\beta_0 + \beta_1 \mathfrak{X}_1 + \beta_2 \mathfrak{X}_2 + \cdots + \beta_p \mathfrak{X}_p)}} \quad (1)$$

Here $\beta_0$ - Intercept term, $\beta_1, \beta_2, \ldots \beta_p$ are Coefficients associated with the predictor variables? $\mathfrak{X}_1, \mathfrak{X}_2, \ldots \mathfrak{X}_p$ and also these coefficients quantify the effect of each predictor variable on the log odds of the outcome.

The logistic function $\frac{1}{1+e^{-(\beta_0 + \beta_1 \mathfrak{X}_1 + \beta_2 \mathfrak{X}_2 + \cdots + \beta_p \mathfrak{X}_p)}}$ Maps the linear combination of the predictor variables and coefficients to a value between 0 and 1. This value represents the probability of the positive class (1). The logistic regression model undergoes training to deduce the optimal values for its coefficients $(\beta_0, \beta_1, \ldots \beta_p)$ That aligns best with the provided data. These coefficients are determined through methods like maximum likelihood estimation. Following the model's training phase, it is equipped to make probability predictions for new data points and categorize them into one of two classes, relying on a specified threshold (e.g., 0.5).

For samples labelled as '0', our objective is to estimate β in a manner that maximizes the product of probabilities approaching 0, or in simpler terms, to make (1 - p($\mathfrak{X}$)) as close to 1 as possible, which can be depicted as

$$0 - Labelled\ samples = f_1(\mathfrak{X}) = \prod_{y_i(s)=1} 1 - p(\mathfrak{X}_i) \quad (2)$$
$$1 - Labelled\ samples = f_2(\mathfrak{X}) = \prod_{y_i(s)=1} 1 - p(\mathfrak{X}_i) \quad (3)$$

The optimized likelihood objective function is $Max\ \beta = f_1(\mathfrak{X}) * f_2(\mathfrak{X})$

$$Max\ \beta = \prod_{y_i(s)=1} 1 - p(\mathfrak{X}_i) * \prod_{y_i(s)=1} 1 - p(\mathfrak{X}_i)$$

Take log-likelihood of $\beta$ which is denoted as

$$\mathcal{L}(\beta) = \sum_{i=1}^{n} y_i \beta \mathfrak{X}_i - log\left(1 + e^{\beta \mathfrak{X}_i}\right) \quad (4)$$

Almalki et al. [9] adopted geographic information systems and spatial analysis techniques to research the relationship of COVID-19 cases with food outlets. They applied cluster analysis in conjunction with overlaying spatial extent maps of food outlet clusters on those of infected cases. After mapping, they analyzed the closeness of these clusters and calculated the spatial variations that had been studied, correlated, and explored the probable causal relationships between them.

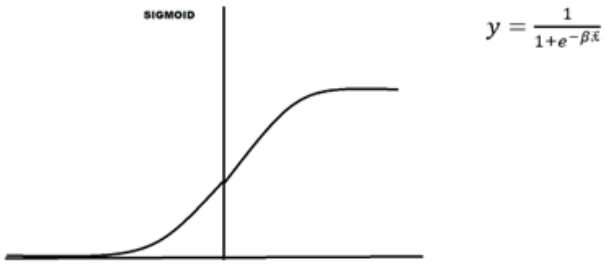


**Figure 1:** Sigmoid Function

Khan et al. [10] proposed the efficiency of machine learning in predicting outbreaks. It evaluates a variety of regression models, including linear, polynomial, ridge, and polynomial ridge regression, as well as support vector regression, based on COVID-19 data obtained from different online sources. A more aggressive experimental approach test of performance for every test is accepted by the various statistical parameters MSE, MAE, RMSE, and the $R^2$ score, which ascertain the precision and reliability of any result. Of these contributions to the field, researchers by Nopour et al. [11] have recently succeeded in developing a standardized model of diagnosis for COVID-19 by using a set of 400 patients from referrals at Ayatollah Talleghani Hospital within the city of Abadan in Iran. The method made use of a chi-square correlation coefficient to determine the choice of feature used in selecting clinical variables to be used within the set. Then, logistic regression analysis was conducted to establish the association of those clinical characteristics and their predictive powers. The study utilized SPSS Version 25.

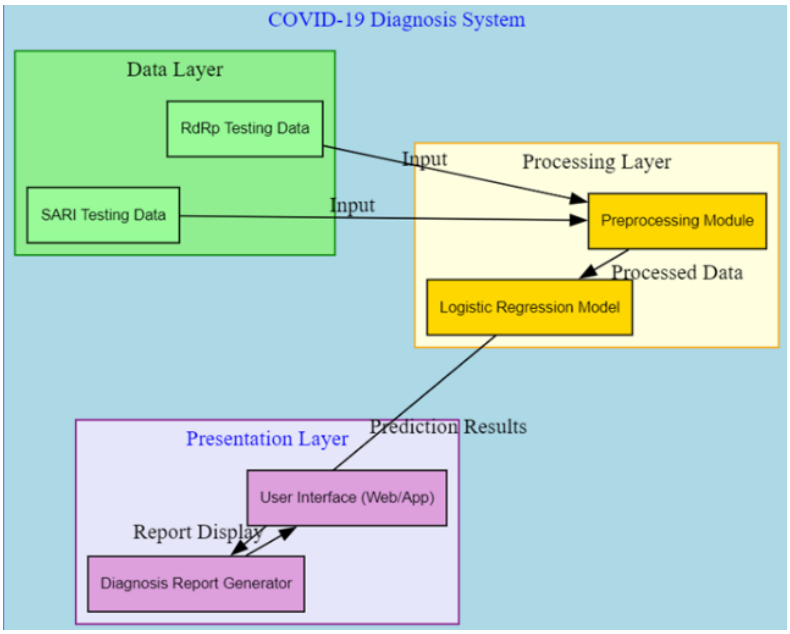


**Figure 2:** Predictive modelling of COVID-19 diagnosis using logistic regression with SARI and RdRp testing parameters

The possible factors under examination comprised various factors, including patients' history, findings from physical examination, laboratory tests, and radiological investigations. It is a more substantial basis for an exact and effective diagnosis model with the multi-dimensional data intended to integrate in this study to identify meaningful predictors of diagnosis for COVID-19. This approach would amalgamate statistical measures and machine learning techniques that emphasize systematic

data analysis to understand the clinical presentation of COVID-19 better, thus underlying potential towards improved diagnostic accuracy and apprising healthcare interventions in appropriate ways tailored accordingly. This integrated approach to feature selection and regression modelling validated clinical and laboratory parameters as potential predictors. It demonstrated the capabilities of incorporating traditional statistical methods and modern computational tools toward addressing diagnostic challenges in the form of complex diseases during such a health crisis. This paper uses logistic regression to analyze COVID-19 test results with SARI and RdRp confirmatory results.

This section utilizes univariate and multivariate logistic regression to develop predictive models of SARI presentation and RdRp gene detection from the COVID-19 testing data. This process starts with chi-square tests to check the variables for association and then moves on to univariate logistic regression. The most significant variables from the univariate analysis are considered for the multivariate logistic regression models. Finally, the equations of the best predictive model, including these variables, are presented. Figure 2 displays a three-tier architecture of the Data, Processing, and Presentation layers. The input sources of the system are SARI and RdRp testing datasets, while green nodes represent raw data storage within the Data Layer. In the Processing Layer, yellow nodes represent a Preprocessing Module and a Logistic Regression Model. These modules will clean, transform, and generate predictions from raw data.

The Logistic Regression Model is given preprocessed data and shall employ statistical techniques to compute the diagnosis result of COVID-19. The Presentation Layer is in lavender, represented as the Diagnosis Report Generator, accompanied by a user interface, whether web-based or application-based, so the user may interact with the system and observe the result. Arrows indicate data flow: testing data goes into preprocessing, the model takes processed data, makes predictions, generates a report, and the report is shown from the interface. The layering of the architecture is built in a way that gives it modularity. All elements are divided into data handling, computation, and user interaction. This ensures that the system is expandable and easy to preserve. A model uses multiple layers for colourization with interconnects that help draw the logical flow of data required to make predictions from an efficient COVID-19 diagnosis.

## 4. Result and Discussion

For this analysis, a data set of 433 patients was used, containing all patients with Severe Acute Respiratory Infection (SARI) and COVID-19-positive results. Chi-Square $\chi^2$) test was run to check the association between the dependent variable (Test Result- COVID-19 Positive/Negative) and independent variables. Variables with a p-value less than 0.05 in the chi-square test are considered for the univariate test. A univariable test is conducted with the variables under control with a reference group, and variables that have significant values less than 0.20 are considered in the Multivariate logistic regression to find the risk factors. Table 1 summarizes the details of the patient categories below.

**Table 1:** Patient Category

| Category Code of Testing Samples | Patient Category Code- Description |
|---|---|
| N1 | All symptomatic (ILI symptoms) cases. |
| N3 | All asymptomatic high-risk individuals. |
| N4 | All symptomatic (ILI symptoms) individuals with a history of international travel in the last 14 days. |
| N9 | All patients of Severe Acute Respiratory Infection (SARI). |
| N10 | All symptomatic (ILI symptoms) patients presenting in a healthcare setting. |

### 4.1. Chi-Square Test for SARI Variable Associations

The Chi-Square test shows the association between the Sociodemographic variables and COVID test results. Hence, the test results show that the age category is highly significant, and the SARI shows statistical significance with a p-value of 0.008 with 95% CI. Gender is insignificant, with a p-value of 0.156 CI (95%), which cannot be taken for further analysis (Table 2).

**Table 2:** Chi-Square Test for SARI

| **Variable** | **COVID-19 Test Result** | | **P** |
|---|---|---|---|
| **Age (years)** | **Positive (%)** | **Negative (%)** | **Value** |
| **Below 14($A_1$)** | 71(3.1) | 2183(96.9) | **0.000*** |
| **15 to 24($A_2$)** | 127(2.9) | 4311(97.1) | |
| **25 to 64($A_3$)** | 644(5.8) | 10477(94.2) | |
| **Above 65($A_4$)** | 140(13) | 934(87) | |

| | | | |
|---|---|---|---|
| **Gender** | | | |
| **Male** | 512(5) | 9641(95) | 0.156 |
| **Female** | 470(5.4) | 8264(94.6) | |
| **SARI** | | | |
| **Positive** | 34(7.9) | 399(92.1) | **0.008*** |
| **Negative** | 948(5.1) | 17506(94.9) | |
| * Statistical significance | | | |

Figure 3 stratifies the distribution by age, gender, and SARI status. The highest percentage of positive results is seen in the above 65 years of age group at 13%, followed by that of the 25-64 age group at 5.8%. The lowest positive rates are below 14 (3.1%) and 15 to 24 (2.9%) age groups. Gender analysis shows a slightly higher positive rate in females at 5.4% compared to males at 5%. SARI status is also significantly associated with the test results, as shown by 7.9% positivity in those with SARI compared to 5.1% in those without SARI.

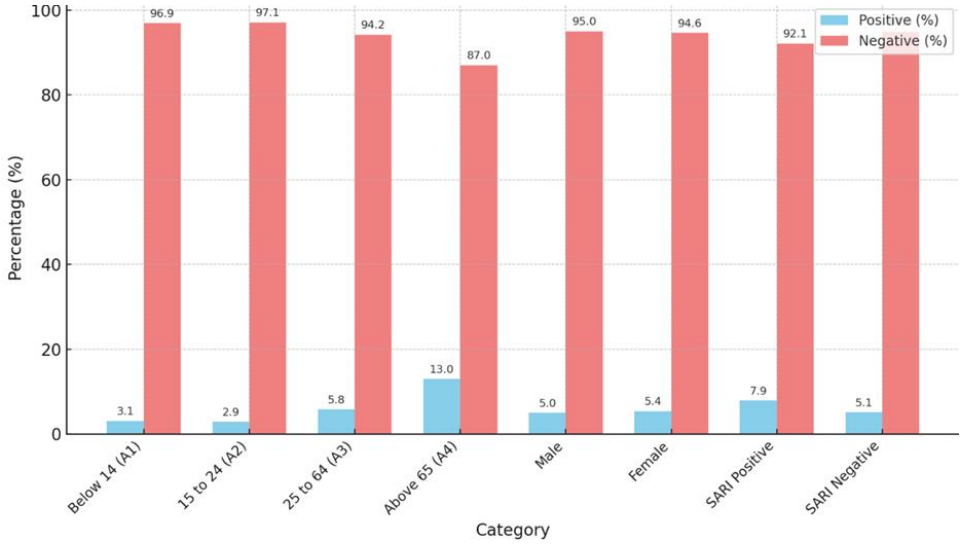


**Figure 3:** COVID-19 test results distribution across demographics and conditions

The table points out statistically significant differences in test positivity by age, p=0.000, and SARI status, p=0.008, which indicate meaningful Association with COVID-19 test results but not for gender, p=0.156. The findings highlight the importance of demographics and clinical variables in assessing the risk of COVID-19.

### 4.2. Predictive Modelling of SARI Analysis

**Univariate Logistic Regression:** Table 3 shows the results of univariate logistic regression. Both age category and SARI show statistical significance so that we can proceed with the multivariate logistic regression.

**Table 3:** Univariate Modeling of SARI Data

| Variable | OR (95%CI) | $p-$ Value |
|---|---|---|
| **Age (years)** | | |
| **Below 14 ($A_1$)** | 1(Reference) | |
| **15 to 24($A_2$)** | 0.91 (0.67 - 1.22) | 0.511 |
| **25 to 64($A_3$)** | 1.89 (1.47 - 2.43) | **0.000*** |
| **Above 65($A_4$)** | 4.61 (3.43 - 6.19) | **0.000*** |
| **SARI (N9)** | | |
| **Negative** | 1(Reference) | |
| **Positive** | 1.57 (1.10 - 2.25) | **0.013*** |
| * Statistical significance **OR**-Odds Ratio | | |

**Multivariate Logistic Regression:** Above 65 years old, people tend to have a higher risk of being affected by COVID-19 with an odds ratio of 4.70 with a CI (3.50 – 6.33) compared with the reference group (below 14 years old) and also shows high statistical significance (Table 4 ). The age groups between 25 and 64 have an odds ratio of 1.92 with a CI (1.50 - 2.47) compared with the reference group (below 14 years old), and they also show high statistical significance.

**Table 4:** Multivariate Modeling of SARI Data

| **Variable** | **AOR (95%CI)** | **p −Value** |
|---|---|---|
| **Age (years)** | | |
| **Below 14 ($A_1$)** | 1(Reference) | |
| **15 to 24($A_2$)** | 0.92 (0.68 - 1.24) | 0.58 |
| **25 to 64($A_3$)** | 1.92 (1.50 - 2.47) | **0.000*** |
| **Above 65($A_4$)** | 4.70 (3.50 - 6.33) | **0.000*** |
| **SARI (N9)** | | |
| **Negative** | 1(Reference) | |
| **Positive** | 1.73 (1.21 - 2.48) | **0.000*** |
| ***** Statistical significance **AOR**- Adjusted Odds Ratio | | |

The age groups between 25 and 64 have an odds ratio of 0.92 with a CI (0.68 - 1.24) compared to the reference group (below 14 years old, but they are not statistically significant since they are the healthy age group people). The people infected by severe acute respiratory infection have an odds ratio of 1.73 with a CI (1.21-2.48), which is also highly statistically significant compared to the reference group that does not have SARI.

Model Equation

$$log\ (p/1-p)\ =\ y \quad (5)$$
$$log\ (p/1-p)\ =\ \beta o\ +\ \beta xi \quad (6)$$
$$y\ =\ 0.315\ +\ (1.92)\ A_3\ +\ (4.70)(A_4)+\ (1.70)\ N9 \quad (7)$$

$A_2$ is not significant; hence, it can be rejected.

**Chi-Square Test for RDRP gene-variable association:** The Chi-Square test indicates the relationship between Sociodemographic variables and test results for COVID-19. Thus, test results suggest that the age category is statistically significant, and the corresponding N4, N10, & N9 have a value. For gender and N1, it is not significant at all, with a p-value >0.05 CI of 95%, and, hence, it cannot be used for further analysis (Table 5).

**Table 5:** Chi-Square test for RDRP gene

| **Variable** | **Final test result** | | **$p$ −Value** |
|---|---|---|---|
| **Age category** | **Negative** | **Positive** | |
| **Below 14 ($A_1$)** | 567(95.13) | 29(4.87) | **0.000*** |
| **15 to 24($A_2$)** | 1659(97.36) | 45(2.64) | |
| **25 to 64($A_3$)** | 4246(94.31) | 256(5.69) | |
| **Above 65($A_4$)** | 371(86.28) | 59(13.72) | |
| **Gender** | | | |
| Male | 3664(94.87) | 198(5.13) | 0.309 |
| Female | 3179(94.33) | 191(5.67) | |
| **N1** | | | |
| Negative | 6473(94.52) | 375(5.48) | 0.122 |
| Positive | 370(96.35) | 14(3.65) | |
| **N4** | | | |
| Negative | 6586(94.45) | 387(5.55) | **0.001*** |
| Positive | 257(99.23) | 2(0.77) | |
| **N10** | | | |
| Negative | 6809(94.66) | 34(87.18) | **0.039*** |
| Positive | 384(5.34) | 5(12.82) | |
| **SARI** | | | |

| | | | |
|---|---|---|---|
| Negative | 6808(94.77) | 376(5.23) | **0.000*** |
| Positive | 35(72.92) | 13(27.08) | |

### 4.3. Predictive Modelling of RDRP Gene Analysis

**Univariate Modelling of RDRP Gene Analysis:** Table 6 shows the outcome of univariate modelling for RdRp, with the odds ratios (ORs) and 95% confidence intervals (CIs) of several factors related to RdRp status. Age group: Compared to the younger age group of 0 to 14 years (Reference group), the group between 15 to 24 years, A_2 is significantly less likely to be positive for RdRp, with an OR of 0.53 (95% CI: 0.33 - 0.85; p = 0.009). The OR for those in the age group 25-64 years (A_3) is 1.18 (95% CI: 0.80 - 1.75), with a p-value of 0.413, where there was no significant relationship with the status of RdRp. Above 65 years old (A_4), chances of having a positive reaction are high, with OR 3.11 (95% CI: 1.96 - 4.94; p < 0.0001).

**Table 6:** Univariate Modeling of RdRp

| **Variable** | **OR (95%CI)** | $p$ **−Value** |
|---|---|---|
| **Age category** | | |
| **Below 14 ($A_1$)** | 1 (reference) | |
| **15 to 24($A_2$)** | 0.53 (0.33 - 0.85) | **0.009*** |
| **25 to 64($A_3$)** | 1.18 (0.80 - 1.75) | 0.413 |
| **Above 65($A_4$)** | 3.11 (1.96 - 4.94) | **0.000*** |
| **N4** | | |
| Negative | 1 (reference) | |
| Positive | 0.13 (0.03 - 0.53) | **0.005*** |
| **N10** | | |
| Negative | 1 (reference) | |
| Positive | 2.61 (1.01 - 6.70) | **0.047*** |
| **SARI** | | |
| Negative | 1 (reference) | |
| Positive | 6.73 (3.53 - 12.82) | **0.000*** |

Regarding N4, those who test positive have significantly reduced odds of testing positive for RdRp, with an OR of 0.13 (95% CI: 0.03-0.53; p = 0.005). For N10, a positive test result is associated with a positive likelihood of RdRp, with an OR of 2.61 (95% CI: 1.01-6.70; p = 0.047). Lastly, for SARI, those with a positive SARI status are much more likely to be positive for RdRp, OR 6.73 (95% CI: 3.53 - 12.82; p < 0.0001). Significant factors influencing positivity for RdRp include age, especially over 65 years, N4 and N10 statuses, and SARI positivity.

### 4.4. Multivariate Modelling of RDRP Gene Analysis

**Table 7:** Multivariate Modelling of RdRp

| **Variable** | **AOR (95%CI)** | $p$ **−Value** |
|---|---|---|
| **Age category** | | |
| **Below 14 ($A_1$)** | 1 (reference) | |
| **15 to 24($A_2$)** | 0.55 (0.34 - 0.88) | 0.013* |
| **25 to 64($A_3$)** | 1.24 (0.83 - 1.84) | 0.296 |
| **Above 65($A_4$)** | 3.10 (1.94 - 4.95) | 0.000* |
| **N4** | | |
| **Negative** | 1 (reference) | |
| **Positive** | 0.13 (0.03 - 0.53) | 0.004* |
| **N10** | | |
| **Negative** | 1 (reference) | |
| **Positive** | 1.89 (0.72 - 5.00) | 0.198 |
| **SARI (N9)** | | |
| **Negative** | 1 (reference) | |
| **Positive** | 6.16 (3.20 - 11.87) | 0.000* |

N10 doesn't show statistical significance with the p-value of 0.198 with 95%CI

Model equation:

$$\text{Log}\,(P/1-P) = y$$
$$\text{Log}\,(P/1-P) = \beta o + \beta xi$$
$$y = 0.49 + (0.55)A_2 + (3.10)A_4 + (0.13)N4 + (6.16)N9$$

Age in $\boldsymbol{A_3}$ It is not significant, and N10 is also not significant, so we reject them from the equation. Above 65 years old, people tend to have a higher risk of being affected by COVID-19 with an odds ratio of 3.10 with a CI (1.94 – 4.95) compared with the reference group (below 14 years old) and also shows high statistical significance. The age groups between 25 and 64 have an odds ratio of 1.24 with a CI (0.83 – 1.84) compared with the reference group (below 14 years old). Still, they are not statistically significant since they are the healthy age group people (Table 7). The age groups between 15 and 24 have an odds ratio of 0.55 with a CI (0.34–0.88) compared with the reference group (below 14 years old), and they also show high statistical significance. Logistic regression predictive modelling allows us to effectively predict patients' risk levels based on their categories, which will help in early diagnosis.

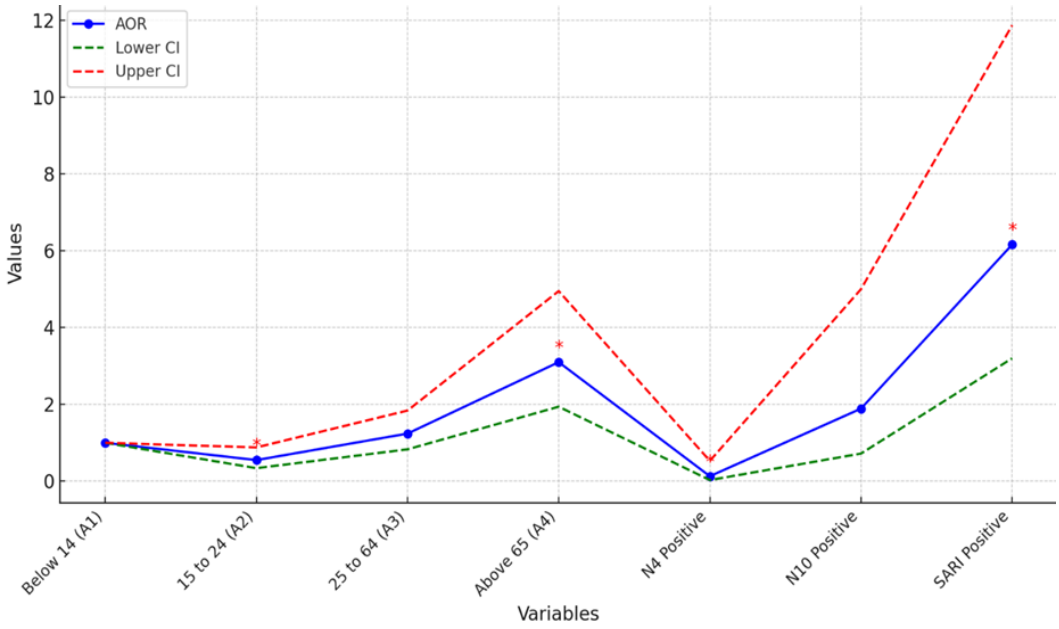


**Figure 4:** Adjusted odds ratios highlighting significant COVID-19 risk factors.

Figure 4 presents the Adjusted Odds Ratios (AOR) with 95% Confidence Intervals (CI) for various variables affecting COVID-19 outcomes. Age categories show a significant increase in odds for individuals above 65 years (AOR: 3.10, CI: 1.94–4.95, p=0.000), while the 15 to 24 group has a reduced odds ratio (AOR: 0.55, CI: 0.34–0.88, p=0.013), both statistically significant. However, the 25 to 64 categories are insignificant (AOR: 1.24, p=0.296). For N4 status, positivity significantly reduces odds (AOR: 0.13, CI: 0.03–0.53, p=0.004), whereas N10 positivity has no significant effect (AOR: 1.89, p=0.198). SARI positivity substantially increases odds (AOR: 6.16, CI: 3.20–11.87, p=0.000), highlighting its strong association with outcomes. This data underscores the influence of age, specific test results, and SARI status on COVID-19 risk, with significant variables emphasized by their p-values, offering critical insights into demographic and clinical risk factors.

## 5. Conclusion

This study demonstrates age-related differences in susceptibility to COVID-19, and it also throws open the SARI and RdRp gene analysis to assess the risk for COVID-19. The study portrays an age-dependent risk gradient, with individuals older than 65 being most susceptible, showing 4.70 times higher odds of COVID-19 compared with children aged under 14 years. The working age population (25-64 years) also had increased risks, with odds of 1.92 times compared to the reference group, whereas adolescents and young adults of 15-24 years had similar risk profiles to children. SARI was the most important risk factor. The odds of infection of the person affected with COVID-19 were 1.73 times more than that of the reference group. The analysis of the RdRp gene further supported this age-related risk pattern, except that the magnitude varied, especially in that the risk was far lower for the 15 to 24 age group than for children less than 14 years. The clinical and public health implications are important. Logistic regression models developed here give practising health professionals practical tools to apply to risk stratification and patient management. These models, including age and SARI status variables, allow for a better estimation of

risk in COVID-19, thus allowing earlier interventions and proper resource utilization within healthcare facilities. Further work should be done on validation across other populations and investigating further risks. This knowledge may improve COVID-19 screening and targeted prevention based on specific groups, such as the elderly and SARI patients.

**Acknowledgment:** We extend our heartfelt gratitude to our co-authors for their invaluable guidance and support.

**Data Availability Statement:** The data supporting the findings of this study are available upon request from the corresponding author.

**Funding Statement:** This manuscript and research paper were prepared without any financial support or funding.

**Conflicts of Interest Statement:** The authors declare no conflicts of interest related to this study.

**Ethics and Consent Statement:** This study was approved by the institutional review board, with informed consent obtained from all participants.

## References

1. V. Yury, D. A. Kistenev, E. E. Vrazhnov, and H. Shnaider, "Predictive models for COVID-19 detection using routine blood tests and machine learning," Heliyon, vol. 8, no. 10, pp. 1–11, 2022.
2. M.A. Callejon-Leblic, R. Moreno-Luna, A. Del Cuvillo, I.M. Reyes-Tejero, M.A. Garcia-Villaran, M. Santos-Pena, J.M. Maza-Solano, D.I. Martín-Jimenez, ~ J.M. Palacios-Garcia, C. Fernandez-Velez, et al., "Loss of smell and taste can accurately predict COVID-19 infection: a machine-learning approach," J. Clin. Med. vol. 10, no. 4, p. 570, 2021.
3. I. Arpaci, S. Huang, M. Al-Emran, M.N. Al-Kabi, M. Peng, "Predicting the COVID-19 infection with fourteen clinical features using machine learning classification algorithms," Multimed. Tool. vol. 80, no. 1, pp. 11943–11957, 2021.
4. Y. Wan, H. Zhou, and X. Zhang, "An interpretation architecture for deep learning models with the application of COVID-19 diagnosis," Entropy (Basel), vol. 23, no. 2, p. 204, 2021.
5. A. Imran, I. Posokhova, H.N. Qureshi, U. Masood, M.S. Riaz, K. Ali, C.N. John, M.I. Hussain, M. Nabeel, "AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app," Inform. Med, Elsevier publications, Amsterdam, Netherlands, 2020.
6. S. Sharma, I. Alsmadi, R. S. Alkhawaldeh, and B. Al-Ahmad, "Data-driven analysis and predictive modeling on COVID-19," Concurr. Comput., vol. 34, no. 28, p. 7390, 2022.
7. Y. Zhang, S. Tang, and G. Yu, "An interpretable hybrid predictive model of COVID-19 cases using autoregressive model and LSTM," Sci. Rep., vol. 13, no. 1, p. 6708, 2023.
8. X. Deng et al., "Building a predictive model to identify clinical indicators for COVID-19 using machine learning method," Med. Biol. Eng. Comput., vol. 60, no. 6, pp. 1763–1774, 2022.
9. A. Almalki, B. Gokaraju, Y. Acquaah, and A. Turlapaty, "Regression analysis for COVID-19 infections and deaths based on food access and health issues," Healthcare (Basel), vol. 10, no. 2, p. 324, 2022.
10. M. A. Khan, R. Khan, F. Algarni, I. Kumar, A. Choudhary, and A. Srivastava, "Performance evaluation of regression models for COVID-19: A statistical and predictive perspective," Ain Shams Eng. J., vol. 13, no. 2, p. 101574, 2022.
11. R. Nopour, M. Shanbehzadeh, and H. Kazemi-Arpanahi, "Using logistic regression to develop a diagnostic model for COVID-19: A single-center study: A single-center study," J. Educ. Health Promot., vol. 11, no. 1, p. 153, 2022.
12. A. N. Hussein, S. V. A. Makki, and A. Al-Sabbagh, "Comprehensive study: Machine learning approaches for COVID-19 diagnosis," International Journal of Power Electronics and Drive Systems/International Journal of Electrical and Computer Engineering, vol. 13, no. 5, p. 5681, 2023.
13. J. -B. Lee, S. C. Kim, J. -H. Lee and H. -Y. Jo, "A prognostic model for classification of COVID-19 severity based on clinical and laboratory testing data," 2023 IEEE International Conference on Big Data and Smart Computing Big, Jeju, South Korea, 2023.
14. P. Wegner, G. M. Jose, V. Lage-Rupprecht, S. G. Khatami, B. Zhang, S. Springstubbe, M. Jacobs, T. Lindén, C. Ku, B. Schultz, M. Hofmann-Apitius, and A. T. Kodamullil, "Common data model for COVID-19 datasets," Bioinformatics, vol. 38, no. 12, pp. 5466-5468, 2022.
15. L. Böttcher, M. R. D'Orsogna, and T. Chou, "A statistical model of COVID-19 testing in populations: effects of sampling bias and testing errors," medRxiv, vol. 380, no. 11, pp. 1-14, 2021.
16. A. Lekham, Y. Wang, E. Hey, and M. T. Khasawneh, "Multi-criteria text mining model for COVID-19 testing reasons and symptoms and temporal predictive model for COVID-19 test results in rural communities," Neural Computing & Applications, vol. 34, no. 10, pp. 7523–7536, 2022.