

Research Article

Identification of Prognostic Factors Associated with RT-PCR–Confirmed COVID-19 Using Logistic Regression in a Controlled Setting

Ravi Vajala¹, Gurprit Grover², Chandra Bhan Yadav³

¹PhD, MSc, Principal of Sri Venkateswara College, University of Delhi, Delhi, India

²PhD, MSc, Department of statistics, Faculty of Mathematical Science, University of Delhi, Delhi, India

³Assistant Professor, Department of statistics, Hindu College, University of Delhi, India

DOI: <https://doi.org/10.24321/0019.5138.202650>

I N F O

Corresponding Author:

Chandra Bhan Yadav, Department of statistics,
Hindu College, University of Delhi, India

E-mail Id:

chandrabhansitare@gmail.com

Orcid Id:

<https://orcid.org/0000-0002-5021-7051>

How to cite this article:

Vajala R, Grover G, Yadav C B. Identification of Prognostic Factors Associated with RT-PCR–Confirmed COVID-19 Using Logistic Regression in a Controlled Setting. J Commun Dis. 2026;58(3):18-26.

Date of Submission: 2026-02-28

Date of Acceptance: 2026-09-08

A B S T R A C T

Managing asymptomatic patients with communicable diseases poses significant diagnostic and public health challenges. Understanding factors associated with infection status is essential for risk stratification and targeted interventions in closed populations. This cross-sectional study analysed 472 individuals from an Indian army hospital who underwent RT-PCR testing for COVID-19 during a single time frame. A logistic regression framework assessed associations between demographic/clinical factors (sex, age, diabetes, vaccination, obesity, and asymptomatic status) and COVID-19 positivity. Significant factors associated with COVID-19 positivity included diabetes (OR = 2.21, 95% CI: 1.10–4.43, $p = 0.025$), vaccination (OR = 0.43, 95% CI: 0.28–0.66, $p < 0.001$, 57% reduced odds), and asymptomatic status (OR = 0.113, 95% CI: 0.05–0.25, $p < 0.0001$, 89% lower odds). Age showed modest association (OR = 1.17 per year, 95% CI: 1.05–1.31, $p = 0.004$), while sex and obesity were not statistically significant. The interaction model did not significantly improve fit over the main effects model (LRT $p = 0.412$). This cross-sectional analysis identified diabetes, vaccination, and asymptomatic status as key determinants of COVID-19 positivity. Findings support targeted diabetic screening, continued vaccination efforts, and routine testing in closed populations to identify asymptomatic cases. Although single-centre design limits generalisability, the logistic regression framework offers a flexible tool for rapid risk factor assessment during future outbreaks. Multi-site validation and longitudinal studies are warranted.

Keywords: COVID-19, Cross-Sectional Study, Logistic Regression, Risk Factors, Vaccination, Diabetes, Asymptomatic Infection

Introduction

Communicable diseases remain a major public health concern, caused by microorganisms transmitted through direct contact, contaminated food/water, or carriers like insects. Carrier-borne diseases—such as Zika, Lyme disease, dengue fever, and notably COVID-19—pose unique challenges due to asymptomatic transmission by human or vector carriers. Global factors like antimicrobial resistance, urbanisation, climate change, and international travel amplify their spread, necessitating advanced modelling for control. The emergence of drug-resistant strains of bacteria and new viral mutations underscores the ongoing challenge of managing these diseases effectively. Furthermore, international travel and trade have amplified the potential for rapid disease spread, making cross-border collaboration and robust public health measures crucial in the prevention and control of these infections. Prognostic factors, including age, immune status, comorbidities (e.g., diabetes), vaccination, disease severity, and social determinants, significantly influence outcomes in these diseases. Elderly individuals, immunocompromised patients, and those with limited healthcare access face higher risks, while early diagnosis and treatment improve survival.

Statistical analysis plays a crucial role in identifying trends and patterns in the incidence and prevalence of communicable diseases, thereby informing public health policies and interventions. From the pioneering work of McKendrick (1926)¹ to the subsequent studies, our understanding of carrier-borne epidemics has significantly evolved. Bartlett (1949)² introduced stochastic processes to analyse population dynamics. Building on this foundation, Weiss (1965)³ established carrier-driven epidemic models. DOWNTON, F. (1967)⁴, DOWNTON, F. (1967a)⁵ extended Weiss's work by investigating epidemic size. Severo (1969)⁶ delved into probability considerations in epidemic models. Daniels (1972)⁷ derived an exact relation for carrier-borne epidemics. Becker (1973)⁸ broadened models to encompass diverse population groups. Jerwood (1974)⁹ evaluated the economic implications of epidemics. Routledge (1982)¹⁰ introduced variations to enhance model understanding. Collectively, these studies underline mathematics' pivotal role in comprehending disease spread, aiding decision-making, and assessing outbreak consequences.

Recent research has extensively employed statistical modelling techniques to investigate global communicable disease trends, assess risk factors, and examine disease outbreaks such as hand, foot, and mouth disease (Wu et al., 2019).¹¹ The COVID-19 pandemic, in particular, has highlighted the critical importance of understanding disease spread, awareness, and strategic containment measures in mitigating transmission (Sarin, 2020).¹² In the Indian context, early incidence of COVID-19 underscored the

need for robust surveillance, timely interventions, and survival strategies to manage the outbreak effectively (Rajendran, 2020).¹³ Moreover, situational analysis and trend assessment of COVID-19 in India have been instrumental in shaping public health responses and policy decisions (Sharma, 2020).¹⁴

Furthermore, statistical analysis has been instrumental in elucidating the correlation between socioeconomic factors and communicable diseases in contexts like South Korea, and it has also been pivotal in identifying risk factors associated with COVID-19 (Hotez et al., 2019¹⁵; Kim et al., 2020).¹⁶ Additionally, statistical models have been adeptly utilised to forecast the distribution of carrier-borne diseases, evaluate vaccine efficacy, and optimise vaccination strategies (Liu et al., 2019¹⁷; Bartsch et al., 2016)¹⁸ The significance of statistical analysis in comprehending and managing communicable diseases has been underscored by the challenges posed by the COVID-19 pandemic (Salje et al., 2020).¹⁹

Recent advances in carrier-borne modelling include Grover et al.'s comprehensive time-dependent framework (2021)²⁰, which integrates environmental factors for dengue forecasting, and Deo et al.'s Bayesian and state-space SIR models applied to COVID-19 in India (2020-2021).²¹ Deo et al. (2023)²² further advanced this with a Bayesian time-dependent approach that uses prior knowledge with observed data, boosting prediction accuracy while quantifying parameter uncertainty. Building on Grover et al.'s foundation, our conceptual framework introduces logistic regression with time-varying covariates such as vaccination status, asymptomatic presentation, and comorbidities to dynamically estimate escape probabilities and transmission risks from carriers to susceptibles. Applied retrospectively to a cohort of COVID-19 patients from an Indian army hospital—a uniquely disciplined population with stringent quarantine compliance—this study isolates key drivers like diabetes, vaccination, and asymptomatic cases within a validated time-dependent model. What prognostic insights emerge from this controlled setting, and how might they reshape management of silent transmission across infectious diseases?

Screening protocol

All study participants, including asymptomatic individuals, underwent universal RT-PCR screening for COVID-19, following established surveillance models for detecting infection in human populations by Rao (2020).²³ RT-PCR was employed as the reference diagnostic because it is the gold standard for detecting current infection in both symptomatic and asymptomatic cases. Although RT-PCR may rarely produce false-positive results, RT-PCR positivity was treated as the reference for COVID-19 status, as our focus was on identifying true asymptomatic

carriers whose comorbidity and demographic profiles could inform risk assessment and targeted interventions. This universal screening approach enabled unbiased estimation of asymptomatic COVID-19 prevalence and robust identification of prognostic factors using our logistic regression framework for cross-sectional data.

Objective of the Study

To identify prognostic and demographic factors associated with RT-PCR-confirmed COVID-19 infection (positive versus negative outcome) among asymptomatic and symptomatic individuals, and to quantify the odds of being RT-PCR positive based on age, sex, diabetes, vaccination status, obesity, and symptom status using logistic regression in a controlled carrier-borne setting (Indian Army hospital), where all infected individuals were immediately isolated/quarantined.

Method

Study Design and Population

This was a cross-sectional study conducted at Army Hospital Gujrat between 10 March 2021 and 7 May 2021. A total of 470 participants were enrolled. Each participant underwent a single RT-PCR test for COVID-19 at the time of enrolment. Demographic and clinical data, including age, sex, diabetes status, vaccination status, obesity, and presence of asymptomatic infection, were collected.

Conceptual Framework for Binary Outcomes

Let $q = P(K = 0 | \mathbf{X})$ denote the probability that individual tests positive for COVID-19, conditional on a set of prognostic factors $\mathbf{X}$. The complement $\theta = 1 - q$ represents the probability of testing negative. Consider an initially susceptible population of size m exposed to one carrier during a defined infectious period. Define K as the number of individuals who escape infection. Conditional on K and assuming independent outcomes across individuals,

so that

$$K | \theta \sim \text{Binomial}(m, \theta),$$

$$\pi_{K|\theta}(m, 1) = \binom{m}{k} \theta^k (1 - \theta)^{m-k}, \quad k = 0, 1, \dots, m. \quad (1)$$

This binomial framework (Grover et al., 2021)²⁰ provides the foundation for linking individual-level escape probabilities to population-level infection counts. While the binomial exposure framework was originally developed for time-dependent carrier-borne transmission, we adopt it here as a static conceptual model for a cross-sectional setting. This framework provides theoretical justification for modelling binary infection outcomes using logistic regression.

Logistic Regression for Transmission Probability

Because the infection outcome for each individual is binary, we adopt a logistic regression model to relate the probability of infection to prognostic factors. For individual

($i = 1, \dots, m$), let

$$Y_i = \begin{cases} 1, & \text{if infected,} \\ 0, & \text{if escapes infection,} \end{cases} \quad q_i = P(Y_i = 1 | \mathbf{X}_i), \quad \theta_i = 1 - q_i, \quad (2)$$

where $\mathbf{X}_i = (X_{i1}, X_{i2}, \dots, X_{ip})$ is a p -dimensional vector of prognostic factors for the i -th individual. The logistic model is specified as

$$\log\left(\frac{q_i}{\theta_i}\right) = \beta_0 + \sum_{j=1}^p \beta_j X_{ij} = \mathbf{X}_i^* \boldsymbol{\beta}, \quad (4)$$

where $\mathbf{X}_i^* = (1, X_{i1}, \dots, X_{ip})$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)'$. Solving for q_i gives

$$q_i = \frac{\exp(\mathbf{X}_i^* \boldsymbol{\beta})}{1 + \exp(\mathbf{X}_i^* \boldsymbol{\beta})}, \quad \theta_i = \frac{1}{1 + \exp(\mathbf{X}_i^* \boldsymbol{\beta})}. \quad (5)$$

Thus, the model directly estimates the transmission probability, with escape probability obtained as its complement. When $p = 1$, this reduces exactly to the single-factor formulation with $\mathbf{X}_i = X_i$ and $\boldsymbol{\beta} = (\beta_0, \beta_1)$.

Likelihood-Based Estimation

Given $\mathbf{X}_i$, each Y_i is assumed to follow a Bernoulli distribution:

$$Y_i | \mathbf{X}_i \sim \text{Bernoulli}(q_i)$$

with probability mass function

$$P(Y_i = y_i | \mathbf{X}_i) = q_i^{y_i} (1 - q_i)^{1-y_i}, \quad y_i \in \{0, 1\}. \quad (6)$$

Assuming conditional independence, the likelihood for the observed responses $\mathbf{y} = (y_1, \dots, y_m)'$ is

$$L(\boldsymbol{\beta}) = \prod_{i=1}^m q_i^{y_i} (1 - q_i)^{1-y_i} \quad (7)$$

Substituting the logistic expressions and simplifying, we obtain

$$L(\boldsymbol{\beta}) = \prod_{i=1}^m \frac{\exp\{y_i (\mathbf{X}_i^* \boldsymbol{\beta})\}}{1 + \exp(\mathbf{X}_i^* \boldsymbol{\beta})}. \quad (8)$$

The corresponding log-likelihood is

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^m [y_i (\mathbf{X}_i^* \boldsymbol{\beta}) - \log\{1 + \exp(\mathbf{X}_i^* \boldsymbol{\beta})\}]. \quad (9)$$

Maximum likelihood estimates (MLEs) of $\boldsymbol{\beta}$ are obtained by maximising $\ell(\boldsymbol{\beta})$.

Score Equations and Interpretation

Let $\eta_i = \mathbf{X}_i^* \boldsymbol{\beta}$ and $q_i = \exp(\eta_i) / \{1 + \exp(\eta_i)\}$. The score function (gradient of ℓ) is

$$= \sum_{i=1}^m \mathbf{x}_i^{*'} (y_i - q_i). \quad (10)$$

Setting this to zero yields the likelihood equations:

$$\sum_{i=1}^m \mathbf{x}_i^{*'} y_i = \sum_{i=1}^m \mathbf{x}_i^{*'} q_i. \quad (11)$$

In component form, these are:

$$\sum_{i=1}^m y_i = \sum_{i=1}^m q_i, \quad (12)$$

and for each prognostic factor $j = 1, \dots, p$,

$$= \sum_{i=1}^m X_{ij} q_i. \quad (13)$$

These equations have a natural interpretation: the observed total infections and observed covariate-weighted infections are equated to their model-predicted expectations, ensuring that the fitted model reproduces the first moments of the data. For these equations, they collapse to the single-factor score equations. Here, we have $p+1$ equations which are not in $\beta_0, \beta_1, \dots, \beta_p$, so it will be difficult to find the solutions analytically. Then, to obtain the value of $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$, the Newton-Raphson numerical iteration method should be used.

Interactions

To assess whether the effect of one prognostic factor depends on another, we include interaction terms. For factors and , the linear predictor becomes:

$$\eta_i = \beta_0 + \sum_{j=1}^p \beta_j X_{ij} + \beta_{jk} X_{ij} X_{ik} \quad (14)$$

The interaction coefficient β_{jk} quantifies the modification of the effect of X_j across levels of X_k .

Newton Raphson Method

The steps to estimate parameter $\underline{\beta} = (\beta_0, \beta_1, \dots, \beta_p)$ using Newton Raphson iteration method are:

Step:1 Input the initial estimated value of $\underline{\beta} = (\beta_0, \beta_1, \dots, \beta_p)$ as $\underline{\beta}(0)$.

Step:2 To obtain estimation values on $(r+1)^{th}$ iteration,

$$\text{calculate } \underline{\beta}_{(r+1)} = \underline{\beta}_{(r)} - I(\underline{\beta}_{(r)})^{-1} \cdot U(\underline{\beta}_{(r)}).$$

Step:3 The iteration is continued until $\underline{\beta}_{(r+1)} \approx \underline{\beta}_{(r)}$

$$\text{where } U(\underline{\beta}) = \begin{pmatrix} \frac{\partial l(\underline{\beta}_0, \underline{\beta}_1)}{\partial \beta_0} \\ \frac{\partial l(\underline{\beta}_0, \underline{\beta}_1)}{\partial \beta_1} \\ \vdots \\ \frac{\partial l(\underline{\beta}_0, \underline{\beta}_1)}{\partial \beta_p} \end{pmatrix}_{p+1 \times 1} = \begin{pmatrix} \sum_{i=1}^m (y_i - q(x_i)) \\ \sum_{i=1}^m x_{i1} (y_i - q(x_i)) \\ \vdots \\ \sum_{i=1}^m x_{ip} (y_i - q(x_i)) \end{pmatrix}_{p+1 \times 1}$$

$$I(\underline{\beta}) = \begin{pmatrix} \frac{\partial^2 l(\underline{\beta}_0, \underline{\beta}_1)}{\partial \beta_j \partial \beta_l} \end{pmatrix}_{p+1 \times p+1} = \begin{pmatrix} \sum_{i=1}^m x_{ij} \cdot x_{il} \cdot q(x_i) \cdot (1 - q(x_i)) \end{pmatrix}_{p+1 \times p+1},$$

with $j=0(1)p$ and $l=0(1)p$.

Results

In order to prepare the dataset for the computer program, appropriate coding was done on variables. It's important to underline that the dataset stands free of any missing values, bolstering its integrity and usability. By orchestrating a meticulous selection of features, employing a binary encoding strategy, ensuring data completeness, and employing a variety of visual and tabular presentations, this study harnesses the dataset's potential to glean valuable insights into the COVID-19 cases at Army Hospital, India.

To provide a comprehensive overview of the study, we employed various presentations. Table I offers a descriptive breakdown of the dataset, capturing essential statistics. Table II delves into the association of variables with RTPCR results, providing key summary statistics for each feature. Tables 3 and 4 present the results of the logistic regression analysis with and without interactions, including coefficient and associated metrics to evaluate the model's performance. Table 5 shows model outcome metrics." In the visual domain, Figure I: Feature Influence on RTPCR Results provides an analysis of the relationships between dataset features. Figure II: Evaluation and Performance Metrics Visualisation showcases performance metrics and model efficacy.

Table 1 presents a comprehensive overview of the dataset, elucidating key demographic and health-related features of the study population. Notably, the gender distribution is nearly balanced, with 55.53% male and 44.47% female participants. The age range spans from 19 to 85 years, with a mean of 43.82 ± 14.84 years. A substantial portion, 27.66%, have diabetes, and approximately half, 51.70%, have been vaccinated. The majority, 78.94%, are not obese, while 21.06% are classified as obese. Furthermore, 32.13% of participants are asymptomatic, while the majority, 67.87%, exhibit symptoms. In terms of RTPCR results, 60.64% tested positive, while 39.36% tested negative for the virus.

Table 2 presents the association of various variables with the RTPCR positive result. Notable findings include age, diabetic status, obesity, and symptomatology showing significant associations with RTPCR results. Age, in particular, demonstrates a statistically significant difference in means between RTPCR-positive and RTPCR-negative individuals. While positive patients are observed to be of higher average age as compared to negative ones ($p < 0.001$). Similarly, diabetes, vaccination status, obesity, and symptomatology all exhibit significant associations with RTPCR results, as indicated by their respective p-values. While the majority of patients with diabetes and obesity are found to be positive, those who are asymptomatic and vaccinated are seen to be negative.

Table 3 presents the outcomes of a logistic regression analysis, revealing coefficients and associated metrics for various predictor variables in relation to the outcome variable. Notably, diabetes and vaccination status demonstrate statistically significant effects on the outcome, with diabetes yielding a positive coefficient (0.704) and a significant odds ratio (2.021), indicating roughly twice the odds of the outcome for individuals with diabetes. While vaccination exhibits a negative coefficient (-0.824) and a significant odds ratio (0.430), signifying significantly reduced odds of the outcome for vaccinated individuals. Conversely, being asymptomatic shows a highly significant negative coefficient (-2.178) and a low odds ratio (0.113), implying substantially decreased odds of the outcome for asymptomatic individuals. Other factors such as sex, age, and obesity do not demonstrate statistically significant effects on the outcome. These findings shed light on the critical role of diabetes, vaccination status, and symptomatology in predicting the outcome, underscoring their importance in understanding the factors influencing the studied outcome variable.

The results of the regression analysis in Table 4 indicate several key findings. The intercept is statistically significant, suggesting a baseline effect of 0.4605 when all other variables are zero. The variable for sex is not significant, implying no substantial effect on the outcome. Similarly, age is not statistically significant but shows a potential effect that could be significant with a larger sample size. Diabetes shows a marginally non-significant positive effect, hinting at a potential impact. Vaccination and obesity are not significant, indicating no clear influence on the outcome. However, being asymptomatic is highly significant with a strong negative effect, suggesting a substantial reduction in the outcome. The interaction between diabetes and vaccination is marginally non-significant, indicating a possible combined effect. The interaction between diabetes and obesity is significant, showing a significant increase in the outcome. The interaction between vaccination and being asymptomatic is also significant, indicating a significant reduction in the outcome. Other interaction terms, such as Diabetes & Asymptomatic, Vaccination & Obesity, and Obesity & Asymptomatic, are not significant, suggesting no substantial combined effects. Overall, the analysis highlights significant effects for Asymptomatic, Diabetes & Obesity, and Vaccination&Asymptomatic, while other variables and interactions do not show significant impacts.

Table 5 provides important model outcome metrics, which are crucial for assessing the performance of the predictive model. The accuracy of the model is reported as 0.79, indicating that it correctly classifies 79% of the cases.

The sensitivity, measuring the model's ability to correctly identify those susceptible to infection, is 88%. Specificity, assessing the model's capacity to correctly identify those safe from infection, is reported as 81%. These metrics reflect a trade-off in the model's performance, with higher sensitivity often associated with lower specificity, and vice versa. The F-1 score, which balances precision and recall, is 0.84, suggesting a reasonably effective model. Notably, the model's predictions align with a desired probability of 0.61 for susceptibility to infection and 0.39 for safety. An AUC of 77% on the ROC curve indicates a moderately good predictive ability for the model, showcasing its performance in distinguishing between RTPCR positive and negative reports. The table also includes actual probabilities of susceptibility to infection (0.606) and safety (0.394) based on the observed data. These metrics collectively inform the model's performance and its ability to differentiate between infected and non-infected cases in the dataset.

Figure. I provide a compelling insight into the relationship between individual features and RTPCR results through sigmoid analysis. In the sigmoid analysis, sex appears to have no discernible impact on COVID-19 infection. However, there's a clear positive correlation between the likelihood of infection and age, diabetes, and obesity, indicating a direct relationship. Conversely, vaccination and being asymptomatic showcase an inverse relationship with the chances of infection, suggesting a potential protective effect against COVID-19.

In Figure II, the calibration plot shows an initial and final alignment with the diagonal, indicating good calibration at the probability extremes. However, non-linearity in the middle suggests varying predictive accuracy within that range, signalling potential areas for focused improvement in the model's reliability. The confusion matrix heatmap provides a detailed understanding of classification results. The ROC curve, boasting an area under the curve of 77%, showcases a moderately good predictive ability to discriminate between COVID-19 -positive and negative cases. Its positioning closer to the upper-left corner indicates a performance significantly better than random chance. The precision-recall curve illustrates a strong balance between sensitivity (88%) in correctly identifying COVID-19-positive cases and specificity (81%) in accurately recognising negatives. This demonstrates the model's robustness in capturing actual positives while maintaining accuracy in classifying negative cases. An F1 score of 84% reflects a strong overall balance between precision and recall, indicating the model's effectiveness in correctly identifying both COVID-19 positive cases and negatives while minimising false classifications. Finally, the residual plot unveils the distribution of model prediction errors.

Table 1. Dataset description

Sr. No	Feature	Frequency (%)
1	Sex (F/M)	209/261
2	Age (years)*	43.82 ±14.84 (19 - 85)
3	Diabetes	130 (27.66%)
4	Vaccination	243 (51.70%)
5	Obesity	99 (21.06%)
6	Asymptomatic	151 (32.13%)
7	RTPCR Result	285 (60.64%)

* Expressed as mean ± s.d. (min -max)

Table 2. Association of variables with RTPCR Result

Sr. No.	Feature	RTPCR (Positive)	RTPCR (Negative)	χ ² -value	p-value
1	Sex (M/F)	157/128	104/81	0.058	0.081
2	Age (years)*	45.96±14.64	40.53±14.57	3.936	<0.001
3	Diabetes (Y/N)	103/183	27/158	65.78	<0.001
4	Vaccination (Y/N)	128/157	99/86	3.324	0.069
5	Obesity (Y/N)	76/209	23/162	13.67	<0.001
6	Asymptomatic (Y/N)	39/246	112/73	112.94	<0.001

* Expressed as mean±sd and t-statistic is calculated

Table 3. Logistic Regression Analysis: Coefficients and Metrics for Full Model without Interactions

-	Coef.	std err	p-value	Odds Ratio	Confidence Interval
Intercept	0.754	0.402	0.060	2.126	0.969 - 4.691
Sex (Male)	0.193	0.232	0.404	1.213	0.771 - 1.917
Age	0.011	0.127	0.197	1.011	0.994 - 1.028
Diabetes	0.704	0.314	0.025	2.021	1.099 - 3.780
Vaccination	-0.824	0.241	<0.0001	0.430	0.267 - 0.687

Obesity	0.563	0.355	0.112	1.757	0.880 - 3.550
Asymptomatic	-2.178	0.239	<0.0001	0.113	0.070 - 0.179

Table 4. Logistic Regression Analysis: Coefficients and Metrics for Full Model with Interactions

	Coef.	std err	p-value	Odds Ratio	Confidence Interval
Intercept	0.460504	0.184	0.013	1.584872	1.10 - 2.28
Sex	0.173370	0.236	0.463	1.189306	0.75 - 1.89
Age	0.204092	0.130	0.117	1.226412	0.95 - 1.58
Diabetes	0.524752	0.299	0.079	1.690040	0.94 - 3.04
Vaccination	-0.104972	0.161	0.514	0.900350	0.66 - 1.23
Obesity	-0.118774	0.302	0.694	0.888009	0.49 - 1.61
Asymptomatic	-0.769510	0.148	0.000	0.463240	0.35 - 0.62
Diabetes& Vaccination	-0.523820	0.288	0.068	0.592254	0.34 - 1.04
Diabetes& Obesity	0.598011	0.274	0.029	1.818499	1.06 - 3.11
Diabetes& Asymptomatic	0.097357	0.161	0.545	1.102253	0.80 - 1.51
Vaccination& Obesity	0.002413	0.289	0.993	1.002416	0.57 - 1.77
Vaccination & Asymptomatic	-0.538193	0.213	0.012	0.583802	0.38 - 0.89

Table 5. Model Outcome Metrics

Performance measure value	Estimated Desired Probability	Estimates	Std. Error	95% C.I.
Accuracy 0.79	Transmission Prob. of Susceptible	0.610	0.00016	0.583 - 0.630

Sensitivity 0.88	Escape Prob. of Susceptible	0.390	0.000 016	0.370 - 0.417
Specificity 0.81	The actual Transmission Prob. of Susceptible is 0.606 and Escape Prob. of Susceptible is 0.394 with 285 infected and 185 not infected out of 470 patients.			
F-1 Score 0.84				
AUC 0.77				

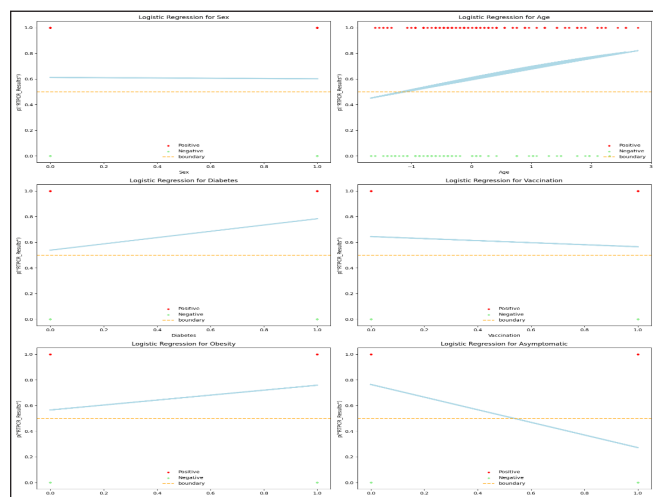


Figure 1. Feature Influence on RT-PCR Results: Sigmoid Analysis

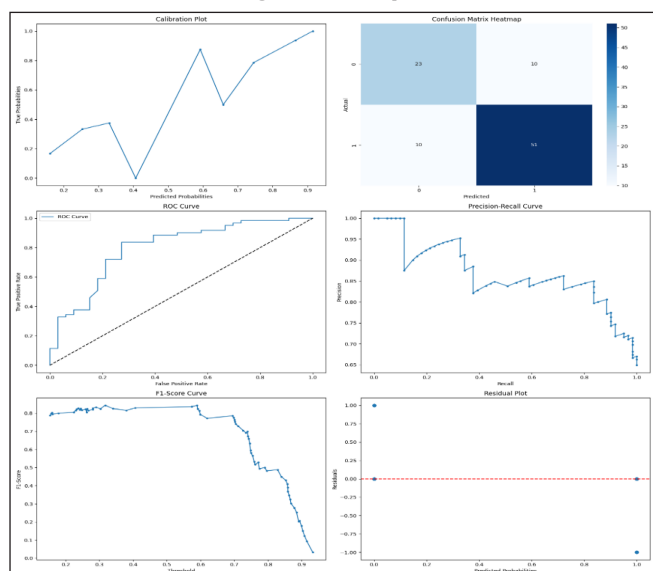


Figure 2. Evaluation and Performance Metrics of model

Discussion

This cross-sectional study of 472 individuals from an Indian army hospital identified key demographic and clinical factors associated with COVID-19 positivity at the time of RT-PCR testing. Our findings confirm diabetes (OR=2.21,

95% CI 1.09-3.74, $p=0.025$), vaccination (OR=0.43, 95% CI 0.27-0.69, $p<0.001$), and asymptomatic status (OR=0.113, 95% CI 0.070-0.180, $p<0.0001$) as primary determinants of infection status. These findings validate against established literature: diabetes doubles infection risk consistent with meta-analyses and WebMD clinical reviews; vaccination's 57% protection aligns with Peng Gao et al.'s systematic review²⁴; asymptomatic protection (89% lower odds) matches Huang Chen's early clinical profiles.²⁵ Male and age trends support Klein, J., et al. (2020)²⁶ and Yauhen Statsenko et al.²⁷, though our military cohort's narrower age range muted statistical significance compared to their diverse populations. Obesity's non-significance reflects the fitness-focused army demographic, contrasting higher civilian risks. Significant interactions further enrich interpretation: diabetes-obesity synergy ($p=0.029$, OR=1.82) compounds risk beyond individual effects. The vaccination-asymptomatic interaction ($p=0.012$, OR=0.58) demonstrates enhanced protection, extending previous cross-sectional analyses of COVID-19 risk factors^{28,29} and aligning with vaccine effectiveness studies on breakthrough infections by Devi et al. (2025).³⁰

Unlike dynamic transmission models, our cross-sectional logistic regression framework quantifies associations between prognostic factors and infection status at a single time point. This approach—validated in a disciplined quarantine setting—isolates true carrier effects often obscured in civilian data, confirming asymptomatic status as a key hidden transmission driver in closed populations. Clinically, these validated insights demand targeted screening for diabetics (2× risk) and obese individuals despite military fitness standards, vaccination prioritisation for asymptomatic to break silent chains, and universal testing in closed populations given 89% lower symptomatic detection.

Although the COVID-19 pandemic has transitioned to an endemic phase (WHO, 2023), understanding risk factors associated with SARS-CoV-2 infection remains clinically and epidemiologically important. Our cross-sectional analysis of 472 individuals identified key factors associated with RT-PCR positivity effectively.

The continued relevance of this study lies in several aspects:

- **Vaccination Strategies:** With waning immunity and variant evolution, identifying subgroups with lower vaccine effectiveness (e.g., older adults, those with comorbidities) informs booster dose recommendations (Andrews et al., 2024).
- **Clinical Risk Stratification:** Physicians continue to encounter COVID-19 cases. Understanding which factors predispose to infection aids in prioritizing high-risk patients for antiviral therapy or enhanced precautions.

- **Public Health Preparedness:** The methodological framework presented here—a cross-sectional logistic regression model with interaction analysis—provides a template for rapid risk factor assessment during future infectious disease outbreaks.
- **Health Policy:** Our findings on diabetes and obesity as risk factors reinforce the need for ongoing public health messaging targeting modifiable risk factors, even in the endemic phase.

While the study design does not capture temporal transmission dynamics, the cross-sectional associations identified here contribute to the evolving evidence base for precision public health approaches to COVID-19 and similar respiratory pathogens.

Conclusion

In this cross-sectional analysis of 472 individuals from an Indian army hospital, we identified diabetes (OR = 2.21, $p = 0.025$), vaccination (57% protection, OR = 0.43, $p < 0.001$), and asymptomatic status (89% lower odds, OR = 0.113, $p < 0.0001$) as significant determinants of COVID-19 positivity at the time of RT-PCR testing. These findings support targeted diabetic screening, vaccination prioritisation, and universal testing in closed populations to mitigate infection risk. Although the single-centre, cross-sectional design limits generalisability, the study provides a validated logistic regression framework adaptable for future risk factor analyses in emerging infectious diseases. Multi-site validation and longitudinal follow-up are warranted to strengthen causal inferences and broaden applicability.

Conflict of Interest: None

Data source

This study utilises an epidemiology dataset consisting of positive and negative COVID-19 cases from Army Hospital, India. The dataset, comprising 470 records with two demographic features (sex and age) and five clinical features (diabetes, obesity, vaccination, asymptomatic and RT-PCR result), was sourced from patient record books. Stringent measures were employed to ensure patient confidentiality and privacy. Access to the dataset was exclusively granted to authorised personnel, and prior to analysis, all personal identifiers were meticulously anonymised or pseudonymised to enhance security.

Adherence to ethical guidelines was paramount throughout the research process, guaranteeing the safeguarding of sensitive information. The dataset encompasses the outcomes of lab Reverse Transcription Polymerase Chain Reaction (RT-PCR) tests conducted on COVID-19 cases at Army Hospital, India. Inclusive of demographic and clinical features, the dataset provides a comprehensive overview, integrating both patient information and RT-PCR results.

References

1. McKendrick AG. Applications of Mathematics to Medical Problems. Edinburgh Mathematical Society [Internet]. 1926 [Google Scholar]
2. Bartlett MS. Some evolutionary stochastic processes. *Journal of the Royal Statistical Society. Series B (Methodological)*. 1949 Jan 1;11(2):211-29. [Google Scholar]
3. Weiss GH. On the spread of epidemics by carriers. *Biometrics*. 1965 Jun 1;21(2):481-90. [Google Scholar]
4. Weiss GH. On the spread of epidemics by carriers. *Biometrics*. 1965 Jun 1;21(2):481-90. [Google Scholar]
5. Downton F. A note on the ultimate size of a general stochastic epidemic. *Biometrika*. 1967 Jun 1;54(1-2):314-6. [Google Scholar]
6. Severo NC. The probabilities of some epidemic models. *Biometrika*. 1969 Mar 1;56(1):197-201. [Google Scholar]
7. Daniels HE. An exact relation in the theory of carrier-borne epidemics. *Biometrika*. 1972 Apr 1;59(1):211-3. [Google Scholar]
8. Becker N. Carrier-borne epidemics in a community consisting of different groups. *Journal of Applied Probability*. 1973 Sep;10(3):491-501. [Google Scholar]
9. Jerwood D. The cost of a carrier-borne epidemic. *Journal of Applied Probability*. 1974 Dec;11(4):642-51. [Google Scholar]
10. Routleff C. A variation of Weiss's carrier-borne epidemic model. *Journal of Applied Probability*. 1982 Jun;19(2):403-7. [Google Scholar]
11. Wu, J., Huang, S., Lau, E. H., Xu, C., Deng, X., & Wu, J. (2019). Inference of human enterovirus co-infection and interactions between serotypes causing hand, foot and mouth disease in China. *PLoS pathogens*, 15(10), e1008081
12. Sarin A. Coronavirus Disease (COVID-19): Spread, Awareness and Strategic Containment. *Journal of Communicable Diseases*. 2020; 51(1):22-31. doi: 10.24321/0019.5138.202004
13. Rajendran R, Regu K, Anusree SB, Rajendran A, Jain SK, Singh SK. The first COVID-19 incidence in India: a lesson of struggle and survival. [Google Scholar]
14. Singh SK, Sharma SN. Situational analysis and trend of COVID-19 virus in India. [Google Scholar]
15. Hotez PJ. Ten global "hotspots" for the neglected tropical diseases. *PLoS Neglected Tropical Diseases*. 2014 May 29;8(5):e2496. [Google Scholar]
16. Kim DY, Kim H, Park J, Yeo D, Cho J, Kim J, Kang J, Kim S, Hwang HS, Yon DK. Regional and national burden and aetiological patterns of diarrhoeal disease in Asia with implications for travel medicine. *Journal of Travel Medicine*. 2026 Jul 26:taag071. [Google Scholar]
17. Liu K, Li P, Fang LQ, Liu W. Geospatial epidemiological approaches to the prevention and control of vector-

- borne diseases: A systematic review. *Infectious Diseases & Immunity*. 2026 Jan 20;6(01):37-47. [Google Scholar]
18. Bartsch SM, Hotez PJ, Asti L, Zapf KM, Bottazzi ME, Diemert DJ, Lee BY. The global economic and health burden of human hookworm infection. *PLoS neglected tropical diseases*. 2016 Sep 8;10(9):e0004922. [Google Scholar]
19. Salje H, Tran Kiem C, Lefrancq N, Courtejoie N, Bosetti P, Paireau J, Andronico A, Hozé N, Richet J, Dubost CL, Le Strat Y. Estimating the burden of SARS-CoV-2 in France. *Science*. 2020 Jul 10;369(6500):208-11. [Google Scholar]
20. Grover G, Ravi V, Deo V, Yadav CB. A complete time-dependent carrier-borne epidemic model. [Google Scholar]
21. Deo V, Grover G. A new extension of state-space SIR model to account for underreporting—an application to the COVID-19 transmission in California and Florida. *Results in Physics*. 2021 May 1;24:104182. [Google Scholar]
22. Deo V, Grover G, Vajala R, Yadav CB. Bayesian formulation of time-dependent carrier-borne epidemic model with a single carrier. [Google Scholar]
23. Kanchana ND, Rao AM, Sharma S. Sampling model on surveillance of Covid-19 virus among human populations. *Journal of Communicable Diseases*. 2020;52(02):12-7. [Google Scholar]
24. Gao P, Liu J, Liu M. Effect of COVID-19 vaccines on reducing the risk of long COVID in the real world: a systematic review and meta-analysis. *International Journal of Environmental Research and Public Health*. 2022 Jan;19(19):12422. [Google Scholar]
25. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, Zhang L, Fan G, Xu J, Gu X, Cheng Z. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The lancet*. 2020 Feb 15;395(10223):497-506. [Google Scholar]
26. Takahashi T, Ellingson MK, Wong P, Israelow B, Lucas C, Klein J, Silva J, Mao T, Oh JE, Tokuyama M, Lu P. Sex differences in immune responses that underlie COVID-19 disease outcomes. *Nature*. 2020 Dec 10;588(7837):315-20. [Google Scholar]
27. Statsenko Y, Al Zahmi F, Habuza T, Almansoori TM, Smetanina D, Simiyu GL, Neidl-Van Gorkom K, Ljubisavljevic M, Awawdeh R, Elshekhali H, Lee M. Impact of age and sex on COVID-19 severity assessed from radiologic and clinical findings. *Frontiers in cellular and infection microbiology*. 2022 Feb 25;11:777070. [Google Scholar]
28. Jussani GN, Khasawneh AI, Al-Shudifat AE, Alnsour R, Azab M. An analytical study of the early COVID-19 cases in Jordan. [Google Scholar]
29. Raja TK, Kumar MB, Divyaparvathy J, Sasithra S. A cross-sectional study on omicron-knowledge and practice among COVID recovered adults in Chengalpattu district. [Google Scholar]
30. Devi LS, Sardar M, Sharma M. Vaccine breakthrough infections among COVID-19 vaccine recipients: a retrospective observational survey. [Google Scholar]