ANALYSIS AND ESTIMATION OF LENGTH OF IN-HOSPITAL STAY USING DEMOGRAPHIC DATA OF COVID-19 RECOVERED PATIENTS IN SINGAPORE

A PREPRINT

Jaya Sreevalsan-Nair* **Reddy Rani Vangimalla** **Pritesh Rajesh Ghogale**
Graphics-Visualization-Computing Lab *and* E-Health Research Center
International Institute of Information Technology Bangalore, Karnataka 560100, India
http://www.iiitb.ac.in/gvcl http://ehrc.iiitb.ac.in

June 6, 2020

ABSTRACT

Singapore is one of the countries which has taken early, systematic, and rigorous nationwide responses to slowing the COVID-19 contagion down. By February 4, 2020, the government of Singapore restricted the mobility of the vulnerable age-groups. In the current study, we study the influence of the age-based vulnerability of the population with respect to COVID-19 conditions on the recovery of COVID-19 patients. We study 245 patients in Singapore recovered and discharged during January 23–April 01. We first study the descriptive statistics of the length of in-hospital stay (LOS) of the COVID-19 patients based on demographic variables, namely age, and gender. Then, we determine the distribution of LOS, using local and generalized linear regression models. We take the approach of periodization based on critical changes in the disease transmission model. Even though the overall recovery rate has reduced drastically after a sudden spike in daily confirmations, our analysis shows that there is a considerable shift in the COVID-19 confirmations to the population in the non-vulnerable age-groups. We show that the LOS of the non-vulnerable age group is considerably lower at 9 days, as opposed to 15 or 20 days in the existing literature.

***K*eywords** COVID-19, Length of in-hospital stay, Descriptive statistics, Hypothesis testing, Regression

1 Introduction

The viral contagion, named COVID-19, was declared as a global pandemic by the World Health Organization (WHO) on March 11, 2020[2]. The pandemic, characterized by atypical pneumonia, is caused by a virus from the coronavirus family, namely SARS-CoV2 (Severe Acute Respiratory Syndrome Coronavirus-2), which is a positive-sense single-stranded RNA virus. As of April 2, there were 827,419 positively confirmed cases, and 40,777 deaths, spread across 206 countries [3], and the ratio of the number of the recovered[4] to the infected patients, r_{ri} was $\sim$0.2. On the other hand, in Singapore, 245 of 1000 COVID-19 positive cases had clinically recovered by April 1, giving r_{ri}=0.245. Aside from the recovered/hospital-discharged cases, the first two deaths in Singapore owing to the COVID-19 complications were only three by April 1. The count of the recovered was $\sim$79.5% of closed cases (i.e., recovery and death) worldwide, whereas $\sim$98.8% in Singapore. Another Southeast Asian country with similar COVID-19 response, South Korea crossed its first 1000 COVID-19 confirmations[5] during January 20-February 26, i.e., within 38 days, compared to 70 days in Singapore, giving $r_{ri} = 0.019$ with the recovered being only 66.67% of closed cases.

*jnair@iiitb.ac.in

[2] https://www.who.int/dg/speeches

[3] https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports

[4] This is based on the unofficial counts are, as retrieved on April 2, 2020, from https://www.worldometers.info/coronavirus/coronavirus-cases

[5] https://www.cdc.go.kr/board/board.es?mid=a30402000000&bid=0030&tag=&act=view&list_no=366352

Singapore has fared well owing to its effective public healthcare system that implemented a quick and efficient pandemic response since its first case got confirmed on January 23. The government enforced strict quarantine, isolation, hospital surveillance, large-scale contact tracing, and testing. The centralization of the system has facilitated responsive data gathering and anonymized patient-wise reporting to the public. The first cases of local transmission were determined on February 4, after which there were concerted efforts to protect the age-groups, namely, (0-20) and (60+) years, considered vulnerable to the COVID-19 conditions. South Korea has implemented similar measures but, did not focus on the early containment of local transmission, unlike Singapore. Thus, the overall slow transmission of COVID-19 in Singapore indicates that its early measures have been effectively slowed local transmission.

In this study, we focus on the influence of the age-based vulnerability of the population on the clinical recovery of the patients in Singapore during January 23-April 1. We hypothesize that these containment measures have indirectly reduced the length of in-hospital stay (LOS), Δt_{LOS}, for COVID-19. We use the open data available through the official press releases of the Ministry of Health (MoH), Government of Singapore, for this study. Thus, we emphasize on the credibility of our study owing to the *reliability*, *accessibility*, and *availability* of the data [1]. A demographic analysis of the discharged patients complements similar demographic-related observations in differential disease transmission, e.g., higher fatality in older males with comorbidities [2, 3]. We perform piecewise analyses for estimating time-varying Δt_{LOS} using periodization, similar to the estimation of reproduction number [4, 5]. We study the temporal changes in the age-gender distribution of the COVID-19 patients and discharged ones to estimate Δt_{LOS}. We then determine the statistical significance of the demographic variables and period on Δt_{LOS}, and find appropriate distributions of the LOS and fitting multivariate regression models. The novel contribution of our work is in linking the temporal reduction in Δt_{LOS} to the demographic structure.

2 Methods

The data for our work has been collated from the case summaries in the public press releases made by the MoH[6]. We use the age, gender, positive confirmation date, and discharge date from these anonymized summaries. We define the *length of in-hospital stay* Δt_{LOS} as the time elapsed between positive and two consecutive negative COVID-19 confirmations using the real-time reverse transcription polymerase chain reaction (RT-PCR) tests. Owing to the strict protocols followed by the Singapore healthcare system, the COVID-19 patients are hospitalized upon positive confirmation. Hence, recovery-period here is equivalent to the *virus shedding period*, which is estimated to be 15 days [6] or 20 days [7].Overall, we study the Δt_{LOS} of 245 patients who were discharged by April 1 (§Figure 1). (*Additional data along with supporting details of the pandemic response in Singapore until April 1 are given in Section 1, Supplementary File.*)

Age-gender distribution: Using age-groups of 10 years, we observe that the age-gender distribution of the confirmed/infected, and the recovered are different (§Figure 1(a)(i)), where the gender ratios (male to female) are 1.36 and 1.27 respectively. (*The age-gender data is given as percentages in Table S1, Supplementary File.*) The highest counts of the confirmed and the recovered are in the age-groups (20-29) years (§Figure 1(a)(ii)) and (30-39) (§Figure 1(a)(iii)) years, respectively, demonstrating a conspicuous shift of the affected population to the age groups (20-39) years.

Periodization: The first imported and local transmission cases were confirmed on January 23 and February 4, respectively. There is a sudden spike in the daily confirmations on March 17. We periodize using these events for piecewise analysis of the pandemic progression (§Figure 1(b)). Given the cumulative counts of positive confirmed cases, N_i, discharged/recovered cases, N_r, and deceased, N_d, we get:

1. Period P_1 *(period of imported cases)* during January 23-February 3, (N_i=18, N_r=0, N_d=0).
2. Period P_2 *(period of local transmission)* during February 4-March 16, (N_i=243, N_r=109, N_d=0).
3. Period P_3 *(period of community transmission)* during March 17-April 1, (N_i=1000, N_r=245, N_d=3).

We observe from the periodwise age-gender distribution that P_1 is different from both P_2 and P_3 (§Figure 1(a)(ii),(a)(iii)). The ratio of recovered to infected, r_{ri}, i.e., $\frac{N_r}{N_i}$, has increased from 0.00 to 0.45, again dipped to 0.245, through P_1 to P_3. However, since the trend in r_{ri} does not explicitly show the influence of shifts in the age-gender distribution (§Figure 1(a)(ii)) of N_i and N_r on Δt_{LOS}, we analyse the time-varying Δt_{LOS} using descriptive statistics and regression.

LOS analysis: Descriptive statistical analysis of Δt_{LOS} in groups based on gender, age, and period give the preliminary findings of the variables on Δt_{LOS}. We use two appropriate regression models to estimate Δt_{LOS}. Firstly, we use the time-series of Δt_{LOS} and fit a line using the loess model. Loess is a non-parametric local regression model for smoothening empirical time-series data [8] and scatterplots [9]. Secondly, we consider Δt_{LOS} as an observed count

[6] https://www.moh.gov.sg/news-highlights/

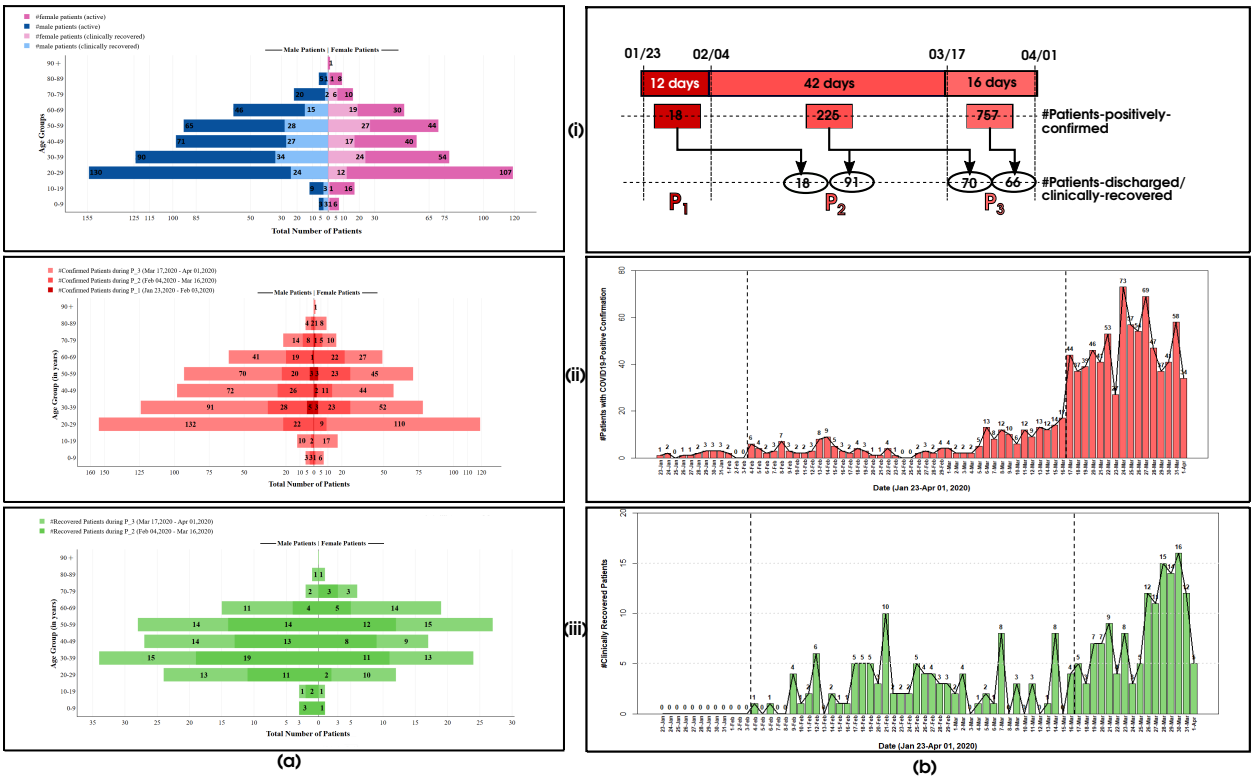


Figure 1: (a) Age-gender distribution observed using stacked population pyramids of COVID-19 infected and recovered, and (b) periodization of January 23-April 01, 2020, using timelines, for COVID-19 patients in Singapore, for (i) general population trends, (ii) COVID-19 positive confirmed patients, and (iii) recovered and discharged patients. The population size is (i) 1000 (576M, 424F), (ii) 1000 (576M, 424F), and (iii) 245 (137M, 108F).

variable and fit the frequency distribution of Δt_{LOS} into a multivariate linear regression models. We use semi-parametric generalized linear models (GLM) for regression, with age, gender, and period are predictor variables, and Δt_{LOS} as the dependent variable. For the period-wise analysis of Δt_{LOS}, we group the discharged patients based on their positive confirmation date, $\mathbb{G}_{\mathbf{+cfrmDt}}$. An alternative grouping is also possible based on their discharge date, $\mathbb{G}_{\mathbf{-cfrmDt}}$. Since the protocols followed in the hospital are similar for patients with closer admission dates, the $\mathbb{G}_{\mathbf{+cfrmDt}}$ grouping shows more cohesive descriptive statistics of Δt_{LOS} than the $\mathbb{G}_{\mathbf{-cfrmDt}}$ grouping. Hence, we use the $\mathbb{G}_{\mathbf{+cfrmDt}}$ grouping here. (*The descriptive statistics of* Δt_{LOS} *with* $\mathbb{G}_{\mathbf{-cfrmDt}}$ *grouping is given in Table S2, Supplementary File.*)

We use the loess model [8] to smooth the conditional means in the observations of Δt_{LOS} for getting its trendline. Loess uses linear least squares regression model, $y = g(x) + \varepsilon$, with independent and dependent variables, x and y, respectively, a smooth function g (we use the Gaussian function), and a 0-mean constant-scale random variable as the error value, ε. The loss function $\varepsilon_w(x) = w(x).(y - g(x))^2$ is minimized using a weight function $w(x)$. Loess uses a key parameter, span α, that is the control parameter for the size of the local neighborhood. When $\alpha < 1$, $w(x) = (1 - |d|^3)^3$, the tricube weight function with d is the normalized distance from data points to the fitted curve, and otherwise, $w(x) = 0$.

Given that school children and senior citizens are considered vulnerable groups, we implement age binning for performing hypothesis testing and regression modeling. These age-bins (0-19), (20-59), and (60-89) are referred to as A_1, A_2, and A_3, respectively. Thus, we use the binning of {male, female}, $\{P_1, P_2, P_3\}$, and $\{A_1, A_2, A_3\}$ for gender, period, and age, respectively, for hypothesis testing and determining appropriate GLM's. We use the Kruskal-Wallis H (KWH) test of Δt_{LOS} with respect to age-bins, and with the period, separately. The KWH test is a rank-based non-parametric test to determine if two or more groups of an independent ordinal/continuous variable (age or period) have statistically significant differences with the continuous dependent variable (Δt_{LOS}). We use the Mann-Whitney-Wilcoxon (MWW) test of Δt_{LOS} with respect to gender. The MWW test is a non-parametric two-sample test to verify the null hypothesis that the distribution of a continuous variable (Δt_{LOS}), which is not normal, is the same in two or more independent groups formed based on a nominal/categorical independent variable (gender).

We determine the appropriate GLM by experimenting with Poisson (PRM) and negative binomial (NBM) distributions for Δt_{LOS}. These distributions are commonly used for LOS, owing to its naturally skewed distribution [10], and for over-dispersed data [11], such as LOS. Here, we fit the model using frequency distribution of LOS, and model has LOS as dependent variable Y, and age, gender, and period of infection as predictor variables X_1, X_2, and X_3, respectively. They are used as numerical, nominal, and ordinal type variables, respectively. We assume the frequency distribution of Y to be Poisson with $E(Y) = Var(Y) = \mu$, or negative binomial, where $Var(Y) = \mu$ is relaxed. Thus, the GLM with intercept β_0, regression coefficients, β_i for $i = 1, 2, 3$, and error term ε is given as:

$$g(\mu) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon.$$

3 Results

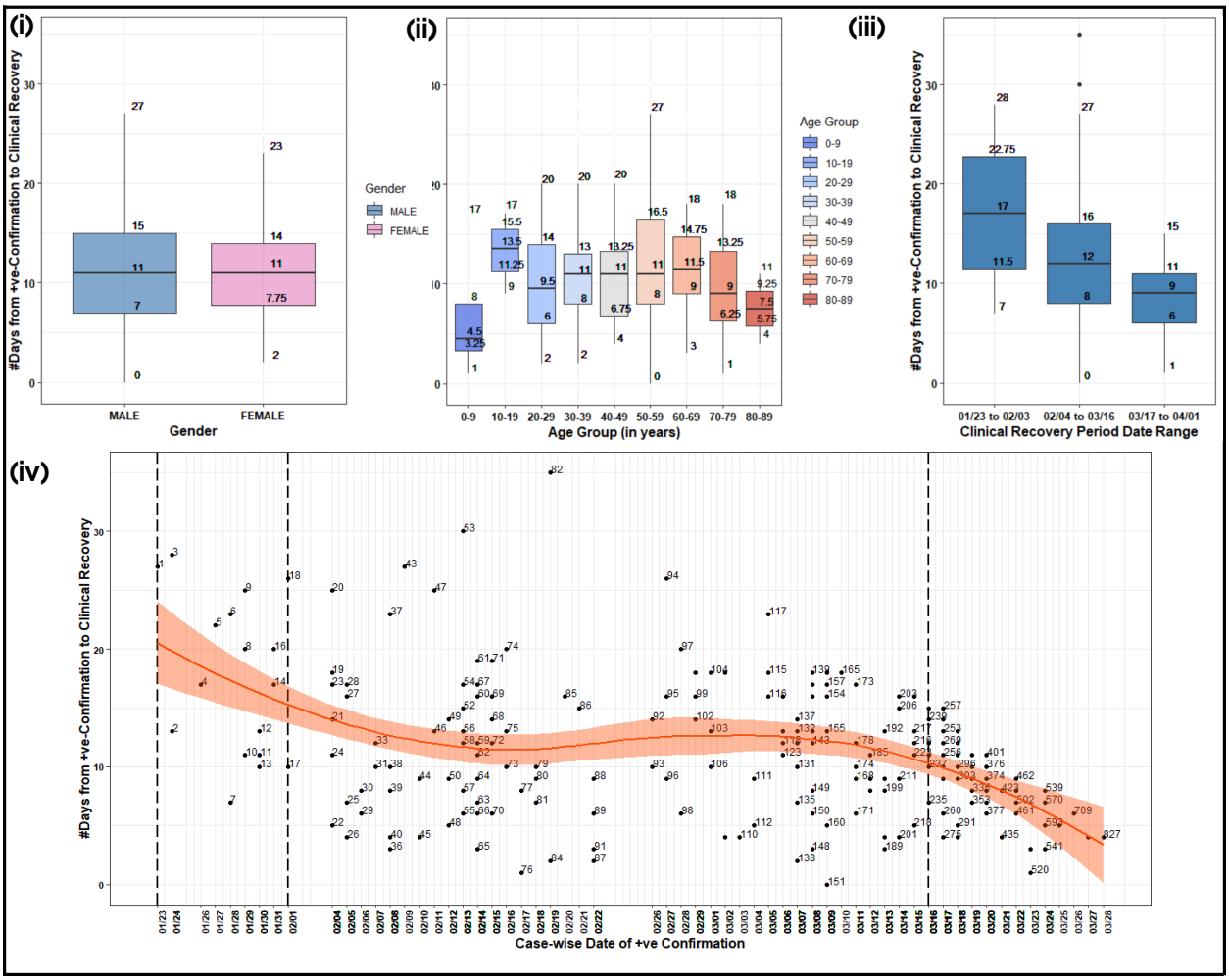


Figure 2: Box-and-whisker plots of LOS of 245 discharged COVID-19 patients during January 23-April 01, 2020, in Singapore, based on (i) gender, (ii) age, and (iii) period, with $\mathbb{G}_{\textbf{+cfrmDt}}$ grouping. (iv) Scatter plot of the length of in-hospital stay for 245 discharged patients with the timeline of their COVID-19 positive confirmation, with a trend-line estimated using a loess model. The fitted loess curve and its error ribbon show a decrease in LOS during the period.

Descriptive statistical analysis: Δt_{LOS} observed during the entire period, January 23-April 01 show the descriptive statistics (§Figure 2 (i)-(iii)) with the following (#patients, mean, standard deviation, median, interquartile range), (N, M, SD, Mdn, IQR), values:

- Overall: (N=245, M=11.4, SD=5.8, Mdn=11, IQR=[7, 14]) days.
- Gender-wise (§Figure 2(i)): (N=108, M=11.28, SD=5.85, Mdn=11, IQR=[7.75, 14]) days for females and (N=137, M=11.42, SD= 5.77, Mdn=11, IQR=[7, 15]) days for males.

Table 1: Hypothesis testing results of 245 discharged patients using the Kruskal-Wallis H test of Δt_{LOS} with respect to age and period, and using the Mann-Whitney-Wilcoxon test of Δt_{LOS} with respect to gender.

Δt_{LOS} by Age-bins Using Kruskal-Wallis H Test					
All	P_1	P_2	P_3	Male	Female
(N=245, df=2, H=0.20, p=0.91)	(N=18, df=1, H=0.01, p=0.94)	(N=161, df=2, H=0.10, p=0.95)	(N=66, df=2, H=2.74, p=0.25)	(N=137, df=2, H=1.83, p=0.40)	(N=108, df=2, H=1.63, p=0.44)

Δt_{LOS} by Period [$\mathbb{G}_{\mathbf{+cfrmDt}}$] Using Kruskal-Wallis H Test					
All	A_1	A_2	A_3	Male	Female
(N=245, df=2, H=36.03, p=**0.000**)	(N=8, df=1, H=1.80, p=0.18),	(N=193, df=2, H=29.64, p=**0.000**,	(N=44, df=2, H=4.15, p=0.13)	(N=137, df=2, H=23.88, p=**0.000**)	(N=108, df=2, H=13.02, p=**0.000**)

Δt_{LOS} by Gender Using Mann-Whitney-Wilcoxon Test						
All	P_1	P_2	P_3	A_1	A_2	A_3
(N=245, W=7562, p=0.77, r=-0.019)	(N=18, W=50.5, p=0.40, r=-0.21)	(N=161, W=3275, p=0.73, r=-0.027	(N=66, W=512, p=0.72, r=0.045)	(N=8, W=8, p=0.62, r=-0.24)	(N=193, W=4320, p=0.60, r=0.038)	(N=44, W=302.5, p=0.10, r=-0.25)

N: *#Samples* H: *H-statistic* df: *Degree of freedom* W: *W-statistic* r: *Effect size statistic*

Table 2: Results of the generalized linear models with age, gender, and period as independent variables, using Poisson distribution and negative binomial distribution of Δt_{LOS} with $\mathbb{G}_{\mathbf{+cfrmDt}}$ grouping.

Model Results	M_1 (age:(0-89))	M_2(age:(0-19))	M_2 (age:(20-59))	M_2(age:(60-89))
#Samples (N)	245	8	193	44
#Degree of freedom	244	7	192	43
Poisson Regression Model (PRM)				
#Parameters (k); $\frac{N}{k}$	4; 61	4; 1.75	4; 48	4; 10.75
Coefficients				
(Intercept)	(p=**0.000**, CI=[2.95, 3.37])	(p=**0.000**, CI=[2.58, 5.98])	(p=**0.000**, CI=[2.27, 3.28])	(p=**0.000**, CI=[2.41, 4.74])
Age	(p=0.13, CI=[-0.001, 0.004])	(p=0.59, CI=[-0.05, 0.03])	(p=**0.012**, CI=[0.001, 0.009])	(p=0.24, CI=[-0.03, 0.01])
Gender	(p=0.49, CI=[-0.10, 0.05])	(p=0.81, CI=[-0.69, 0.49])	(p=0.64, CI=[-0.11, 0.06])	(p=**0.021**, CI=[0.03, 0.38])
Period	(p=**0.000**, CI=[-0.43, -0.30])	(p=0.09, CI=[-1.75, 0.10])	(p=**0.000**, CI=[-0.42, -0.27])	(p=**0.000**, CI=[-0.56, -0.18])
#F. S. Iter.†	4	4	4	4
AIC; AIC^{c}	1629.4; NA	NA; 74.41	1280; NA	NA; 289.11
Deviance residuals	(Mdn=0.012, IQR=[2.19, 10.31])	NA	(Mdn=**0.05**, IQR=[**2.13, 9.98**])	(Mdn=**0.02**, IQR=[**2.16, 7.17**])
MSE$^{\diamond}$	0.32	0.80	0.28	0.34
RMSE$^{\diamond}$	0.56	0.89	0.53	0.58
Max. likelihood value	9.9% at $\Delta t_{\text{LOS}} = 10$	8.6% at $\Delta t_{\text{LOS}} = 11$	**9.8% at $\Delta t_{\text{LOS}} = 10$**	**9.8% at $\Delta t_{\text{LOS}} = 10$**
Negative Binomial Regression Model (NBM)				
#Parameters (k); $\frac{N}{k}$	5; 48.8	5; 1.4	5; 38.4	5; 8.6
Coefficients				
(Intercept)	(p=**0.000**, CI=[2.82, 3.48])	(p=**0.001**, CI=[1.72, 6.88])	(p=**0.000**, CI=[2.59, 3.47])	(p=**0.000**, CI=[1.90, 5.32])
Age	(p=0.38, CI=[-0.002, 0.006])	(p=0.78, CI=[-0.08, 0.06])	(p=0.15, CI=[-0.002, 0.01])	(p=0.37, CI=[-0.04, 0.01])
Gender	(p=0.76, CI=[-0.14, 0.10])	(p=0.86, CI=[-1.06, 0.93])	(p=0.68, CI=[-0.16, 0.11])	(p=0.15, CI=[-0.07, 0.47])
Period	(p=**0.000**, CI=[-0.47, -0.25])	(p=0.27, CI=[-2.24, 0.58])	(p=**0.000**, CI=[-0.46, -0.22])	(p=**0.02**, CI=[-0.64, -0.07])
#F. S. Iter.†	1	1	1	1
AIC; AIC$_{c}$	1499; NA	NA; 88.21	NA; **1180.28**	NA; **275.19**
Deviance residuals	(Mdn=0.02, IQR=[1.41, 6.74]	NA	(Mdn=**0.04**, IQR=[**1.35, 6.58**])	(Mdn=**-0.01**, IQR=[**1.44, 5.04**])
MSE$^{\diamond}$	0.32	0.79	0.28	0.34
RMSE$^{\diamond}$	0.56	0.89	0.53	0.58
Max. likelihood value	7.6% at $\Delta t_{\text{LOS}} = 9$	6.8% at $\Delta t_{\text{LOS}} = 5$	**7.5% at $\Delta t_{\text{LOS}} = 9$**	**7.8% at $\Delta t_{\text{LOS}} = 9$**

CI: *95% Confidence Interval* Mdn: *Median* † *Fisher Scoring Iterations* $^{\diamond}$ *Error with respect to observed data* NA *Not Applicable*

- Age-wise, for age-groups in years given in parentheses (§Figure 2(ii)):
  - A_1: (N=4, M=6.75, SD=7.04, Mdn=4.5, IQR=[3.25, 8]) days for (0-9), (N=4, M=13.25, SD=3.5, Mdn=13.5, IQR=[11.25, 15.5]) days for (10-19),
  - A_2: (N=36, M=10.44, SD=4.88, Mdn=9.5, IQR=[6, 14]) days for (20-29), (N=58, M=11.33, SD=5.64, Mdn=11, IQR=[8, 13]) days for (30-39), (N=44, M=10.64, SD=4.73, Mdn=11, IQR=[6.75, 13.25]) days for (40-49), (N=55, M=12.78, SD=7.21, Mdn=11, IQR=[8, 16.5]) days for (50-59),
  - A_3: (N=34, M=12, SD=5.53, Mdn=11.5, IQR=[9, 14.75]) days for (60-69), (N=8, M=9.38, SD= 5.55, Mdn=9, IQR=[6.25, 13.25]) days for (70-79), (N=2, M=7.5, SD=4.95, Mdn=7.5, IQR=[5.75, 9.25] days for (80-89).
- Period-wise, based on $\mathbb{G}_{\mathbf{+cfrmDt}}$ grouping (§Figure 2(iii)):
  - (N=18, M=17.61, SD=6.6, Mdn=17, IQR=[11.5, 22.75]) days for P_1,
  - (N=161, M=11.88, SD=5.81, Mdn=12, IQR=[8, 16]) days for P_2,
  - (N=66, M=8.58, SD=3.35, Mdn=9, IQR=[6, 11]) days for P_3.

We observe that the Δt_{LOS} is overall lesser in this dataset than reported in early analyses of COVID-19 patients, i.e., 15 days [6] and 20 days [7]. The overall median of Δt_{LOS} and the gender-weighted median is 11 days, while the age-weighted and period-weighted medians are 10.7 and 11.6 days, respectively. Thus, the overall descriptive statistical analysis gives a conservative estimate of Δt_{LOS} to be $\sim$11 days.

There is a monotonous decrease in Δt_{LOS} with time, demonstrated from P_1 to P_3. To analyse this observation further, we look at the period-wise descriptive statistics for Δt_{LOS} with respect to gender, using $\mathbb{G}_{\mathbf{+cfrmDt}}$ grouping.

- For females: (N=9, M=16.33, SD=6.44, Mdn=13, IQR=[11, 22]) days for P_1, (N=69, M=11.8, SD=6.1, Mdn=12, IQR=[7, 15]) days for P_2, and (N=30, M=8.57, SD=3.36, Mdn=9, IQR=[6.25, 11]) days for P_3.
- For males: (N=9, M=18.89, SD=6.88, Mdn=20, IQR=[17, 23]) days for P_1, (N=92, M=11.93, SD=5.61, Mdn=12, IQR=[8, 16]) days for P_2, and (N=36, M=8.22, SD=3.37, Mdn=8, IQR=[5.75, 11]) for P_3.

We observe that Δt_{LOS} has similar descriptive statistics across both genders in P_3 than in P_1 or P_2. The IQR of Δt_{LOS} decreases from P_1 to P_3 for both genders, indicating the decrease in the *measure of the spread* of Δt_{LOS} with time. (*The age-groupwise breakdown of descriptive statistics is given in Table S2, with corresponding box-and-whisker plots in Figure S2 in the Supplementary File.*)

There is a strong influence of age as well as period on Δt_{LOS}, unlike gender. Δt_{LOS} is relatively lower for the age group of (20-29) years, specifically. This result is significant as the age group of (20-29) years contributes the highest (27.3%) to the infected population and relatively high (14.6%) to the recovered population. At the same time, the number of COVID-19 cases has increased in this age group over time, as 88.6% of the patients in this age group have been confirmed positive in P_3. Thus, our key conclusion is that since the most susceptible group of people has lower Δt_{LOS}, the overall Δt_{LOS} is influenced by age to decrease with time progressively.

Loess model: We smooth the time-series of the Δt_{LOS}, represented as a scatter plot based on hospital admission dates (§Figures 2 (iv)) using loess, implemented using `loess`[7] with default span $\alpha = 0.75$ in R [12]. The loess model for discharged patients during January 23-April 01 gives the following degrees of freedom (df) and residual standard error (RSE) for N patients: (N=245, df=5.37, RSE=145.4) (§Figure 2(iv)). Δt_{LOS} has a slope of -25° in P_3. Thus, the key conclusion from the local regression model on the time-series is the negative slope, i.e., a *downward* trend in Δt_{LOS} in P_3. (*Smoothing of periodwise observations using separate loess models is given in Figure S1 in Supplementary File.*)

Multivariate linear regression model: The regression model can be made piecewise with respect to a predictor variable in the case of non-monotonous behavior of the observed variable with respect to the predictor variable. We observe that the median Δt_{LOS} monotonously increases from age-bins (20-29) to (50-59), and then further monotonously decreases with age (§Figure 2 (ii)). There is anomalous behavior in age-bin (0-19). Unlike age, we observe that Δt_{LOS} shows monotonous behavior in the case of gender and period (§Figure 2 (i), (iii)). Hence, we use the piecewise GLMs for regression with respect to age-bins.

We have implemented the hypothetical tests using `kruskal.test` and `wilcox.test` in Stats package in R [12]. The null hypothesis H_0 is that "all groups are statistically similar," and the alternative H_1 hypothesis is that "all groups are not statistically similar." We reject the H_0 when p-value <5%. Our results (§Table 1) show that groups based on the period with $\mathbb{G}_{\mathbf{+cfrmDt}}$ grouping show statistically significant differences, thus indicating that period influences Δt_{LOS}.

[7] The loess model is a default local regression model used for a sample with less than 1000 observations in Stats package in R.

With respect to our study on vulnerable populations, we see that A_2 sub-population with the $\mathbb{G}_{+\mathbf{cfrmDt}}$ grouping shows statistical significance with respect to the period, unlike A_1 and A_3. We observe that both male and female separately have statistical differences with respect to period. Overall, these results confirm the use of age-binning for piecewise GLMs.

We consider two modeling scenarios for GLM, where we use age, gender, and period as independent variables: M_1 which is a single GLM for all 245 patients; and M_2 with three piecewise linear GLM's for A_1, A_2, and A_3, respectively. (*The count distribution and approximation of distribution functions are shown in Figure S3, Supplementary File.*) We have implemented these models using the `glm` in Stats package in R [12]. For GLM with Poisson and binomial families, the dispersion is fixed at 1.0, and the number of parameters (k) is the same as the number of coefficients in the regression model [12]. The negative binomial distribution has an additional parameter to model over-dispersion in the data. We use the Akaike Information Criterion (AIC) and its corrected version for a small sample size (AIC_c) for determining the *goodness of fit* of our proposed models. For the number of samples (N) in the data, we use AIC when $\left(\frac{N}{k} > 40\right)$, and AIC_c, otherwise [13]. We infer the following from our results (§Table 2):

- The "period" variable is significant for both models and all scenarios, except M_2 (age:(0-19)), with p-value $<5\%$ for corresponding coefficients.
- The "age" and "gender" variables are significant only in the PRM for M_2 (age:(20-59)) and M_2 (age:(60-89)) respectively, with p-value of corresponding coefficients being $<5\%$, confirming the age-bin (20-59) getting progressively affected with time, and co-morbidities in older males, respectively.
- The NBM shows lower medians and the variances (observable from range and IQR) of deviance residuals, and AIC/AIC_c and more consistent Δt_{LOS} for maximum likelihood value than the PRM. Thus, NBM is a better fit than PRM.
- Overall, the small sample size in M_2 (age:(0-19)) makes the modeling erroneous. For both PRM and NBM, the remaining piecewise models in scenario M_2 are a better fit than the single model, M_1, owing to the lower error measures (AIC/AIC_c and deviance residuals) as well as the consistency in the maximum likelihood value. We also find that the errors, MSE and RMSE in M_1, are reduced when we exclude the samples of (age:(0-19)) from the observations.

Overall, choosing NBM and M_2, we estimate $\Delta t_{\text{LOS}} = 9$ days with a maximum likelihood of 7.5%. Since we do not have a large number of variables to discard, we retain the "gender" and "age" variables in the model despite its relative insignificance compared to "period." Our additional experiments show that the error values increase in the absence of the period as an independent variable. The key conclusion from the hypothesis testing and multivariate regression analysis is that a GLM with NBM for M_2 is the best model to estimate Δt_{LOS}.

4 Discussion

In summary, Δt_{LOS} can be estimated to have decreased to 9 days during January 23-April 01, 2020, in Singapore. We have shown that the recovery in the vulnerable groups, such as the children and the older population, has become insignificant with time. Our work shows that this temporal shift in the demographic structure has led to the overall decrease in Δt_{LOS}, which is reflected in two results. Firstly, we observe "period" causes significant differences in the distribution of Δt_{LOS} in both the entire population, as well as, specifically, in age-bin (20-59) (A_2) subpopulation. Secondly, we find that "age" is a significant variable for the multivariate regression model with PRM for the population of the age-bin (20-59) years in M_2. These indicate that our age-binning and periodization are significant in analyzing and estimating Δt_{LOS}. Periodwise piecewise GLM's of observations in A_2, with age and gender as predictor variables, show that the estimate of Δt_{LOS} reduces from $\sim$16 days in P_1 to $\sim$9 days in P_2, and further to $\sim$8 days in P_3. (*Details of the periodwise GLMs are in Table S3 in Supplementary File.*) Thus, our findings of the progressive decrease in Δt_{LOS} are confirmed using these models. The overall reduction in Δt_{LOS} is an outcome of the significance of Δt_{LOS} in A_2, which is the non-vulnerable population.

Our findings can be attributed to the early response of restricted mobility of the vulnerable groups. The vulnerable population becoming insignificant can be an outcome of the increase in positive confirmations in the non-vulnerable group. However, our work has the limitation of modeling recovery exclusively, without the interaction with the distribution of daily positive confirmations. With $N_i = 1000$ in Singapore by April 1, but $N_r = 245$, we have not analyzed the outcome of the remaining 755 patients. Our study can be enriched by doing a similar analysis on all the cases confirmed by April 1, closed by recovery and death. Our regression model can be improved to include interaction effects, which could be effective if we have additional data on other variables, e.g., co-morbidities, whose data is not available in the public domain. Nevertheless, our estimated Δt_{LOS} based on the first 245 discharged patients is useful for assessing the hospital load in terms of required resources for the in-hospital stay of the COVID-19 cases.

Acknowledgements

The authors would like to thank IIIT Bangalore, and particularly, the Graphics-Visualization-Computing Lab and the E-Health Research Center, for supporting this work. The authors appreciate the constructive review from Srujana Merugu in shaping this work. This work has been partially supported by the IBM Shared University Grant awarded to J. Sreevalsan-Nair and the Visvesvaraya Ph.D. Fellowship by the Ministry of Electronics and IT (MeitY), Government of India, awarded to R. R. Vangimalla.

References

[1] Rachael Pung, Calvin J. Chiew, Barnaby E. Young, Sarah Chin, Mark I-C Chen, Hannah E. Clapham, Alex R. Cook, Sebastian Maurer-Stroh, Matthias P. H. S. Toh, Cuiqin Poh, Mabel Low, Joshua Lum, Valerie T. J. Koh, Tze M. Mak, Lin Cui, Raymond V. T. P. Lin, Derrick Heng, Yee-Sin Leo, David C. Lye, and Vernon J. M. Lee. Investigation of three clusters of COVID-19 in Singapore: implications for surveillance and response measures. *The Lancet*, 395:1039–46, 2020.

[2] Nanshan Chen, Min Zhou, Xuan Dong, Jieming Qu, Fengyun Gong, Yang Han, Yang Qiu, Jingli Wang, Ying Liu, Yuan Wei, and et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. *The Lancet*, 395(10223):507–513, 2020.

[3] Xiaobo Yang, Yuan Yu, Jiqian Xu, Huaqing Shu, Jia'an Xia, Hong Liu, Yongran Wu, Lu Zhang, Zhui Yu, Minghao Fang, Ting Yu, Yaxin Wang, Shangwen Pan, Xiaojing Zou, Shiying Yuan, and You Shang. Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study. *The Lancet Respiratory Medicine*, 2020.

[4] Adam J Kucharski, Timothy W. Russell, Charlie Diamond, Yang Liu, John Edmunds, Sebastian Funk, and Rosalind M. Eggo. Early dynamics of transmission and control of COVID-19: a mathematical modelling study. *The Lancet Infectious Diseases*, 2020.

[5] Qun Li, Xuhua Guan, Peng Wu, Xiaoye Wang, Lei Zhou, Yeqing Tong, Ruiqi Ren, Kathy S.M. Leung, Eric H.Y. Lau, Jessica Y. Wong, Xuesen Xing, Nijuan Xiang, Yang Wu, Chao Li, Qi Chen, Dan Li, Tian Liu, Jing Zhao, Man Liu, Wenxiao Tu, Chuding Chen, Lianmei Jin, Rui Yang, Qi Wang, Suhua Zhou, Rui Wang, Hui Liu, Yinbo Luo, Yuan Liu, Ge Shao, Huan Li, Zhongfa Tao, Yang Yang, Zhiqiang Deng, Boxi Liu, Zhitao Ma, Yanping Zhang, Guoqing Shi, Tommy T.Y. Lam, Joseph T. Wu, George F. Gao, Benjamin J. Cowling, Bo Yang, Gabriel M. Leung, and Zijian Feng. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia. *New England Journal of Medicine*, 382(13):1199–1207, 2020.

[6] Cleo Anastassopoulou, Lucia Russo, Athanasios Tsakris, and Constantinos Siettos. Data-based analysis, modelling and forecasting of the COVID-19 outbreak. *PloS one*, 15(3):e0230405, 2020.

[7] Fei Zhou, Ting Yu, Ronghui Du, Guohui Fan, Ying Liu, Zhibo Liu, Jie Xiang, Yeming Wang, Bin Song, Xiaoying Gu, Lulu Guan, Yuan Wei, Hui Li, Xudong Wu, Jiuyang Xu, Shengjin Tu, Yi Zhang, Hua Chen, and Bin Cao. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *The Lancet*, 395(10229):1054–62, 2020.

[8] William S Cleveland and Susan J Devlin. Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American statistical association*, 83(403):596–610, 1988.

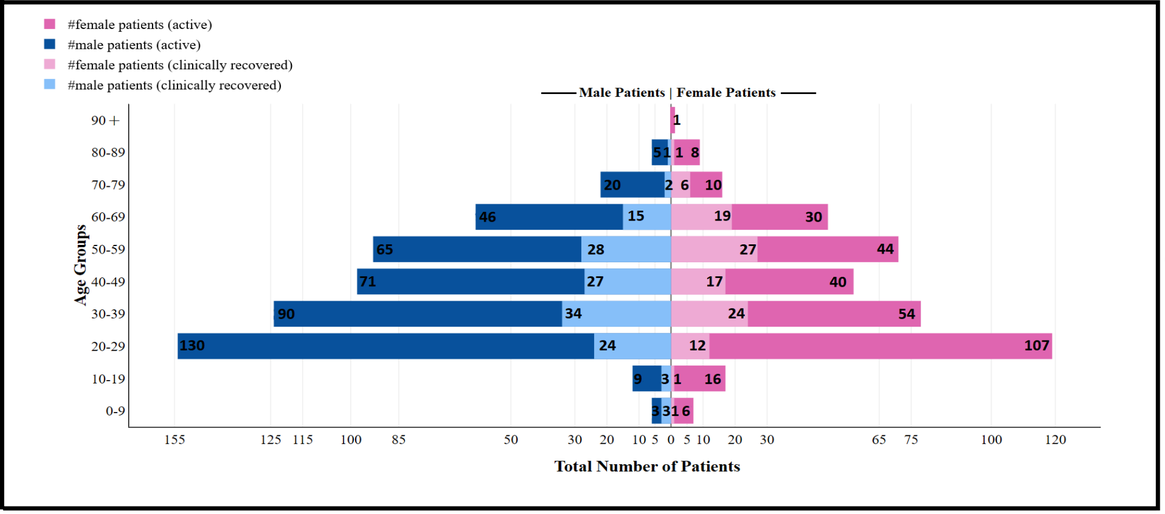
[9] William G Jacoby. Loess:: a nonparametric, graphical tool for depicting relationships between variables. *Electoral Studies*, 19(4):577–613, 2000.

[10] Evelene M Carter and Henry WW Potts. Predicting length of stay from an electronic patient record system: a primary total knee replacement example. *BMC medical informatics and decision making*, 14(1):26, 2014.

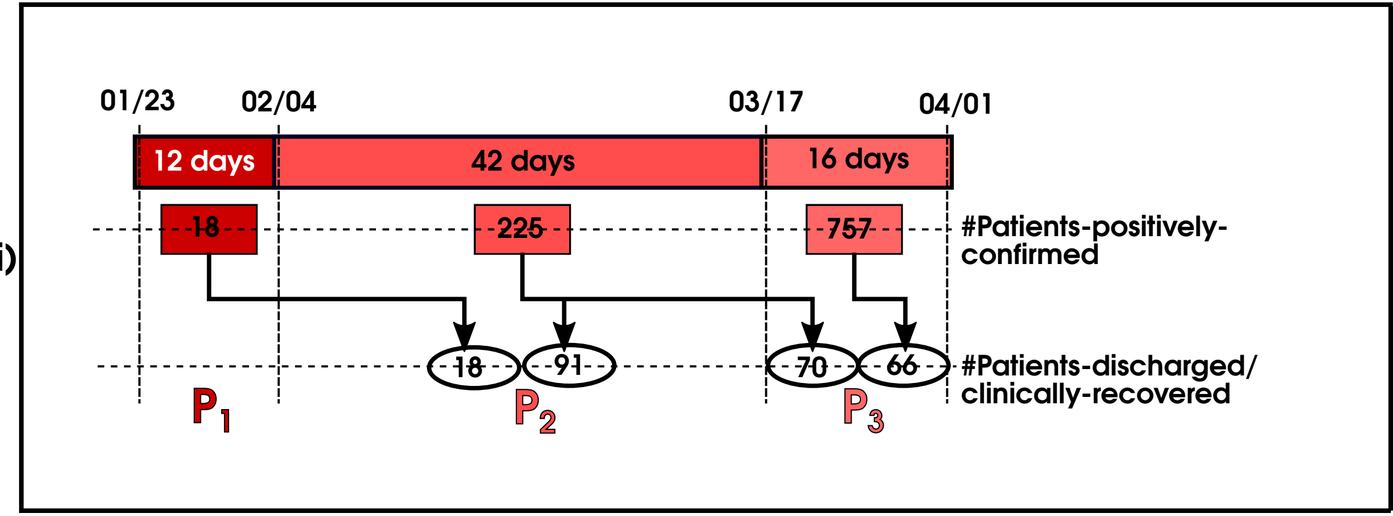
[11] Liming Xiang, Andy H Lee, Kelvin KW Yau, and Geoffrey J McLachlan. A score test for overdispersion in zero-inflated poisson mixed regression model. *Statistics in medicine*, 26(7):1608–1622, 2007.

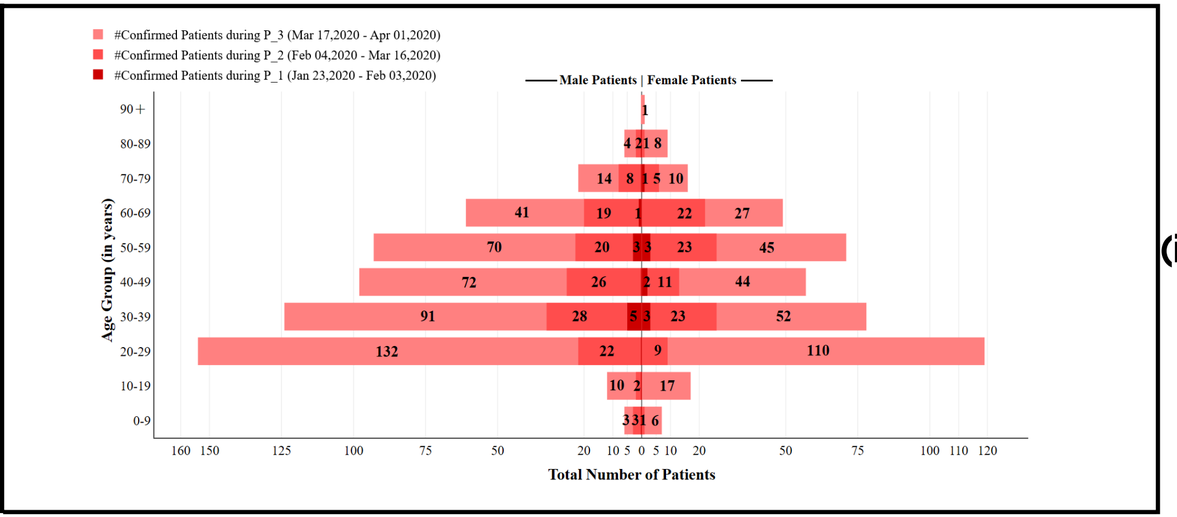
[12] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. ISBN 3-900051-07-0.

[13] K. P. Burnham and D. R. Anderson. *Model selection and inference: a practical information-theoretic approach*. Springer Verlag, 1998.

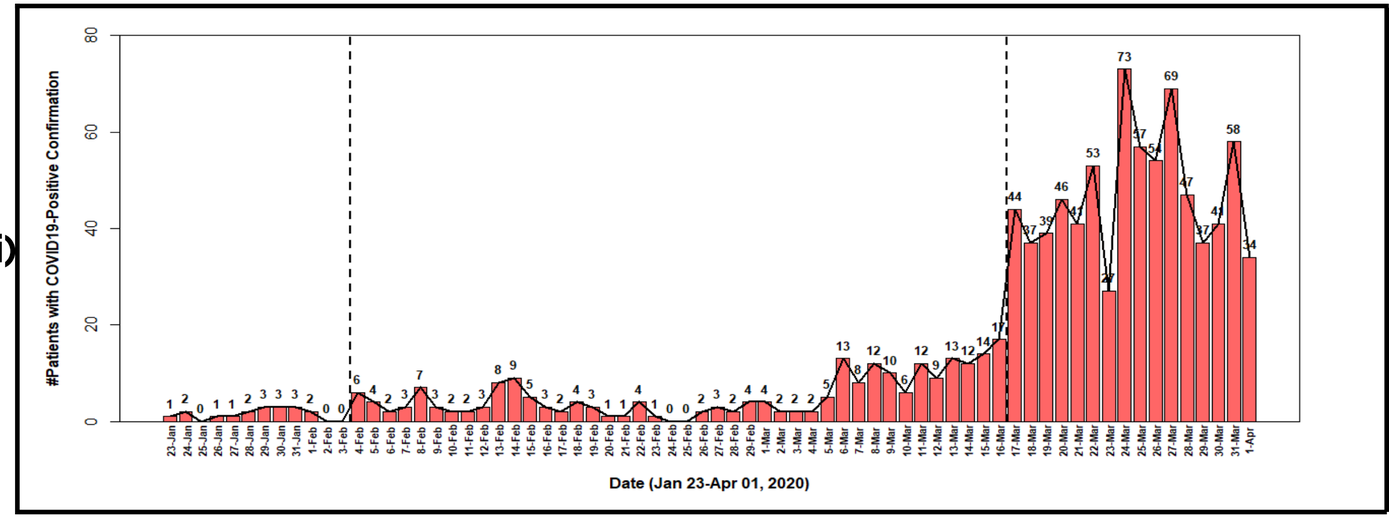

#female patients (active)
#male patients (active)
#female patients (clinically recovered)
#male patients (clinically recovered)
Male Patients | Female Patients
Age Groups
Total Number of Patients

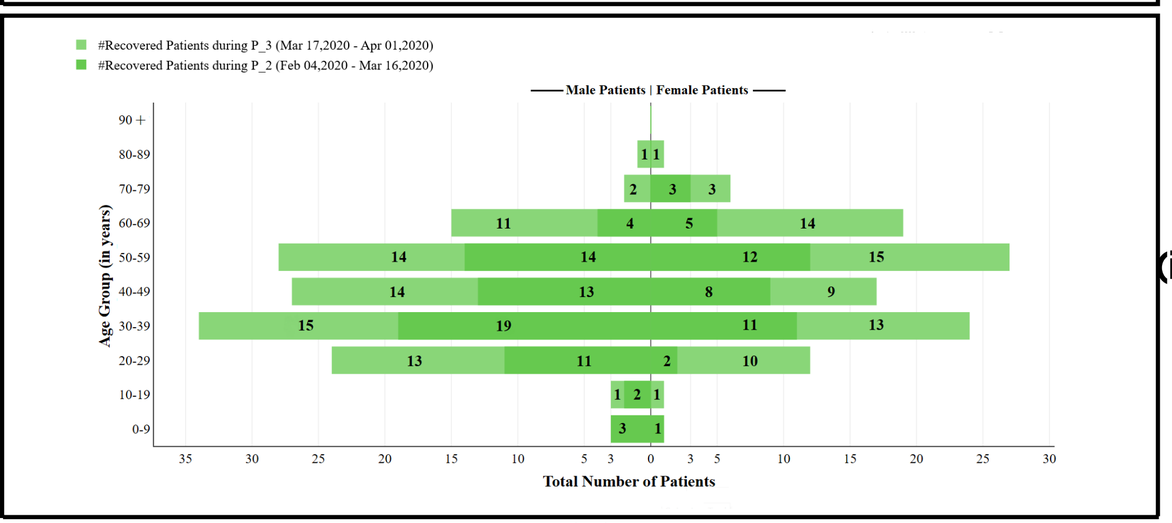
(i)


01/23
02/04
03/17
04/01
12 days
42 days
16 days
18
225
757
#Patients-positively-confirmed
18
91
70
66
#Patients-discharged/ clinically-recovered
P1
P2
P3

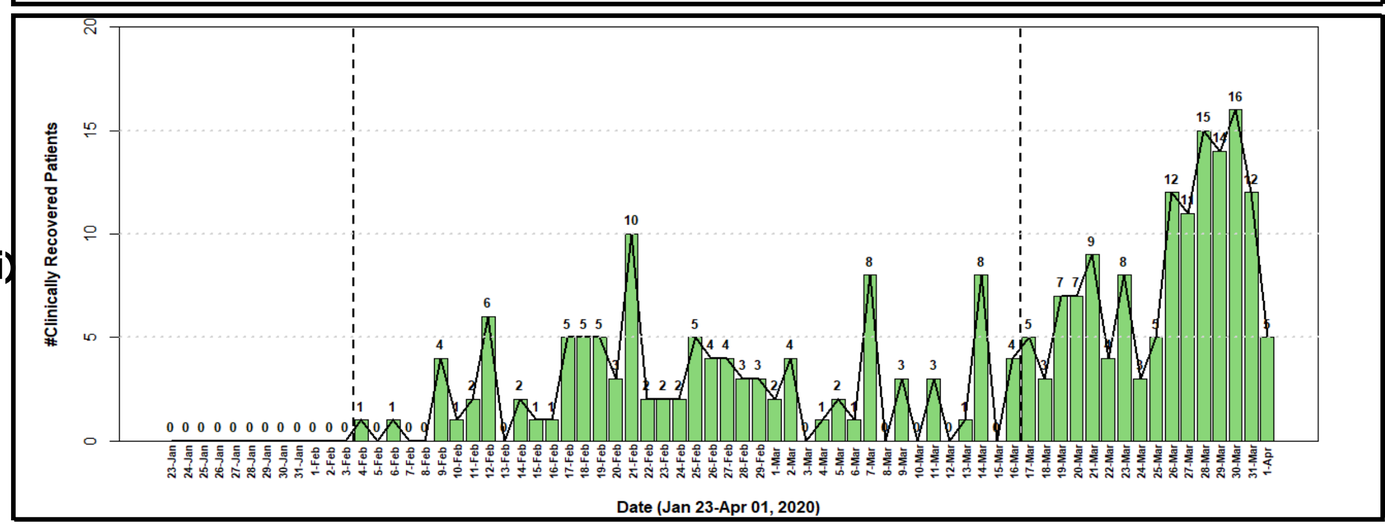

#Confirmed Patients during P_3 (Mar 17,2020 - Apr 01,2020)
#Confirmed Patients during P_2 (Feb 04,2020 - Mar 16,2020)
#Confirmed Patients during P_1 (Jan 23,2020 - Feb 03,2020)
Male Patients | Female Patients
Age Group (in years)
Total Number of Patients

(ii)


#Patients with COVID19-Positive Confirmation
Date (Jan 23-Apr 01, 2020)

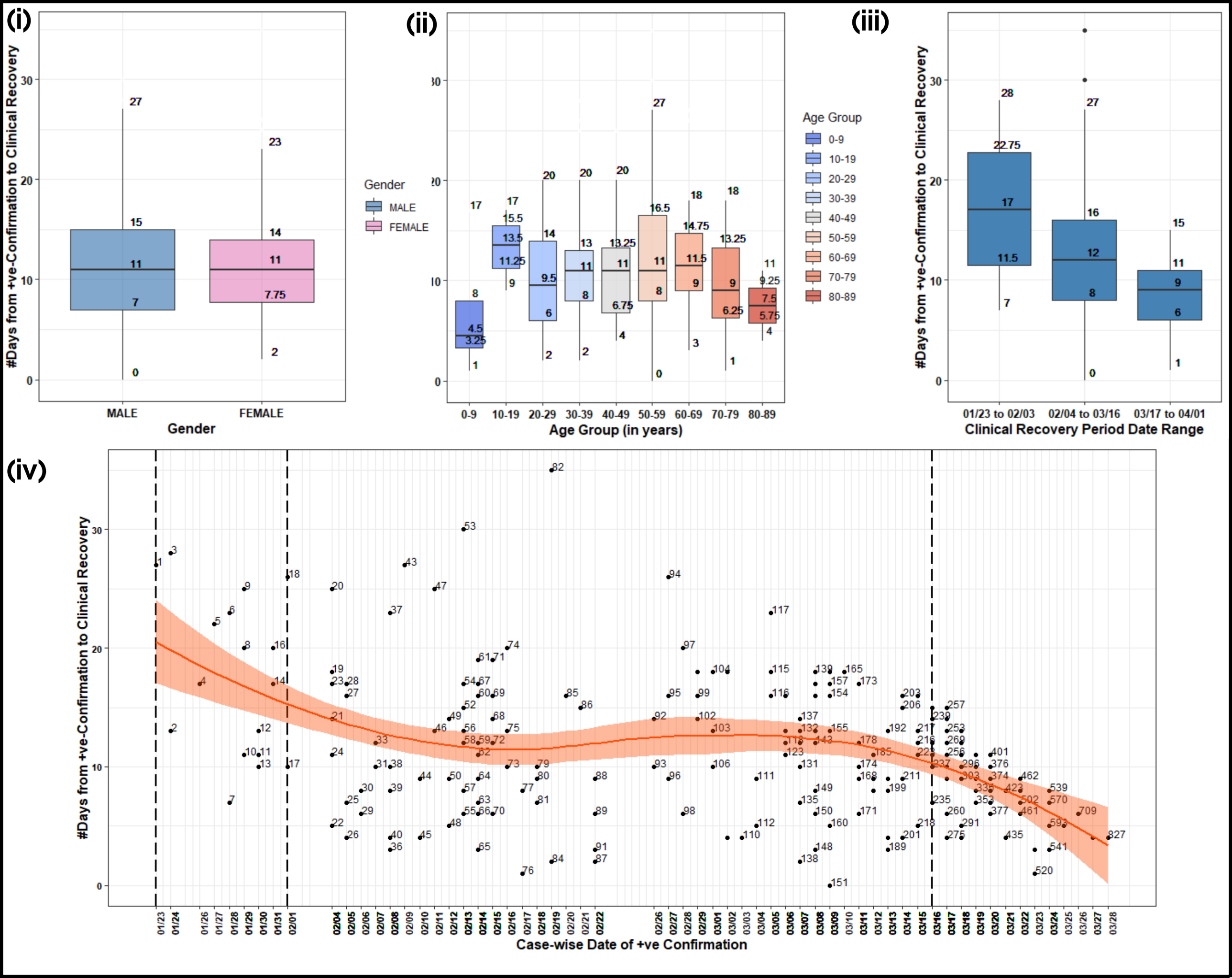

#Recovered Patients during P_3 (Mar 17,2020 - Apr 01,2020)
#Recovered Patients during P_2 (Feb 04,2020 - Mar 16,2020)
Male Patients | Female Patients
Age Group (in years)
Total Number of Patients

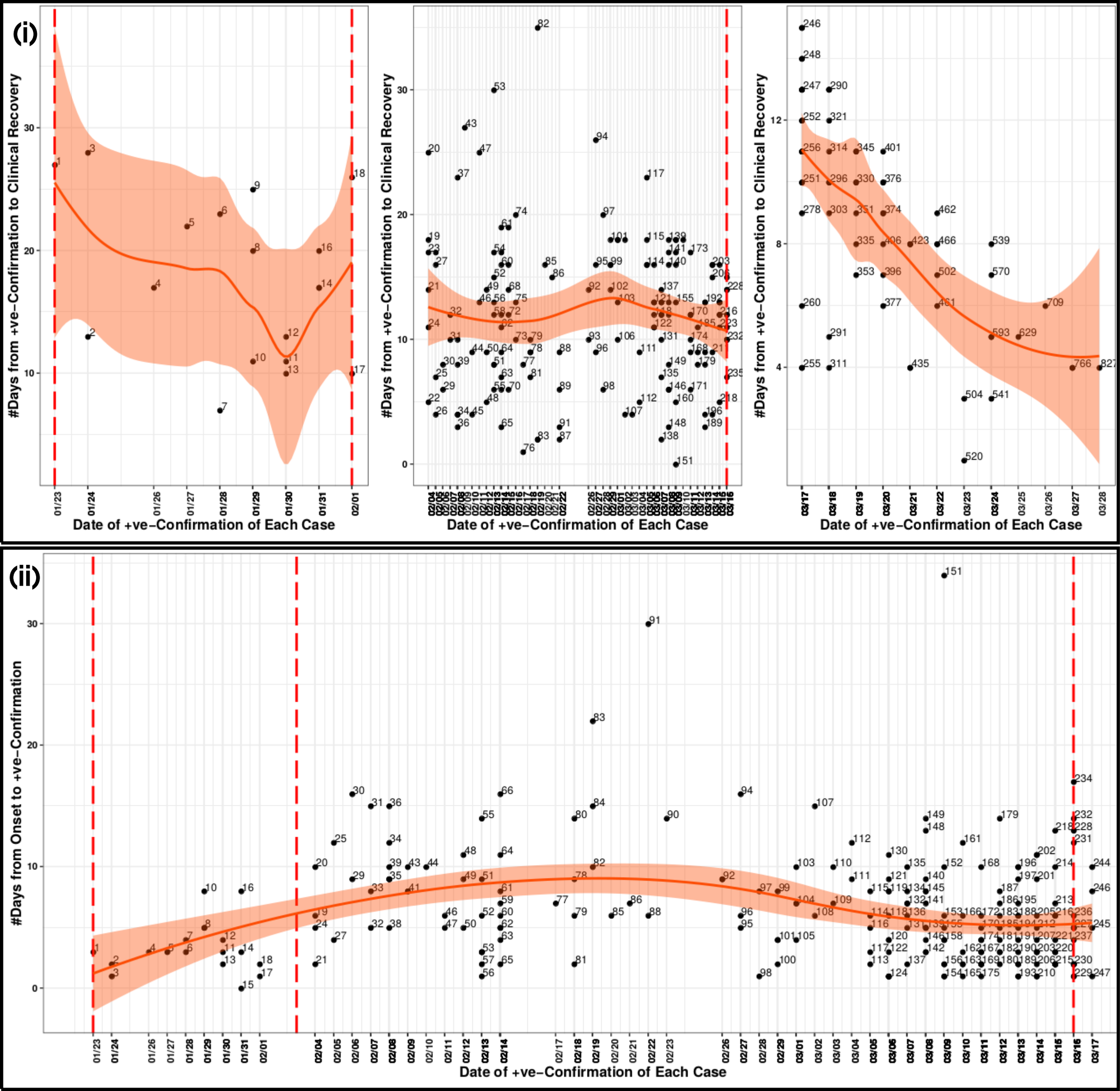
(iii)

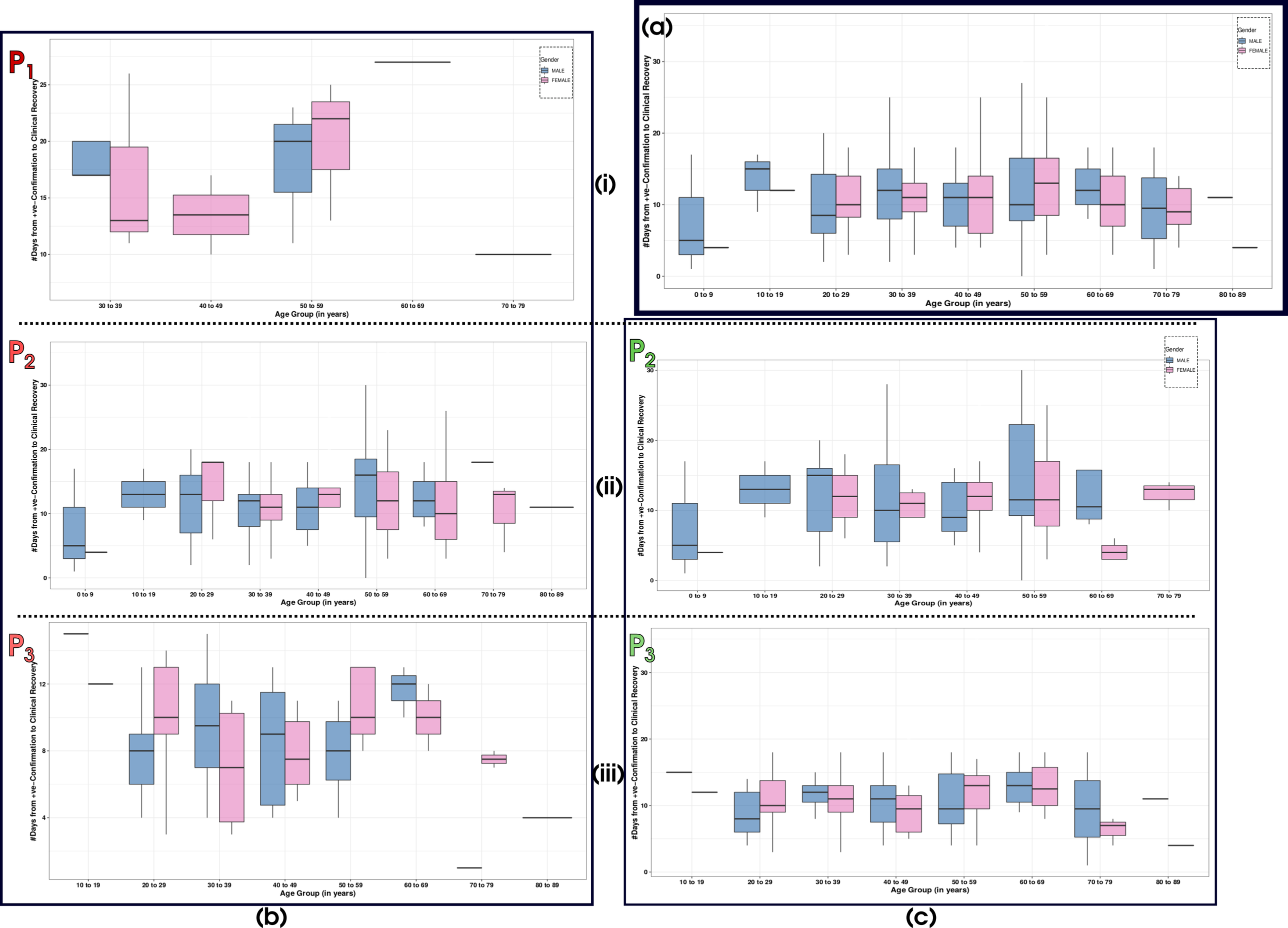

#Clinically Recovered Patients
Date (Jan 23-Apr 01, 2020)

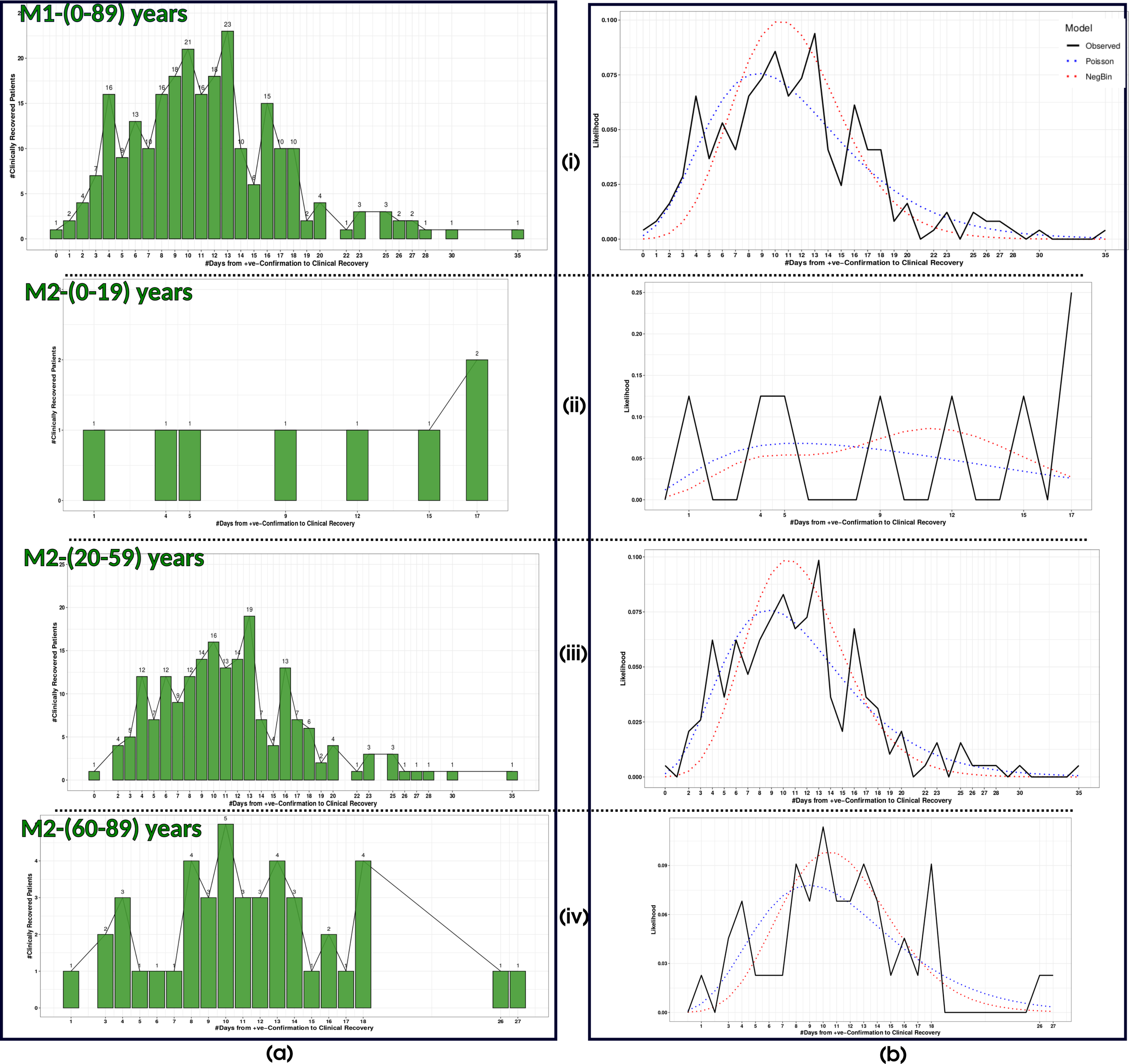
(a)

(b)


(i)
#Days from +ve-Confirmation to Clinical Recovery
MALE
FEMALE
Gender
(ii)
Gender
MALE
FEMALE
0-9
10-19
20-29
30-39
40-49
50-59
60-69
70-79
80-89
Age Group (in years)
Age Group
(iii)
#Days from +ve-Confirmation to Clinical Recovery
01/23 to 02/03
02/04 to 03/16
03/17 to 04/01
Clinical Recovery Period Date Range
(iv)
#Days from +ve-Confirmation to Clinical Recovery
Case-wise Date of +ve Confirmation


(i)
#Days from +ve-Confirmation to Clinical Recovery
Date of +ve-Confirmation of Each Case
(ii)
#Days from Onset to +ve-Confirmation
Date of +ve-Confirmation of Each Case


(a)
(i)
(ii)
(iii)
(b)
(c)
P1
P2
P3
Gender
MALE
FEMALE
#Days from +ve–Confirmation to Clinical Recovery
Age Group (in years)
0 to 9
10 to 19
20 to 29
30 to 39
40 to 49
50 to 59
60 to 69
70 to 79
80 to 89


M1-(0-89) years
M2-(0-19) years
M2-(20-59) years
M2-(60-89) years
#Clinically Recovered Patients
#Days from +ve-Confirmation to Clinical Recovery
Likelihood
Model
Observed
Poisson
NegBin
(i)
(ii)
(iii)
(iv)
(a)
(b)