ANALYSIS OF CLINICAL RECOVERY-PERIOD AND RECOVERY RATE ESTIMATION OF THE FIRST 1000 COVID-19 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

April 19, 2020

ABSTRACT

COVID-19 has been declared as a global pandemic by the World Health Organization (WHO) on March 11, 2020. In this paper, we investigate various aspects of the clinical recovery of the first 1000 COVID-19 patients in Singapore, spanning from January 23 to April 01, 2020. This data consists of 245 clinically recovered patients. The first part of the paper studies the descriptive statistics and the influence of demographic parameters, namely age and gender, in the clinical recovery-period of COVID-19 patients. The second part of the paper is on identifying the distribution of the length of the recovery-period for the patients. We identify a piecewise analysis of three different periods, identified based on trends of both positive confirmation and clinical recovery of COVID-19. As expected, the overall recovery rate has reduced drastically during the exponential increase of incidences. However, our in-depth analysis shows that there is a shift in the age-group of incidences to the younger population, and the recovery-period of the younger population is considerably lower. Here, we have estimated the recovery rate to be 0.125. Overall, the prognosis of COVID-19 indicates an improvement in recovery rate owing to the government-mandated practices of restricted mobility of the older population and aggressive contact tracing.

Keywords COVID19, Clinical recovery

1 Introduction

The viral contagion, named COVID-19, has been declared as a global pandemic by the World Health Organization 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, 2020, there are 827,419 positively confirmed cases, and 40,777 deaths, spread across 206 countries [3]. The total number of recovered patients in an unofficial count is 193,989 out of 935,197 positively confirmed patients, which implies that the ratio of the recovered to the infected patients, r_{ri} is $\sim$0.2, as of April 1, 2020 [4].

In this paper, we analyze the statistics of hospital recovery of patients tested positive for COVID-19 infection in Singapore [1]. The case study of Singapore has been carefully chosen owing to the reliable, accessible, and available data from the official press releases of the Ministry of Health (MoH), Government of Singapore. The healthcare system of Singapore has been unique in its handling of the widespread contagion in terms of imposing strict lockdown, quarantine, and isolation, and aggressive, large-scale contact tracing and testing. 1000 individuals have been confirmed

*jnair@iiitb.ac.in
[2] https://www.who.int/dg/speeches
[3] https://www.who.int/emergencies/diseases/novel-coronavirus-2019
[4] As retrieved on April 02, 2020, from https://www.worldometers.info/coronavirus/

positive for COVID-19 during January 23-April 01, 2020, of which 245 have been discharged after clinical recovery. This gives an overall r_{ri} higher than that of the world, as r_{ri}=0.245. The positive confirmation has been made using the real-time reverse transcription polymerase chain reaction (RT-PCR) tests on respiratory samples (sputum or nasal/throat/nasopharyngeal swabs), based on experiential learning from the outbreak of SARS in 2003 [5]. Similarly, the protocol for clinical recovery or hospital discharge has been based on the results on the RT-PCR tests of two consecutive samples being negative over two days [2].

Since the SARS outbreak in 2003, Singapore has systematically strengthened the system of managing the spread of infectious diseases [2]. The measures include opening dedicated facilities (the National Center for Infectious Diseases (NCID), National Public Health Laboratory, and more biosafety level-3 laboratories), increasing capacity in the public healthcare system (e.g., negative pressure isolation beds, personal protective equipment, trained health professionals), and deploying formal (digital) platforms for inter-governmental agency cooperation. For containing the spread of contagions, systems have been in place for upscaled, quick-responsive, and aggressive contact tracing at entry points of the country (airports) and through local healthcare providers. There has been a holistic improvement, supported by increased economic investment, in building expertise in infectious disease management. This organized system has thus facilitated a controlled management of the pandemic COVID-19 in Singapore with patient-wise reporting to the public. Hence, a case study in Singapore pertaining to the demographic analysis of clinically recovered patients enables a systematic understanding of the recovery rate (γ) of the pandemic.

One of the significant benefits of aggressive contact tracing has been hospital isolation within 5 days from the onset of symptoms [3]. However, 12.6% of transmission has been found to be presymptomatic [4], as analysed for seven clusters found in Singapore. While there is an innate uncertainty from the onset of symptoms in the symptomatic cases to a positive confirmation, the hospital stay from positive confirmation to discharge upon negative results exhibits more cohesive statistics, as can be observed with similar studies in hospital stay [5]. Hence, in this work, we perform statistical analysis of the *recovery-period*, i.e., length of hospital stay, of COVID-19 patients in the hospital. Also, the motivation behind studying the clinical recovery of patients is to assess the overall load on the healthcare system in terms of patient occupancy as the hospital stays determines the load. The clinical recovery studied in this paper corresponds to the hospitalization period for each patient. The demographic analysis of the recovered patients gives insight to shifts in gender and age-groups in the recovery of patients. This analysis further complements the observation in the transmission rate is higher in older males with comorbidities [6]. Fitting the data of hospitalization period to a statistical distribution is essential for estimating γ.

The epidemiological models are generally used to simulate the progression of a disease. The proportions of population being "susceptible," "infected," and "removed" are used in these models. "Removed" implies both "recovered" and "deceased." The first two deaths in Singapore owing to the COVID-19 contagion occurred on March 21, 2020. The number increased to 3 deaths by April 01. Owing to relatively low number of deaths in Singapore due to COVID-19 contagion during January 23-April 01, 2020, we have assumed death/mortality/fatality rate to be 0 in our work. Thus, here, "recovery" implies the state of "clinically recovered and discharged from hospital."

Since the contagion has time-varying reproduction number (R_t) with characteristic trends in specific time-periods [7], we split the time-period of January 23-April 01, to perform a piecewise analysis of the timeline [7, 8]. We perform two analyses on the periodized timeline. Firstly, we study the age-gender distribution of the patients who have been confirmed positive of COVID-19 and those who have clinically recovered. Secondly, we extract the distribution of clinical recovery-periods and fit regression models. In both analyses, we discuss the observable period-wise shifts in trends and their influencing factors, thus estimating γ. The novel contribution of our work is an in-depth analysis of the clinical recovery of COVID-19 patients to estimate the recovery rate γ, which is a key parameter in the SIR (susceptible-infected-recovered) model for the disease [9].

2 Methods

The data for our work has been collated from the public press releases made by the MoH, the Government of Singapore [6]. This dataset includes the case-ID's, age, gender, positive confirmation date, discharge date, and date of onset of symptoms[7]. The data has been cross-verified with dashboard [8] for case details. We have analyzed this patient-wise data pertaining to age, gender, and timeline of the disease progression.

[5] https://www.cdc.gov/mmwr/preview/mmwrhtml/mm5218a1.htm

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

[7] The press releases did not carry patient-wise information after March 17, 2020, owing to which the date of onset of symptoms is not available beyond that date.

[8] https://experience.arcgis.com/experience/7e30edc490a5441a874f9efe67bd8b89

We define *recovery-period* Δt_r as the time elapsed between positive COVID-19 confirmation using RT-PCR test, and the discharge date from hospital after two consecutive negative results, using RT-PCR tests. Owing to the strict protocols followed in the Singapore healthcare system, the recovery-period can be considered equivalent to the *virus shedding period*. Δt_r is estimated to be 15 days [10] or 20 days [11]. We consider Δt_r as an *observed* count variable.

Age-gender distribution: There has been early evidence of the influence of both age and gender in susceptibility of COVID-19 infection [6, 12]. Hence, we look at the influence of age and gender in clinical recovery of patients, with respect to the recovery-period.

There are 1000 patients (576M, 424F) in the population confirmed COVID-19 positive, of which 245 patients (137M,108F) have clinically recovered (§Figure 1(i)). The age distribution of the positively confirmed patients is (13 in [0-9], 29 in [10-19], 273 in [20-29], 202 in [30-39], 155 in [40-49], 164 in [50-59], 110 in [60-69], 38 in [70-79], 15 in [80-89], 1 in [100-]) patients in age groups given in years in square brackets. Similarly the age distribution of the clinically recovered in (4 in [0-9], 4 in [10-19], 36 in [20-29], 58 in [30-39], 44 in [40-49], 55 in [50-59], 34 in [60-69], 8 in [70-79], 2 in [80-89]). The preliminary counts indicate that the restrictions used for slowing the contagion down in the susceptible group of the population, namely the older males [12], have led to the shifts in distribution of contagion with respect to age and gender. There is a conspicuous shift to the younger population, namely, in the age groups [20-29], and [30-39], and more evenly across both genders. Here, we investigate the change in recovery rate γ owing to these shifts.

Periodization: As the pandemic progresses, the evolution needs to be studied piecewise in different time periods [7, 8]. From the timeline of disease progression in Singapore, we have identified the following significant dates:

- On January 23, the first patient was confirmed COVID-19 positive.
- On February 4, the first clinically recovered patient was discharged from the hospital.
- On March 17, the total/cumulative number of positively confirmed patients started increasing exponentially with 44 new cases on a single day.
- On March 21, the first two deaths owing to complications from COVID-19 were recorded.
- On April 1, the number of confirmed patients reached 1000.

These trends can be observed in the daily profile of patient counts (§Figure 1 (ii),(iii)). Assuming a zero death rate owing to the low number of deaths in Singapore from COVID-19, we consider the following three periods:

1. Period P_1 during January 23-February 3, which is the *period with no clinically recovered cases*.
2. Period P_2 during February 4-March 16, which is the period of slow growth in the total/cumulative number of positively COVID-19 confirmed cases, N_i, with an increase in the total/cumulative number of clinically recovered cases, N_r, and zero deaths.
3. Period P_3 during March 17-April 1, which is the period of exponential growth in N_i, reaching $N_i = 1000$, slow growth in N_r, and having the first 3 deaths.

Recovery rate γ: The governing differential equations in the simplest SIR model, also known as Kermack-McKendrick model [9], are given as follows:

$$\begin{aligned} \frac{dN_s}{dt} &= -\beta.\frac{N_s}{N_p}.N_i, \\ \frac{dN_i}{dt} &= \beta.\frac{N_s}{N_p}.N_i - \gamma.N_i, \\ \frac{dN_r}{dt} &= \gamma.N_i \end{aligned} \tag{1}$$

β is the rate at which an infected individual infects others, γ the transition rate in SIR model[9], N_p the size of the population, N_i the number of infected persons (i.e., with positive COVID-19 confirmation), N_r the number of recovered persons (i.e., clinically recovered), and N_s is the number of susceptible people. The basic reproduction number or reproduction rate, $\mathcal{R}_0 = \frac{\beta}{\gamma}$, characterizes an infection. $\mathcal{R}_0 > 1$ implies the infection will continue to spread, and $\mathcal{R}_0 < 1$ implies that the spread is limited and under control. Currently, $\mathcal{R}_0$ for COVID-19 is estimated to be (0.8-5.0)[8, 13]. γ is estimated as the reciprocal of the recovery-period Δt_r, which implies that $\gamma \sim$ (0.05-0.067), based on estimates of Δt_r [10, 11].

[9]The transition rate must include both recovery and deceased. However, since we assume a zero death rate, the transition rate is equivalent to recovery rate, in our work.

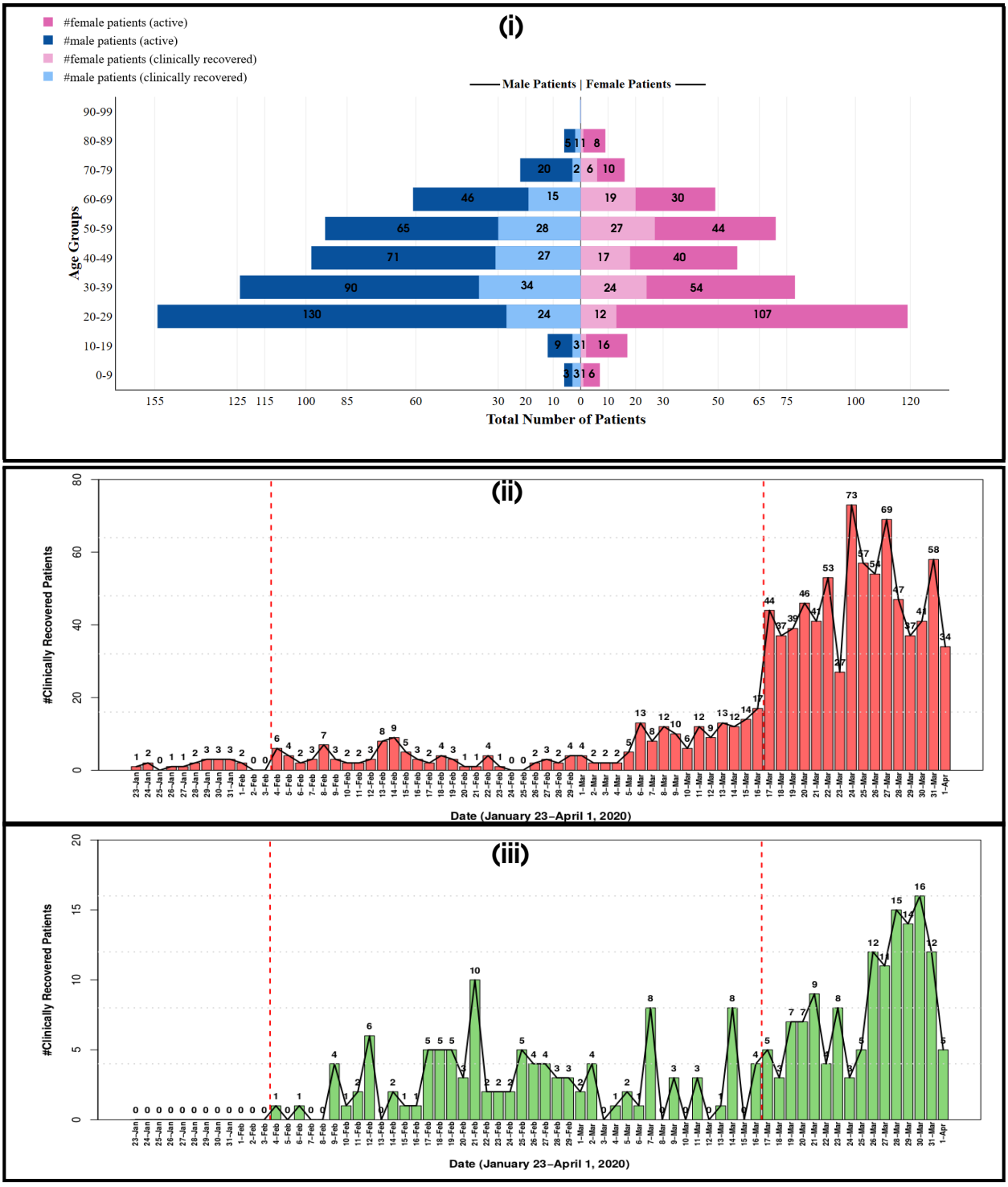


Figure 1: (i) Stacked population pyramid of age-gender distribution of population of the 245 discharged and 754 active (i.e., COVID-19 positive but not discharged, including the deceased) during the same period (1 active patient who is a 102 years old female and confirmed positive on April 01, 2020, has been excluded in this population pyramid). Daily profile of count of patients in Singapore during January 23-April 01, 2020, (ii) who were confirmed positive for COVID-19, with a total of 1000, and (iii) who got clinically recovered and were discharged from the hospitals, with a total of 245. The red dotted lines indicate the three periods we have introduced in this work.

The absolute numbers indicate that r_{ri}, i.e., $\frac{N_r}{N_i}$, has increased from 0.00 at the end of P_1 to 0.45 $\left(=\frac{109}{18+225}\right)$ at the end of P_2, and again dipped to 0.245 $\left(=\frac{245}{1000}\right)$ at the end of P_3. The dip is unfavorable for the scenario, given that the number of positive COVID-19 confirmations has increased exponentially, starting the beginning of P_3. Since r_{ri} and γ are positively correlated, it implies a further decrease in γ.

However, the absolute counts N_i and N_r do not explicitly show the shifts in the age-gender distribution of the infected population (§Figure 2(i)). Thus, it is difficult to demonstrate the influence of this shift in the recovery rate γ. Social distancing reduces β, thus decreasing $\mathcal{R}_0$. At the same time, increasing γ also favours a decrease in $\mathcal{R}_0$. In this work, we hypothesize that restricted mobilization and aggressive contact tracing would have indirectly increased γ. Thus, we propose computing the time-varying Δt_r using the shifts in the age-gender distribution in the periodized timeline.

Recovery-period Δt_r analysis: Our goal is to study the period-wise changes in Δt_r owing to the shift in the demographic structure, in order to determine the period-wise change in γ. We performed descriptive statistical analysis using median and interquartile range (IQR), followed by fitting appropriate regression models. We consider two types of regression models. Firstly, we use the time-series of Δt_r and fit a line using the loess model. Loess is a non-parametric local regression model for smoothening empirical time-series data [14] and scatterplots [15]. Secondly, we use the number of patients recovering for a specific Δt_r as a count variable and fit multivariate linear regression models considering age and gender as independent variables, and Δt_r as the dependent variable. Since we are using a combination of a categorical variable (gender) and numerical variable (age), we use generalized linear models (GLM) for regression, which is semi-parametric. Length of hospital stay (LoS) is a *naturally skewed* distribution, for which GLM's such as the Poisson regression model (PRM) and negative binomial regression model (NBM) have been used [5]. Hence, we propose the use of PRM and NBM for modeling Δt_r.

For the period-wise analysis of Δt_r, we group the clinically recovered patients using two strategies.

1. Grouping based on recovery date, $\mathbb{G}_{\mathbf{-cfrmDt}}$: The two groups of clinically recovered patients can be obtained as per the period in which their discharge/recovery date falls, namely, P_2 and P_3 (§Figure 2 (ii),(a)). $\mathbb{G}_{\mathbf{-cfrmDt}}$ corresponds to a group of 109 patients who were discharged in P_2, and 136 patients in P_3.
2. Grouping based on positive confirmation date, $\mathbb{G}_{\mathbf{+cfrmDt}}$: However, we can perform a finer-grain analysis of the groups of patients based on the period in which their date of positive confirmation/hospital admission falls (§Figure 2 (ii),(b)). This gives us groups of patients who got tested positive in a period and got clinically recovered during the entire period of our study here. This gives us 18 patients in P_1, 161 in P_2, and 66 in P_3.

Symptom-onset period Δt_{so} analysis: We additionally have data of the date of onset of symptoms for 227 of the 245 clinically recovered patients, who were confirmed positive during P_1 and P_2. We define the *symptom-onset period*, Δt_{so}, as the number of days between the onset of symptoms and positive confirmation of COVID-19. We use the time-series of Δt_{so} to fit a loess model, similar to Δt_r.

3 Results

Table 1 gives the percentage values of the data presented in Figure 2.

Gender\Age	(0-9)	(10-19)	(20-29)	(30-39)	(40-49)	(50-59)	(60-69)	(70-79)	(80-89)	(90+)	Total
Positive COVID-19 Confirmation (in %-age of 1000 patients)											
Male	0.6	1.2	15.4	12.4	9.8	9.3	6.1	2.2	6	0.0	57.6
Female	0.7	1.7	11.9	7.8	5.7	7.1	4.9	1.6	0.9	0.1	42.4
Total	1.3	2.9	27.3	20.2	15.5	16.4	11	3.8	1.5	0.1	100
Clinically Recovery (in %-age of 245 patients)											
Male	1.2	1.2	9.7	13.8	11	11	6	0.8	0.4	0	55.9
Female	0.4	0.4	4.9	9.8	6.9	11	7.7	2.4	0.4	0	44.1
Total	1.6	1.6	14.6	23.7	17.9	22.4	13.9	3.3	0.8	0	100

Table 1: Percentage values of the age-gender structure of population confirmed positive with COVID-19 during January 23-April 01, 2020 in Singapore (§Figures 1(i), and 2(i)).

Descriptive statistical analysis: We present the descriptive statistics either as a Δt_r five-number summary[10] and as "median [IQR]" of the observed count variable, i.e., Δt_r.

Δt_r observed during the entire period, January 23-April 01, has the following median and IQR values:

[10] A five-number summary is (minimum, first quartile, median, third quartile, maximum) values of a (count) variable.

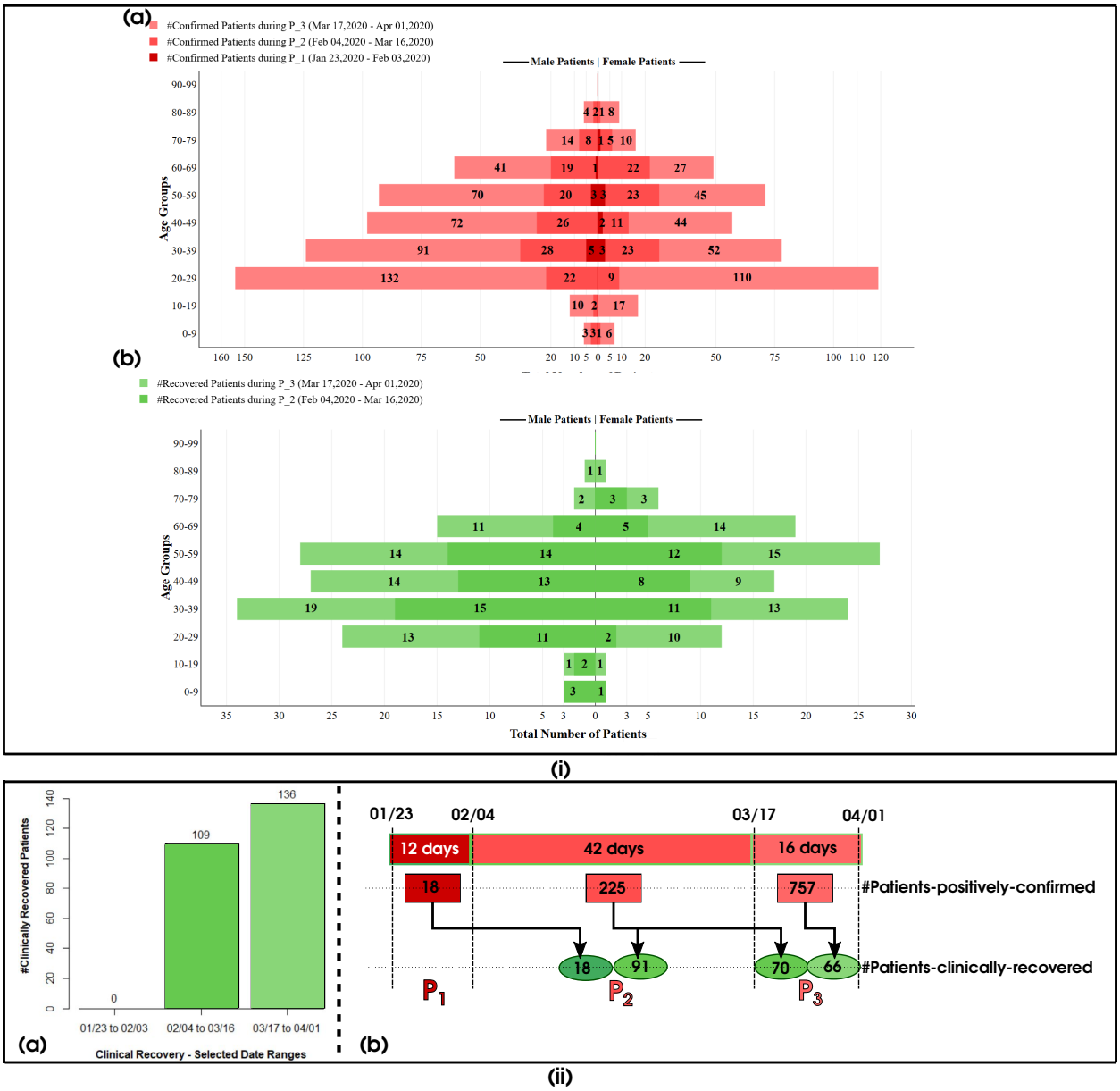


Figure 2: The entire period of our study is split into three periods: January 23-February 03 [P_1], February 04-March 16 [P_2], and March 17-April 01 [P_3]. (i) Period-wise stacked population pyramids to show the age-gender distribution of patients: ((i),(a)) 999 who were confirmed positive (excluding one 102-year old female patient who tested positive on April 01), and ((i),(b)) 245 who were clinically recovered. (ii) Period-wise counts of clinically recovered patients: ((ii),(a)) a barplot of the counts, and ((ii),(b)) a schematic diagram to show the counts of patients grouped based on periods when they were confirmed positive *and* clinically recovered.

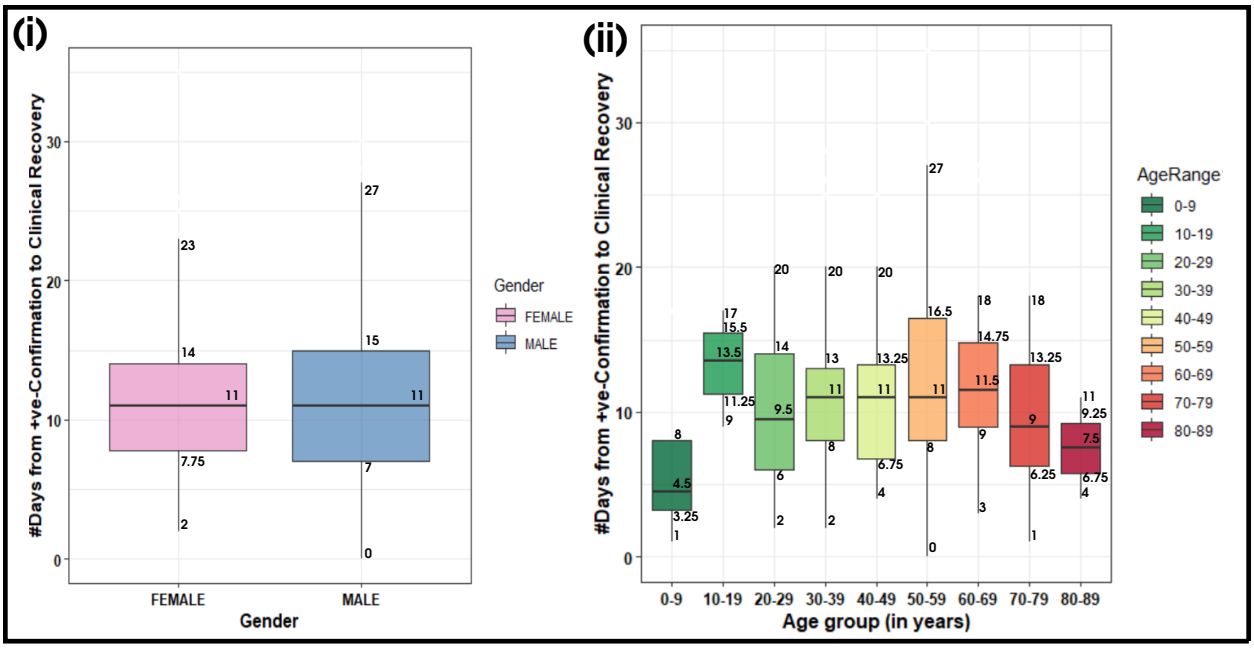


Figure 3: 245 clinically recovered patients of 1000 positively confirmed COVID-19 patients during January 23-April 01, 2020, in Singapore, presented in box and whisker plots based on (i) gender, (ii) age-groups, and (iii) gender and age-groups combined.

- Overall: 11 [7] days.
- Gender-wise: 11 [6.25] days for females and 11 [8] days for males (§Figure 3(i)).
- Age-wise: 4.5 [4.75] days for (0-9), 13.5 [4.25] days for (10-19), 9.5 [8] days for (20-29), 11 [5] days for (30-39), 11 [6.5] days for (40-49), 11 [8.5] days for (50-59), 11.5 [5.75] days for (60-69), 9 [7] days for (70-79), and 7.5 [3.5] days for (80-89), for age groups in paranthesis given in years (§Figure 3(ii)).

We observe that the Δt_r is overall lesser in this dataset than reported in early analyses of COVID-19 patients, i.e., 15 days [10] and 20 days [11]. The overall median of Δt_r is 11 days, which is the same as the gender-weighted median, and the age-group-weighted median is 10.7 days. Thus, the overall descriptive statistical analysis gives a conservative estimate of γ to be $\sim \frac{1}{11}$. The remaining work is to estimate γ more precisely based on the influence of gender and age.

There is a stronger influence of age than gender on Δt_r, as the median values are similar for both genders. In contrast, it is relatively lower for the age groups of (0-9), (20-29), and (80-89) years, specifically. These age groups comprise of 1.6%, 14.6%, and 0.8% of the clinically recovered patients (§Table 1). This result is significant as the age group of (20-29) years contributes the highest (27.3%) to the infected population. 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_r, the recovery rate γ is bound to increase further in the period after April 01 compared to the value estimated in our work.

The period-wise five-number summaries for Δt_r, using $\mathbb{G}_{\mathbf{+cfrmDt}}$ grouping, are:

- (7, 11.5, 17, 22.75, 28) for P_1, (0, 8, 12, 16, 27) for P_2, and (1, 6, 9, 11, 15) for P_3.
- For females: (10, 11, 13, 22, 26) for P_1, (0, 7, 12, 15, 26) for P_2, and (3, 6.25, 9, 11, 14) for P_3.
- For males: (11, 17, 20, 23, 29) for P_1, (0, 8, 12, 16, 27) for P_2, and (1, 5.75, 8, 11, 15) for P_3.

We observe similar trends when considering $\mathbb{G}_{\mathbf{-cfrmDt}}$ grouping. The range, the IQR, and the median of Δt_r decrease from P_1 to P_3, when we look at the data for each gender as well as the data without the gender information. This indicates that irrespective of gender, the *measure of the spread* of Δt_r decreases with time, similar to the trend in the value of Δt_r. The age-wise minima reduce sharply from P_1 to P_2, and increase slightly further to P_3, indicating an overall trend of decrease in minima. This supports the overall decrease in Δt_r value from P_1 to P_3.

Table 2 shows the fine-grained five-number summaries of the age-gender based box and whisker plots in Figure 4. We observe that there is a strong influence of age on Δt_r. While the overall median values for females and males are

Age-group	Jan/23-Apr/01 §Figure 4(a)		P_1 [$\mathbb{G}_{\mathbf{+cfrmDt}}$] §Figure 4((b),(i))	
	F	M	F	M
(0-9 years)	(4, 4, 4, 4, 4)	(1, 3, 5, 11, 17)	-	-
(10-19 years)	(12, 12, 12, 12, 12)	(9, 12, 15, 16, 17)	-	-
(20-29 years)	(3, 8.25, **10**, 14, 18)	(2, 6, **8.5**, 14.25, 20)	-	-
(30-39 years)	(3, 9, 11, 13, 18)	(2, 8, 12, 15, 25)	(11, 12, 13, 19.5, 26)	(17, 17, 17, 20, 20)
(40-49 years)	(4, 6, 11, 14, 25)	(4, 7, 11, 13, 18)	(10, 11.75, 13.5, 15.25, 17)	-
(50-59 years)	(3, 8.5, 13, 16.5, 25)	(0, 7.75, 10, 16.5, 27)	(13, 17.5, 22, 23.5, 25)	(11, 15.5, 20, 21.5, 23)
(60-69 years)	(3, 7, 10, 14, 18)	(8, 10, 12, 15, 18)	-	(27, 27, 27, 27, 27)
(70-79 years)	(4, 7.25, 9, 12.25, 14)	(1, 5.25, 9.5, 13.75, 18)	(10, 10, 10, 10, 10)	-
(80-89 years)	(4, 4, 4, 4, 4)	(11, 11, 11, 11, 11)	-	-

Age-group	P_2 [$\mathbb{G}_{\mathbf{+cfrmDt}}$] §Figure 4((b),(ii))		P_3 [$\mathbb{G}_{\mathbf{+cfrmDt}}$] §Figure 4((b),(iii))	
	F	M	F	M
(0-9 years)	(4, 4, 4, 4, 4)	(1, 3, 5, 11, 17)	-	-
(10-19 years)	-	(9, 11, 13, 15, 17)	(12, 12, 12, 12, 12)	(15, 15, 15, 15, 15)
(20-29 years)	(6, 12, 18, 18, 18)	(2, 7, 13, 16, 20)	(3, 9, 10, 13, 14)	(4, 6, 8, 9, 13)
(30-39 years)	(3, 9, 11, 13, 18)	(2, 8, 12, 13, 18)	(3, 3.75, 7, 10.25, 11)	(4, 7, 9.5, 12, 15)
(40-49 years)	(11, 11, 13, 14, 14)	(5, 7.5, 11, 14, 18)	(5, 6, 7.5, 9.75, 11)	(4, 4.75, 9, 11.5, 13)
(50-59 years)	(3, 7.5, 12, 16.5, 23)	(0, 9.5, 16, 18.5, 30)	(8, 9, 10, 13, 13)	(4, 6.25, 8, 9.75, 11)
(60-69 years)	(3, 6, 10, 15, 26)	(8, 9.5, 12, 15, 18)	(8, 9, 10, 11, 12)	(10, 11, 12, 12.5, 13)
(70-79 years)	(4, 8.5, 13, 13.5, 14)	(18, 18, 18, 18, 18)	(7, 7.25, 7.5, 7.75, 8)	(1, 1, 1, 1, 1)
(80-89 years)	-	(11, 11, 11, 11, 11)	(4, 4, 4, 4, 4)	-

Age-group	P_2 [$\mathbb{G}_{\mathbf{-cfrmDt}}$] §Figure 4((c),(ii))		P_3 [$\mathbb{G}_{\mathbf{-cfrmDt}}$] §Figure 4((c),(iii))	
	F	M	F	M
(0-9 years)	(4, 4, 4, 4, 4)	(1, 3, 5, 11, 17)	-	-
(10-19 years)	-	(9, 11, 13, 15, 17)	(12, 12, 12, 12, 12)	(15, 15, 15, 15, 15)
(20-29 years)	(6, 9, 12, 15, 18)	(2, 7, 15, 16, 20)	(3, 9, 10, 13.75, 18)	(4, 6, 8, 12, 14)
(30-39 years)	(9, 9, 11, 12.5, 13)	(2, 5.5, 10, 16.5, 28)	(3, 9, 11, 13, 18)	(8, 10.5, 12, 13, 15)
(40-49 years)	(4, 10, 12, 14, 17)	(5, 7, 9, 14, 16)	(5, 6, 9.5, 11.5, 13)	(4, 7.5, 11, 13, 18)
(50-59 years)	(3, 7.75, 11.5, 17, 25)	(0, 9.25, 11.5, 22.25, 30)	(4, 9.5, 13, 14.5, 17)	(4, 7.25, 9.5, 14.75, 18)
(60-69 years)	(3, 3, 4, 5, 6)	(8, 8.75, 10.5, 15.75, 16)	(8, 10, 12.5, 15.75, 18)	(9, 10.5, 13, 15, 18)
(70-79 years)	(10, 11.5, 13, 13.5, 14)	-	(4, 5.5, 7, 7.5, 8)	(1, 5.25, 9.5, 13.75, 18)
(80-89 years)	-	-	(4, 4, 4, 4, 4)	(11, 11, 11, 11, 11)

Table 2: Five-number summaries of age-gender based box and whisker plots in Figure 3 and 4.

similar, we observe that there are variations across different age groups. In particular, for the age group of (20-29) years, the median [IQR] value of males of 8.5 [8.25] days is lower than the median value of females of 10 [5.75] days. This also shows that there is a higher spread (IQR) of Δt_r values in males. We further observe that the gender difference in IQR arises in P_3 in the $\mathbb{G}_{\mathbf{+cfrmDt}}$ grouping, which happens in P_2 in the $\mathbb{G}_{\mathbf{-cfrmDt}}$ grouping. The minima and median values are overall lower in P_3 in the $\mathbb{G}_{\mathbf{+cfrmDt}}$ than the $\mathbb{G}_{\mathbf{-cfrmDt}}$ grouping. Since the protocols followed in the hospital can be perceived to be similar for patients with closer hospital admission dates, the Δt_r has more cohesive descriptive statistics in the $\mathbb{G}_{\mathbf{+cfrmDt}}$ grouping than the $\mathbb{G}_{\mathbf{-cfrmDt}}$ one. Hence, we use the $\mathbb{G}_{\mathbf{+cfrmDt}}$ grouping exclusively in the regression analysis.

Since the symptom-onset dates have been studied [7, 8], we report the five-number summaries of Δt_{so}, for which data is available for P_1 and P_2 only.

- Overall: (0, 3, 6, 9, 16) for P_1 and (1, 2, 5, 8, 17) for P_2.
- For females: (0, 2, 5, 7, 12) for P_1, and (1, 3, 5, 8, 15) for P_2.
- For males: (1, 4.75, 8, 9, 15) for P_1, and (1, 2, 5, 8, 17) for P_2.

We observe that Δt_{so} has similar IQR and median across P_1 and P_2, irrespective of gender. We attribute this to the continuous monitoring of susceptible cases in Singapore leading to lesser delays in positive confirmations, thus showing a lower measure of spread and overall lower values for Δt_{so}. Overall, we do not emphasize on analysing Δt_{so} owing to the clinical uncertainties involved [4], which do not reflect in the timeline data that we are using here.

Loess model: We now use the loess model to confirm the trend in the change in Δt_r. The loess model has been estimated on the time-series of the Δt_r and Δt_{so} values, represented as scatter plots (§Figure 5). The loess model has been implemented using the stats.loess[11] in R [16]. We have considered the scatter plots based on the positive

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

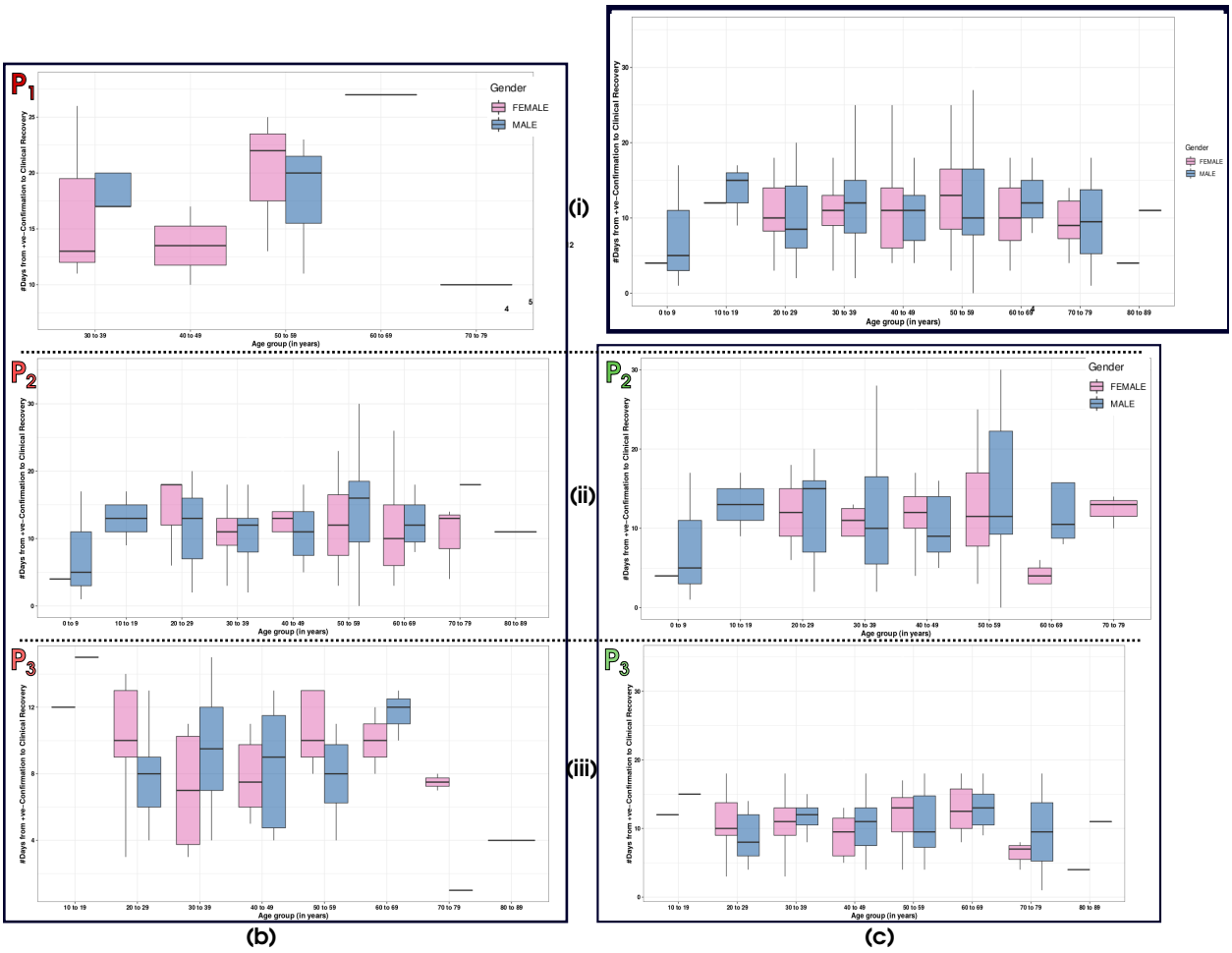


Figure 4: Gender-wise and age-wise box and whisker plots of 245 clinically recovered patients grouped in periods during (i) January 23-February 03 [P_1] (ii) February 04-March 16 [P_2], and (iii) March 17-April 01 [P_3]. Counterclockwise from top-right: (a) during January 23-April 01, 2020 [entire period], (b) period-wise grouping of patients based on COVID-19 positive confirmation date ($\mathbb{G}_{\mathbf{+cfrmDt}}$), and (c) period-wise grouping of patients based on clinical recovery/hospital discharge dates ($\mathbb{G}_{\mathbf{-cfrmDt}}$). The number of patients considered is (a) 245, ((b),(i)) 18, ((b),(ii)) 161, ((b),(iii)) 66, ((c),(ii)) 109, ((c),(iii)) 136.

COVID-19 confirmation date, akin to $\mathbb{G}_{\mathbf{+cfrmDt}}$ grouping. The loess model for Δt_r, considered during January 23-April 04, 2020, uses 245 observations, with 5.37 degrees of freedom (DoF) and residual standard error (RSE) of 145.4 (§Figure 5(i)). The loess models done on period-wise data (§Figure 5(ii)) have the following details: for P_1 uses 18 observations, with 4.66 DoF and RES of 5.817; for P_2 uses 161 observations, with 5.23 DoF and RSE of 61.74; and for P_3 uses 66 observations, with 4.91 DoF and RSE of 109.8. The loess model for Δt_{so} uses 227 observations, with 5.03 DoF and RSE of 70 (§Figure 5(iii)).

The degrees of freedom roughly corresponds to the degree of the polynomial used to generate the fitting curve. Thus, both Δt_r and Δt_{so} can be modeled using a 5th degree polynomial. Higher degree polynomial implies less bias but larger variance. Δt_r has a slope of -25^o in P_3 when using the loess model for the entire time period (§Figure 5(i)) and that of -6^o when using loess model for P_3 (§Figure 5(ii)). Thus, the key conclusion from the local regression model on the time-series is the negative slope, i.e., a *downward* trend in Δt_r in P_3, which is favorable in improving recovery rate γ.

Multivariate (linear) regression model: Now that we have observed and concluded from both the descriptive statistical analysis and loess model that Δt_r is decreasing during the period of January 23-April 01, our next step is to predict the value of Δt_r. We experiment with the generalized linear model (GLM) for a multivariate linear regression model for Δt_r using Poisson (PRM) and negative binomial (NBM) distributions. Our choice of model and distributions are

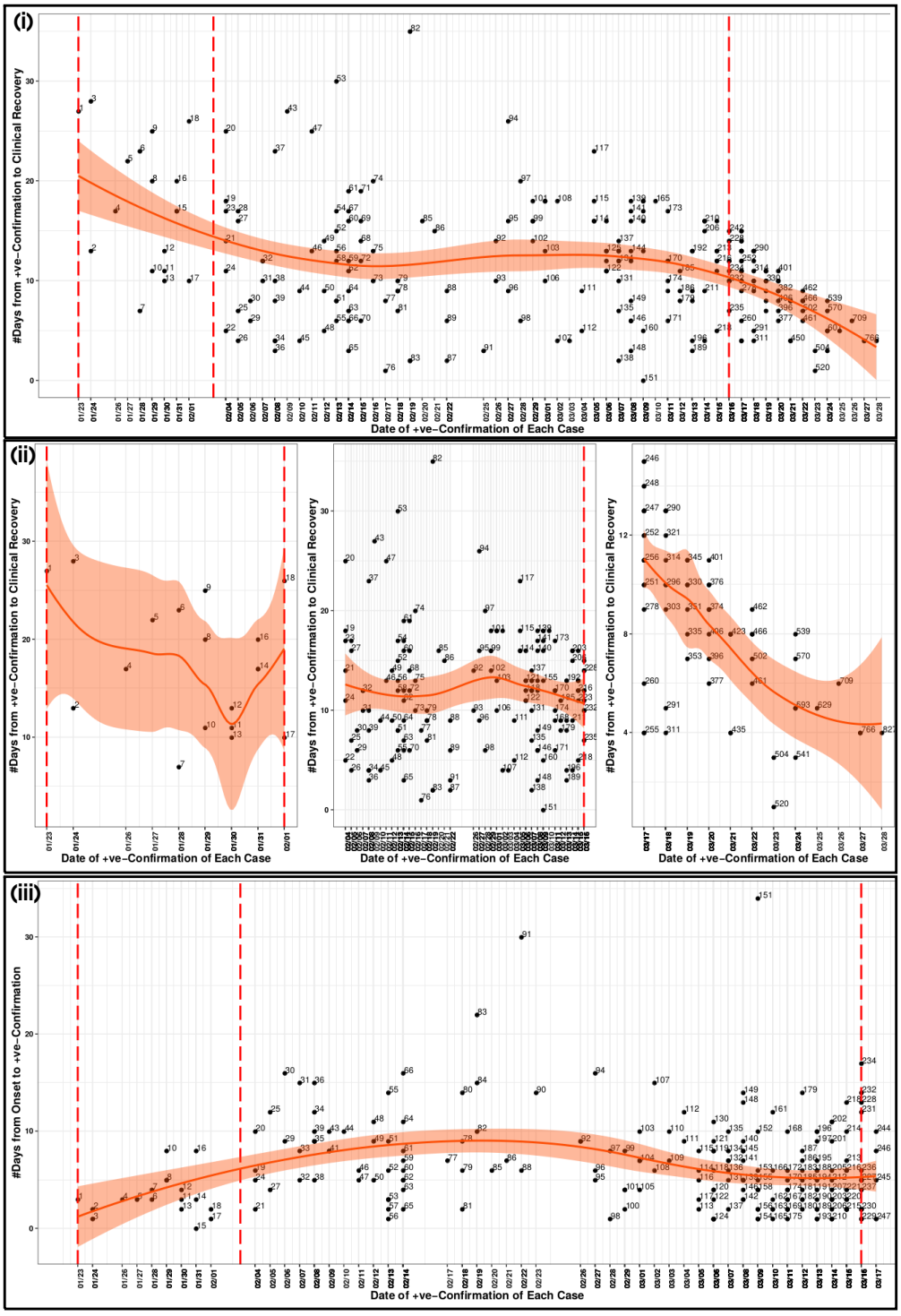


Figure 5: Scatter plot of the number of days with the timeline of positive confirmation of COVID-19 in Singapore during January 23-April 01, 2020, estimated using a Loess model. The scatter points using patient (case) ID, fitted loess curve and the error ribbon are shown for (i) clinical recovery-period, Δt_r, for the entire period, (ii) Δt_r, for each period, and (ii) Δt_{so}, from the onset of symptoms to positive confirmation of COVID-19 for P_1 and P_2.

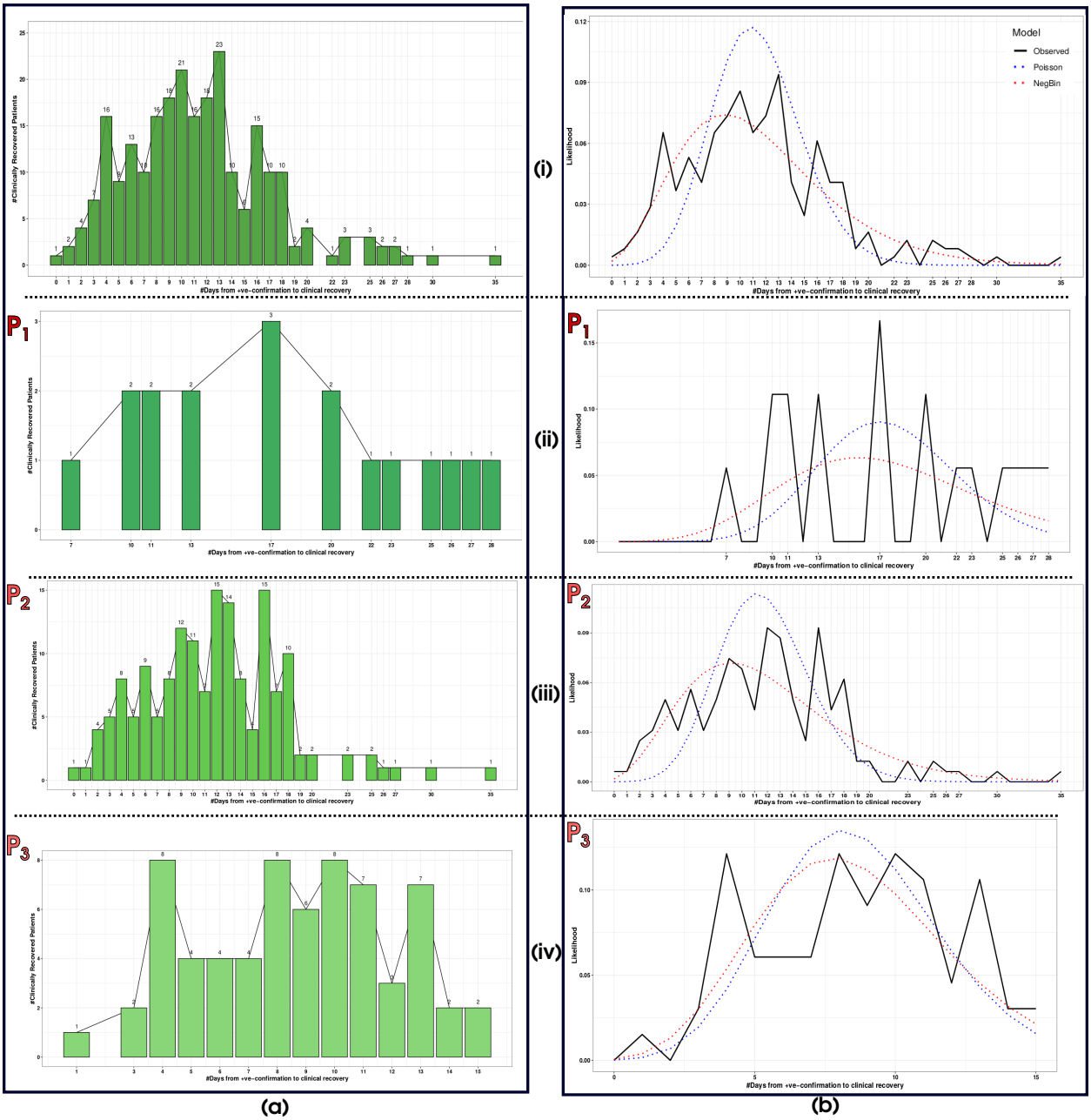


Figure 6: Multivariate linear regression model of the recovery-period Δt_r for 245 clinically recovered patients amongst 1000 COVID-19 confirmed cases in Singapore during (i) January 23-April 01 [entire-period], (ii) January 23-February 03 [P_1], (iii) February 04-March 16 [P_2], and (iv) March 17-April 01 [P_3]; uses (a) patient-count distribution and (b) fitting Poisson and negative binomial distributions. The number of patients considered is (i) 245, (ii) 18, (iii) 161, and (iv) 66, using $\mathbb{G}_{\mathbf{+cfrmDt}}$ grouping.

commonly used for count data [5], and hospital length of hospital stay (LoS) is commonly over-dispersed data [17]. For each model, we use four *scenarios*, namely, for the entire period and for each period. 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. We have implemented these models using the stats.glm in R [16].

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 [16]. The negative binomial distribution has an additional parameter to model over-dispersion in the data. For the number of samples (n) in the data, AIC is used if $\left(\frac{n}{k} > 40\right)$, and AIC_c is used otherwise [18]. Thus, we use AIC for scenarios of the entire period and P_2, and AIC_c for P_1 and P_3, owing to the relatively lesser samples (§Table 3).

Model results \Scenario (Time-period)	January 23-April 01	P_1	P_2	P_3
#Samples (n)	245	18	161	66
#Degree of freedom	242	15	158	63
Poisson Regression Model (PRM)				
#Parameters (k)	3	3	3	3
Coefficients*				
(Intercept)	**0.000** [2.057-2.412]	**0.000** [1.934-3.080]	**0.000** [1.950-2.413]	**0.000** [1.990-2.650]
Age	**0.007** [0.001-0.006]	0.552 [(-0.006)-0.011]	**0.003** [0.002-0.008]	0.292 [(-0.007)-0.002]
Gender	0.438 [(-0.046)-0.106]	0.181 [(-0.07)-0.374]	0.298 [(-0.044)-0.145]	0.675 [(-0.200)-0.130]
#Fisher Scoring Iterations	4	4	4	4
AIC	1735.7	-	1152.6	-
AIC_c	-	**132.68**	-	**358.10**
Deviance residuals**	-0.15 [2.33, 10.36]	-0.07 [2.28, 5.43]	0.04 [2.21, 10.3]	**0.00 [1.74, 5.08]**
MSE (with model data)	0.369	**0.157**	0.362	**0.250**
RMSE (with model data)	0.607	**0.397**	0.602	**0.500**
Maximum likelihood value at $argmax$	11.7% at $\Delta t_r = 11$	9.0% at $\Delta t_r = 17$	11.4% at $\Delta t_r = 11$	**13.5% at $\Delta t_r = 8$**
Negative Binomial Regression Model (NBM)				
#Parameters (k)	4	4	4	4
Coefficients*				
(Intercept)	**0.000** [1.928-2.537]	**0.000** [1.674-3.383]	**0.000** [1.779-2.581]	**0.000** [1.938-2.703]
Age	0.105 [(-0.001)-0.008]	0.723 [(-0.011)-0.016]	0.082 [(-0.001)-0.010]	0.364 [(-0.008)-0.003]
Gender	0.667 [(-0.102)-0.160]	0.39 [(-0.189)-0.485]	0.565 [(-0.116)-0.212]	0.719 [(-0.225)-0.156]
AIC	1535.8	-	1023.2	-
AIC_c	-	**128.00**	-	**357.67**
#Fisher Scoring Iterations	1	1	1	1
Deviance residuals**	-0.08 [1.34, 6.34]	-0.044 [1.54, 3.66]	0.02 [1.29, 6.3]	**0.00 [1.51, 4.51]**
MSE (with model data)	0.368	**0.157**	0.362	**0.250**
RMSE (with model data)	0.607	**0.397**	0.602	**0.500**
Maximum likelihood value at $argmax$	7.4% at $\Delta t_r = 9$	6.0% at $\Delta t_r = 16$	7.2% at $\Delta t_r = 9$	**11.8% at $\Delta t_r = 8$**

* *p-value [95% confidence interval (CI)]*
** *median [IQR, range]*

Table 3: Results of the generalized linear models using Poisson distribution and negative binomial distribution of recovery-period Δt_r using $\mathbb{G}_{\mathbf{+cfrmDt}}$ grouping.

Table 3 gives the results of our models. We infer the following:

- The "age" variable is significant only in the PRM, and only for the scenarios of the entire period and P_2, with a p-value for the coefficients corresponding to the variable being less than 5%.
- The NBM shows lower values for the median and the variance (observable from range and IQR) of deviance residuals, and AIC/AIC_c than the PRM. Thus, we conclude that NBM is a better fit than PRM. Also, for both PRM and NBM, the models for the scenarios of P_1 and P_3 are a better fit than those of the entire period and P_2.

These observations may be attributed to the relatively small sample size for P_1 and P_3. Since we do not have a large number of variables to discard, we retain the "gender" variable in the model despite its insignificance.

The key conclusion from the multivariate regression analysis is that a GLM with NBM for P_3 is the best model for us to estimate Δt_r. This helps us to estimate the value of Δt_r to be 8 days, with the maximum likelihood of 11.8%. The expected value of Δt_r of NBM for P_3 is 8.05 days. Hence, overall, we conclude that the estimated value of γ is $\frac{1}{8}$.

4 Discussion

The improved γ, as per our estimate, is an outcome of the existing protocols in Singapore. The approach of containment of the contagion undertaken by Singapore has been government-mandated, which has ensured delays in the spread of the disease. While the spread got contained, there has been a shift in the age-group of the population getting infected. This shift has brought about the decrease in the recovery-period, Δt_r.

Our work has two specific limitations. Firstly, our study is short of using a non-zero death/mortality/fatality rate of the disease. The number of deaths will continue to increase, warranting its consideration in the SIR model. Secondly, we have modeled recovery isolated from the infection. Since the number of infected persons in Singapore has increased exponentially since March 17, 2020, the infection rate, β, and consequently, basic reproduction number $\mathcal{R}_0$ have to be re-estimated. Nevertheless, our estimated γ is thus applicable for improving estimation of $\mathcal{R}_0$ until April 01, and for simulating/predicting disease progression using SIR model beyond April 01.

In summary, we have looked at the demographic data and timeline of the first 1000 COVID-19 patients in Singapore during January 23-April 01, 2020. We have closely investigated the data on the positive confirmation and discharge/clinical recovery dates of 245 patients who recovered during this time period. We have used regression analysis, subsequent to a descriptive statistical analysis, to get an estimate of recovery-period Δt_r (i.e., hospital length of stay (LoS)). We have found that the Δt_r is time-varying, after performing periodization to find three significant periods, namely P_1 (January 23-February 03), P_2 (February 04-March 16), and P_3 (March 17-April 01). The estimates of Δt_r varied from $\sim$17 days in P_1 to $\sim$10 days in P_2 to $\sim$8 days in P_3. We have used the loess model for time-series data to demonstrate the negative slope of the regression curve of Δt_r in P_3, in particular. We then estimated period-wise Δt_r using generalized linear models for multivariate (linear) regression with Poisson and negative binomial distributions for count data. This shows an improvement in the values published for Δt_r, i.e., 20 days [11] and 15 days [10]. This has led us to estimate the current recovery rate γ in the SIR model to be $\frac{1}{8} = 0.125$ from March 17, 2020 onwards.

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.

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] John E. L. Wong, Yee Sin Leo, and Chorh Chuan Tan. COVID-19 in Singapore—Current Experience: Critical Global Issues That Require Attention and Action. *JAMA*, 02 2020.

[3] Yixiang Ng, Zongbin Li, Yi Xian Chua, Wei Liang Chaw, Zheng Zhao, Benjamin Er, Rachael Pung, Calvin J. Chiew, David C. Lye, Derrick Heng, and Vernon J. Lee. Evaluation of the Effectiveness of Surveillance and Containment Measures for the First 100 Patients with COVID-19 in Singapore — January 2–February 29, 2020. *Morbidity and Mortality Weekly Reports (MMWR)*, 69(11):307–311, 2020.

[4] Wycliffe E. Wei, Zongbin Li, Calvin J. Chiew Chiew, Sarah E. Yong, Matthias P. Toh, and Vernon J. Lee Lee. EPresymptomatic Transmission of SARS-CoV-2 — Singapore, January 23–March 16, 2020. *Morbidity and Mortality Weekly Reports (MMWR)*, 69(14):411–415, 2020.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] William Ogilvy Kermack and Anderson G McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character*, 115(772):700–721, 1927.

[10] 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.

[11] 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.

[12] 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.

[13] J. Riou and C. L. Althaus. Pattern of early human-to-human transmission of Wuhan 2019 novel coronavirus (2019-nCoV), December 2019 to January 2020. *Euro Surveillance*, 25(4,pii:2000058), 2020. published correction appears in Euro Surveill. 2020 Feb;25(7):.

[14] 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.

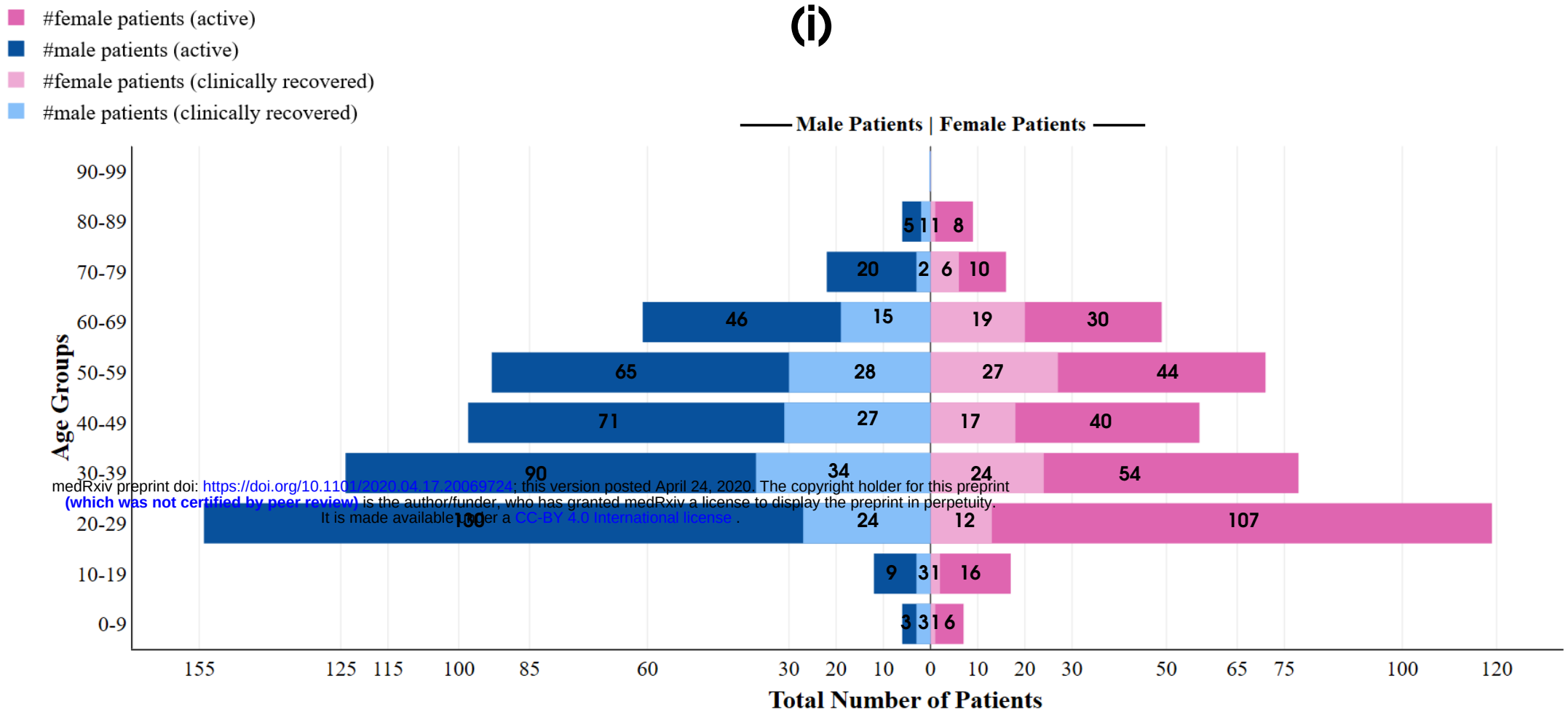
[15] William G Jacoby. Loess:: a nonparametric, graphical tool for depicting relationships between variables. *Electoral Studies*, 19(4):577–613, 2000.

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

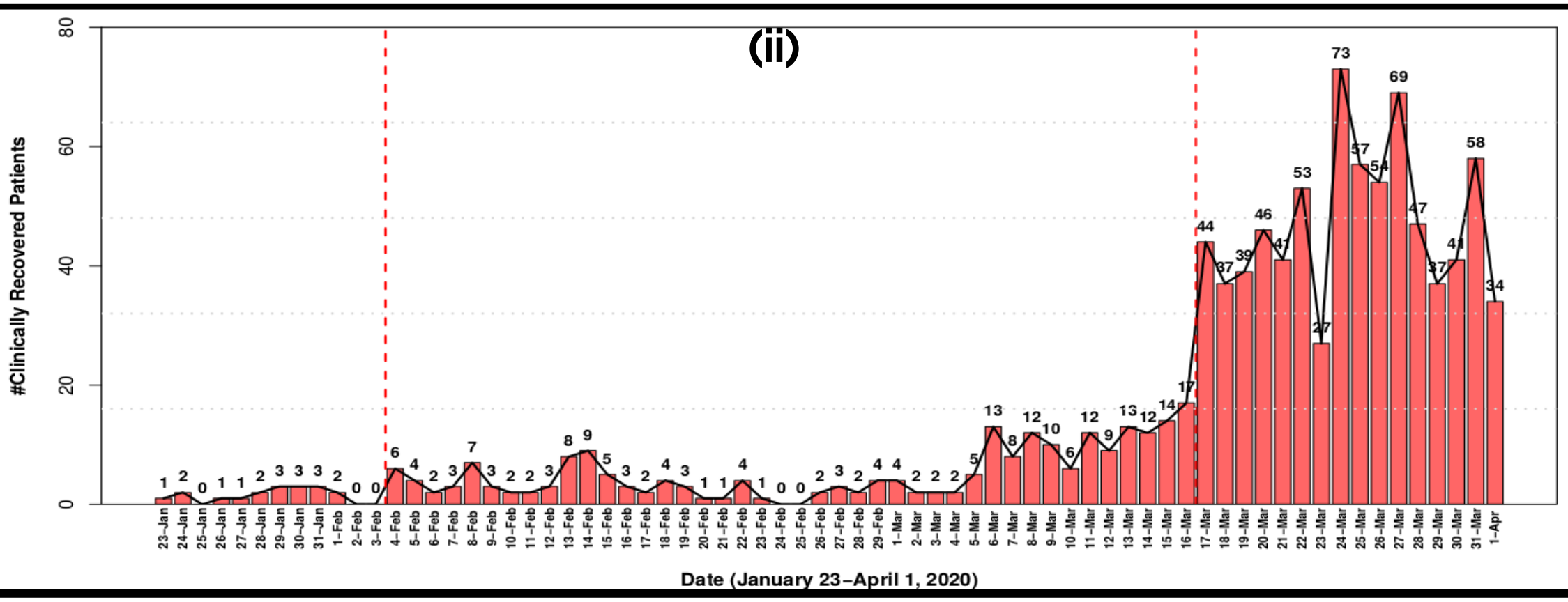
[17] Liming Xiang, Andy H Lee, Kelvin KW Yau, and Geoffrey J McLachlan. A score test for overdispersion in zero-inflated poisson mixed regression model. *Statistics in medicine*, 26(7):1608–1622, 2007.

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

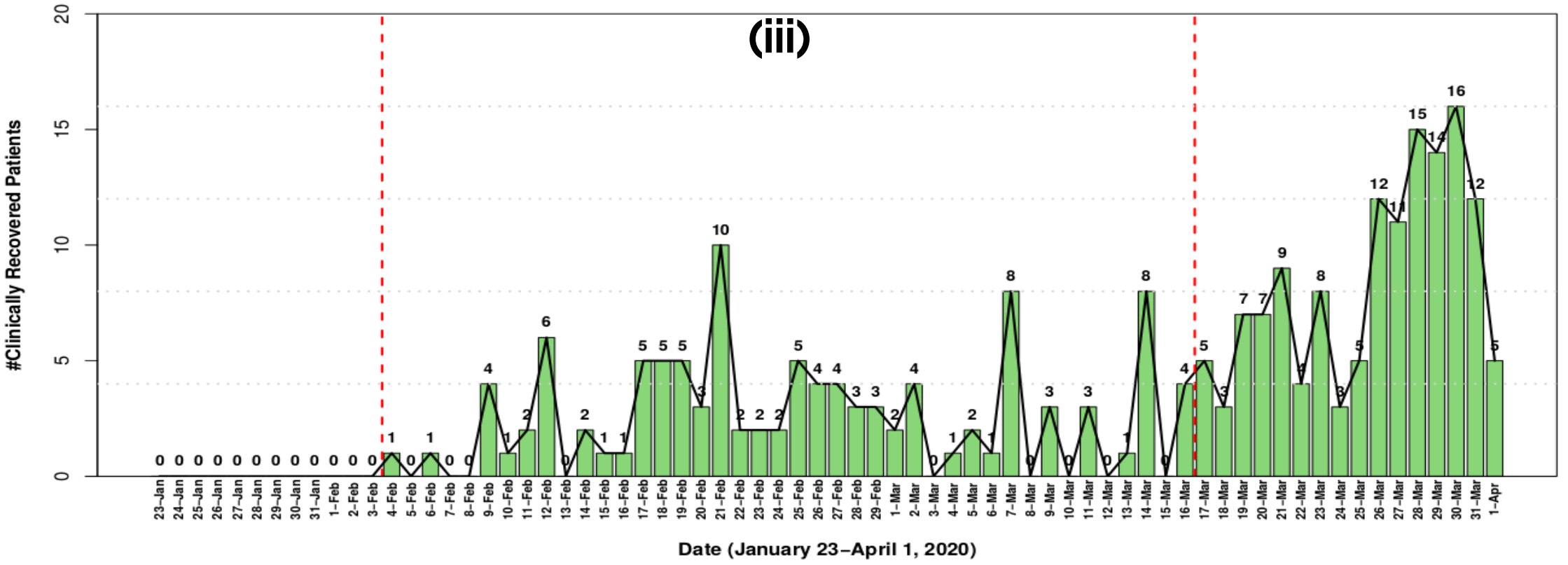
(i)



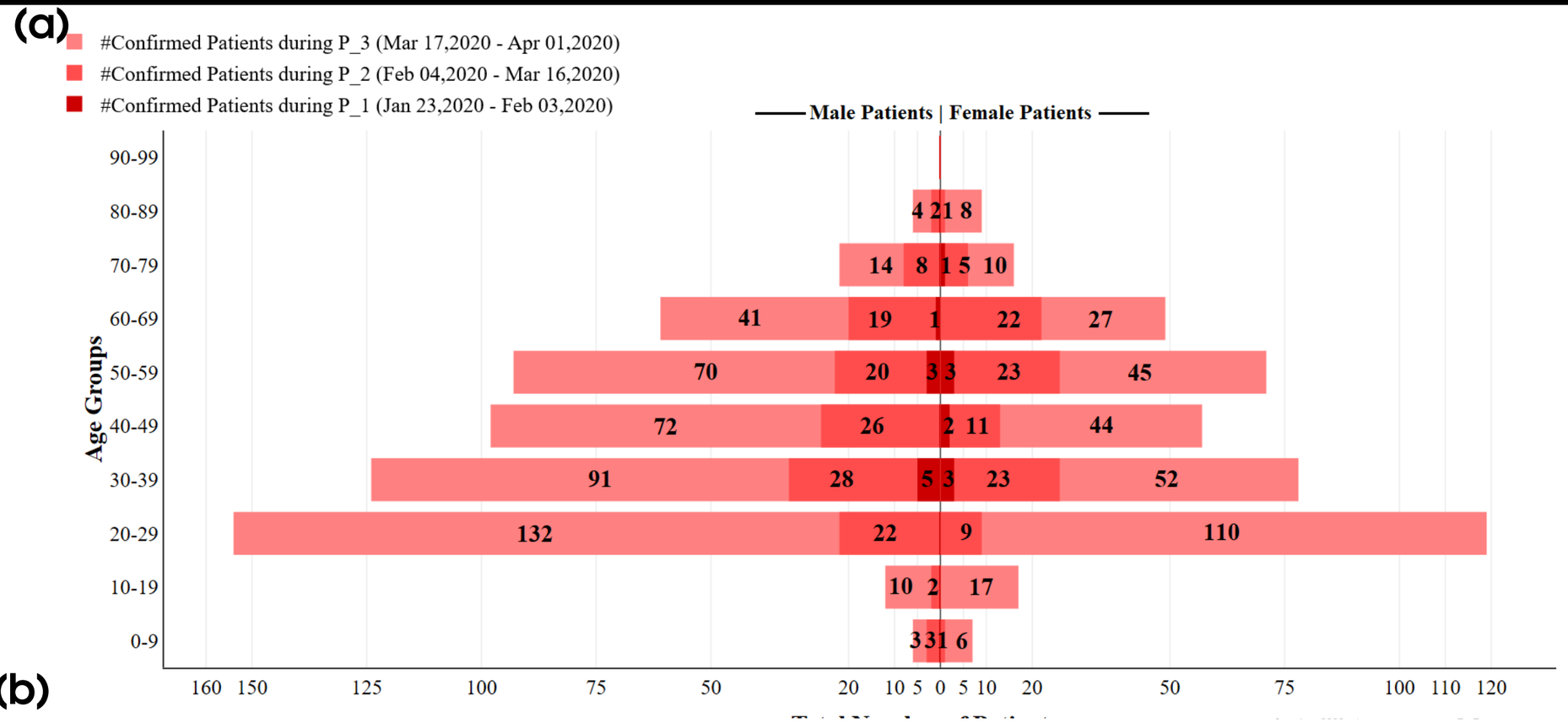
(ii)



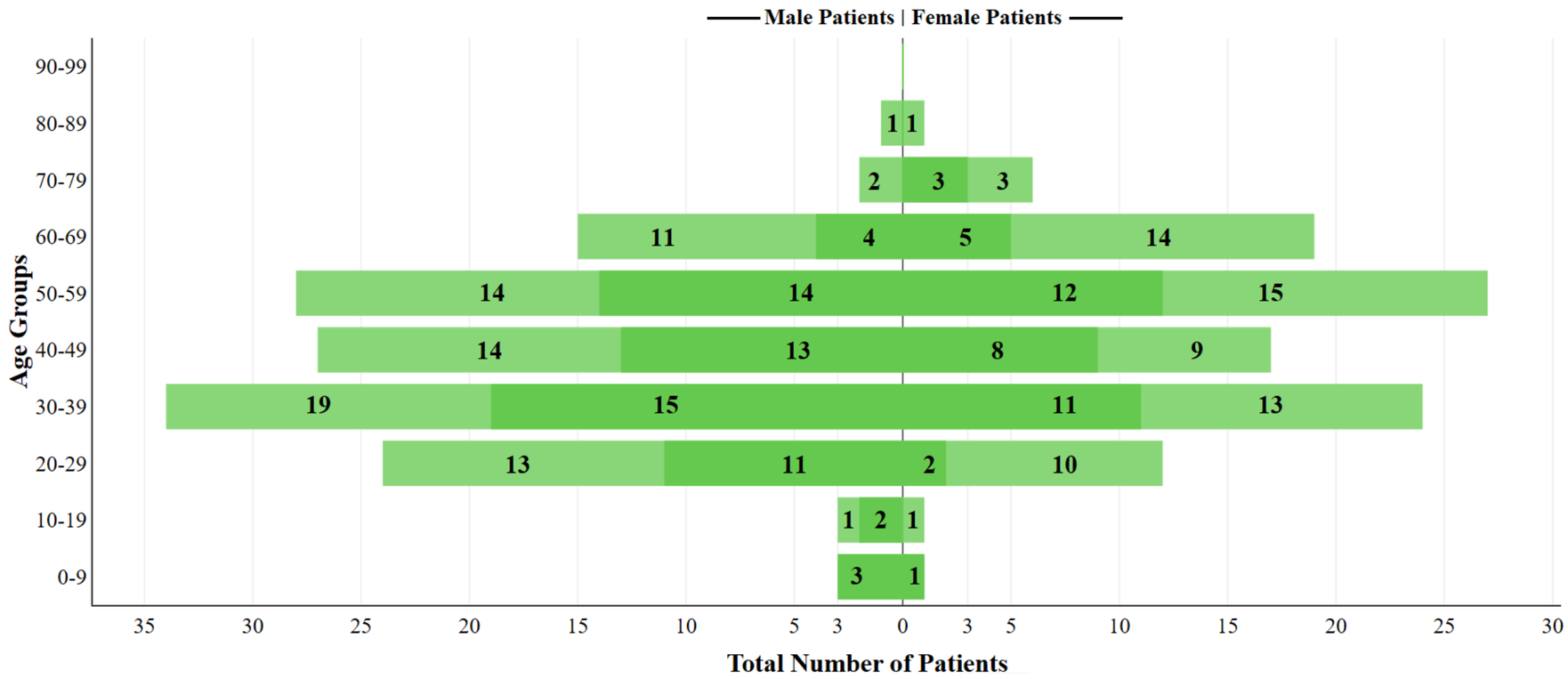
(iii)



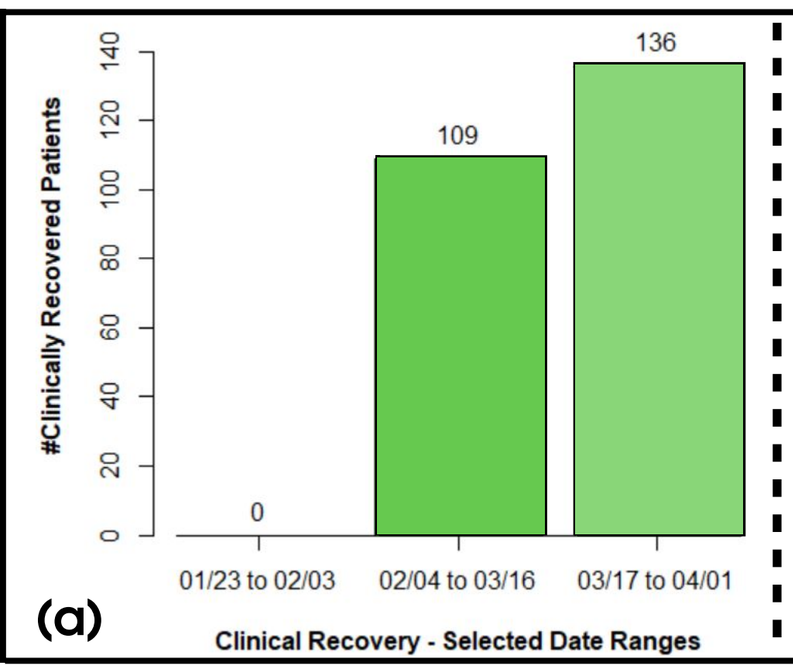
(a)



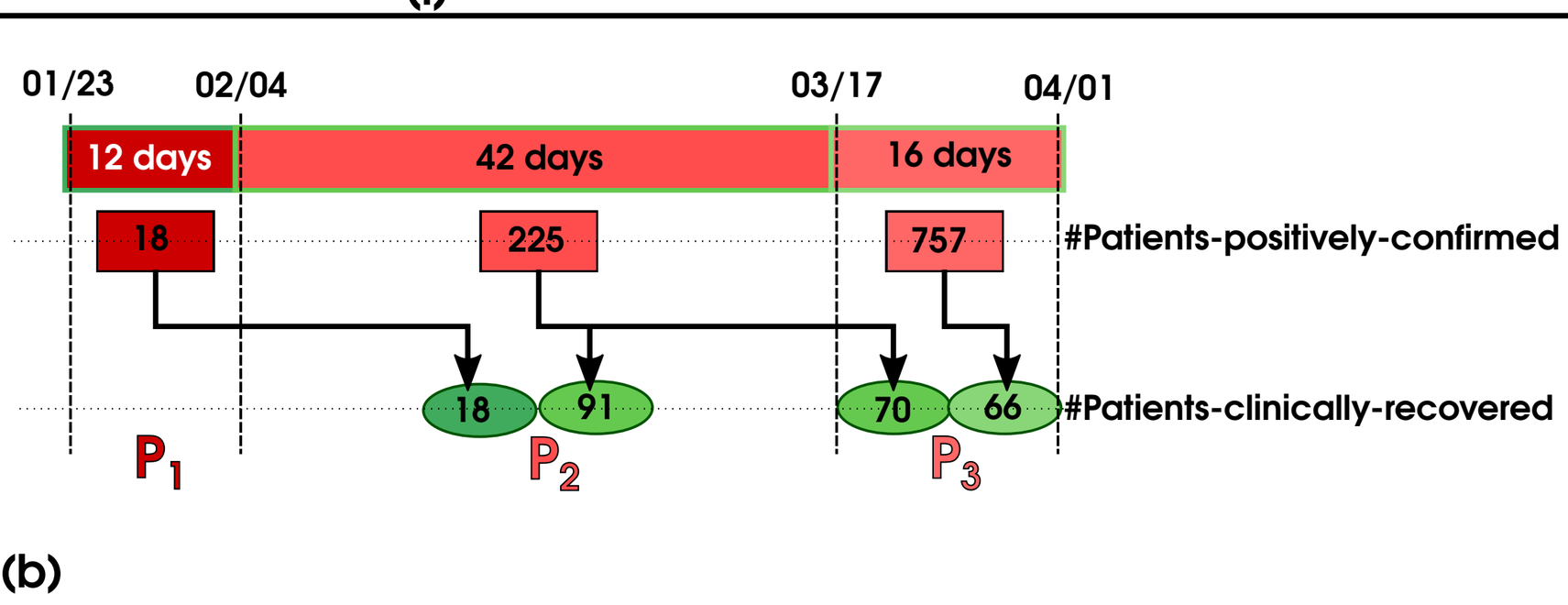
(b)



(i)

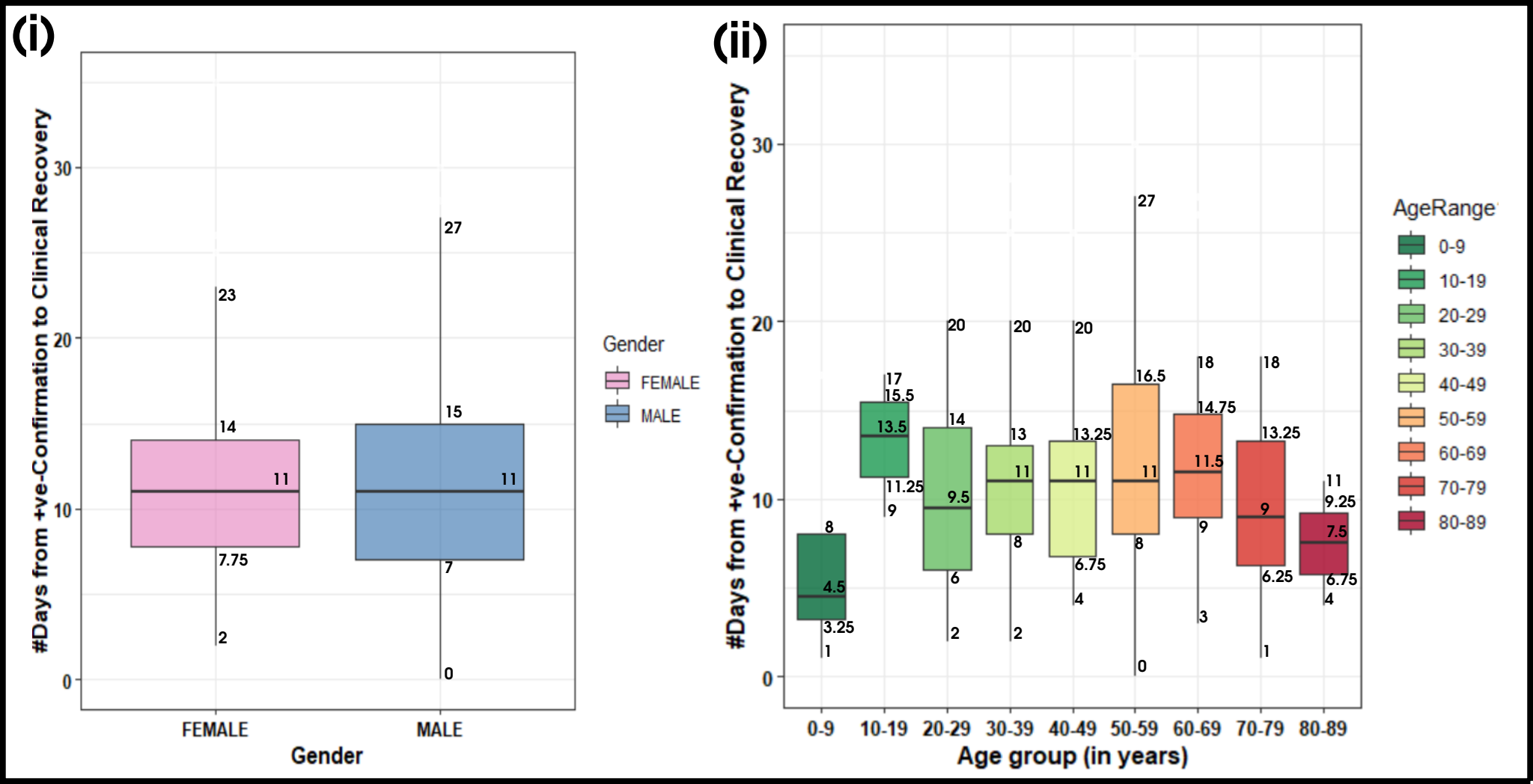


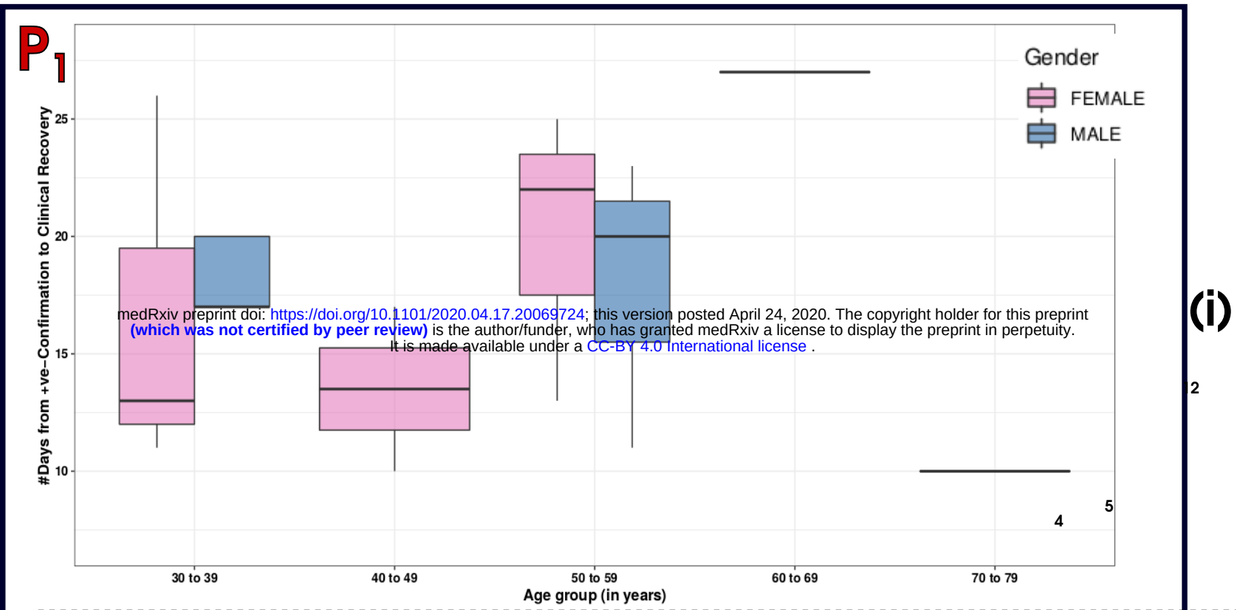
(a)



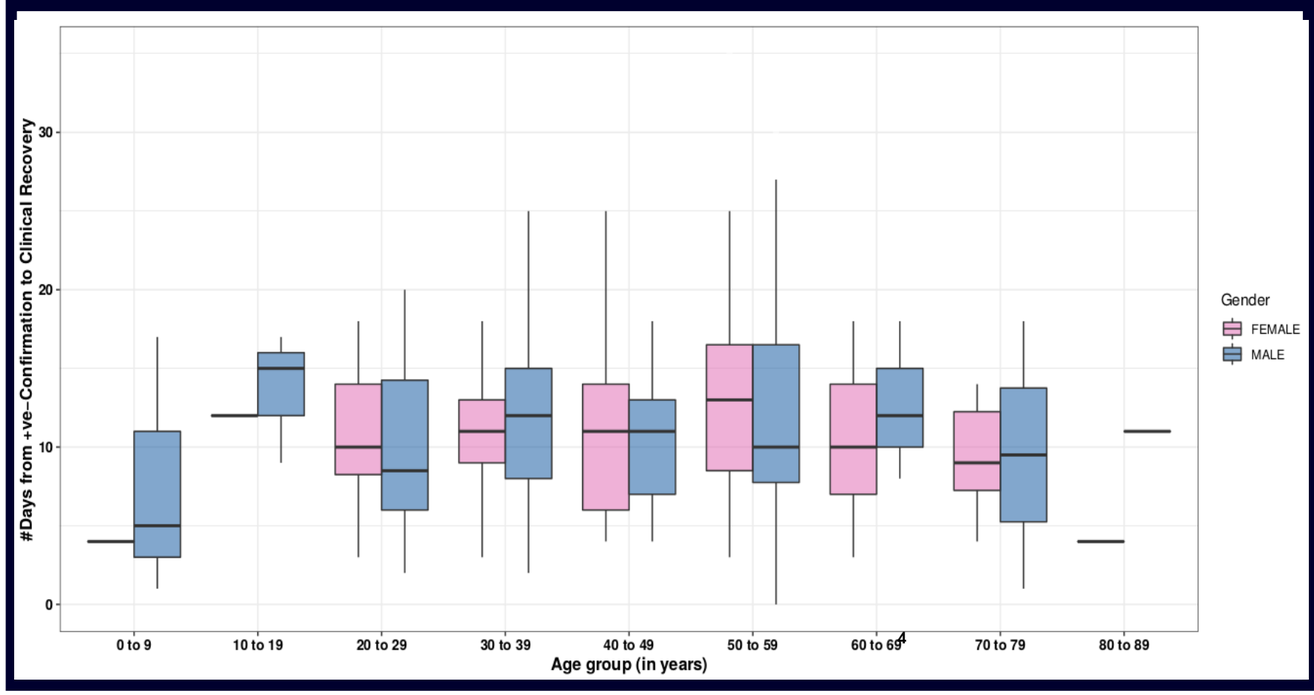
(b)

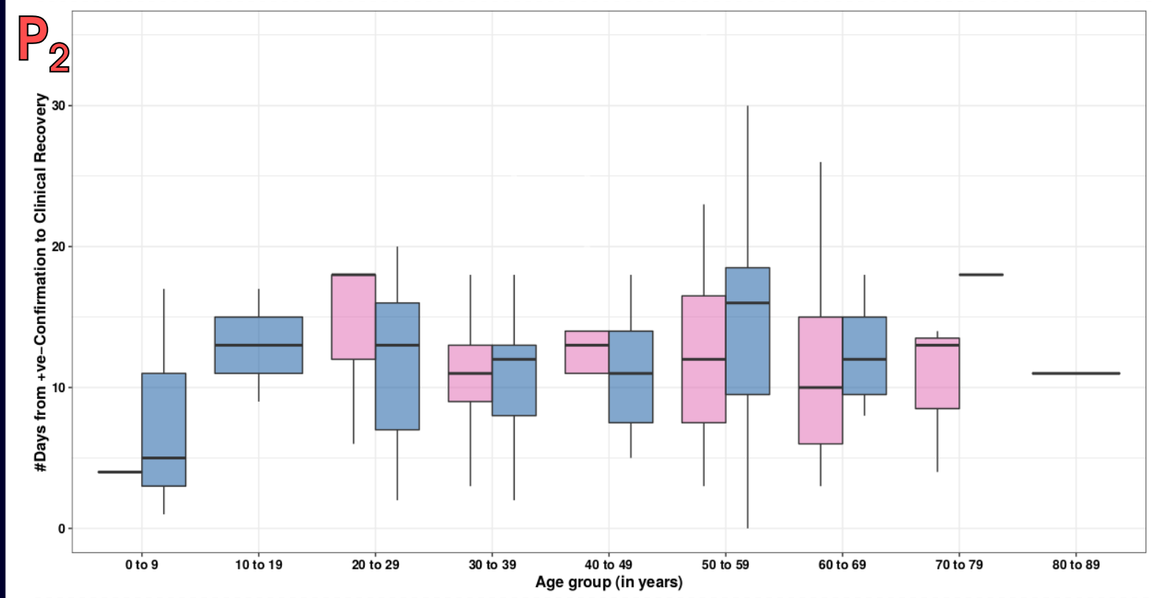
(ii)


(i)
#Days from +ve-Confirmation to Clinical Recovery
0
10
20
30
23
14
11
7.75
2
27
15
11
7
0
FEMALE
MALE
Gender
Gender
FEMALE
MALE
(ii)
#Days from +ve-Confirmation to Clinical Recovery
0
10
20
30
8
4.5
3.25
1
17
15.5
13.5
11.25
9
20
14
9.5
6
2
20
13
11
8
2
20
13.25
11
6.75
4
27
16.5
11
8
0
18
14.75
11.5
9
3
18
13.25
9
6.25
1
11
9.25
7.5
6.75
4
0-9
10-19
20-29
30-39
40-49
50-59
60-69
70-79
80-89
Age group (in years)
AgeRange
0-9
10-19
20-29
30-39
40-49
50-59
60-69
70-79
80-89

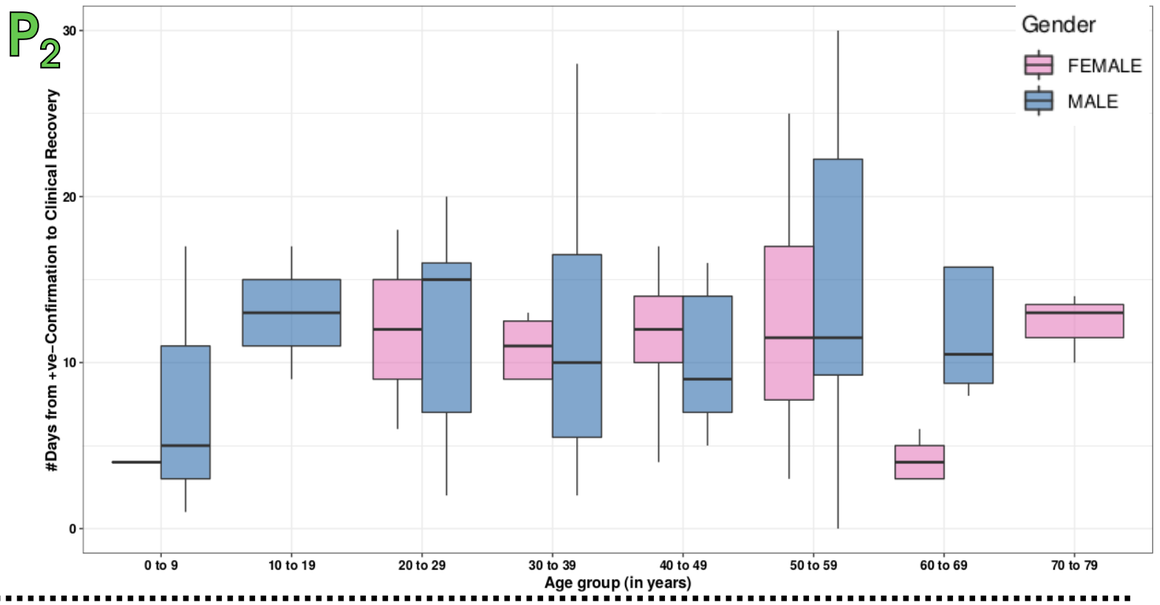


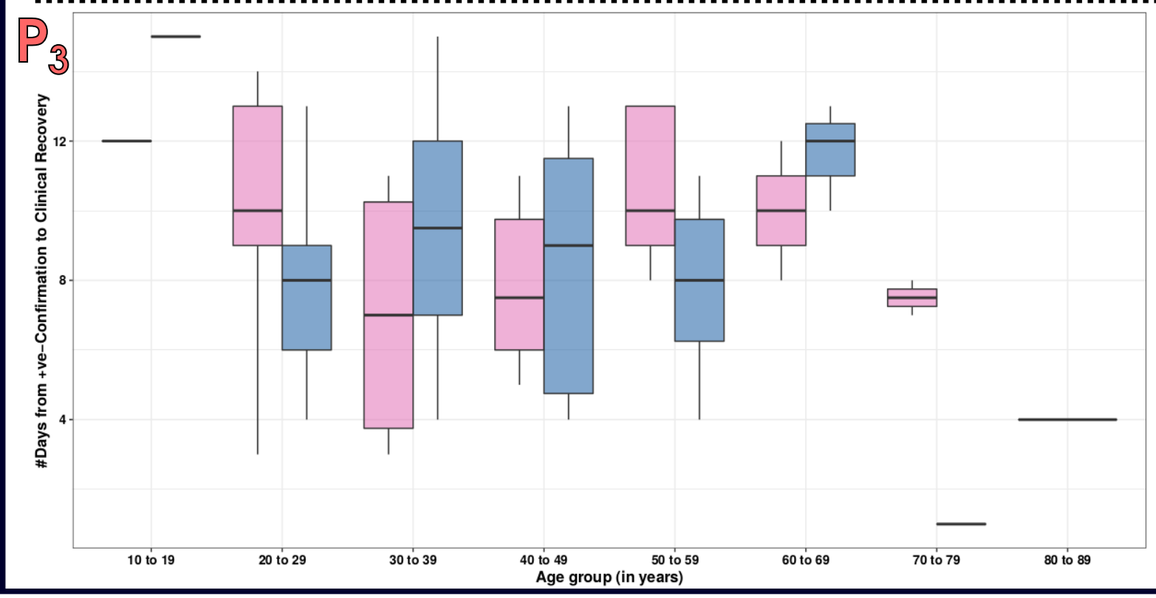
(i)



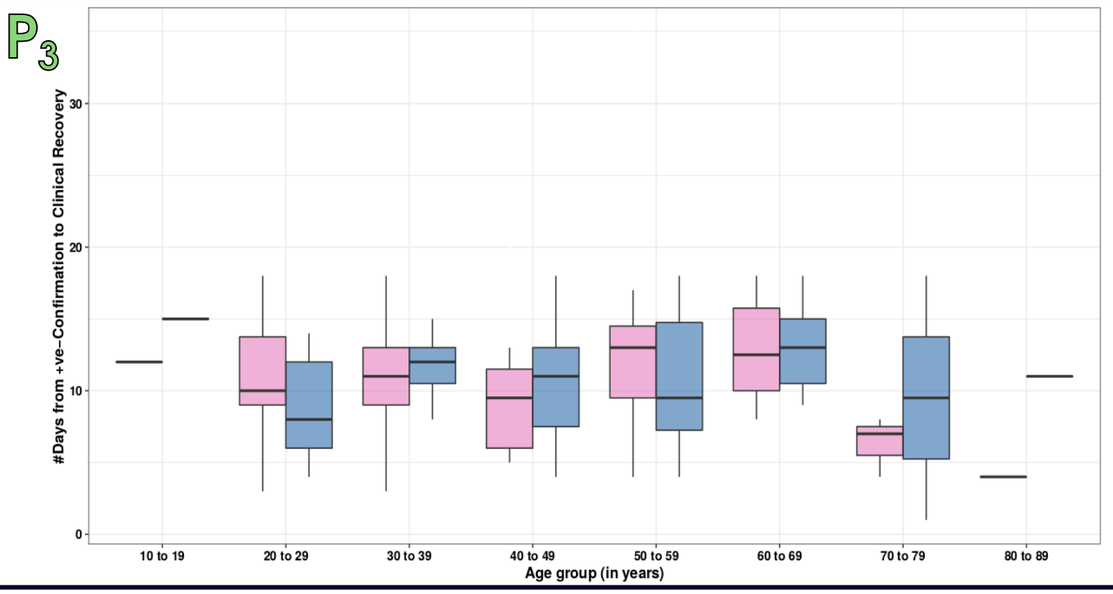


(ii)



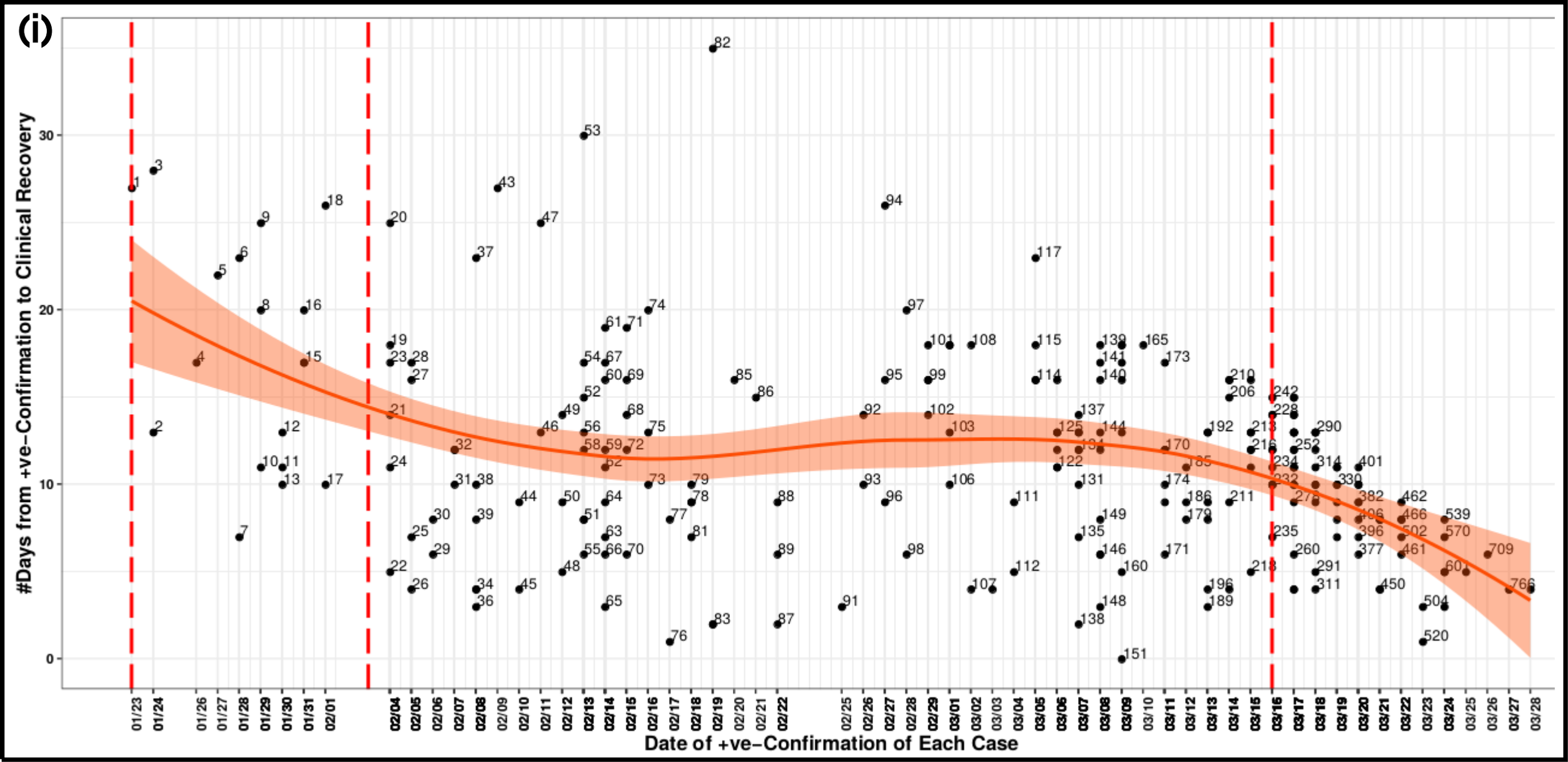


(iii)

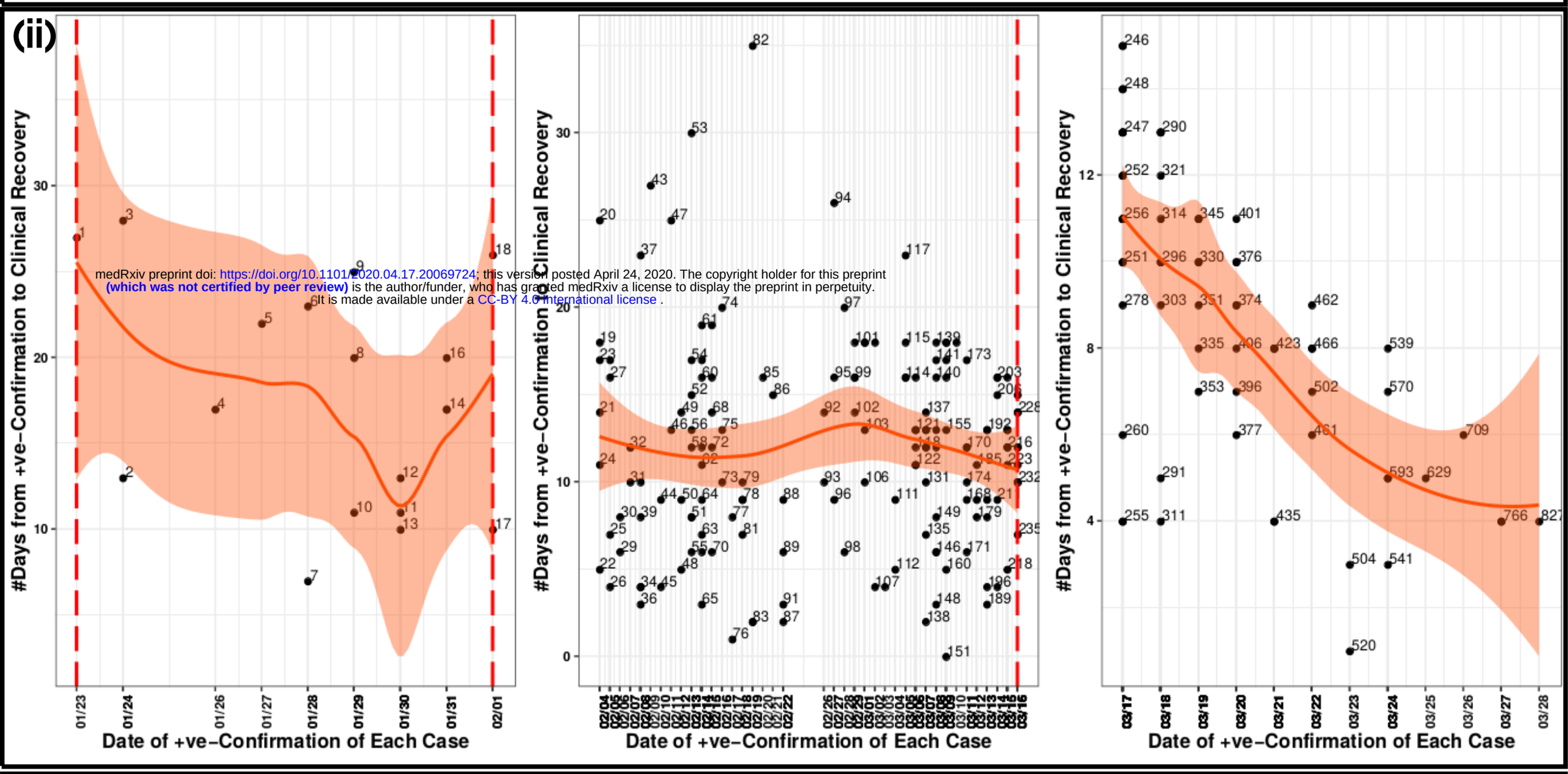


(b) (c)

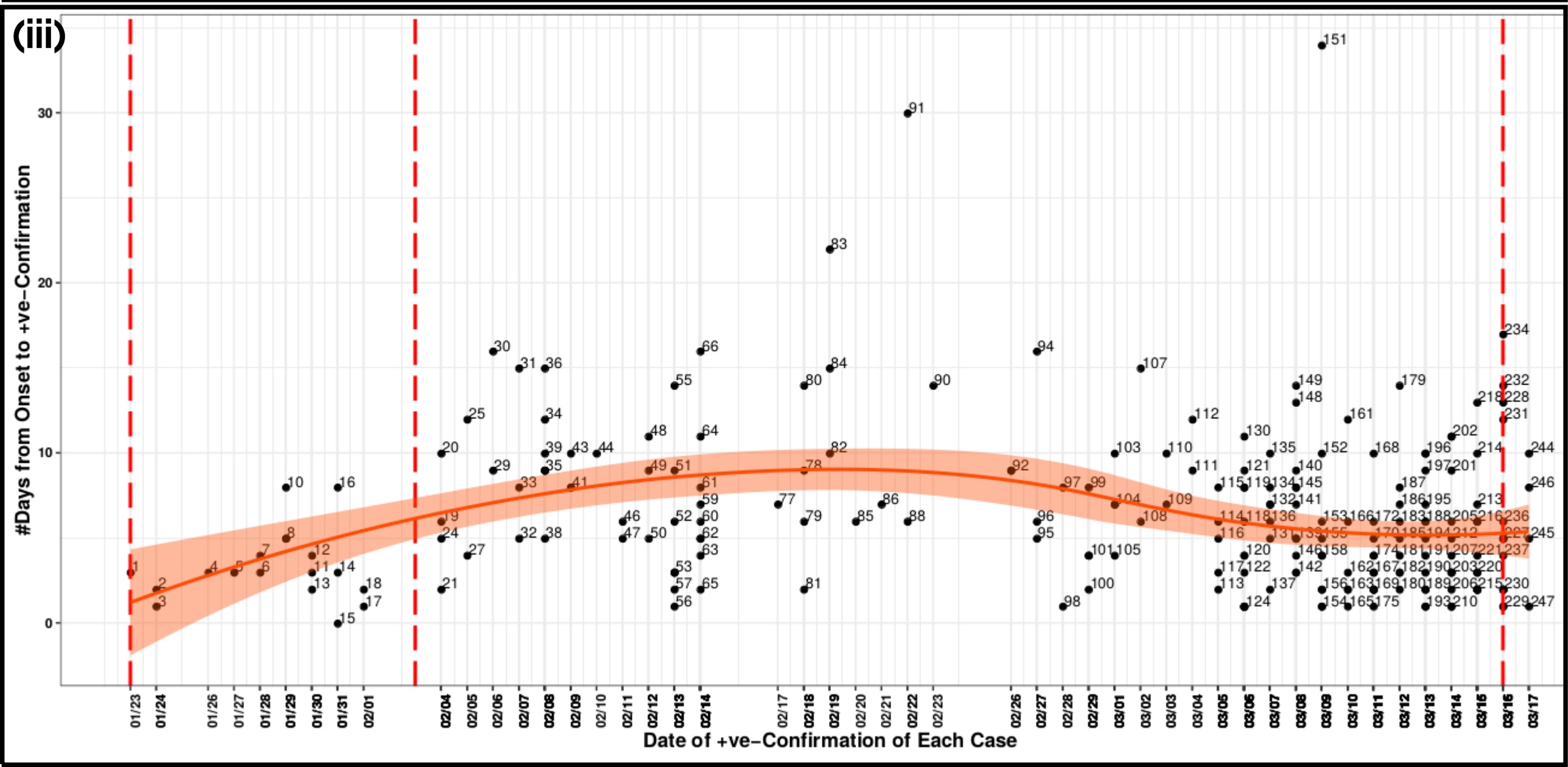
(i)

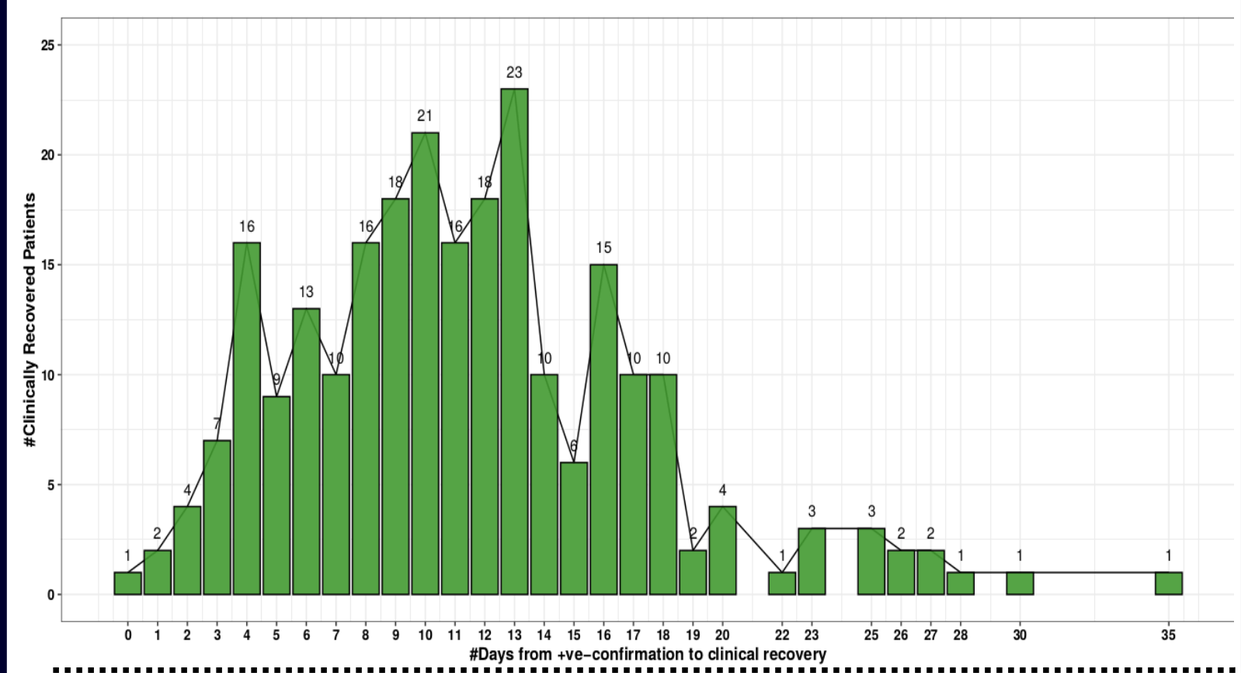


(ii)

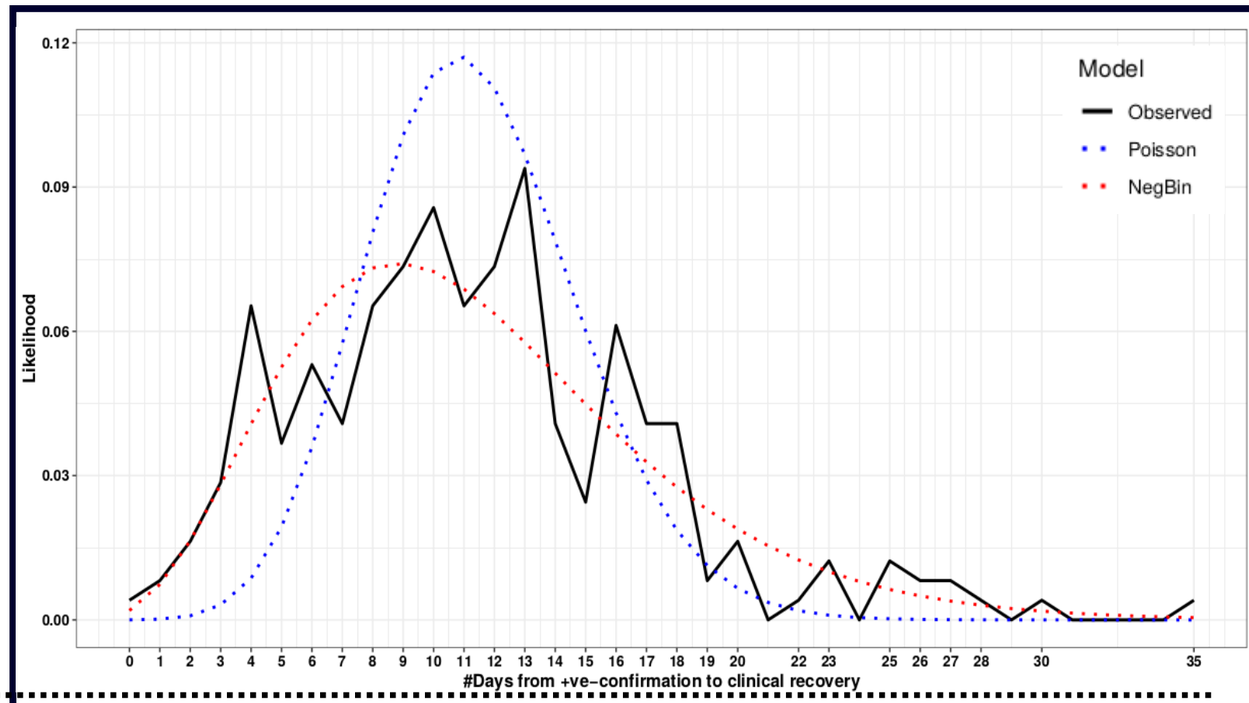


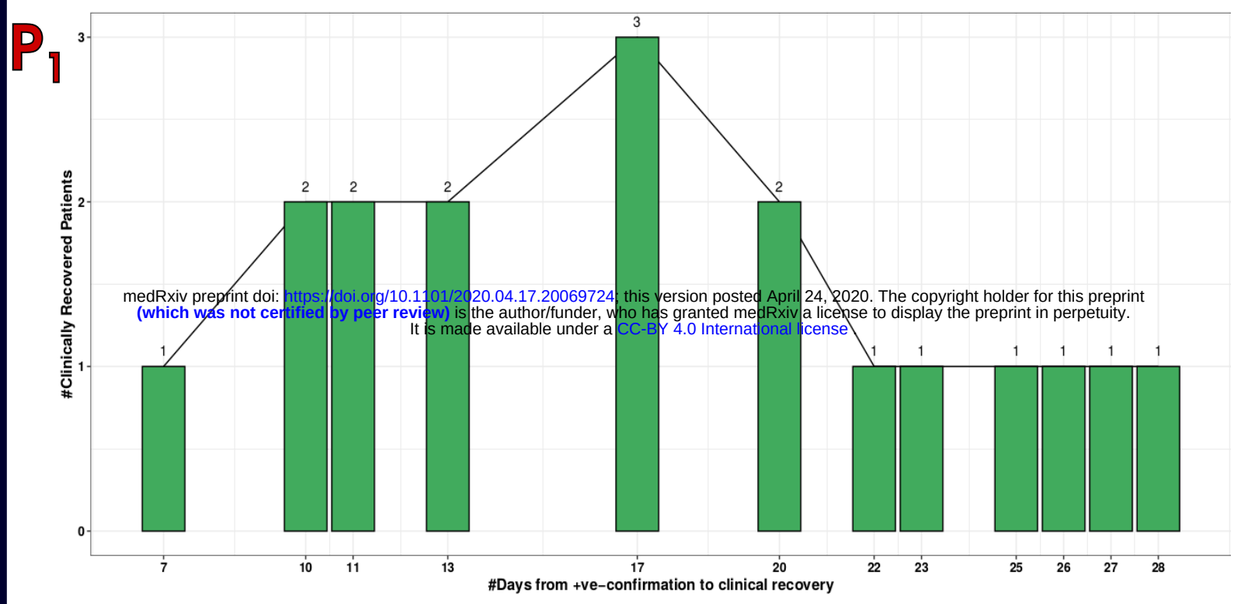
(iii)



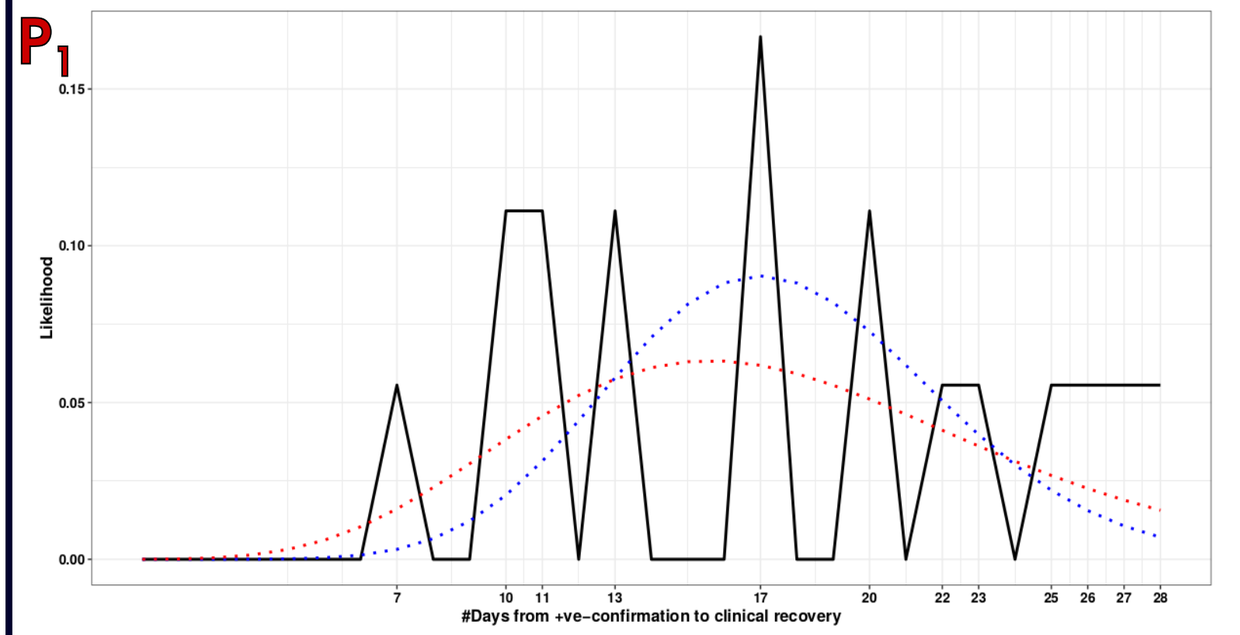


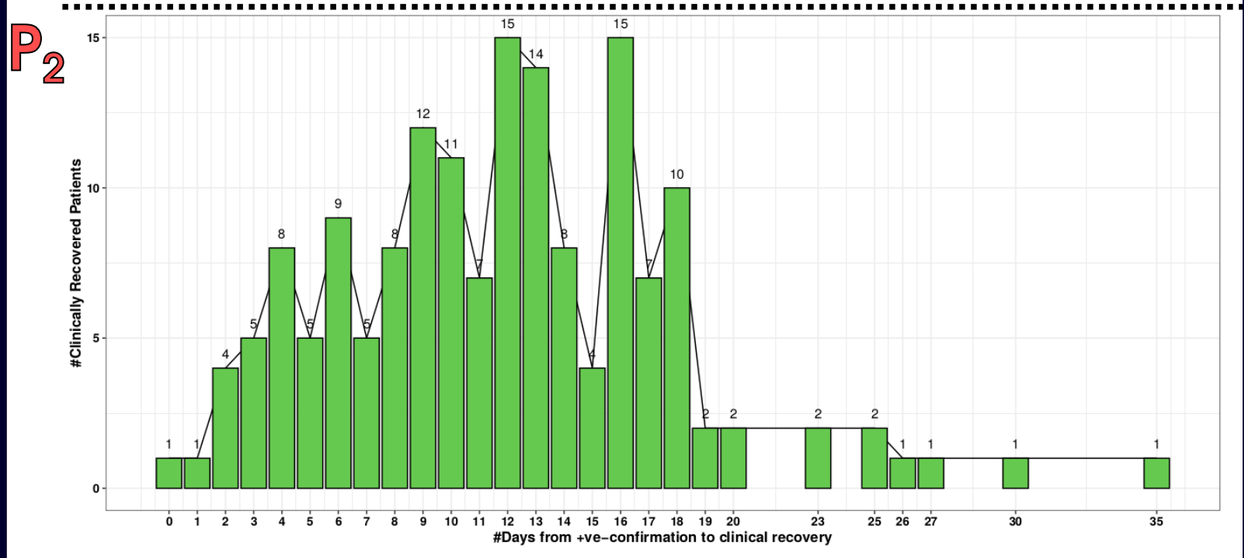
(i)



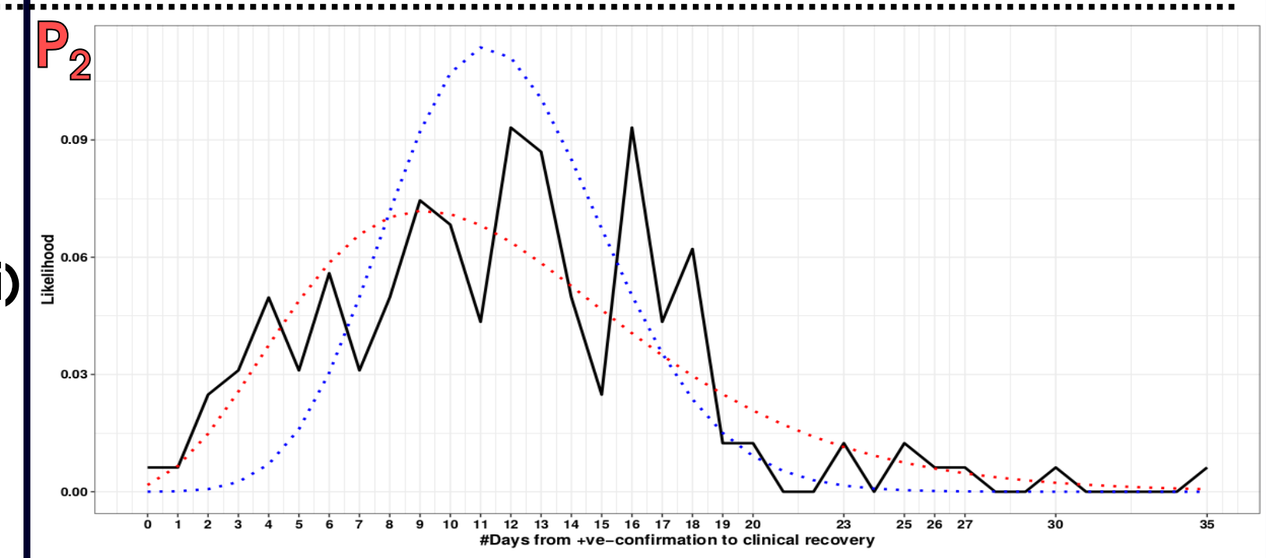


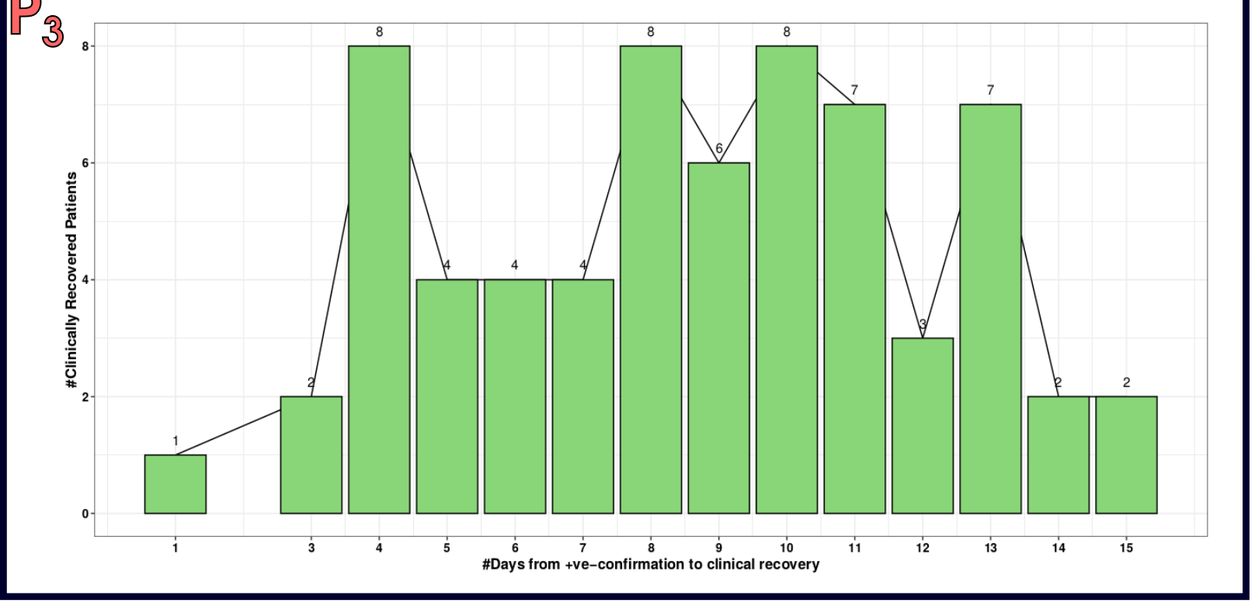
(ii)



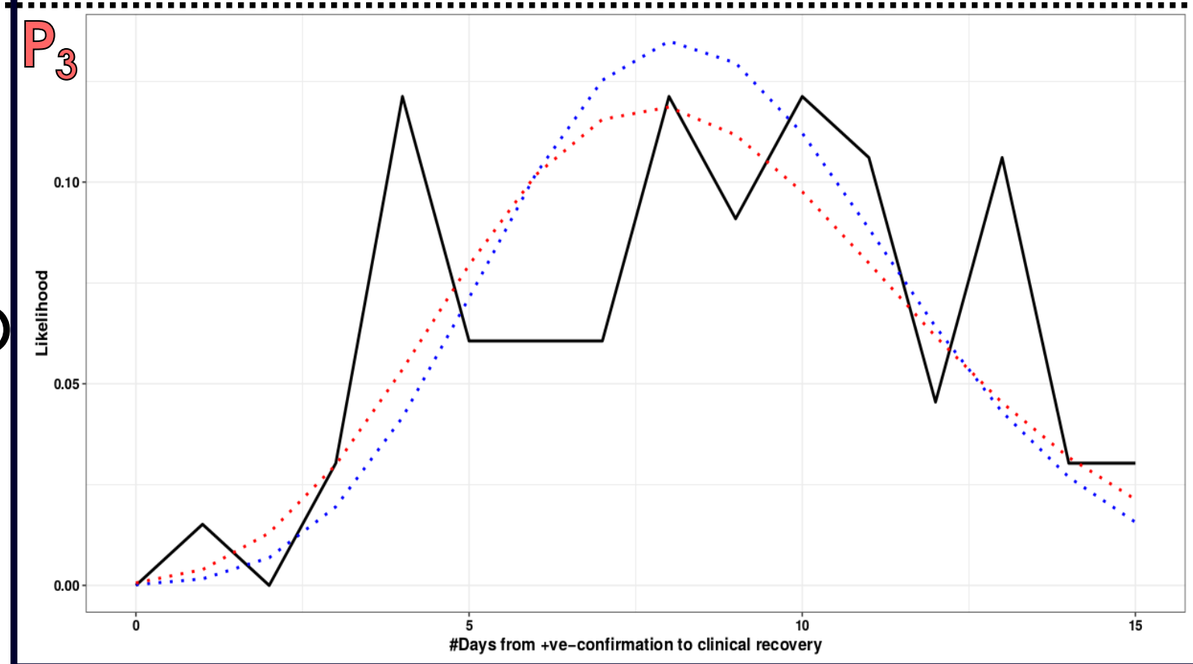


(iii)





(iv)



(a)

(b)