Journal of Data Science xx (xx), 1–10 DOI: xx.xxxx/xxxxxxxxx
July 2020

Data Visualization and Descriptive Analysis for Understanding Epidemiological Characteristics of COVID-19: A Case Study of a Dataset from January 22, 2020 to March 29, 2020

Yasin Khadem Charvadeh *[1] and Grace Y. Yi†[1, 2]

[1]Department of Statistical and Actuarial Sciences, University of Western Ontario, London, ON N6A 3K7, Canada

[2]Department of Computer Science, University of Western Ontario, London, ON N6A 3K7, Canada

Abstract

COVID-19 is a disease caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) that was reported to spread in people in December 2019. Understanding epidemiological features of COVID-19 is important for the ongoing global efforts to contain the virus. As a complement to the available work, in this article we analyze the Kaggle novel coronavirus dataset of 3397 patients dated from January 22, 2020 to March 29, 2020. We employ semiparametric and nonparametric survival models as well as text mining and data visualization techniques to examine the clinical manifestations and epidemiological features of COVID-19. Our analysis shows that: (*i*) the median incubation time is about 5 days and older people tend to have a longer incubation period; (*ii*) the median time for infected people to recover is about 20 days, and the recovery time is significantly associated with age but not gender; (*iii*) the fatality rate is higher for older infected patients than for younger patients.

Keywords *Incubation time; Recovery time; Risk factors; Symptom onset; Survival analysis; Text mining.*

1 Introduction

SARS-CoV-2 (Lai et al., 2020) is a member of coronaviruses family which causes a transmittable infectious respiratory disease known as COVID-19. The novel coronavirus was first reported in December 2019 in the city of Wuhan, China (Zhang et al., 2020). On March 11, 2020, the World Health Organization (WHO) upgraded the status of the COVID-19 outbreak from epidemic to a global pandemic and now almost all countries have reported confirmed cases with the USA having the highest number of confirmed cases (Worldometers, 2020). As of May 21, 2020, the WHO reported 4,893,186 confirmed cases with 323,256 deaths.

Estimating the incubation period is crucial for the disease control. Having a sensible estimate of the median incubation time helps the government and healthcare sector decide a rationale quarantine time. Estimating recovery times for infected patients is of great importance for healthcare workers to effectively allocate the limited medical resources to cope with the COVID-19 crisis. Moreover, understanding the relationships of demographic factors, such as age and gender, with COVID-19 is essential as it helps healthcare professionals prioritize treatment of patients with different characteristics. While various efforts have been made to

*Email: ykhademc@uwo.ca
†Corresponding author, Email: gyi5@uwo.ca

Received April, 2020; Accepted May, 2020

study the behaviour of SARS-CoV-2 since the outbreak of COVID-19, the understanding of COVID-19 has been constantly enhanced as more COVID-19 data become available. Extensive evidence-based studies from multiple angels are required to comprehensively unveil the clinical characteristics of COVID-19 by examining the data coming from different sources as the pandemic evolves.

To this end, here we study the Kaggle novel coronavirus dataset from January 22, 2020 to March 29, 2020, to be described in detail in Section 2, to preliminarily examine the following questions: (1). What is the average time of symptom onset? (2). How long does it take for infected patients to recover? While each of these questions warrants in-depth research when more data become available with the evolvement of COVID-19, here we focus on providing an exploratory analysis using the techniques of data visualization and text mining as well as modeling of survival data. We hope such a study will offer intuitive insights into future in-depth research of each topic.

The remainder of this article is organized as follows. In Section 2 we describe the data and examine different features of COVID-19 by data visualization. In Section 3 we employ survival analysis techniques to estimate the distribution of recovery times for infected patients. In Section 4 we estimate the average time of symptom onset. We conclude the manuscript with discussion in the last section.

2 Data Visualization

2.1 Data Description

In this study, we use the Kaggle novel coronavirus dataset from January 22, 2020 to March 29, 2020. The dataset, available as Google spreadsheet at https://www.kaggle.com, has been updated automatically every five minutes based on Johns Hopkins Center for System Science and Engineering (CSSE) data (https://github.com/CSSEGISandData/COVID-19). The dataset consists of measurements of 3397 people with the novel coronavirus from 39 countries including those in Europe, Asia, and Africa. There are 14 variables representing the *summary, location, country, gender, age, symptom onset, hospital visit date, exposure start, exposure end, visiting Wuhan, from Wuhan, death, recovery status,* and *symptoms* of the infected cases. Using the information given in the *summary*, *exposure start*, *exposure end*, *symptom onset*, and *recovery status*, we further extract more specific information from the original dataset, including *infection source, travel history, time gap between exposure to symptom onset,* and *time gap between symptom onset to recovery.* A copy of the dataset is available at https://github.com/YasinKhc/Covid-19. Among 3397 patients, only 1449 of them have the information of age which ranges from 3 months to 96 years. In Table 1 we present the age distribution of infected cases separately for females and males.

2.2 Descriptive Analysis

Among the 3397 patients, we found that older people have a higher fatality rate comparing to younger people. The mean and median of age for deceased cases were found to be 71.5 and 73.5, respectively. The left graph in Figure 1 displays the side by side barplots for the counts of deceased cases for males and females divided into six age groups, and the right graph in Figure 1 records the fatality rate for men and women in the six age groups, where the fatality rate is calculated as the ratio of the number of deaths in an age group with a given gender to the

Table 1: Age distribution of infected cases by gender: The entries display the number and the percentage (in parentheses) for each cohort.

	Age range (in year)					
	0-19	20-39	40-59	60-79	80-96	Total
Male	28 (3%)	193 (24%)	313 (38%)	242 (30%)	41 (5%)	817
Female	25 (4%)	168 (27%)	212 (34%)	186 (29%)	41 (6%)	632

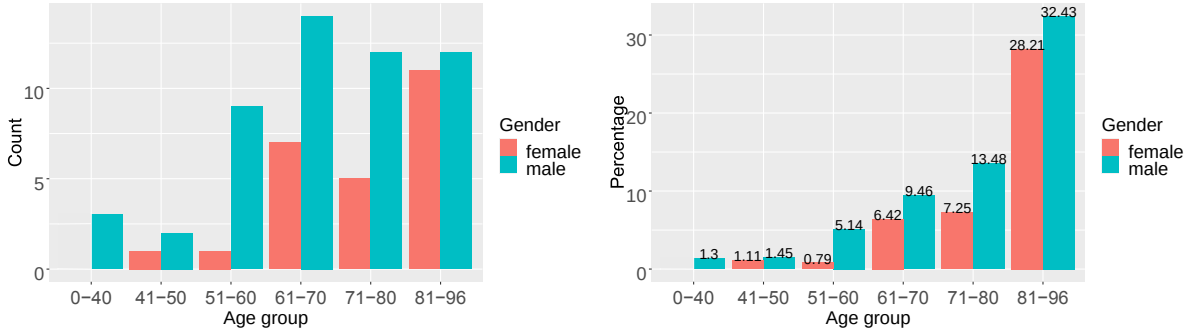

Figure 1: Barplots for the number of deceased cases and fatality rate.

number of infected cases in that group. It is clear that the fatality rate increases with age, and the fatality rate for men in each age group appears higher than that for women. These results are consistent with those reported by Jin et al. (2020).

We further perform the Chi-square test of independence (Pearson, 1900) to determine whether there is a statistically significant association between age/gender and fatality. For the null hypothesis that the fatality rate is identical for all the age groups, we obtain the p-value of the Chi-square test to be 0.0005. For the null hypothesis that the fatality rate is identical for males and females, we obtain the p-value of the Chi-square test to be 0.0748.

The left plot in Figure 2 shows that around 28% of the infected people had a recent travel history. The right plot in Figure 2 reports that 13% of the cases had a close contact with other infected people, and the source for the rest large portion (87%) of infections remains unknown, which is very likely due to undetected community transmissions.

To understand what symptoms are most related to infected cases with COVID-19, we per-

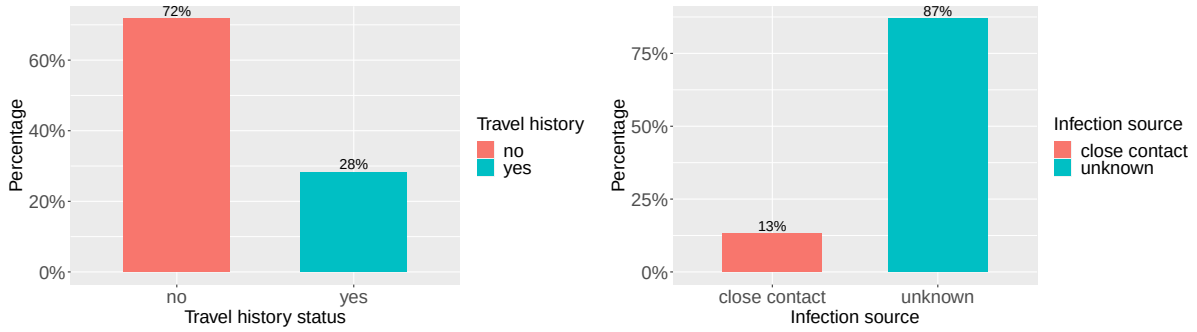

Figure 2: Barplots for the recent travel history and infection source

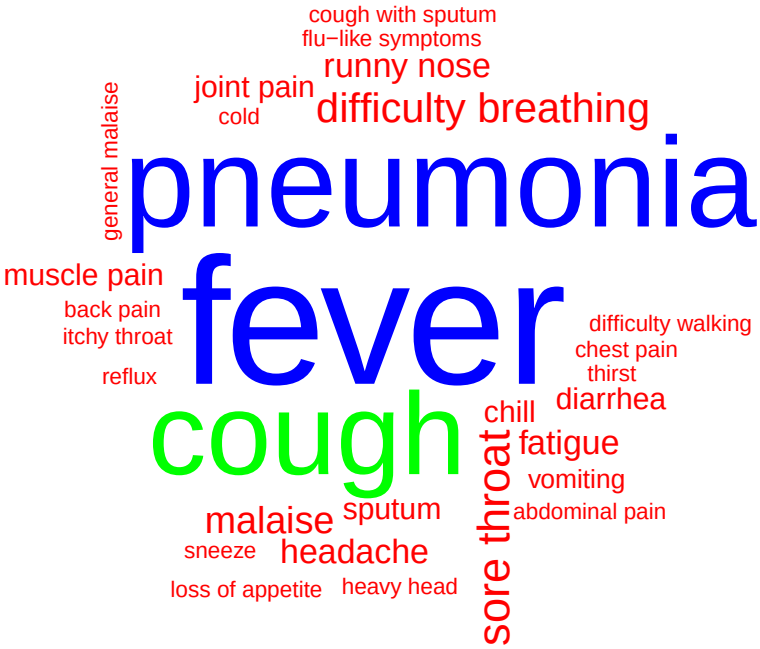

Figure 3: Word cloud of the symptoms.

form a text analysis using *word clouds* (Viégas and Wattenberg, 2008) which typically visualize word frequencies by using different sizes of words. The more common a term appears in a text dataset, the larger and bolder it appears in the word cloud. Word clouds are an intuitive tool in visualizing and highlighting words with greater prominence. To generate a word cloud for symptoms of COVID-19, we first collapse the *summary* into a single text document and extract the terms and words of describing the symptoms of infected patients, and then store them in a new text document. Thereafter, different medical words and terms that represent a specific symptom are summarized as a single unique word or term. For example, in the *summary*, besides *difficulty breathing*, three other terms were alternatively used to describe the same symptom related to breathing: *shortness of breath*, *dyspnea*, and *respiratory distress*. In our text analysis here, we classify them as the same description for the symptom of breathing and then unify them with the term "difficulty breathing". Next, using the obtained text document and the word cloud generator in the package *wordcloud* of R, we summarize the symptoms for 652 COVID-19 infected patients in Figure 3. It is clearly seen that fever, cough and pneumonia are the most frequent symptoms reported by those patients.

3 Examination of Recovery Time

To help the government and health authorities prepare for major spikes of the number of new COVID-19 infected cases, it is important to estimate the time for infected patients to recover. In this section, we use survival analysis techniques to study recovery times of infected patients. Here the recovery time of an infected patient, denoted as T, is taken as the time-to-event, or survival time, using the terminology in survival analysis (e.g., Lawless, 2003). In other words, the *event* is defined to be *recovered*, and hence, patients who die from COVID-19 are treated as censored.

Table 2: Median recovery time for male and female.

Gender	The number of infected patients	The number (percentage) of recovery	Median	95% Confidence interval
Female	52	43 (83%)	20	(17, 21)
Male	89	58 (65%)	20	(19, 23)

Table 3: Median recovery time (in day) for different age groups.

Age group	The number of infected patients	The number (percentage) of recovery	Median	95% Confidence interval
0-40	47	45 (96%)	18	(16, 20)
41-60	50	45 (90%)	20	(17, 22)
61-96	43	10 (23%)	26	(21, 30)

First, we use the distribution-free Kaplan-Meier approach to examine the survivor function $S(t) = P(T > t)$ for the recovery time, where $t \in [0, 45]$ with $[0, 45]$ representing the study period of 45 days, and 0 is defined as the time of symptom onset for an infected patient.

We examine the recovery time from three angels. First, we do not distinguish infected cases; secondly, we classify the infected cases into two groups by gender; thirdly, we divide the infected cases into three age groups: (0, 40], (40, 60], and (60, 96]. The corresponding Kaplan-Meier estimates are reported in Figure 4. The top panel of Figure 4 illustrates the Kaplan-Meier time-to-recovery survival curve for all the infected cases, where the red curve represents the estimated probabilities, the red shaded areas stand for the 95% confidence region, and patients who are censored are marked with + signs. The dashed dark lines indicate the survivor probability at the median recovery time, saying that with 50% of the probability an infected patient takes more than 20 days to recover (if they would recover). A 95% confidence interval for the median recovery time is (19, 21).

The middle panel of Figure 4 shows the Kaplan-Meier survival curves of recovery times for men and women, which are not considerably different. Furthermore, applying the log-rank test (Harrington, 2005) to assess whether or not the difference between the two curves is statistically significant, we obtain that the p-value is 0.5, clearly showing no evidence that the recovery time differs for men and women. The details of median recovery times and their corresponding 95% confidence intervals for men and women are reported in Table 2.

The bottom panel of Figure 4 displays the Kaplan-Meier survival curves for the three age groups. It can be visually concluded that people of older age are more likely to have longer recovery times. The corresponding log-rank test yields the p-value to be 10^{-4}, supporting that the differences in recovery times for different age groups are statistically significant. Median recovery times and their corresponding 95% confidence intervals for the three age groups are summarized in Table 3.

Next, we quantify how the recovery time is associated with age and gender by employing the semiparametric accelerated failure time (AFT) model:

$$\log T = \beta_0 + \beta_1 \times \text{gender} + \beta_2 \times \text{age} + \epsilon,$$

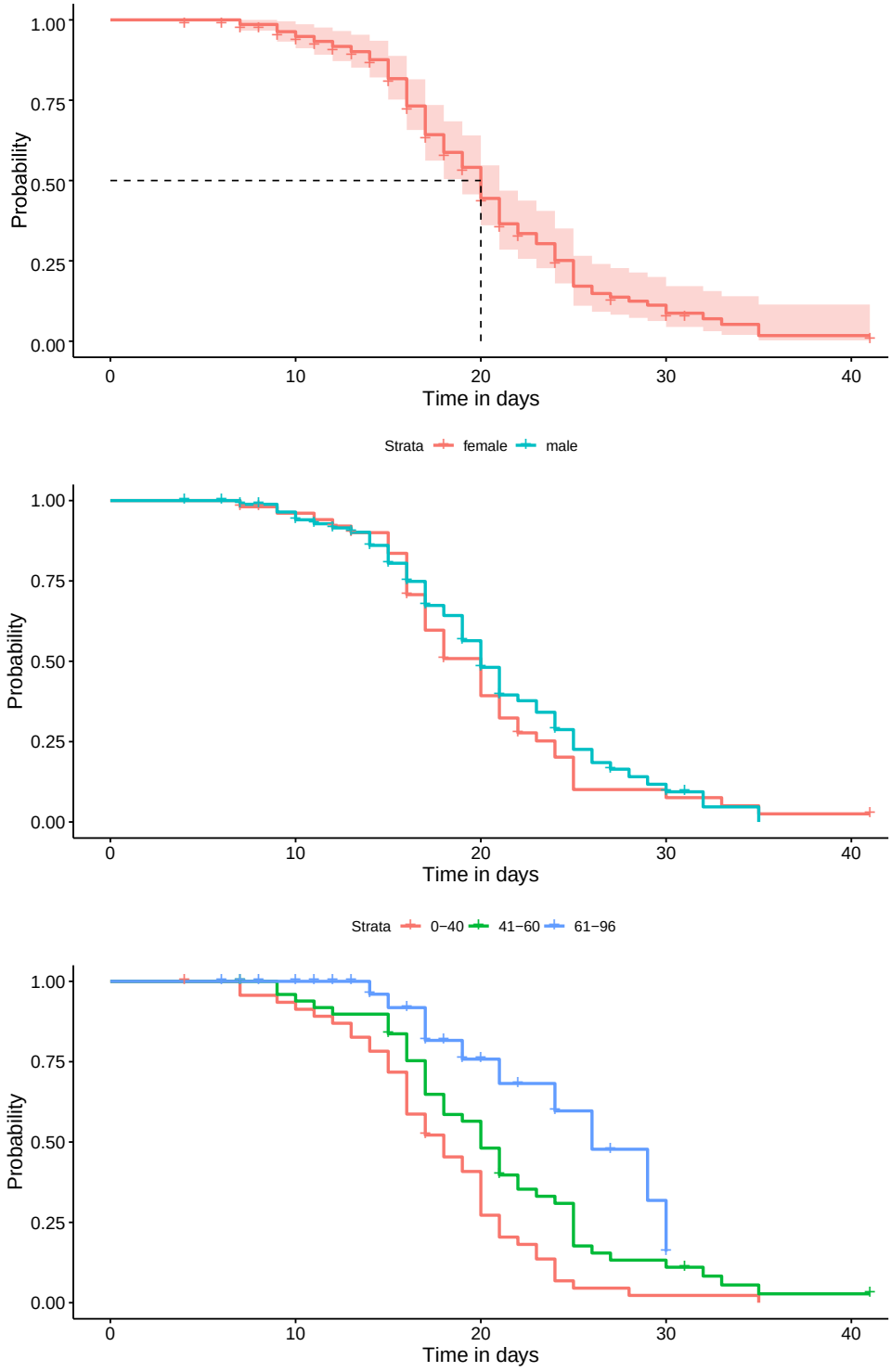

Figure 4: Kaplan-Meier time-to-recovery survival curves

Table 4: Analysis results of recovery times under the semiparametric AFT model.

Parameters	Estimate	Standard error	95% Confidence interval
Intercept (β_0)	2.498	0.119	(2.265, 2.731)
gender (β_1)	0.066	0.069	(-0.069, 0.201)
age (β_2)	0.094	0.022	(0.051, 0.137)

where β_0 is the intercept, β_1 and β_2 are regression parameters, and ϵ is the error term with mean zero and an unspecified probability distribution. For ease of interpretation, we use ten years as the unit of age, as suggested by the editor. Estimation of the parameters can be obtained using the generalized least squares approach (e.g., Chiou et al., 2014); the results are reported in Table 4.

The analysis results show no evidence that recovery times differ in women and men. Age is found to be significantly related to the recovery time. Older infected patients need a longer time to recover from COVID-19. Exponentiating the estimate of β_2, we quantify the age effect on the recovery time. With the gender effect adjusted, ten years older in age would extend the recovery time by 9.9%.

4 Gap Time between Exposure and Symptom Onset

One of the major concerns that healthcare workers and the government have been trying to address is on *stealthy transmissions* of COVID-19. Researchers in Columbia University's Mailman School of Public Health used a computer model to show how undetected cases may boost the spread of the COVID-19 outbreak in China. They showed that the virus spread was rapid and its containment was challenging (Li et al., 2020). Understanding the average gap time between the time of exposure to the virus and symptom onset for infected patients is useful for healthcare workers and the government to make effective measures to curb the spread of the virus.

Among the 3397 infected people, 207 reported both the time for exposure and the symptom onset time. The time of exposure is taken as an approximate time a patient contracted the virus by having a close contact with someone who was already infected or travelling to infected areas. The symptom onset date is based on the time when an infected patient experienced flu-like symptoms such as fever, sore throat, in more severe cases, difficulty breathing. Eighty-five patients reported a time interval for exposure spanning from 1 to 27 days. We treat those exposure intervals with a length less than one day as a single time point. To understand the underlying incubation times for infected cases who reported different types of information on infection, we estimate the median and average incubation times for the cohort of 3397 infected patients using the following three methods:

- Method 1: the time period between the start time of exposure and symptom onset;
- Method 2: the time period between the end time of exposure and symptom onset;
- Method 3: we use the middle point of the time interval to approximate the exposure time, and take the time period between the approximated exposure time and symptom onset.

For 140 patients who reported only a single time point for exposure, these three methods yield the same values. For the cohort of 3397 infected cases, Method 1 yields that the mean and median incubation times to be 8.4 and 6 days, respectively; Method 2 outputs a lot smaller mean and median incubation times which are respectively 3.3 and 2 days; and Method 3 gives that the

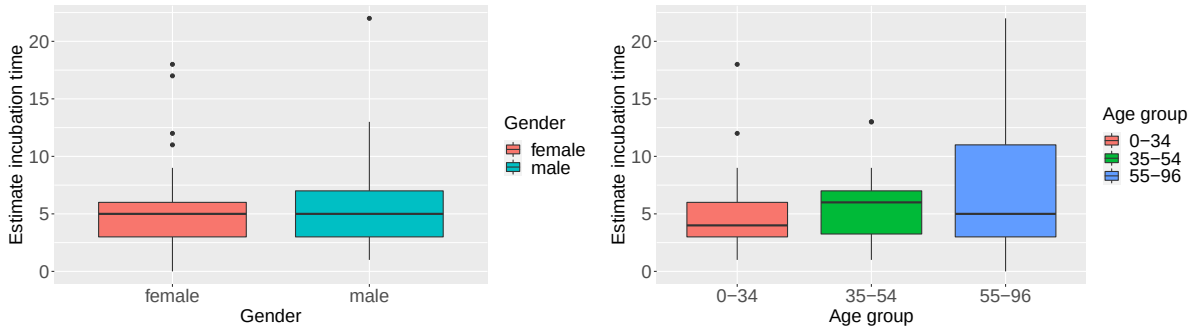

Figure 5: Boxplots of estimated incubation times by gender and three age groups.

mean and median of the incubation period are 5.8 and 5 days, respectively. The estimates of Method 3 are similar to those reported by Lauer et al. (2020) and Han (2020). Lauer et al. (2020), by conducting a pooled analysis of 181 infections reported between January 4, 2020 and February 24, 2020, found that median incubation period to be 5.1 days. Han (2020) used a chain-of-infection data collected from 10 regions in China to estimate the median incubation period. They employed different statistical approaches such as Monte-Carlo simulations as well as non-parametric methods and estimated that the mean and median of incubation times are 5.8 and 5 days, respectively.

To show how incubation times may differ between females and males, in the left panel of Figure 5 we report the boxplots of the incubation times obtained from Method 3 for 31 females and 49 males. To see possible age effects, in the right panel of Figure 5 we graph the incubation times for three age groups, where 21, 30, and 25 patients are included in the age groups of 0-34, 35-54, and 55-96, respectively. The median incubation period for patients aged within 35-54 is the largest, and the median incubation period for patient over 55 years of age is slightly longer than that of the age 0-34 group. However, incubation times for older patients have more variability than those for younger infected cases.

5 Discussion

In this article we explore epidemiological characteristics of COVID-19 by studying a Kaggle novel coronavirus dataset, dated from January 22, 2020 to March 29, 2020, which includes 3397 infected cases and 83 deaths from COVID-19. We find that the median incubation time of COVID-19 is about 5 days. Our text analysis shows that the most dominant symptoms of COVID-19 are fever, cough, and pneumonia. The non-parametric Kaplan-Meier method yields a median recovery time of 20 days for infected patients who are not stratified by their characteristics. Our findings further suggest that the recovery time increases as the age increases, and there is no significant gender-difference in recovery times.

As discussed by He et al. (2020), while many studies examined epidemiological characteristics of COVID-19, those studies do not necessarily reveal the same findings or similar estimates of the same measure. For instance, regarding the estimate of the average incubation times, He et al. (2020) reviewed five studies conducted between December 31, 2019 and February 24, 2020, and those studies reported varying estimates of the average incubation time, ranging from 4.9 days to 6.4 days. In addition, we note that our estimate of the median incubation time differs from the estimate, 8.1 days, provided by Qin et al. (2020). The discrepancies in estimating the

same quantity are primarily attributed to the heterogeneity in different studies, including the differences in the time window, the study subjects, the study design, the model assumptions, and the measures of controlling the virus spread by different regions.

We point out that the validity of the analysis results here relies on the quality of the Kaggle data we use. In our analysis we ignore missing observations, which is basically driven by the perception that missingness arises completely at random. However, when such an assumption is not feasible, proper adjustments of missingness effects are generally expected. On the other hand, as commented by a referee, reporting bias and recall bias should be aware of when analyzing the COVID-19 data. If the degree of such biases are not mild, then proper de-biasing adjustments should be introduced in inferential procedures to yield valid or nearly valid analysis results. Methods of addressing effects of error-in-variables can be employed for this purpose. For detail, see Carroll et al. (2006) and Yi (2017).

Finally, we note that our analysis results are obtained from using the reported information for those patients who were assessed by medical personnel. The information for infected patients with mild symptoms or asymptomatic infections was often not available for being included in the dataset, because those patients did not go to hospital for assessment. As a result, when interpreting the results, care is needed for the target population.

Acknowledgements

The authors thank the Editor Jun Yan and the review team for their useful comments on the initial submission. This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC). Yi is Canada Research Chair in Data Science (Tier 1). Her research was undertaken, in part, thanks to funding from the Canada Research Chairs Program.

References

Carroll, R. J., D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu (2006). *Measurement Error in Nonlinear Models* (2nd ed.). Chapman & Hall/CRC.

Chiou, S. H., S. Kang, J. Kim, and J. Yan (2014). Marginal semiparametric multivariate accelerated failure time model with generalized estimating equations. *Lifetime Data Analysis 20*(4), 599–618.

Han, H. (2020). Estimate the incubation period of coronavirus 2019 (COVID-19). medRxiv preprint: https://doi.org/10.1101/2020.02.24.20027474.

Harrington, D. (2005). Linear rank tests in survival analysis. *Encyclopedia of Biostatistics 4*, 2802–2812.

He, W., G. Y. Yi, and Y. Zhu (2020). Estimation of the basic reproduction number, average incubation time, asymptomatic infection rate, and case fatality rate for COVID-19: Meta-analysis and sensitivity analysis. *Journal of Medical Virology*. Forthcoming. https://doi.org/10.1002/jmv.26041.

Jin, J.-M., P. Bai, W. He, F. Wu, X.-F. Liu, D.-M. Han, S. Liu, and J.-K. Yang (2020). Gender differences in patients with COVID-19: Focus on severity and mortality. *Frontiers in Public Health 8*, 152.

Lai, C. C., T. P. Shih, W. C. Ko, H. J. Tang, and P. R. Hsueh (2020). Severe acute respira-

tory syndrome coronavirus 2 (SARS-CoV-2) and corona virus disease-2019 (COVID-19): The epidemic and the challenges. *International Journal of Antimicrobial Agents 55*(3), 105924.

Lauer, S. A., K. H. Grantz, Q. Bi, F. K. Jones, Q. Zheng, H. R. Meredith, A. S. Azman, N. G. Reich, and J. Lessler (2020). The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: estimation and application. *Annals of Internal Medicine 172*(9), 577–582.

Lawless, J. F. (2003). *Statistical Models and Methods for Lifetime Data* (2nd ed.). John Wiley & Sons, Inc., Hoboken, New Jersey.

Li, R., S. Pei, B. Chen, Y. Song, T. Zhang, W. Yang, and J. Shaman (2020). Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). *Science 368*(6490), 489–493.

Pearson, K. (1900). X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 50*(302), 157–175.

Qin, J., C. You, Q. Lin, T. Hu, S. Yu, and X.-H. Zhou (2020). Estimation of incubation period distribution of COVID-19 using disease onset forward time: A novel cross-sectional and forward follow-up study. medRxiv preprint: https://doi.org/10.1101/2020.03.06.20032417.

Viégas, F. B. and M. Wattenberg (2008). Timelines tag clouds and the case for vernacular visualization. *Interactions 15*(4), 49–52.

Worldometers (2020). Reported cases and deaths by country, territory, or conveyance. https://www.worldometers.info/coronavirus/. Accessed: 2020-05-21.

Yi, G. Y. (2017). *Statistical Analysis with Measurement Error or Misclassification: Strategy, Method and Application*. Springer Science+Business Media LLC, New York.

Zhang, T., Q. Wu, and Z. Zhang (2020). Probable pangolin origin of SARS-CoV-2 associated with the COVID-19 outbreak. *Current Biology 30*(7), 1346–1351.