Assessing risk factors for the incubation period of COVID-19

Abstract

In December 2019, there was a cluster of pneumonia cases of unknown etiology detected in Wuhan, Hubei Province, China. It was later designated as coronavirus disease 2019 (COVID-19) and soon becomes a global pandemic, which has extremely complex behaviors regarding to its transmission, symptoms and incubation period. A better understanding of these key characteristics is needed for scientists and policy makers to monitor and control the pandemic as well as to set up proper quarantine procedures. In our study, we analyze data from 463 Wuhan-exported cases who left Wuhan before the travel ban on January 23, 2020. The data set includes the dates of beginning stay in Wuhan, ending stay in Wuhan, and symptom onset, as well as variables age and gender. We perform regression analysis under the Cox proportional hazards model to evaluate the association of age and gender with the incubation period of COVID-19, defined as the duration between infection and symptom onset. Since the infection time is not observed, we develop three different methods to handle the incubation period. The multiple imputation method that imputes the unobserved infection time by assuming an exponential epidemic growth properly accounts for the data structure and fits the data better than the other two methods. It indicates that younger people have a longer incubation period than the older, while there is no significant difference in incubation time between male and female.

I. Introduction

In December 2019, a cluster of viral pneumonia cases of unknown etiology was detected in Wuhan, Hubei Province, China, and it was initially traced to one seafood market in Wuhan [7]. However, many cases were reported later to have no association with this market, meaning that there exists a human-to-human transmission of the virus. The virus that causes this disease, now referred to as SARS-CoV-2, is a novel coronavirus and it bears significant resemblance to SARS, MERS and other previous respiratory pathogens [12]. The coronavirus disease caused by SARS-CoV-2 is now referred to as COVID-19. Originally from Wuhan, COVID-19 quickly spread out to mainland China and many other Asian countries and turned into a global pandemic later. As of 20th January 2020, there were 282 confirmed cases reported from China, Thailand, Japan and Republic of Korea [12]. As of 7th April 2021, more than 133 million people were infected, leading to more than 2.9 million deaths all over the world [2]. These high numbers of infections are compatible with the epidemic doubling time about 2 to 3 days found by [8], [10], and [14] at the end of January 2020.

A better understanding of the characteristics of COVID-19, including its transmission, symptoms and incubation period, is urgently needed for scientists and policy makers to monitor and control the pandemic as well as to set up proper quarantine procedures. The incubation period of COVID-19 is defined as the duration between infection of SARS-CoV-2 and symptom onset. Since the outbreak of COVID-19, there are numerous studies of its incubation period. The mean incubation period estimated by Linton et al. [6] was 5.2 days (95% CI: 4.1, 7.0). As an initial investigation of infections in the earliest phrase, this study had several limitations including that data collected from various sources are not uniformly distributed and that the variance is likely to be biased due to the limited sample size. In addition, as noted by Zhao et al. [15], the incubation period estimated by Backer et al. [1] using a log-normal distribution and by Lauer et al. [3] using three commonly used incubation period distributions (Gamma, Weibull and Erlang) were biased as well since these anaylses did not start from a generative model and could not correctly adjust for sample selection bias in their statistical inferences. Compared to earlier studies, Zhao et al. [15] collected a more reliable study sample and accounted for sample selection in their likelihood. By using a nonparametric Bayesian analysis, they concluded that about 5% of COVID-19 patients develop the symptoms after 14 days after contracting the virus. Although the incubation period of COVID-19 has been studied by many authors, to the best of our knowledge, there is no existing work assessing its risk factors. In our study, we consider regression analysis of the incubation period of COVID-19 under the Cox proportional hazards model. Our findings could provide additional insights on shortening or lengthening the standard quarantine duration of 14 days for certain subgroups to help better control the pandemic.

We consider the dataset collected by Zhao et al. [15] that consists of 463 Wuhan-exported cases (from 14 locations in and outside China) who left Wuhan before the travel ban on January 23, 2020. The dataset includes three key dates: beginning stay in Wuhan (B), ending stay in Wuhan (E), and date of symptom onset (S). It also includes age (median: 46 and IQR: [33.25, 56]; see Figure 1 for histogram) and gender (213 female and 250 male). As noted by Zhao et al. [15], the biases from under-ascertainment and non-random sample selection are minimized in this dataset. In this study, we are interested in assessing the association of age and gender with the incubation period of COVID-19 based on this dataset. Note that the infection time is not available and only observed to fall between the beginning and ending stay in Wuhan, and the symptom onset time is subject to right-censoring (6 cases had not shown symptoms at diagnosis and thus were censored at the time of diagnosis). The existing statistical methods cannot handle these issues. Therefore, we develop three methods, which will be described below, to analyze the data.

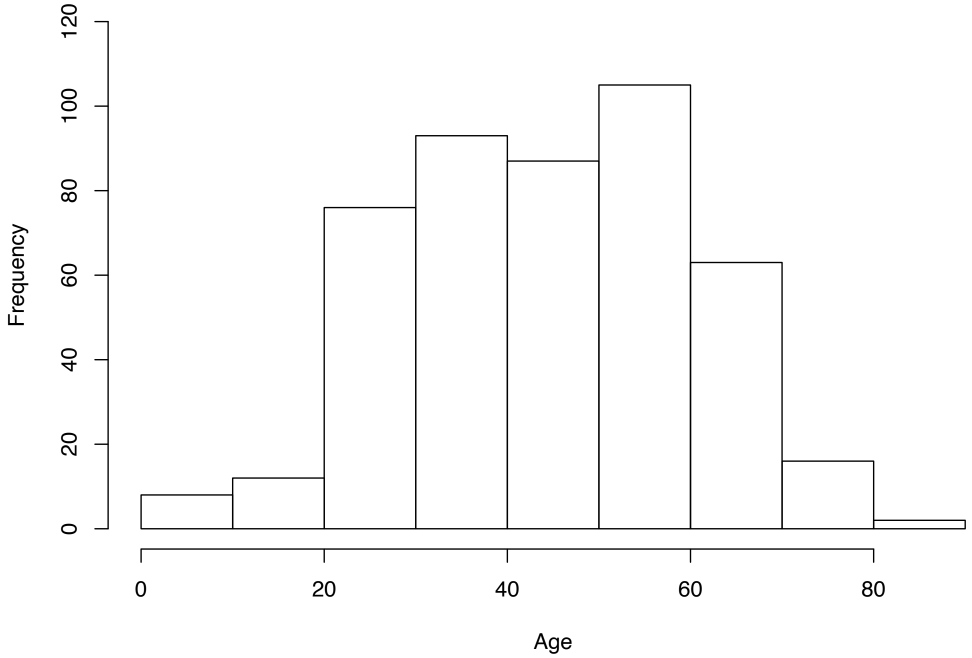

Figure 1: Histogram of Age

The remainder of the paper is organized as follows. In Section II, we develop three methods for regression analysis of the incubation period of COVID-19 under the Cox proportional hazards model in order to account for the unobserved and truncated infection time and the censored symptom onset time. In Section III, we analyze the dataset described above using three proposed methods to evaluate the association of age and gender with the incubation period of COVID-19. Section IV includes some discussion.

II. Methods

Our dataset consists of $n = 463$ Wuhan-exported cases who left Wuhan before the travel ban on 23th January 2020. It includes the date of beginning stay in Wuhan, the date of ending stay in Wuhan, and the date of symptom onset, as well as age and gender. The infection time is only observed to fall between the beginning and ending stay in Wuhan, while the symptom onset time is subject to right-censoring since six subjects had not shown symptoms at diagnosis and were considered as right-censored at the time of diagnosis. Thus, our observed data are given by

$$\big\{B_i,\, E_i,\, S_i = \min\{\tilde{S}_i, C_i\},\, \Delta_i = I(\tilde{S}_i \le C_i),\, X_i\big\}, \quad \text{for } \ i = 1, 2, ..., n,$$

where B_i is the time of beginning stay in Wuhan, E_i is the time of ending stay in Wuhan, $\tilde{S}_i$ is the true symptom onset time, C_i is the censoring time, and X_i is the covariate (age or gender) for subject i. Here the time origin is chosen as December 1, 2019 as in other studies of COVID-19.

Method 1: Midpoint imputation for infection time

First, we consider a naive method by imputing the infection time using the midpoint of the duration staying in Wuhan. Specifically, we first calculate the midpoint time (denoted by M_i) between the begin time staying in Wuhan B_i and the end time staying in Wuhan E_i as follows:

$$M_i = \frac{B_i + E_i}{2}, \quad \text{for } \ i = 1, 2, ..., n.$$

We then estimate the incubation period (denoted by T_i) as the difference between the symptom onset time S_i and the midpoint time M_i:

$$T_i = S_i - M_i, \quad \text{for } \ i = 1, 2, ..., n.$$

Note that 4 subjects are removed due to negative estimates of the incubation period.

We employ the Cox proportional hazards model which is commonly used for the analysis of failure time data. Under this model, the conditional hazard function of the incubation time T_i given the covariate X_i has the form

$$\lambda(t|X_i) = \lambda_0(t)e^{\beta X_i}, \tag{1}$$

where $\lambda_0(t)$ is the unspecified baseline hazard function and β is the unknown regression coefficient.

The Cox model has the following desirable features: (i) it is flexible by leaving the baseline hazard function $\lambda_0(t)$ completely unspecified; (ii) it has good interpretability since the regression coefficient β can be interpreted as the log hazard ratio, i.e.,

$$\beta = \log\left\{\frac{\lambda(t|X_i = x+1)}{\lambda(t|X_i = x)}\right\}; \tag{2}$$

if $\beta > 0$, then higher values of the covariate are associated with higher risk of failure; if $\beta < 0$, then higher values of the covariate are associated with lower risk of failure; (iii) it is mathematically tractable and its statistical properties can be established using the elegant Martingale theory; (iv) it is widely applicable in practice, particularly in biomedical studies; (v) it can be implemented easily using the statistical software.

Besides, the nonparametric Kaplan-Meier method is used to estimate the survival function $S(t)$ of the incubation time T_i, i.e., the probability of not having symptoms at time t since infection. In particular, the Kaplan-Meier estimate is given by

$$\widehat{S}(t) = \prod_{i:\, t_i < t}\left(1 - \frac{d_i}{n_i}\right), \tag{3}$$

where t_i's are the observed incubation times, d_i is the number of subjects who have symptom onset at time t_i, and n_i is the number of subjects at risk at time t_i, for $i = 1, \ldots, n$.

Method 2: Treat incubation time as interval-censored

Another simple method is to treat the incubation time T_i as interval-censored, since T_i is not exactly observed but known only to fall within an interval $(L_i, U_i]$, where

$$L_i = S_i - E_i \quad \text{and} \quad U_i = S_i - B_i,$$

for $i = 1, \dots, n$. If T_i is right-censored, then we let $U_i = \infty$. We then perform regression analysis under the Cox model with interval-censored data given above to evaluate the risk factors associated with the incubation time. In particular, we employ the nonparametric maximum likelihood method proposed by [13] for the analysis.

Method 3: Multiple imputations for infection time

We now consider a more sophisticated method for handling the unobserved infection time. We propose a multiple imputation procedure [9] as follows:

1. Estimate the distribution of the infection time based on the observed data $\big\{[B_i, E_i] : i = 1, \dots, n\big\}$;
2. Generate or impute M ($M = 100$) infection times according to the above estimated distribution restricted to $[B_i, E_i]$ for subject i, for $i = 1, \dots, n$;
3. Calculate the incubation time T_{im} for subject i as $S_i - W_{im}$, where S_i is the observed symptom onset time and W_{im} is the imputed infection time, for $m = 1, \dots, M$ and $i = 1, \dots, n$;
4. For the mth imputed data set, perform regression analysis of the incubation time under the Cox model and obtain the estimate of the regression coefficient $\widehat{\beta}_m$ and its estimated variance $\widehat{\sigma}^2_m$ as well as the Breslow estimate $\widehat{\Lambda}_{0m}(t)$ of the cumulative baseline hazard function [5],

$$\widehat{\Lambda}_{0m}(t) = \sum_{i=1}^{n} \frac{I(T_i \le t)\Delta_i}{\sum_{j \in \mathcal{R}_i} e^{\widehat{\beta}_m X_j}}, \quad \text{for } m = 1, \dots, M;$$

5. Calculate the final regression coeffcient estimate $\widehat{\beta}$ and its estimated variance $\widehat{\sigma}^2$ as [4]

$$\widehat{\beta} = \frac{1}{M} \sum_{m=1}^{M} \widehat{\beta}_m \tag{4}$$

   and

$$\widehat{\sigma}^2 = \left(1 + \frac{1}{M}\right) \frac{\sum_{m=1}^{M} (\widehat{\beta}_m - \widehat{\beta})^2}{M-1} + \frac{1}{M} \sum_{m=1}^{M} \widehat{\sigma}^2_m; \tag{5}$$

6. Estimate the survival function of the incubation time for subject i as

$$\widehat{S}(t|X_i) = \exp\big\{ - \widehat{\Lambda}_0(t) e^{\widehat{\beta} X_i} \big\}, \tag{6}$$

   where

$$\widehat{\Lambda}_0(t) = \frac{1}{M} \sum_{m=1}^{M} \widehat{\Lambda}_{0m}(t).$$

In the following, we consider three methods for estimating the distribution of the infection time in Step #1 of the above multiple imputation procedure.

Method 3.1: Multiple imputations for infection time based on NPMLE

We first calculate the nonparametric maximum likelihood estimate (NPMLE) of the distribution of infection time based on interval-censored data $\big\{[B_i, E_i] : i = 1, \dots, n\big\}$ using the self-consistency algorithm [11]. The NPMLE has positive mass on the maximal intersections $\big\{[s_j, t_j] : j = 1, \dots, d\big\}$, where s_j's are from $\{B_i : i = 1, \dots, n\}$ and t_j's are from $\{E_i : i = 1, \dots, n\}$ such that $[s_j, t_j] \cap [B_i, E_i]$ is either $[s_j, t_j]$ or an empty set for every $j = 1, \dots, d$ and $i = 1, \dots, n$. We then generate the infection time for subject i as follows: (i) look for the maximal intersections included in $[B_i, E_i]$; (ii) reweight the masses on those maximal intersections so that they sum up to 1; (iii) randomly select one of the maximal intersections according to their probability masses; (iv) generate the infection time from the uniform distribution on the selected maximal intersection.

Method 3.2: Multiple imputations for infection time based on uniform distribution

We assume that the infection time for subject i is uniformly distributed on $[B_i, E_i]$, for $i = 1, \ldots, n$.

Method 3.3: Multiple imputations for infection time based on exponential epidemic growth

Following [15], we assume that the probability of contracting the virus in Wuhan was increasing exponentially before the quarantine, that is, the probability density function (pdf) of the infection time on $[B, E]$ has the form of

$$f(t) = \frac{re^{rt}}{e^{rE} - e^{rB}} \tag{7}$$

and the cumulative distribution function (cdf) is given by

$$F(t) = \int_B^t \frac{re^{rs}}{e^{rE} - e^{rB}} ds$$

for $t \in [B, E]$, where r is the growth exponent. We estimate r using $\log(2)/2.1$ in the data analysis below, where 2.1 is the doubling time estimated by [15].

We employ the probability integral transform to generate the infection time from $F(t)$ as follows: (i) generate a random number u from the uniform distribution on $[0, 1]$; (ii) generate the infection time W as

$$W = F^{-1}(u) = \frac{\log\left[(e^{rE} - e^{rB})u + e^{rB}\right]}{r}.$$

III. Results

We analyze our dataset of 463 Wuhan-exported cases described above using three proposed methods and evaluate the association of age and gender with the incubation period of COVID-19. For gender, we define an indicator variable with value 1 for male and 0 for female. We also consider an indicator of age group defined as 1 if age > 46 (old) and 0 if age ≤ 46 (young), where 46 is the median age. For each variable, we report the hazard ratio and its 95% confidence interval under the Cox model (1). The hazard ratio is estimated by $e^{\hat{\beta}}$ according to (2) and its standard error is calculated as $e^{\hat{\beta}}\hat{\sigma}$ by the Delta method, where $\hat{\sigma}$ is the standard error of $\hat{\beta}$. The confidence interval is obtained based on the normal approximation to the distribution of $\hat{\beta}$. We also report the p-value for testing $H_0 : \beta = 0$ vs $H_a : \beta \neq 0$. Furthermore, we plot the estimates of the survival function $S(t)$, i.e., the probability of not having symptoms at time t since infection, based on gender and age group, respectively.

Method 1 - Results

The estimation results for hazard ratio using Method 1 are given in Table 1. These results are obtained by using the function "coxph" in the package "survival" in R that implements the partial likelihood method for fitting the Cox model (1) to right-censored data. The Kaplan-Meier estimates of the survival function given by (3) are calculated using the function "survfit" in the package "survival" and are plotted in Figures 2 and 3 for gender and age group, respectively. According to the results in Table 1, older people tend to have higher risk of symptom onset and shorter incubation period, while gender is not significantly associated with the incubation time.

Table 1: Estimation results for hazard ratio based on Method 1

Variables	p-value	Hazard Ratio (HR)	95% CI of HR
Gender (Male)	0.850	1.019	(0.837, 1.241)
Age group (Old)	0.008	1.311	(1.075, 1.598)
Age	0.004	1.009	(1.003, 1.016)

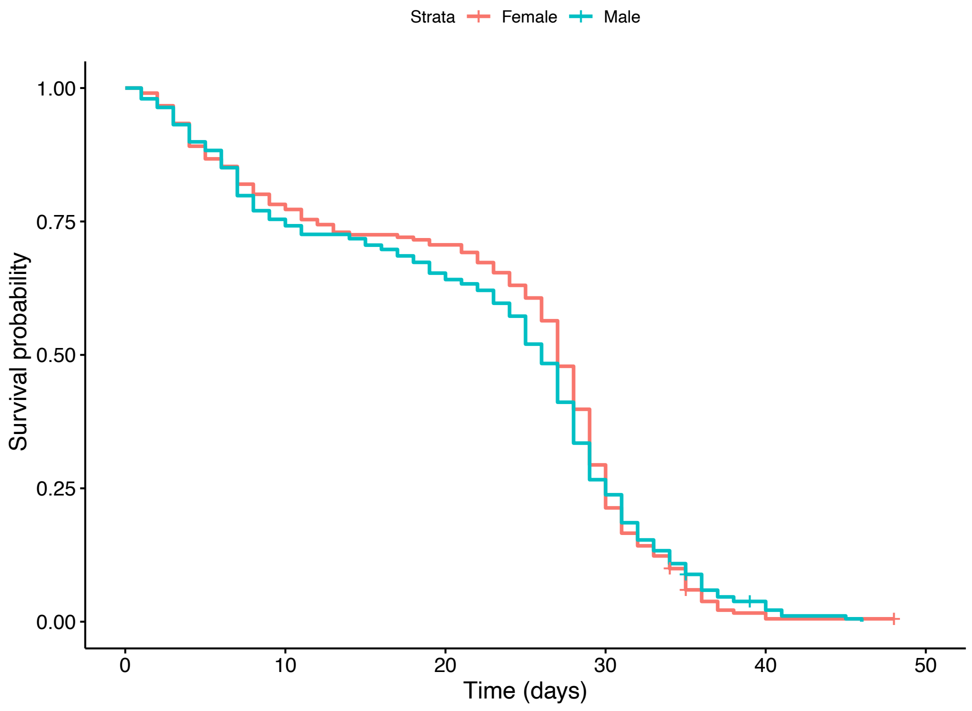

Figure 2: Estimated survival function for incubation period based on gender using Method 1

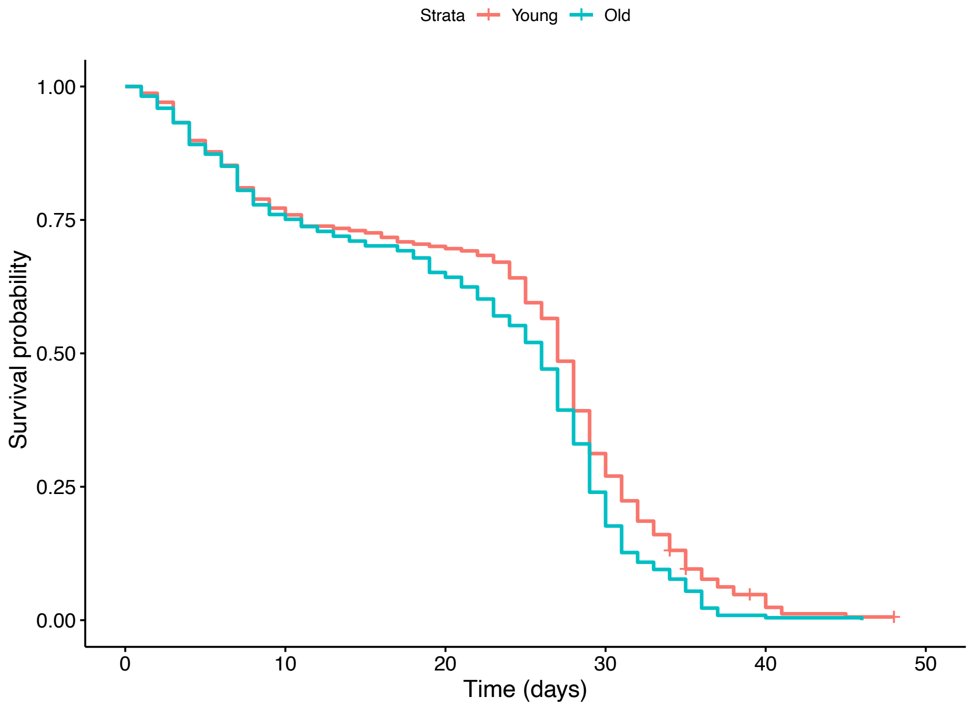

Figure 3: Estimated survival function for incubation period based on age group using Method 1

Method 2 - Results

The estimation results for hazard ratio using Method 2 are presented in Table 2. These results are obtained by using the function “unireg” in the package “IntCens” in R that implements nonparametric maximum likelihood estimation for a class of semiparametric regression models, including the Cox model (1), with interval-censored data [13]. The estimate of the survival function is obtained from the nonparametric maximum likelihood estimate of the cumulative baseline hazard function given by [13]. Figures 4 and 5 plot the estimates of the survival function based on gender and age group, respectively. The results in Table 2 suggest that neither gender nor age is significantly associated with the incubation time.

Table 2: Estimation results for hazard ratio based on Method 2

Variables	p-value	Hazard Ratio (HR)	95% CI of HR
Gender (Male)	0.899	0.975	(0.585, 1.365)
Age group (Old)	0.278	1.240	(0.758, 1.722)
Age	0.191	1.009	(0.996, 1.021)

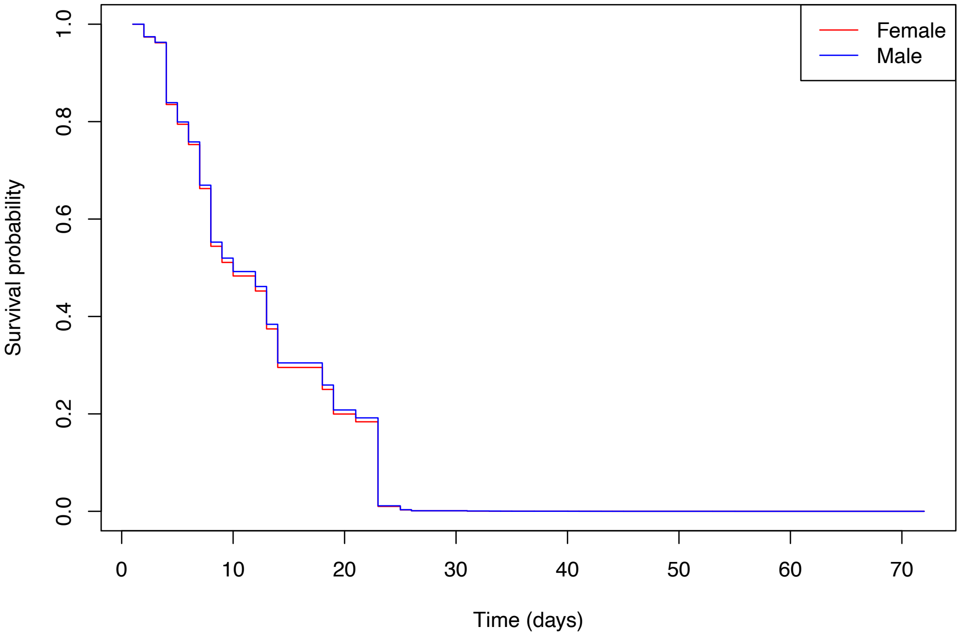

Figure 4: Estimated survival function for incubation period based on gender using Method 2

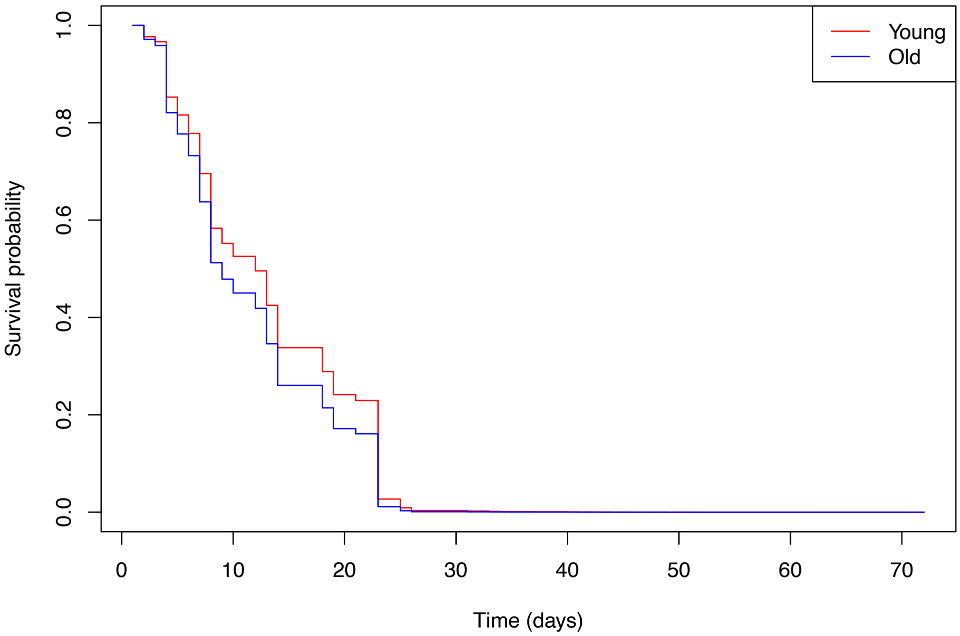

Figure 5: Estimated survival function for incubation period based on age group using Method 2

Method 3 - Results

We now present the estimation results for hazard ratio and survival function obtained by using three multiple imputation methods. The hazard ratio is estimated by $e^{\hat{\beta}}$ based on (2) and (4). The standard error is calculated as $e^{\hat{\beta}}\hat{\sigma}$ by the Delta method, where $\hat{\sigma}$ is given by (5). The confidence interval is obtained based on the normal approximation to the distribution of $\hat{\beta}$. The survival function is estimated by (6). These methods are programmed by the authors using R.

Method 3.1

The estimation results for hazard ratio using Method 3.1 are given in Table 3 and the estimates of the survival function based on gender and age group are plotted in Figures 6 and 7, respectively. The results in Table 3 imply that neither gender nor age is significantly associated with the incubation time.

Table 3: Estimation results for hazard ratio based on Method 3.1

Variables	p-value	Hazard Ratio (HR)	95% CI of HR
Gender (Male)	0.979	0.997	(0.762, 1.232)
Age group (Old)	0.149	1.184	(0.912, 1.456)
Age	0.128	1.006	(0.999, 1.013)

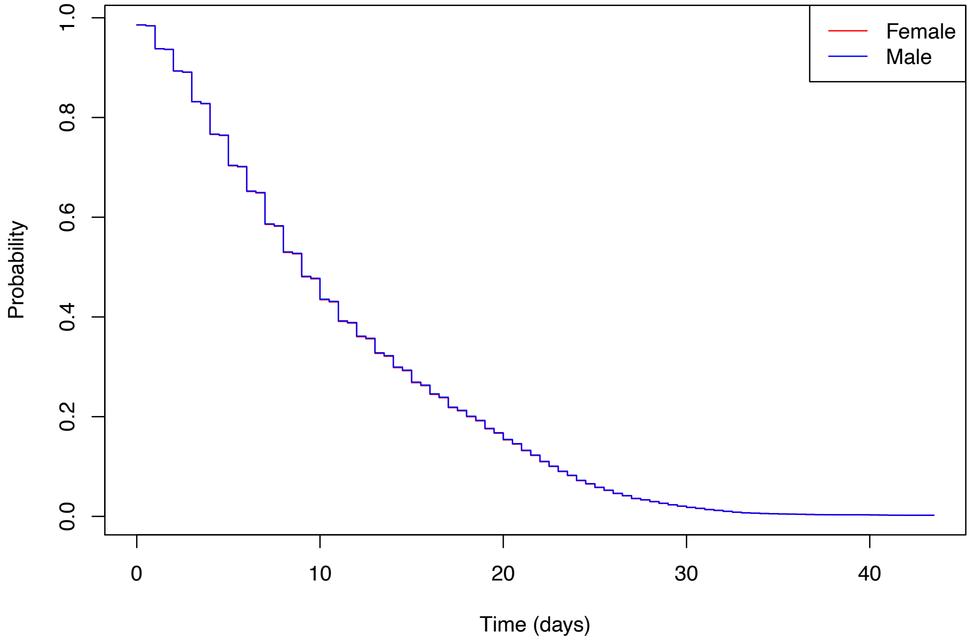

Figure 6: Estimated survival function for incubation period based on gender using Method 3.1

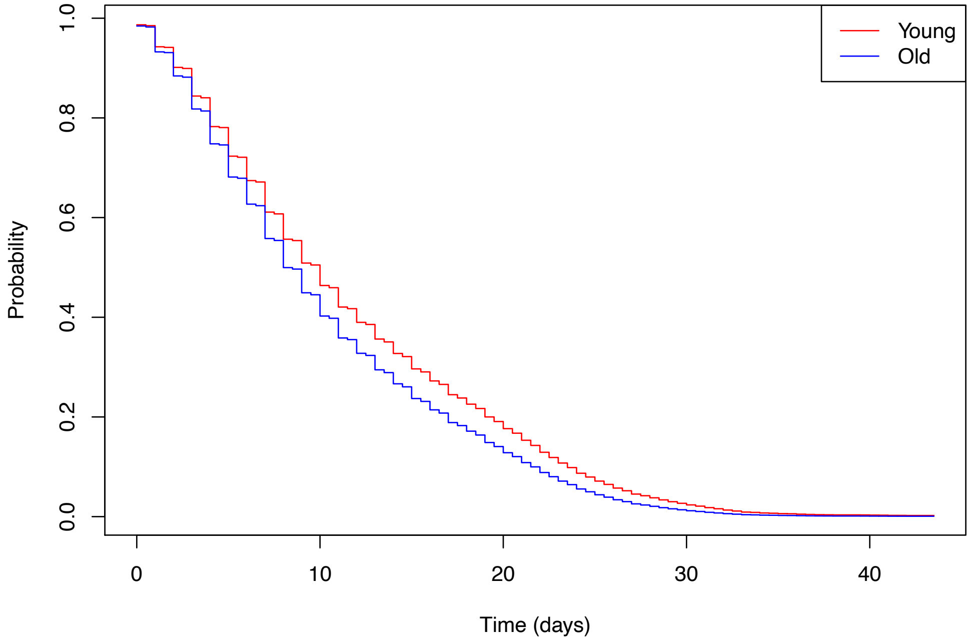

Figure 7: Estimated survival function for incubation period based on age group using Method 3.1

Method 3.2

The estimation results for hazard ratio using Method 3.2 are given in Table 4 and the estimates of the survival function based on gender and age group are plotted in Figures 8 and 9, respectively. The results in Table 4 suggest that neither gender nor age is significantly associated with the incubation time.

Table 4: Estimation results for hazard ratio based on Method 3.2

Variables	p-value	Hazard Ratio (HR)	95% CI of HR
Gender (Male)	0.829	1.024	(0.799, 1.249)
Age group (Old)	0.302	1.126	(0.872, 1.381)
Age	0.312	1.004	(0.996, 1.011)

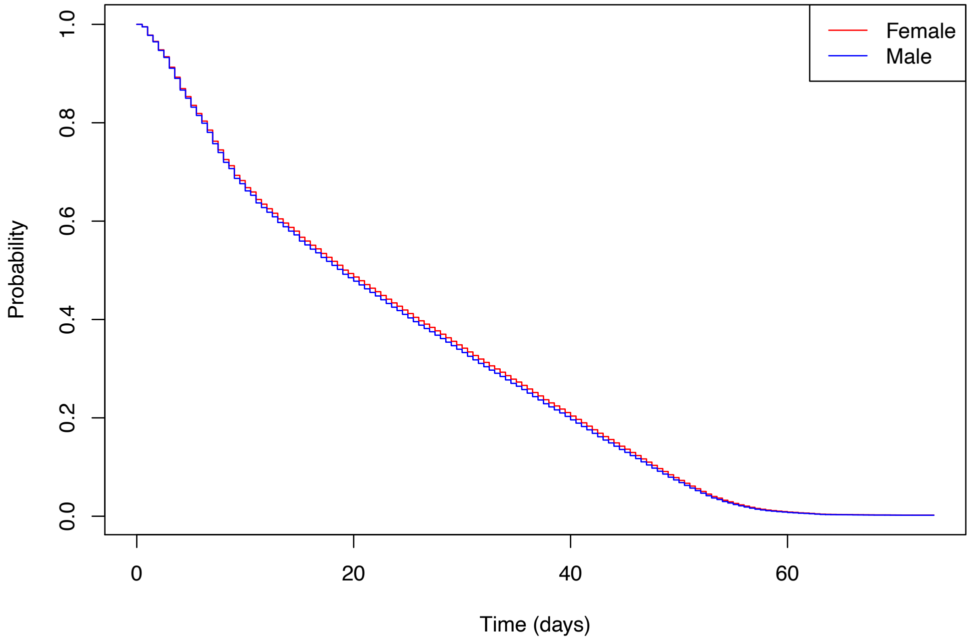

Figure 8: Estimated survival function for incubation period based on gender using Method 3.2

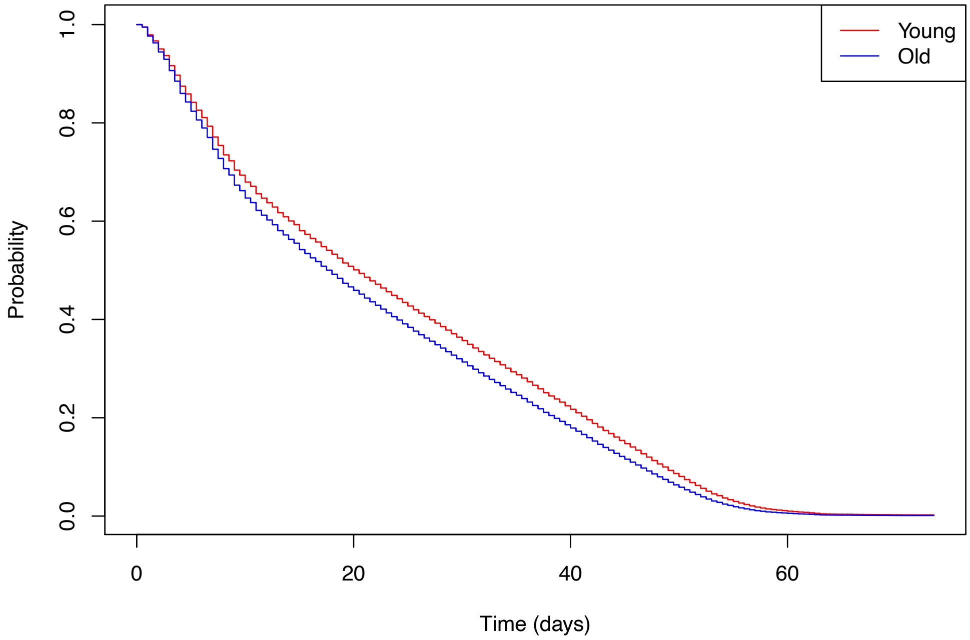

Figure 9: Estimated survival function for incubation period based on age group using Method 3.2

Method 3.3

The estimation results for hazard ratio using Method 3.3 are given in Table 5 and the estimates of the survival function based on gender and age group are plotted in Figures 10 and 11, respectively. According to the results in Table 5, older people tend to have higher risk of symptom onset and shorter incubation period, while gender is not significantly associated with the incubation time.

Table 5: Estimation results for hazard ratio based on Method 3.3

Variables	p-value	Hazard Ratio (HR)	95% CI of HR
Gender (Male)	0.433	0.92	(0.723, 1.111)
Age group (Old)	0.027	1.267	(1.002, 1.532)
Age	0.036	1.007	(1.001, 1.014)

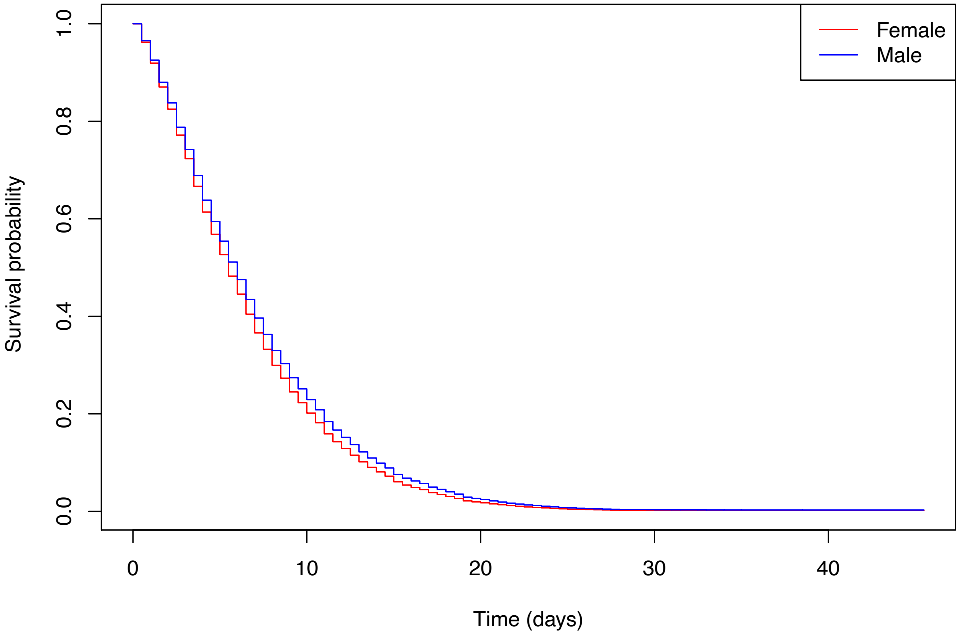

Figure 10: Estimated survival function for incubation period based on gender using Method 3.3

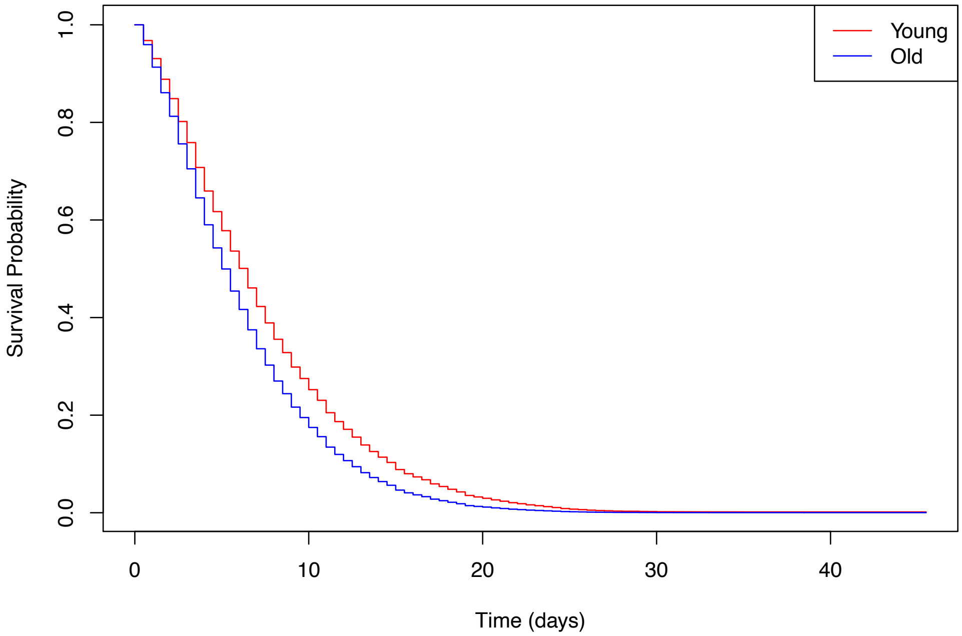

Figure 11: Estimated survival function for incubation period based on age group using Method 3.3

IV. Discussion

In this study, we are interested in assessing the association of age and gender with the incubation period of COVID-19 based on a dataset of 463 Wuhan-exported cases. In this dataset, the infection time is not available and known only to fall within an interval, while the symptom onset time is subject to right-censoring. To evaluate the association of age and gender with the incubation period, we consider regression analysis of the incubation period under the Cox model with age and gender as covariates. We develop three methods for the analysis to account for our data structure. We include some discussion about these methods below.

For the method of midpoint imputation, we approximate the unobserved infection time using the midpoint between the beginning and ending time of stay in Wuhan. The results of this method presented in Table 1 show that the incubation period has no significant difference between male and female, while younger people seem to have lower risk of symptom onset and longer incubation period. From Figures 2 and 3, one can see that the survival probabilities of the incubation time, i.e., the probability of not having symptom onset at a certain time since infection, are highly overestimated. Particularly, the estimated survival probability at 14 days is about 0.73 and the incubation period is likely to reach up to 40 days. This is quite different from the results of existing studies in the literature [3], [6], [15]. The estimation bias of this method is expected due to naive approximation of the infection time.

In an effort to correct for the estimation bias, we consider a simple method based on interval-censored data of the incubation period. This method indicates that neither gender nor age is significantly associated with the incubation period (Table 2). One can see from Figures 4 and 5 that the survival probabilities of the incubation time are still overestimated but much more reasonable than those given by the the midpoint imputation method. In particular, the estimated survival probability at 14 days is about 0.25. Although this method is better than the midpoint imputation method in terms of accounting for our data structure, there are some other features of the dataset that cannot be accounted for by this method. For example, there is some truncation issue with the infection time, because the dataset only includes Wuhan-exported cases who left Wuhan before the travel ban on January 23, 2020. However, this interval-censoring method assumes that the dataset consists of a simple random sample and that all subjects were followed since time zero (December 1, 2019). Thus, it does not address the truncation problem. In addition, this method is not efficient because it does not fully utilize the information about the infection time.

To better account for the data structure and make use of the available information, we develop a method of multiple imputations by imputing the unobserved infection time multiple times according to an estimated distribution of the infection time. This method is intended to reduce the estimation bias and address the truncation problem. It also has the flexibility of using different methods to estimate the distribution of infection time for imputation. Specifically, we consider three methods for estimating the infection time distribution. First, we employ the nonparametric maximum likelihood estimation (NPMLE) based on interval-censored data of the infection time. Second, we assume that the infection time is uniformly distributed, i.e., the risk of being infected is constant during the stay in Wuhan. Lastly, we assume that the epidemic in Wuhan grows exponentially before the travel ban and model the infection time as in (7). The first method suggests that neither age nor gender is significantly associated with the incubation period (Table 3), and the estimated survival probability at 14 days is about 0.25 (Figures 6 and 7). The results are similar to those of the interval-censoring method, because they both treat the data as interval-censored and ignore the truncation problem. The second method implies that neither age nor gender is significant (Table 4), and the estimated survival probability at 14 days is about 0.75 (Figures 8 and 9), which is highly overestimated as with the midpoint imputation method. This overestimation is expected because the uniform distribution does not properly model the epidemic growth and it tends to underestimate the infection time and thereby overestimate the incubation time. The last method not only accounts for the truncation problem of the infection time, but also properly model the exponential growth of the early-stage epidemic in Wuhan before the travel ban. This method suggests that gender is not significantly associated with the incubation period, while younger people tend to have lower risk of symptom onset and longer incubation period (Table 5). From Figures 10 and 11, the estimated survival probability at 14 days is about 0.1 for both male and female, while it is about 0.15 for young people (≤ 46 years old) and 0.1 for old people (> 46 years old). This is consistent with the findings of existing work in the literature [15].

References

[1] Backer, J.A., Klinkenberg, D. and Wallinga, J. 2020. Incubation period of 2019 novel coronavirus (2019-nCoV) infections among travellers from wuhan, china, 20–28 january 2020. *Eurosurveillance.* 25, 5 (2020), 2000062.

[2] COVID-19 dashboard by the center for systems science and engineering (csse) at johns hopkins university: *https://coronavirus.jhu.edu/map.html.*

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

[4] Li, S., Sun, J., Tian, T. and Cui, X. 2020. Semiparametric regression analysis of doubly censored failure time data from cohort studies. *Lifetime Data Analysis.* 26, 2 (2020), 315–338.

[5] Lin, D. 2007. On the breslow estimator. *Lifetime Data Analysis.* 13, 4 (2007), 471–480.

[6] Linton, N.M., Kobayashi, T., Yang, Y., Hayashi, K., Akhmetzhanov, A.R., Jung, S.-m., Yuan, B., Kinoshita, R. and Nishiura, H. 2020. Incubation period and other epidemiological characteristics of 2019 novel coronavirus infections with right truncation: A statistical analysis of publicly available case data. *Journal of Clinical Medicine.* 9, 2 (2020), 538.

[7] Novel coronavirus (2019-nCoV) situation report - 1, 21 january 2020: *https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200121-sitrep-1-2019-ncov.pdf.*

[8] Read, J.M., Bridgen, J.R., Cummings, D.A., Ho, A. and Jewell, C.P. 2020. Novel coronavirus 2019-nCoV: Early estimation of epidemiological parameters and epidemic predictions. *MedRxiv.* (2020).

[9] Rubin, D.B. 2004. *Multiple imputation for nonresponse in surveys.* John Wiley & Sons.

[10] Sanche, S., Lin, Y.T., Xu, C., Romero-Severson, E., Hengartner, N. and Ke, R. 2020. High contagiousness and rapid spread of severe acute respiratory syndrome coronavirus 2. *Emerging Infectious Diseases.* 26, 7 (2020), 1470–1477.

[11] Turnbull, B.W. 1976. The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society: Series B (Methodological).* 38, 3 (1976), 290–295.

[12] WHO statement regarding cluster of pneumonia cases in wuhan, china: *https://www.who.int/china/news/detail/09-01-2020-who-statement-regarding-cluster-of-pneumonia-cases-in-wuhan-china.*

[13] Zeng, D., Mao, L. and Lin, D. 2016. Maximum likelihood estimation for semiparametric transformation models with interval-censored data. *Biometrika.* 103, 2 (2016), 253–271.

[14] Zhao, Q., Chen, Y. and Small, D.S. 2020. Analysis of the epidemic growth of the early 2019-nCoV outbreak using internationally confirmed cases. *MedRxiv.* (2020).

[15] Zhao, Q., Ju, N., Bacallado, S. and Shah, R.D. 2021. BETS: The dangers of selection bias in early analyses of the coronavirus disease (covid-19) pandemic. *The Annals of Applied Statistics.* 15, 1 (2021), 363–390.