medRxiv preprint doi: https://doi.org/10.1101/2020.04.11.20062190; this version posted April 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Estimating the real-time case fatality rate of COVID-19 using Poisson mixtures model

Running head: Case fatality rate of COVID-19

Paul H. Lee, PhD

School of Nursing, Hong Kong Polytechnic University

Correspondence to: Dr Paul H. Lee (email: paul.h.lee@polyu.edu.hk, phone: +852-3400 8275, fax: +852-2364 9663), School of Nursing, GH527, Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong

Manuscript word count: 1,321

Keywords: Bootstrap; computational statistics; emerging infectious disease; parameter estimation

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2020.04.11.20062190; this version posted April 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

ABSTRACT

We proposed using Poisson mixtures model that utilized data of deaths, recoveries, and total confirmed cases in each day since the outbreak. We demonstrated that our CFR estimates for Hubei Province and other parts of China were superior to the simple CFR estimators in the early stage of COVID-19 outbreak.

medRxiv preprint doi: https://doi.org/10.1101/2020.04.11.20062190; this version posted April 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

INTRODUCTION

The case fatality rate (CFR) of a disease refers to the proportion of death within those infected from the disease. The CFR of an emerging infectious disease can only be determined after its outbreak, but the real-time CFR estimation during the disease outbreak, especially in the early stage, has to be made for public health decisions. There are two naive, simple methods to estimate CFR at a time point t using the cumulative number of deaths, recoveries, and confirmed cases at t. Denote $D(t)$, $R(t)$, and $C(t)$ be the observed number of deaths, recoveries, and confirmed cases at day t. The first simple estimator equals $D(t)/C(t)$ and the second simple estimator equals $D(t)/[D(t)+R(t)]$. Note that these two estimators are equivalent after the outbreak when $D(t)+R(t)=C(t)$ [1].

The second estimator usually outperforms the first estimator as there existed a time lag between the diagnosis of disease and death [2-5]. The second estimator at any stage of the outbreak would be unbiased if the proportion of confirmed deaths equals the proportion of unconfirmed deaths [1]. This is the case for severe acute respiratory syndrome (SARS) in 2003 where the time from confirmed diagnosis to death equaled that from confirmed diagnosis to recovery (=23 days [1]). For novel coronavirus disease 2019 (COVID-19), the duration from diagnosis to death was much shorter, with a mean duration of 8-14 days [2-5].

With survivors being hospitalized longer than the deaths, we can expect that the second CFR estimator (hereby referred to as the simple CFR estimator) would overestimate the actual CFR as deaths were more likely to be observed earlier in the outbreak. During the COVID-19 outbreak, many researchers have developed new methods to estimate CFR that replace the nominator or denominator by the relevant figures 8 to 14 days ago to reduce the biasedness of the simple CFR estimator [2-5]. However, this approach ignored the underlying process of how these numbers (deaths, recoveries, and total confirmed cases) were generated and these methods were regarded as

medRxiv preprint doi: https://doi.org/10.1101/2020.04.11.20062190; this version posted April 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

invalid [6].

In fact, a statistical modeling approach was developed 15 years ago using the SARS data [1]. The rationale behind this approach is that deaths and recovery conditional at any time point should both follow a parametric distribution (gamma distribution was used to fit the SARS CFR data). The limitation of this approach lies in the estimation of the parameters of this distribution, which require the often-non-public data of time from confirmed diagnosis to death and time from confirmed diagnosis to recovery. Here, we suggest replacing the gamma distribution by Poisson distribution, as we can see below that the knowledge of time to death and time to recovery is not necessary. The proposed model requires data of deaths, recoveries, and total confirmed cases recorded in each day since the outbreak of a population. These data for COVID-19 are publicly-available [7].

Some researchers proposed that the denominator of CFR should also include both those who are diagnosed with the disease and those who are asymptomatic and never diagnosed. In this current study, the CFR for detected disease (also referred to as the symptomatic CFR, or sCFR) was modeled, as the actual number of asymptomatic cases was unknown and our proposed method could not be evaluated.

METHODS

We assumed that a confirmed case will either die or recover and it is determined *a priori* through a Bernoulli process with a parameter p, which is the CFR. If a case is deemed to die, the time to death (in days) follows a Poisson distribution with a mean of λ_d, and the time to recovery for a case who will survive follows a Poisson distribution with a mean of λ_r. We call this model the Poisson mixtures model.

medRxiv preprint doi: https://doi.org/10.1101/2020.04.11.20062190; this version posted April 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Under this model, the expected number of deaths at day $t+a$ resulting from the cases confirmed at a particular day t should be $\mathrm{CFR}\times\mathrm{C}(t)\times\mathrm{Poi}(a, \lambda_d)$, where $\mathrm{Poi}(x,\lambda)$ is the probability mass function of the Poisson distribution with the parameter λ. Similarly, the expected number of recoveries at day $t+a$ for the cases confirmed at a particular day t should be $(1\text{-}\mathrm{CFR})\times\mathrm{C}(t)\times\mathrm{Poi}(a, \lambda_r)$. Therefore, the expected number of deaths at a particular day t, $\mathrm{D}'(t)$, would be $\Sigma_{k=0}^{(t-1)} \mathrm{CFR}\times\mathrm{C}(t-k)\times\mathrm{Poi}(k, \lambda_d)$, and the expected number of deaths at a particular day t, $\mathrm{R}'(t)$, would be $\Sigma_{k=0}^{(t-1)} (1\text{-}\mathrm{CFR})\times\mathrm{C}(t-k)\times\mathrm{Poi}(k, \lambda_r)$. We estimated the three parameters of this model by minimizing the chi-square goodness-of-fit statistic, i.e., $\Sigma_{k=1}^{t} (\mathrm{D}'(k)-\mathrm{D}(k))^2/\mathrm{D}'(k) + (\mathrm{R}'(k)-\mathrm{R}(k))^2/\mathrm{R}'(k)$.

The 95% confidence interval of the CFR estimates can be estimated using bootstrap method. Both parametric and non-parametric bootstrap can be used, and here we used non-parametric bootstrap that simulated 1000 datasets with $\mathrm{D}(t)$, $\mathrm{R}(t)$, and $\mathrm{C}(t)$ follow multinomial distribution with the observed proportions as the parameters, and the 2.5th percentile and 97.5th percentile of the estimated CFR from these 1000 datasets were the 95% confidence interval of the corresponding CFR estimate.

The COVID-19 data in Hubei Province, China, and other parts of China up to 10th April 2020 were obtained from the Center for Systems Science and Engineering of John Hopkins University,[7] and all statistical analysis was performed in R 3.6.1. The syntax for our proposed method was available as supplementary material.

RESULTS

Figure 1 shows the simple CFR estimates and the estimates obtained using our proposed model by time, for Hubei Province and other parts of China, respectively. The 95% CIs of the CFR estimates were shaded in gray. Our proposed CRF estimates for Hubei Province became stable on 5th

medRxiv preprint doi: https://doi.org/10.1101/2020.04.11.20062190; this version posted April 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

February, and that for the other parts of China were all very close to the actual CFR on the first day of CFR estimation (25th January). The performance of the Poisson mixtures model for Hubei Province started to overestimate the CFR starting from mid-March when the daily confirmed cases dropped to single-digit onwards and the simple CFR estimator began to converge. Similarly, the performance of our model for other parts of China became inferior to the simple CFR estimator starting from late February when the daily confirmed cases dropped from 200-800 to less than 30.

DISCUSSION

Our CFR estimates for Hubei Province and other parts of China were superior to the simple CFR estimators. Making use of all real-time information available for COVID-19, including the number of deaths, recoveries, and total confirmed cases for every single day during the outbreak, our proposed CFR estimates for Hubei Province were acceptable even at the very early stage of the COVID-19 pandemic of mid-February (4.5% to 7%) at which the simple CFR estimators overestimated the actual CFR by more than five folds. Similarly, we achieved CFR estimates for other parts of China at a reasonable level of 1.5% to 1.6% with data up to late January at which the simple CFRs were in the range between 15% and 35%.

Similar to the gamma mixtures model, our proposed model that substituted gamma distribution with Poisson distribution, the CFR estimate was reasonable at the early stage of the disease outbreak, however, the estimate was positively biased in the later stage. We believed that this finding reflected that the actual CFR was changing over time at the later stage of the COVID-19 pandemic, and fitting the data with a constant CFR model may not be appropriate. To test this hypothesis, we can allow the Poisson mixtures model to have one CFR at a time range and another CFR in the other time range, and test the equivalence of these two CFRs. This test of time-varying CFR could also be used to examine the effectiveness of population-level interventions.

medRxiv preprint doi: https://doi.org/10.1101/2020.04.11.20062190; this version posted April 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

The Poisson mixtures model proposed here can be applied to countries in the early and middle stages of the COVID-19 outbreak, for example, the US, Spain, Italy, France, and Germany. However, note that the validity of our model relied on several assumptions, including the constant nature of CFR and the identification of deaths, recoveries, and confirmed cases. The constant CFR assumption may not be appropriate if effective treatments were invented within the disease outbreak.

Conflict of interest: none.

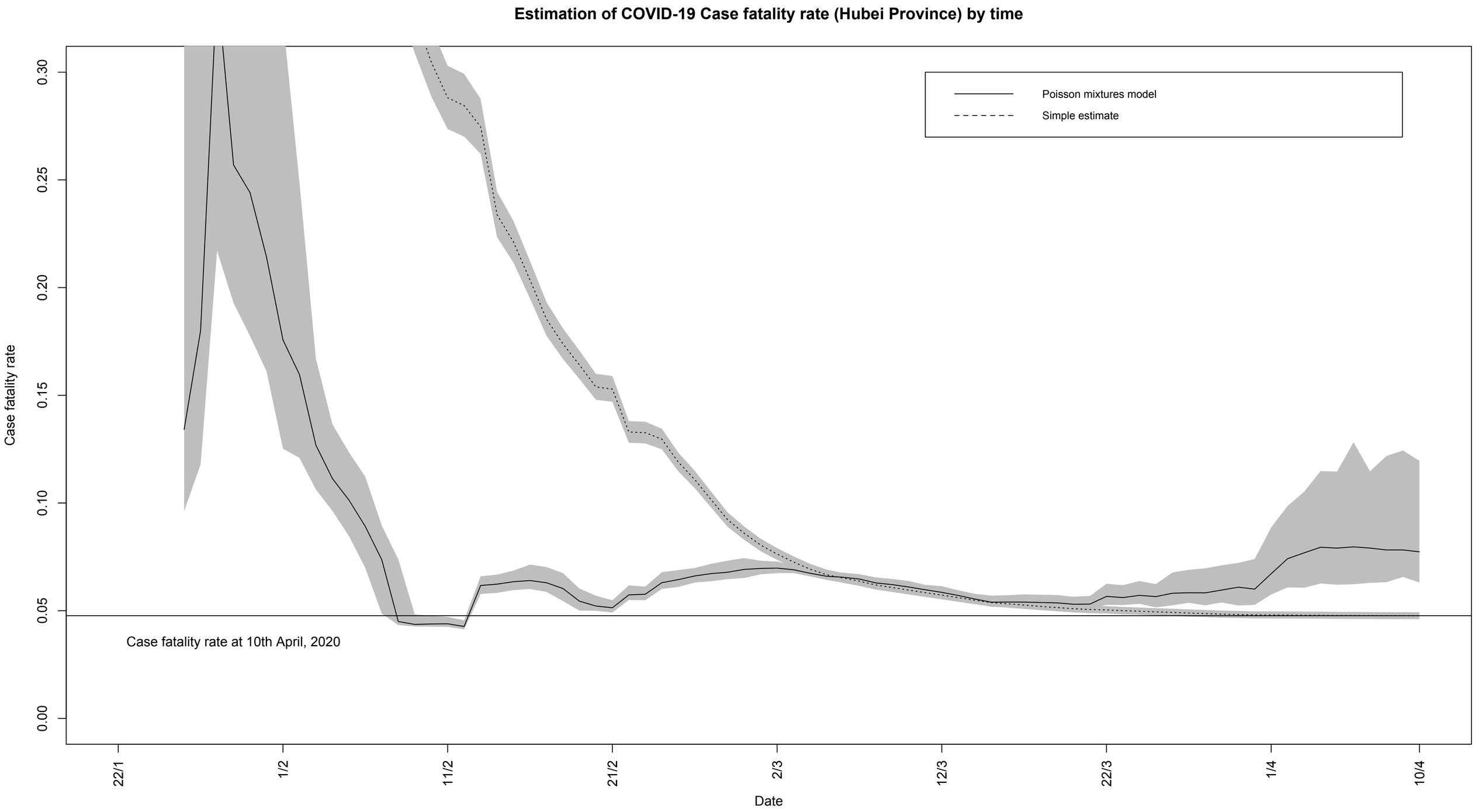
REFERENCES

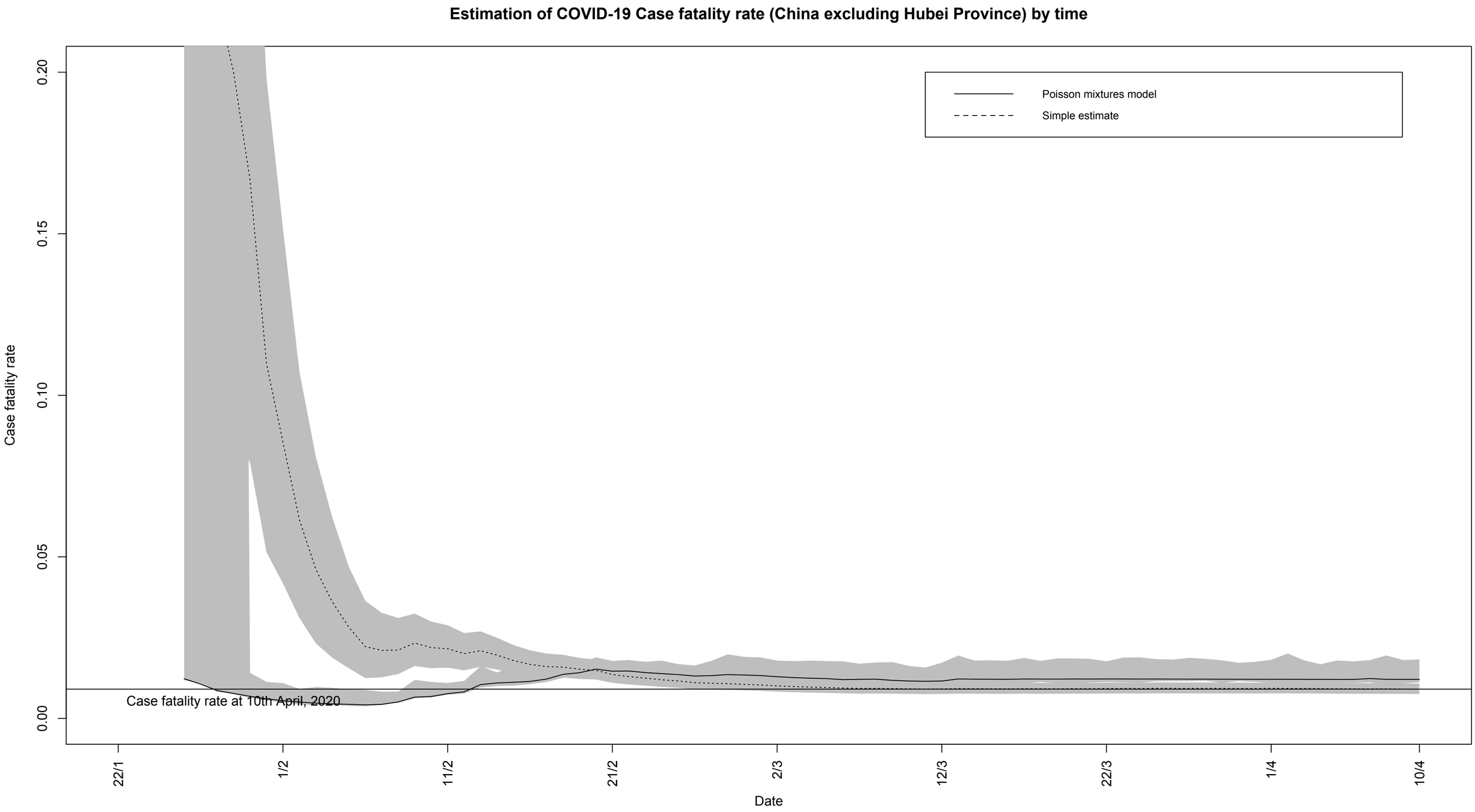
1. Ghani AC, Donnelly CA, Cox DR, et al. Methods for Estimating the Case Fatality Ratio for a Novel, Emerging Infectious Disease. Amer J Epidemiol **2005**; 162: 479-86.
2. Baud D, Qi X, Nielsen-Saines K, Musso D, Pomar L, Favre G. Real estimates of mortality following COVID-19 infection. Lancet Infect Dis **2020**: article in press.
3. Granozio FM. On the problem of comparing Covid-19 fatality rates. arXiv **2020**: arXiv:2004.03377v1.
4. Qi R, Ye C, Qin XR, Yu XJ. Case fatality rate of novel coronavirus disease 2019 in China. medRxiv **2020**: doi.org/10.1101/2020.02.26.20028076.
5. Wilson N, Kvalsvig A, Barnard LT, Baker MG. Case-fatality risk estimates for COVID-19 calculated by using a lag time for fatality. Emerg Infect Dis **2020**: article in press.
6. Lipsitch M. Estimating case fatality rates of COVID-19. Lancet Infect Dis **2020**; (article in press).
7. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis **2020**: article in press.

medRxiv preprint doi: https://doi.org/10.1101/2020.04.11.20062190; this version posted April 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

FIGURE CAPTION

Figure 1. (A) Estimation of COVID-19 Case fatality rate (Hubei Province) by time; (B) Estimation of COVID-19 Case fatality rate (China excluding Hubei Province) by time


Estimation of COVID-19 Case fatality rate (Hubei Province) by time
Poisson mixtures model
Simple estimate
Case fatality rate at 10th April, 2020
Case fatality rate
0.00
0.05
0.10
0.15
0.20
0.25
0.30
22/1
1/2
11/2
21/2
2/3
12/3
22/3
1/4
10/4
Date


Estimation of COVID-19 Case fatality rate (China excluding Hubei Province) by time
Poisson mixtures model
Simple estimate
Case fatality rate at 10th April, 2020
Case fatality rate
0.00
0.05
0.10
0.15
0.20
22/1
1/2
11/2
21/2
2/3
12/3
22/3
1/4
10/4
Date