

A sequential test to compare the real-time fatality rates of a disease among multiple groups with an application to COVID-19 data

Statistical Methods in Medical Research
 2022, Vol. 31(2) 348–360
 © The Author(s) 2021
 Article reuse guidelines:
sagepub.com/journals-permissions
 DOI: 10.1177/09622802211061927
journals.sagepub.com/home/smm



Yuanke Qu¹, Chun Yin Lee² , and KF Lam^{1,3} 

Abstract

Infectious diseases, such as the ongoing COVID-19 pandemic, pose a significant threat to public health globally. Fatality rate serves as a key indicator for the effectiveness of potential treatments or interventions. With limited time and understanding of novel emerging epidemics, comparisons of the fatality rates in real-time among different groups, say, divided by treatment, age, or area, have an important role to play in informing public health strategies. We propose a statistical test for the null hypothesis of equal real-time fatality rates across multiple groups during an ongoing epidemic. An elegant property of the proposed test statistic is that it converges to a Brownian motion under the null hypothesis, which allows one to develop a sequential testing approach for rejecting the null hypothesis at the earliest possible time when statistical evidence accumulates. This property is particularly important as scientists and clinicians are competing with time to identify possible treatments or effective interventions to combat the emerging epidemic. The method is widely applicable as it only requires the cumulative number of confirmed cases, deaths, and recoveries. A large-scale simulation study shows that the finite-sample performance of the proposed test is highly satisfactory. The proposed test is applied to compare the difference in disease severity among Wuhan, Hubei province (exclude Wuhan) and mainland China (exclude Hubei) from February to March 2020. The result suggests that the disease severity is potentially associated with the health care resource availability during the early phase of the COVID-19 pandemic in mainland China.

Keywords

Brownian motion, COVID-19, emerging infectious disease, fatality rates, sequential test

1 Introduction

The incidence of emerging infectious diseases has increased worldwide in recent decades and has posed one of the greatest threats to public health globally.¹ In particular, the ongoing coronavirus pandemic (COVID-19), first identified in Wuhan city of China in December 2019, is affecting 217 countries and territories across the world with a death toll of over 1.7 million out of around 79 million cases by the end of 2020.² The COVID-19 crisis has become a public health emergency and has seriously disrupted every aspect of our life, economies, and societies. For this deadly infectious disease caused by a novel pathogen, its lethality is one of the most important characteristics of the virulence of the disease for evaluating the effectiveness of responding strategies.

The case fatality rate (CFR) is one of the most essential epidemiological quantities to measure the virulence of an infectious disease, which is commonly defined as the proportion of deaths among all confirmed cases. The CFR has been

¹Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong

²Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong

³Centre for Quantitative Medicine, Duke-NUS Medical School, Singapore

Corresponding author:

KF Lam, Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong.
 Email: hrntlkf@hku.hk

adopted by health authorities in the current COVID-19 pandemic as a severity indicator.³ However, it was reported that this simple estimator only performs well at the end of an epidemic when all the cases have been resolved (affected individuals either died or recovered), but may not be a reliable indicator during an ongoing epidemic.^{4,5} Various statistical approaches have been proposed to provide a more accurate estimate for disease severity by adjusting for the reporting delay from illness onset to death during an epidemic.^{6–8} Among others, Yip et al.⁹ suggested that the fatality rate of an emerging epidemic should be time-varying in nature, and a decreasing trend in fatality rate could be a reflection of an effective measure. To provide some critical guidance on developing prompt decisive policies during an outbreak, they proposed to use the real-time fatality rate (RTFR) to measure the severity of an epidemic as opposed to the traditional CFR. Specifically, the RTFR is defined as the probability of death conditioned on a transition to death or recovery based on a counting process approach. Relative to CFR, the RTFR was shown to be more sensitive to capture changes in fatality rate during the course of an epidemic. To detect a change in RTFR statistically, Lam et al.¹⁰ developed a one-sample sequential test for the null hypothesis of constant fatality rate, which is applied to investigate the effectiveness of the interventions in Hong Kong and Beijing during the severe acute respiratory syndrome epidemic in 2003. Therein, the testing procedure starts before the implementation of a potential intervention. Under the null hypothesis, the RTFR remains constant, which means that the intervention is not effective in suppressing the fatality, at least, in a short-term period, say two months. Hence, a significant reduction in RTFR in a short run can be assumed to be attributed to an effective intervention. However, one should be cautious about the test results in the long run, as a progressive reduction in RTFR can be caused by other factors, such as the rise in temperature, improved medical health care, and mutation that the virus becomes less lethal. A more promising test to identify a potential factor that affects the severity is to compare the RTFRs among multiple independent groups over time that the effects of the above-mentioned confounding factors are shared by all groups.

There exists a modest statistical literature for comparing virulence among different subgroups. Reich et al.¹¹ defined the relative CFR as the CFR of one group divided by that of another reference group. They compared the group-specific fatality rates using a generalized linear model framework, which was adopted by Chen et al.¹² for estimating CFR based on the maximum profile-likelihood approach. However, the assumption of time-invariant CFRs in their approach is quite restrictive and is presumably more suitable for chronic diseases rather than novel emerging infectious diseases. Apart from that, most of the contemporary studies for epidemics compared the disease severity among different subgroups in a pre-specified study period and drew a conclusion on the performance of a certain intervention only at the end of the study. This approach fails to assess the efficacy of the implemented measures in a timely manner even if there is strong statistical evidence supporting differences in performance among subgroups during the observation period. Moreover, these conclusions may not be applied directly to future episodes of the same epidemic even if the viruses are of the same strain, because the characteristics of the viruses may change. During the outbreak of a novel infectious epidemic, there is an urgent need to identify an effective intervention at the earliest possible time so that prompt action can be taken to secure public health in an effective way. Also, the collection of complete and complex data is extraordinarily difficult due to various administrative reasons. For instance, information such as the times to death and recovery of patients is non-trivial and hard to assess. On the other hand, it is relatively easy to obtain the summary data on the number of confirmed cases, deaths, and recoveries from countries, such as those compiled daily during the ongoing COVID-19 pandemic. In addressing the aforementioned problems, a statistical test that provides a timely comparison of the fatality rates among multiple groups based on a simple data structure is warranted.

Motivated by the idea of Yip et al.⁹ where they captured the progressive changes in disease severity efficiently using RTFR, we propose a sequential test for the null hypothesis of equal RTFRs among different subgroups over a time period $[0, \tau]$. The null hypothesis can be rejected at time t where $t \in [0, \tau]$ as soon as statistical evidence accumulates. With the proposed method, one can test for the difference in RTFRs among neighboring areas, different age groups, different treatment arms of a clinical trial to inform and formulate public health strategies. For example, a single-arm clinical trial conducted in March 2020 found clinical improvement in patients with severe COVID-19 receiving Remdesivir, the first drug recommended for treating COVID-19.¹³ To further study this potential antiviral agent, a growing number of controlled clinical trials are conducted to judge its efficacy.^{14,15} In this case, a potential usage of a multiple sub-group test is to compare the RTFRs of patients receiving Remdesivir and standard treatment over time. Essentially, a rejection of the null hypothesis before the end of the study indicates the superior effectiveness of one treatment over the other(s). Another example, as will be illustrated in the “Application” section, is to compare the difference in RTFRs across different neighboring areas to identify the target areas that need assistance in medical health care resources during public health emergencies.

The test statistic for two-sample comparison and its asymptotic properties are studied in section 2. The generalization of the two-sample case to the K -sample case ($K > 2$) is delineated in section 3. In section 4, a large-scale simulation study is carried out to evaluate its finite-sample performance in various scenarios. In section 5, the proposed test is applied to the

COVID-19 epidemic data of mainland China to investigate the difference in disease severity among three separate area clusters: Wuhan, Hubei province (exclude Wuhan) and mainland China (exclude Hubei) during the disease outbreak. Discussions and recommendations are given in section 6.

2 A two-sample test for equality of RTFRs

We consider two populations, classified by age, treatment, or any other categories that are of our interest, subject to infection during an epidemic. Some basic epidemiological data are collected in real-time. Very often, public health officials aim to examine the difference in disease severity among two subgroups. Typically, clinicians are interested in tackling the following questions:

- Is the newly proposed treatment more effective than the standard treatment (placebo) in treating the specific infectious disease?
- Compared with area A where no measures have been taken, is the fatality rate lower in area B with effective policies?
- Do patients from resource-poor area A have a higher fatality rate than those from area B?

These questions have primary importance to guide the decision-making process during the outbreak of an infectious disease.

2.1 The test statistic

We set the observation period to be $[0, \tau]$ in the hope that a reliable decision can be made at time τ . Time 0 can be set as the day where a certain intervention, treatment, or response strategy is implemented on a particular group. We partition $[0, \tau]$ into H regular intervals (naturally in days or weeks) and the information regarding the numbers of inpatients, deaths and recoveries for the two subgroups are collected in sequence at the end of the h th interval, $h = 1, \dots, H$. Denote the numbers of deaths and recoveries in group k ($k = 1, 2$) in the h th interval by $n_{k,D}(h)$ and $n_{k,R}(h)$, respectively. We further denote the cumulative numbers of deaths and recoveries in group k at the end of the h th interval by $N_{k,D}(h)$ and $N_{k,R}(h)$, respectively, where $N_{k,D}(h) = \sum_{j=1}^h n_{k,D}(j)$ and $N_{k,R}(h) = \sum_{j=1}^h n_{k,R}(j)$. Let $I_k(h-1)$ be the number of inpatients just before the start of the h th interval for group k . We assume that in the h th interval, each of the $I_k(h-1)$ inpatients will either die, recover or remain in the hospital with respective probabilities $p_{k,D}(h)$, $p_{k,R}(h)$ and $1 - p_{k,D}(h) - p_{k,R}(h)$. Conditional on the past information, we have

$$(n_{k,D}(h), n_{k,R}(h) | I_k(h-1)) \sim \text{Multinomial}(I_k(h-1); p_{k,D}(h), p_{k,R}(h)) \quad (1)$$

Let $\mathcal{F}_{k,h} = \{I_k(j), n_{k,D}(j), n_{k,R}(j), N_{k,D}(j-1), N_{k,R}(j-1), j \leq h\}$ be the filtration or history generated by the observed data, which satisfies the usual regularity conditions.¹⁶ For group k , a discrete RTFR¹⁷ for the h th interval by considering recovery and death as two competing risks is defined as

$$\pi_k(h) = \frac{p_{k,D}(h)}{p_{k,D}(h) + p_{k,R}(h)} \quad (2)$$

which can be treated as the probability of a death conditioned on an event of death or recovery. The maximum likelihood estimator (MLE) of $\pi_k(h)$ can be easily shown to be

$$\hat{\pi}_k(h) = \frac{\hat{p}_{k,D}(h)}{\hat{p}_{k,D}(h) + \hat{p}_{k,R}(h)} = \frac{n_{k,D}(h)}{n_{k,D}(h) + n_{k,R}(h)}$$

where $\hat{p}_{k,D}(h) = [n_{k,D}(h)/I_k(h-1)]$ and $\hat{p}_{k,R}(h) = [n_{k,R}(h)/I_k(h-1)]$ are the respective sample death and recovery proportions for group k in the h th interval. The above framework, together with a smoothed version of the RTFR estimator, was summarized in Yip et al.¹⁷ With the sensitivity in picking up the changes in severity over time, the RTFR in (2) can be used to compare the virulence of the disease between two subgroups. That is, to test

$$H_0 : \pi_1(h) = \pi_2(h) \text{ for all } h \leq H \quad \text{versus} \quad H_1 : \pi_1(h) > \pi_2(h) \text{ for some } h \leq H$$

which can be reformulated as

$$H_0 : \frac{p_{1,D}(h)}{p_{1,R}(h)} = \frac{p_{2,D}(h)}{p_{2,R}(h)} \text{ for all } h \leq H \quad \text{versus} \quad H_1 : \frac{p_{1,D}(h)}{p_{1,R}(h)} > \frac{p_{2,D}(h)}{p_{2,R}(h)} \text{ for some } h \leq H \quad (3)$$

When the null hypothesis of equal RTFRs between two subgroups ($K = 2$) holds true, we expect that the ratios $n_{1,D}(h)/n_{1,R}(h)$ and $n_{2,D}(h)/n_{2,R}(h)$ are similar throughout the whole observation period. Therefore, we propose the following two-sample test statistic:

$$Z_2(h) = \sum_{j=1}^h w(j) \{ n_{1,D}(j)n_{2,R}(j) - n_{2,D}(j)n_{1,R}(j) \} \tag{4}$$

where $w(j)$ is a locally bounded, non-negative \mathcal{F}_{j-1} predictable weight process. The subscript 2 in $Z_2(h)$ corresponds to the 2-sample case discussed in this section. The proposed test statistic has mean zero under H_0 for all $h \leq H$, but has a positive expected value under H_1 for some $h \leq H$. Let $Z_2^\dagger(h)$ be the test statistic in (4) with the typical weights $w(j) = 1$ for $j = 1, \dots, h$, which represents the situation that the contribution from every interval is weighted equally. Presumably, one can introduce different sets of weights to the test statistic $Z_2(h)$ to allow extra flexibility. For example, the choice of weights $w(j) = I_1(j-1) + I_2(j-1)$ allocates a heavier weight to the period with more inpatients in groups, but a lighter weight to the time period with fewer inpatients as the fluctuations can be erratic in these intervals. Another set of intuitive weights is $w(j) = 1/[I_1(j-1) \times I_2(j-1)]$, which makes the changes in fatality rate contribute equally to the test statistic throughout the whole study period regardless of the size of inpatients, and the resulting test statistic is denoted by

$$Z_2^*(h) = \sum_{j=1}^h \widehat{p}_{1,D}(j)\widehat{p}_{2,R}(j) - \widehat{p}_{2,D}(j)\widehat{p}_{1,R}(j)$$

2.2 Asymptotic properties of the test statistic

Denote $\mathcal{T} = \{1, 2, \dots, H\}$ as the time indices for the discrete time process. For $j \in \mathcal{T}$, $\widehat{\mathbf{p}}_j = \{\widehat{p}_{1,D}(j), \widehat{p}_{1,R}(j), \widehat{p}_{2,D}(j), \widehat{p}_{2,R}(j)\}^T$ is the vector of MLEs of $\mathbf{p}_j = \{p_{1,D}(j), p_{1,R}(j), p_{2,D}(j), p_{2,R}(j)\}^T$. Under the multinomial setting in (1), we have

$$\widehat{\mathbf{p}}_j - \mathbf{p}_j \xrightarrow{\mathcal{D}} N(\mathbf{0}, \boldsymbol{\Sigma}_j)$$

where

$$\boldsymbol{\Sigma}_j = \begin{pmatrix} \frac{p_{1,D}(j)(1-p_{1,D}(j))}{I_1(j-1)} & -\frac{p_{1,D}(j)p_{1,R}(j)}{I_1(j-1)} & 0 & 0 \\ -\frac{p_{1,D}(j)p_{1,R}(j)}{I_1(j-1)} & \frac{p_{1,R}(j)(1-p_{1,R}(j))}{I_1(j-1)} & 0 & 0 \\ 0 & 0 & \frac{p_{2,D}(j)(1-p_{2,D}(j))}{I_2(j-1)} & -\frac{p_{2,D}(j)p_{2,R}(j)}{I_2(j-1)} \\ 0 & 0 & -\frac{p_{2,D}(j)p_{2,R}(j)}{I_2(j-1)} & \frac{p_{2,R}(j)(1-p_{2,R}(j))}{I_2(j-1)} \end{pmatrix}$$

Similarly, $g(\widehat{\mathbf{p}}_j) = \{n_{1,D}(j)n_{2,R}(j) - n_{2,D}(j)n_{1,R}(j)\}$ is the MLE of $g(\mathbf{p}_j) = I_1(j-1)I_2(j-1)\{p_{1,D}(j)p_{2,R}(j) - p_{1,R}(j)p_{2,D}(j)\}$, and by the Delta method,¹⁸ we have

$$g(\widehat{\mathbf{p}}_j) - g(\mathbf{p}_j) \xrightarrow{\mathcal{D}} N\left(0, \frac{\partial g(\mathbf{p}_j)}{\partial \mathbf{p}_j^T} \boldsymbol{\Sigma}_j \frac{\partial g(\mathbf{p}_j)}{\partial \mathbf{p}_j}\right)$$

When the null hypothesis of equal RTFRs between the two groups holds true, $g(\mathbf{p}_j) = 0$ for all $j \in \mathcal{T}$, which implies

$$g(\widehat{\mathbf{p}}_j) \xrightarrow{\mathcal{D}} N\left(0, \frac{\partial g(\mathbf{p}_j)}{\partial \mathbf{p}_j^T} \boldsymbol{\Sigma}_j \frac{\partial g(\mathbf{p}_j)}{\partial \mathbf{p}_j}\right)$$

Moreover, for any $s \neq t$, $g(\widehat{\mathbf{p}}_s)$ and $g(\widehat{\mathbf{p}}_t)$ are independent. Consequently, we have

$$Z_2(h) = \sum_{j=1}^h w(j)g(\widehat{\mathbf{p}}_j) = \sum_{j=1}^h w(j)\{n_{1,D}(j)n_{2,R}(j) - n_{2,D}(j)n_{1,R}(j)\} \xrightarrow{\mathcal{D}} N(0, \sigma^2(h)) \tag{5}$$

where $\sigma^2(h) = \sum_{j=1}^h w(j)^2 \partial g(\mathbf{p}_j) / \partial \mathbf{p}_j^T \boldsymbol{\Sigma}_j \partial g(\mathbf{p}_j) / \partial \mathbf{p}_j$ and it can be consistently estimated by

$$\begin{aligned} \widehat{\sigma}^2(h) &= \sum_{j=1}^h w(j)^2 (I_1(j-1) [n_{2,D}^2(j) \widehat{p}_{1,R}(j) + n_{2,R}^2(j) \widehat{p}_{1,D}(j) - \{n_{2,D}(j) \widehat{p}_{1,R}(j) \\ &\quad - n_{2,R}(j) \widehat{p}_{1,D}(j)\}^2] \\ &\quad + I_2(j-1) [n_{1,D}^2(j) \widehat{p}_{2,R}(j) + n_{1,R}^2(j) \widehat{p}_{2,D}(j) - \{n_{1,D}(j) \widehat{p}_{2,R}(j) - n_{1,R}(j) \widehat{p}_{2,D}(j)\}^2]) \end{aligned} \quad (6)$$

With the above definitions, it is easy to obtain the test statistic $Z_2(H)$ and its corresponding variance estimate $\widehat{\sigma}^2(H)$ evaluated at the endpoint τ . It follows that a straightforward test statistic for equal RTFRs between two subgroups is given by

$$V(H) = \frac{Z_2(H)}{\widehat{\sigma}(H)} \quad (7)$$

which is distributed according to a standard normal distribution under the null hypothesis. Therefore, a decision can be made at the end of the observation period τ and one can reject H_0 if $V(H) \geq z_\alpha$ at the α level of significance, where z_α satisfies $\Phi(z_\alpha) = 1 - \alpha$ and $\Phi(\cdot)$ is the distribution function of a standard normal random variable.

Note that $\widehat{\sigma}^2(\cdot)$ in (6) is a non-decreasing function and for $\forall s < t$, we have $\text{cov}\{Z_2(s), Z_2(t)\} = \min\{\sigma^2(s), \sigma^2(t)\} = \sigma^2(s)$. Therefore, $Z_2(h)$ is a Gaussian process with independent increment. Hence, we have

$$Z_2 \circ \widehat{\sigma}^2 \xrightarrow{\mathcal{Q}} W \circ \sigma^2 \text{ on } D[0, \tau]$$

where $W \circ \sigma^2$ is a standard Brownian motion. With these properties, the asymptotic normality of $\{Z_2(h), h \in \mathcal{T}\}$ with variance estimate $\widehat{\sigma}^2(h)$ can be applied to develop a sequential testing procedure, discussed in the next section.

2.3 The sequential testing procedure

Consider the above statistical test conducted over the observation period $[0, \tau]$ with a pre-specified value of H . The idea of the sequential test is to conduct a test at the end of each of the H non-overlapping intervals until a decision is made. We can think of it as a test running for H days or weeks, for example, $H = 50$. To be specific, for $h = 1, 2, \dots, H$, a test statistic $Z_2(h)$ is calculated based on the filtration $\mathcal{F}_{k,h}$ and a simple rule is used to decide when to stop: if $Z_2(h)$ exceeds the corresponding critical value b_h , we reject the null hypothesis and stop the test at the end of the h th interval. To maintain the overall type I error rate in a sequential design, the cumulative type I error is achieved by a non-decreasing function defined for each interval such that $\alpha(0) = 0$ and $\alpha(H) = \alpha$, say $\alpha = 0.05$. Adjustment to the significance level for each interval can be made through the α -spending function approach as proposed in Gordon Lan and DeMets¹⁹ with

$$\alpha(h) = 4 - 4\Phi\left(\frac{z_{\alpha/4}}{\sqrt{h/H}}\right), \quad h = 1, 2, \dots, H \quad (8)$$

Then, the set of rejection boundaries (b_1, b_2, \dots, b_H) is chosen such that

$$1 - \alpha = P[\cap_{h=1}^H \{Z_2(h) < b_h\}]$$

Suppose that H_0 is first rejected at the end of the T th interval, where $T \in \{1, \dots, H\}$, we have

$$P(T = h) = P[\{Z_2(h) > b_h\} \cap \{Z_2(j) < b_j, 1 \leq j \leq h-1\}]$$

Based on the α -spending function in (8), we have the recursive relationship

$$\alpha(1) = P(T = 1) = P\{Z_2(1) > b_1\},$$

$$\alpha(2) - \alpha(1) = P(T = 2) = P[\{Z_2(2) > b_2\} \cap \{Z_2(1) < b_1\}]$$

$$\alpha(h) - \alpha(h-1) = P(T = h) \text{ for } h = 3, \dots, H$$

We have shown in Section 2.2 that the sequence of test statistics $\{Z_2(1), \dots, Z_2(H)\}$ is a Brownian motion with independent increments under the null hypothesis. We have, $Z_2(1) \sim N(0, \sigma^2(1))$, and for each $h = 2, 3, \dots, H$,

$Z_2(h) - Z_2(h - 1) \sim N(0, \sigma^2(h) - \sigma^2(h - 1))$, independent of $Z_2(1), \dots, Z_2(h - 1)$. The calculation of the rejection boundaries can be simplified and obtained recursively by solving

$$P(T = 1) = \alpha(1) = 1 - \Phi\left(\frac{b_1}{\widehat{\sigma}(1)}\right)$$

and, for $h > 1$

$$P(T = h) = \alpha(h) - \alpha(h - 1) = \int_{b_h}^{\infty} \int_{-\infty}^{b_{h-1}} \dots \int_{-\infty}^{b_1} \prod_{j=1}^h \frac{1}{\sqrt{2\pi(\widehat{\sigma}^2(j) - \widehat{\sigma}^2(j - 1))}} \times \exp\left\{-\frac{(u_j - u_{j-1})^2}{2(\widehat{\sigma}^2(j) - \widehat{\sigma}^2(j - 1))}\right\} du_1 \dots du_h$$

The multiple integral is evaluated using a Gaussian quadrature that replaces each integral by a weighed sum. Details regarding the numerical computation for sequential methods are given in Chapter 19 of Jennison and Turnbull.²⁰ Therefore, we can compare the test statistic $Z_2(h)$ with its corresponding rejection boundary b_h ($h = 1, \dots, H$) and one can reject the null hypothesis of equal RTFRs at the end of the h^{th} interval, where $h^* = \min\{j : Z_2(j) > b_j\}$. If $Z_2(j) < b_j$ for $j = 1, \dots, H$, then one may conclude that the null hypothesis is not rejected at the end of the observation period.

3 Generalization of the two-sample test to K -sample test

In the last section, we propose a sequential test for equal RTFRs between two subgroups. This test can be easily generalized to accommodate the K -sample cases ($K > 2$) for handling complicated clinical issues during an epidemic. For instance, the test can be applied to identify the most effective drug among several candidate treatments; the test is useful when the health authority wants to determine whether the disease severity is associated with some continuous or ordinal scale measurements, such as age or the level of hospital-based health care technology. Moreover, it is of epidemiological importance to compare the severity of the disease among different areas or countries. For the ongoing COVID-19, one can compare the RTFRs among different areas in China, or among different countries to exchange information and learn from the experiences of areas with comparatively improved fatality rates.

Analogous to the two-sample test, we consider the observation period $[0, \tau]$ that contains H equally spaced time intervals. We aim to test for the null hypothesis that the RTFRs of a specific disease are equal across K ($K > 2$) independent subgroups against the alternative that the RTFRs increase with age or other measurements or factors of our interest over time. To be more specific, the hypotheses are

$$H_0 : \frac{p_{1,D}(h)}{p_{1,R}(h)} = \dots = \frac{p_{K,D}(h)}{p_{K,R}(h)} \text{ for all } h \leq H \quad \text{versus} \quad H_1 : \frac{p_{1,D}(h)}{p_{1,R}(h)} > \dots > \frac{p_{K,D}(h)}{p_{K,R}(h)} \text{ for some } h \leq H$$

The test statistic is proposed as follows:

$$Z_K(h) = \sum_{j=1}^h \sum_{k=1}^{K-1} w_{k,k+1}(j) \{n_{k,D}(j)n_{k+1,R}(j) - n_{k+1,D}(j)n_{k,R}(j)\} \tag{12}$$

where $n_{k,D}(j)$ and $n_{k,R}(j)$ are the numbers of deaths and recoveries in j th interval for the k th subgroup, respectively. In particular, $w_{k,k+1}(h)$ is a set of predetermined weights regarding the size of population in groups k and $k + 1$. For illustration, we set $K = 3$ in (12) to accommodate the comparison of RTFRs among three subgroups. The proposed test statistic becomes

$$\begin{aligned} Z_3(h) &= \sum_{j=1}^h \sum_{k=1}^2 w_{k,k+1}(j) \{n_{k,D}(j)n_{k+1,R}(j) - n_{k+1,D}(j)n_{k,R}(j)\} \\ &= \sum_{j=1}^h w_{1,2}(j) \{n_{1,D}(j)n_{2,R}(j) - n_{2,D}(j)n_{1,R}(j)\} + w_{2,3}(j) \{n_{2,D}(j)n_{3,R}(j) - n_{3,D}(j)n_{2,R}(j)\} \end{aligned} \tag{13}$$

Similarly, let $\widehat{p}_j^* = \{\widehat{p}_{k,D}(j), \widehat{p}_{k,R}(j); k = 1, 2, 3\}$ be the vector of the MLEs of the probabilities of death and recovery in

the j th interval for the three groups, we can show that the test statistic can be rewritten as

$$\begin{aligned} Z_3(h) &= \sum_{j=1}^h I_1(j-1)I_2(j-1)I_3(j-1) \left[\widehat{p}_{2,D}(j) \{w_{2,3}(j)\widehat{p}_{3,R}(j)/I_1(j-1) - w_{1,2}(j)\widehat{p}_{1,R}(j)/I_3(j-1)\} \right. \\ &\quad \left. - \widehat{p}_{2,R}(j) \{w_{2,3}(j)\widehat{p}_{3,D}(j)/I_1(j-1) - w_{1,2}(j)\widehat{p}_{1,D}(j)/I_3(j-1)\} \right] \\ &= \sum_{j=1}^h I_1(j-1)I_2(j-1)I_3(j-1)g^*(\widehat{\boldsymbol{p}}_j^*) \end{aligned}$$

where $I_k(j-1)$ denotes the number of inpatients for group k at the start of the j th interval, $g^*(\widehat{\boldsymbol{p}}_j^*)$ is a function of $\widehat{\boldsymbol{p}}_j^*$. It follows that the asymptotic variance of the test statistic in (13) can be derived easily by the Delta method, which is given by

$$\widehat{\sigma}^{*2}(h) = \sum_{j=1}^h I_1^2(j-1)I_2^2(j-1)I_3^2(j-1) \frac{\partial g^*(\widehat{\boldsymbol{p}}_j^*)}{\partial \boldsymbol{p}_j^{*T}} \boldsymbol{\Sigma}_j^* \frac{\partial g^*(\widehat{\boldsymbol{p}}_j^*)}{\partial \boldsymbol{p}_j^*} \tag{14}$$

where $\boldsymbol{\Sigma}_j^* = \text{diag}(\boldsymbol{\Sigma}_{1,j}^*, \boldsymbol{\Sigma}_{2,j}^*, \boldsymbol{\Sigma}_{3,j}^*)$ with

$$\boldsymbol{\Sigma}_{k,j}^* = \begin{pmatrix} \frac{p_{k,D}(j)(1-p_{k,D}(j))}{I_k(j-1)} & -\frac{p_{k,D}(j)p_{k,R}(j)}{I_k(j-1)} \\ -\frac{p_{k,D}(j)p_{k,R}(j)}{I_k(j-1)} & \frac{p_{k,R}(j)(1-p_{k,R}(j))}{I_k(j-1)} \end{pmatrix}$$

for $k = 1, 2, 3$ and $j = 1, \dots, h$.

Analogous to the two-sample case, we denote $Z_3^\dagger(h)$ as the test statistic in (13) with a typical set of weights $w_{k,k+1}(h) = 1$ for all $k = 1, 2$ and $h = 1, \dots, H$. Another set of intuitive weights is $w_{k,k+1}(j) = \frac{1}{I_k(j-1) \times I_{k+1}(j-1)}$, which yields the corresponding test statistic:

$$Z_3^*(h) = \sum_{j=1}^h \widehat{p}_{1,D}(j)\widehat{p}_{2,R}(j) - \widehat{p}_{2,D}(j)\widehat{p}_{1,R}(j) + \widehat{p}_{2,D}(j)\widehat{p}_{3,R}(j) - \widehat{p}_{3,D}(j)\widehat{p}_{2,R}(j) \tag{15}$$

We can easily show that the test statistic in (13) enjoys the same asymptotic properties as in the two-sample test. It converges to a Brownian motion under the null hypothesis, and the sequential testing procedure mentioned in Section 2.3 can be readily adopted using (13) and (14). Therefore, the differences of the RTFRs among $K > 2$ independent groups can be identified at the earliest possible time when enough statistical evidence accumulates.

4 Simulation study

A large-scale simulation is carried out to assess the finite-sample performance of the proposed two- and three-sample sequential tests. We assume that surveillance data are routinely reported while the exact death and discharge times are generally unknown. This mimics the real-world epidemiological data that only a summary of aggregated counts is available during the outbreak. We assume a 50-day observation period (i.e. $\tau = 50$), which is divided into $H = 50$ equal intervals and the daily number of inpatients is set to be $I_k(h) = 3000 - 30h$ ($k = 1, 2, 3$ and $h = 1, 2, \dots, 50$). We consider different scenarios that imitate how the RTFRs change over time in practice based on the prespecification of the death and recovery probabilities $p_{k,D}(h)$ and $p_{k,R}(h)$ on day h for group k ($k = 1, 2, 3$), respectively. The daily numbers of deaths and recoveries are then generated under the multinomial setting in (1). Based on the filtration $\mathcal{F}_{k,h}$ on day h , the test statistics Z_2^\dagger and Z_3^\dagger can be calculated and the sequential test can be conducted. The overall level of significance is set at $\alpha = 0.05$, and the α -spending function described in (8) is adopted throughout the simulation.

For each scenario, 10,000 independent simulated data sets were generated. Under H_0 , various scenarios with equal RTFRs among subgroups were considered to evaluate the empirical rejection rates of the sequential test, and the results for two- and three-sample tests are summarized in Tables 1 and 3, respectively. We can see that the empirical sizes for both tests match closely with the nominal level of 0.05 in all cases, suggesting that the proposed tests are empirically unbiased.

We consider 24 scenarios under the alternative hypothesis for the two-sample comparison. The results are summarized in Table 2, where \overline{h}^* in the last column represents the empirical average of h^* , the day at which the null hypothesis is first rejected (among those with H_0 being rejected). In the first eight scenarios, the RTFRs of the two groups are different only between a specific interval [25, 40] or [30, 40]. The second eight scenarios correspond to the situation that the RTFRs of

Table 1. Simulation results for the empirical sizes of the proposed two-sample test under different scenarios when H_0 is true.

$p_{1,D}(h)$	$p_{1,R}(h)$	$p_{2,D}(h)$	$p_{2,R}(h)$	Size(%)
0.005	0.005	0.01	0.01	4.94
0.005	0.01	0.005	0.01	5.09
0.01	$0.02 + 0.02I\{h > 25\}$	0.01	$0.02 + 0.02I\{h > 25\}$	4.89
0.01	$0.02 + 0.02I\{h > 25\}$	$e^{0.01h+0.005}p_{1,D}(h)$	$e^{0.01h+0.005}p_{1,R}(h)$	4.96
0.01	$0.02 + 0.02I\{h > 25\}$	$(0.005h + 0.5)p_{1,D}(h)$	$(0.005h + 0.5)p_{1,R}(h)$	5.02
0.01	$0.02 + 0.02I\{h > 35\}$	0.01	$0.02 + 0.02I\{h > 35\}$	5.01
0.01	$0.02 + 0.02I\{h > 35\}$	$e^{0.01h+0.005}p_{1,D}(h)$	$e^{0.01h+0.005}p_{1,R}(h)$	4.99
0.01	$0.02 + 0.02I\{h > 35\}$	$(0.005h + 0.5)p_{1,D}(h)$	$(0.005h + 0.5)p_{1,R}(h)$	5.39
$0.05 - 0.0005h$	$p_{1,D}(h)$	$0.05 + 0.0005h$	$p_{2,D}(h)$	5.08
$0.05 - 0.0005h$	$2p_{1,D}(h)$	$0.05 + 0.0005h$	$2p_{2,D}(h)$	4.99
$0.08 - 0.001h$	$p_{1,D}(h)$	$0.05 + 0.001h$	$p_{2,D}(h)$	5.11
$0.08 - 0.001h$	$2p_{1,D}(h)$	$0.05 + 0.001h$	$2p_{2,D}(h)$	4.98
$0.08 - 0.001h$	$p_{1,D}(h)$	$0.08 - 0.001h$	$p_{2,D}(h)$	5.06
$0.001 + 0.0005h$	$4p_{1,D}(h)$	$0.001 + 0.0005h$	$4p_{2,D}(h)$	4.99
$0.0008 + 0.0004h$	$p_{1,D}(h)$	$0.001 + 0.0005h$	$p_{2,D}(h)$	5.03
$0.0008 + 0.0004h$	$4p_{1,D}(h)$	$0.001 + 0.0005h$	$4p_{2,D}(h)$	5.12
$0.04e^{-0.04 h-30 }$	$2p_{1,D}(h)$	$e^{0.01h+0.005}p_{1,D}(h)$	$e^{0.01h+0.005}p_{1,R}(h)$	5.07
$0.04e^{-0.04 h-30 }$	$4p_{1,D}(h)$	$e^{0.01h+0.005}p_{1,D}(h)$	$e^{0.01h+0.005}p_{1,R}(h)$	4.89
$0.02e^{-0.01 h-30 }$	$2p_{1,D}(h)$	$e^{0.01h+0.005}p_{1,D}(h)$	$e^{0.01h+0.005}p_{1,R}(h)$	4.93
0.005	$0.0008 + 0.001h$	$e^{0.01h+0.005}p_{1,D}(h)$	$e^{0.01h+0.005}p_{1,R}(h)$	4.88
$0.05e^{-0.02h}$	0.05	$0.05e^{-0.02h}$	0.05	4.96
$0.05e^{-0.02h}$	0.05	$0.05e^{-0.01h}$	$0.05e^{0.01h}$	5.05
$0.05e^{-0.02h}$	0.05	$e^{0.01h+0.005}p_{1,D}(h)$	$e^{0.01h+0.005}p_{1,R}(h)$	4.92
$0.05e^{-0.02h}$	0.05	$(0.005h + 0.5)p_{1,D}(h)$	$(0.005h + 0.5)p_{1,R}(h)$	5.03

Table 2. Simulation results for the empirical powers of the proposed two-sample test under different scenarios when H_1 is true.

$p_{1,D}(h)$	$p_{1,R}(h)$	$p_{2,D}(h)$	$p_{2,R}(h)$	Power(%)	\bar{h}^*
0.01	$0.02 + 0.02I\{h > 40\}$	$p_{1,D}(h)$	$0.02 + 0.02I\{h > 25\}$	99.83	31.9
0.01	$0.04 + 0.08I\{h > 40\}$	$p_{1,D}(h)$	$0.04 + 0.08I\{h > 25\}$	99.96	28.6
0.01	$0.01 + 0.03I\{h > 40\}$	$p_{1,D}(h)$	$0.01 + 0.03I\{h > 25\}$	99.98	28.3
0.01	$0.02 + 0.02I\{h > 40\}$	$p_{1,D}(h)$	$0.02 + 0.02I\{h > 30\}$	94.98	36.4
0.01	$0.04 + 0.08I\{h > 40\}$	$p_{1,D}(h)$	$0.04 + 0.08I\{h > 30\}$	99.88	33.6
0.01	$0.01 + 0.03I\{h > 40\}$	$p_{1,D}(h)$	$0.01 + 0.03I\{h > 30\}$	99.98	33.9
0.01	$0.02 + 0.01I\{h > 40\}$	$p_{1,D}(h)$	$0.02 + 0.01I\{h > 30\}$	50.80	37.9
0.01	$0.02 + 0.02I\{h > 40\}$	$p_{1,D}(h)$	$0.02 + 0.02I\{h > 30\}$	95.56	36.5
0.01	$2p_{1,D}(h)$	$p_{1,D}(h)$	$0.02 + 0.02I\{h > 30\}$	99.87	36.9
$0.05 - 0.0005h$	$2p_{1,D}(h)$	0.01	$0.02 + 0.02I\{h > 30\}$	99.97	36.2
$0.05 - 0.0005h$	$4p_{1,D}(h)$	0.01	$0.04 + 0.04I\{h > 30\}$	99.56	35.6
$0.08 + 0.0004h$	$p_{1,D}(h)$	0.01	$0.01 + 0.03I\{h > 30\}$	99.99	32.4
$0.02e^{-0.01 h-30 }$	$2p_{1,D}(h)$	0.01	$0.02 + 0.02I\{h > 30\}$	99.82	35.3
$0.02e^{-0.01 h-40 }$	$2p_{1,D}(h)$	0.01	$0.02 + 0.02I\{h > 30\}$	99.98	35.9
$0.05 + 0.0005h$	$p_{1,D}(h)$	0.01	$0.01 + 0.02I\{h > 30\}$	99.99	32.9
$0.05 - 0.0005h$	$p_{1,D}(h)$	0.01	$0.01 + 0.02I\{h > 30\}$	99.99	33.8
$0.05 - 0.0005h$	$2p_{1,D}(h)$	$p_{1,D}(h)$	$(2 + 0.01h)p_{2,D}(h)$	94.92	36.9
$0.05 - 0.0005h$	$2p_{1,D}(h)$	$0.05 + 0.0005h$	$(2 + 0.01h)p_{2,D}(h)$	99.48	32.3
$0.05 + 0.0005h$	$2p_{1,D}(h)$	$p_{1,D}(h)$	$(2 + 0.01h)p_{2,D}(h)$	99.48	29.5
$0.05 + 0.0005h$	$2p_{1,D}(h)$	$0.05 - 0.0005h$	$(2 + 0.01h)p_{2,D}(h)$	99.46	32.0
$0.08 - 0.001h$	0.08	$p_{1,D}(h)$	$0.08 + 0.001h$	99.99	21.5
$0.08 + 0.001h$	0.08	$p_{1,D}(h)$	$0.08 + 0.002h$	99.99	14.0
0.08	$0.08 - 0.001h$	$p_{1,D}(h)$	$0.10 - 0.001h$	99.99	11.5
0.08	$0.08 + 0.001h$	$p_{1,D}(h)$	$0.08 + 0.002h$	99.99	20.9

Table 3. Simulation results of the proposed three-sample test under different scenarios when H_0 is true.

$p_{1,D}(h)$	$p_{1,R}(h)$	$p_{2,D}(h)$	$p_{2,R}(h)$	$p_{3,D}(h)$	$p_{3,R}(h)$	Size(%)
0.01	$(1 + 0.005h)p_{1,D}(h)$	0.01	$(1 + 0.005h)p_{2,D}(h)$	0.01	$(1 + 0.005h)p_{3,D}(h)$	4.85
0.01	$(1 + 0.005h)p_{1,D}(h)$	$0.05 + 0.0005h$	$(1 + 0.005h)p_{2,D}(h)$	$0.1 + 0.0008h$	$(1 + 0.005h)p_{3,D}(h)$	4.79
0.01	$e^{0.02h}p_{1,D}(h)$	0.01	$e^{0.02h}p_{2,D}(h)$	0.01	$e^{0.02h}p_{3,D}(h)$	5.05
$0.05 - 0.0005h$	$e^{0.02h}p_{1,D}(h)$	0.01	$e^{0.02h}p_{2,D}(h)$	$0.1 + 0.0008h$	$e^{0.02h}p_{3,D}(h)$	5.02
0.01	$e^{0.02 h-25 }p_{1,D}(h)$	0.01	$e^{0.02 h-25 }p_{2,D}(h)$	0.01	$e^{0.02 h-25 }p_{3,D}(h)$	4.97
0.01	$e^{0.02 h-25 }p_{1,D}(h)$	$0.05 - 0.0005h$	$e^{0.02 h-25 }p_{2,D}(h)$	$0.1 + 0.0008h$	$e^{0.02 h-25 }p_{3,D}(h)$	4.78
$0.5 - 0.0005h$	$p_{1,D}(h)$	$0.05 + 0.0005h$	$p_{2,D}(h)$	$0.1 + 0.0008h$	$p_{3,D}(h)$	4.88
$0.5 - 0.0005h$	$2p_{1,D}(h)$	$0.05 + 0.0005h$	$2p_{2,D}(h)$	$0.1 + 0.0008h$	$2p_{3,D}(h)$	4.94
0.01	$p_{1,D}(h)$	0.01	$p_{2,D}(h)$	0.05	$p_{3,D}(h)$	5.11
0.02	$p_{1,D}(h)$	0.04	$p_{2,D}(h)$	0.04	$p_{3,D}(h)$	4.89
$0.10 - 0.0008h$	$2p_{1,D}(h)$	$0.1 - 0.0008h$	$2p_{2,D}(h)$	$0.1 - 0.0008h$	$2p_{3,D}(h)$	5.04
$0.04e^{-0.04 h-45 }$	$2p_{1,D}(h)$	$0.04e^{-0.04 h-35 }$	$2p_{2,D}(h)$	$0.04e^{-0.04 h-25 }$	$2p_{3,D}(h)$	4.94
$0.04e^{-0.04 h-45 }$	$4p_{1,D}(h)$	$0.04e^{-0.04 h-35 }$	$4p_{2,D}(h)$	$0.04e^{-0.04 h-25 }$	$4p_{3,D}(h)$	4.97

two groups remain the same at first, but the RTFR of group 2 drops suddenly at $h = 30$. One can see that the proposed test is reasonably powerful (with empirical powers over 95%) in detecting a sudden change in RTFR between groups. Also, the null hypothesis can be rejected within a short period of time, say 7 days, since a change has been imposed to the RTFR of group 2. Nevertheless, we may observe a relatively small power in some cases where the change in RTFR in group 2 is modest or small. For example, the scenario in the seventh row of Table 2 only attains a power of 50.80% due to a relatively small jump size in the recovery probabilities in group 2. When we increase the jump size from 0.01 to 0.02 (the next row in Table 2), the empirical power increases from 50.80% to 95.56%, and \bar{h}^* becomes closer to $h = 30$ where a change occurs. The remaining 8 scenarios correspond to the situation that the RTFR of group 1 is uniformly higher than that of group 2, and the empirical powers are high in general. Table 4 demonstrates the good performance of the proposed three-sample test under H_1 . Specifically, the RTFR is always the highest in group 1 and the lowest in group 3 throughout the observation period. The empirical powers are close to 1 in all cases and the null hypothesis can be rejected quickly as soon as there is enough statistical evidence supporting the alternative hypothesis.

In addition to the results reported in Tables 1 to 4, we have tried different sequences of daily number of inpatients in the simulation setup, such as $I_k(h) = 3000 + 30h$ and $I_k(h) = 3000 + 30h - 60(h - 15)_+$ where $u_+ = \max(0, u)$, $k = 1, 2, 3$ and $h = 1, 2, \dots, 50$ as well as small sample size with $I_k(h)$ around 800. We also tried another weight function corresponding to the test statistics Z_2^* and Z_3^* in replacement of Z_2^\dagger and Z_3^\dagger . It is noted that the results obtained in Tables 1 to 4 are quite robust to these changes, hence those findings are not reported here. Moreover, when compared with the non-sequential test $V(H)$ discussed in (7), the sequential test achieves the same level of power in all cases with no additional cost but it allows conclusion to be made at a much earlier time.

Table 4. Simulation results of the three-sample sequential test under the alternative hypothesis when H_1 is true.

$p_{1,D}(h)$	$p_{1,R}(h)$	$p_{2,D}(h)$	$p_{2,R}(h)$	$p_{3,D}(h)$	$p_{3,R}(h)$	Size(%)	\bar{h}^*
0.02	$p_{1,D}(h)$	0.02	$(1 + I(h > 30))p_{2,D}(h)$	0.02	$(1 + 2I(h > 30))p_{3,D}(h)$	99.99	35.0
0.02	$p_{1,D}(h)$	$0.05 + 0.0005h$	$(1 + I(h > 30))p_{2,D}(h)$	$0.1 + 0.0008h$	$(1 + 2I(h > 30))p_{3,D}(h)$	99.99	33.6
0.01	$p_{1,D}(h)$	0.01	$(1 + I(h > 35))p_{2,D}(h)$	0.04	$(1 + 2I(h > 25))p_{3,D}(h)$	99.99	30.1
0.01	$p_{1,D}(h)$	$0.05 + 0.0005h$	$(1 + I(h > 35))p_{2,D}(h)$	$0.1 + 0.0008h$	$(1 + 2I(h > 25))p_{3,D}(h)$	99.99	27.2
$0.1 - 0.0008h$	$p_{1,D}(h)$	$0.05 + 0.0005h$	$(1 + I(h > 30))p_{2,D}(h)$	$0.1 + 0.0008h$	$(1 + 2I(h > 30))p_{3,D}(h)$	99.99	32.3
$0.1 + 0.0005h$	$p_{1,D}(h)$	$0.04 + 0.0005h$	$(1 + I(h > 35))p_{2,D}(h)$	0.04	$(1 + 3I(h > 25))p_{3,D}(h)$	99.99	27.1
$0.05 - 0.0005h$	$2p_{1,D}(h)$	$0.05 + 0.0005h$	$(2 + I(h > 35))p_{2,D}(h)$	$0.1 + 0.0008h$	$(2 + 2I(h > 25))p_{3,D}(h)$	99.99	28.2
$0.04e^{-0.04 h-25 }$	$1.5p_{1,D}(h)$	$0.02e^{-0.02 h-30 }$	$2p_{2,D}(h)$	$0.04e^{-0.04 h-35 }$	$3p_{3,D}(h)$	99.99	7.9
$0.04e^{-0.04 h-25 }$	$1.5p_{1,D}(h)$	$0.02e^{-0.02 h-30 }$	$2.5p_{2,D}(h)$	$0.04e^{-0.04 h-35 }$	$3.5p_{3,D}(h)$	99.99	7.2
$0.08 + 0.001h$	$0.05 + 0.001h$	$0.08 + 0.001h$	$0.05 + 0.002h$	$0.08 + 0.001h$	$0.05 + 0.003h$	99.99	12.4
$0.08 - 0.001h$	$0.05 + 0.001h$	$0.08 - 0.001h$	$0.05 + 0.002h$	$0.08 - 0.001h$	$0.05 + 0.003h$	99.99	13.0
$0.12 - 0.001h$	$0.05 + 0.001h$	$0.12 - 0.0015h$	$0.05 + 0.002h$	$0.12 - 0.002h$	$0.05 + 0.003h$	99.99	13.6
$0.12 - 0.001h$	$0.12 + 0.001h$	$0.12 - 0.0015h$	$0.12 + 0.002h$	$0.12 - 0.002h$	$0.12 + 0.003h$	99.99	15.2
$0.05 - 0.0005h$	$p_{1,D}(h)$	$0.05 + 0.0005h$	$(1 + I(h > 35))p_{2,D}(h)$	$0.1 + 0.0008h$	$3p_{3,D}(h)$	99.99	3.2

In addition, to assess the effect of the choice of τ , and hence the number of intervals H (say, in days or weeks), on the performance of the proposed test, $\tau = 40$ and 60 days were also considered. For the cases with a sudden change in the RTFR under H_1 , the power and \bar{h}^* for different values of τ are virtually identical. For the cases with a gradual increase in the difference in RTFRs over time, it is natural to expect a higher power based on a larger value of τ as more statistical evidence would accumulate over time, but the difference is minimal. On the other hand, a larger value of τ also means that the significance level assigned to each interim analysis is smaller, which will also lead to a slight increase in \bar{h}^* . In practice, we suggest to set τ to be reasonably large to allow accumulation of more statistical evidence, at the expense of a slight delay in the decision if there is a difference.

5 Application

In December 2019, several cases of novel coronavirus infection, now known as COVID-19, were reported in Wuhan, Hubei province, China. Despite the implementation of strict lockdown in Wuhan on 23 January 2020,²¹ this virus had rapidly spread from the epicenter to different regions across China. By the end of February 2020, 79,394 cases including 2838 deaths were reported in mainland China.²² Thereof, 66,337 occurred in the Hubei province with a death toll of 2727, suggesting a CFR of 3.26% at first glance, which contrasts with 0.8% in other areas of mainland China. The study suggested that the accessibility level of health care resources may be the cause of the considerable gap in mortality among different areas.²³ According to the level of medical resource availability during the outbreak, we partition mainland China into three clusters, namely Wuhan city, Hubei province excluding Wuhan city, and mainland China excluding Hubei province. The main objective of the analysis is to explore the difference in disease severity based on the RTFR among these clusters and to investigate the potential effects of medical resource availability (i.e. the numbers of doctors and hospital beds) on the fatality rate in China. The cumulative numbers of confirmed cases, deaths, and recoveries between 1 February and 31 March for each cluster were summarized and extracted from the public domain.²⁴

A smoothed version of the RTFR estimator¹⁷ for the three separate clusters over the observation period is shown in Figure 1. We can see that there exist clear disparities in severity among areas in mainland China during the early phase of the COVID-19 epidemic. In this regard, we provide some explanatory notes to describe the observed pattern. As

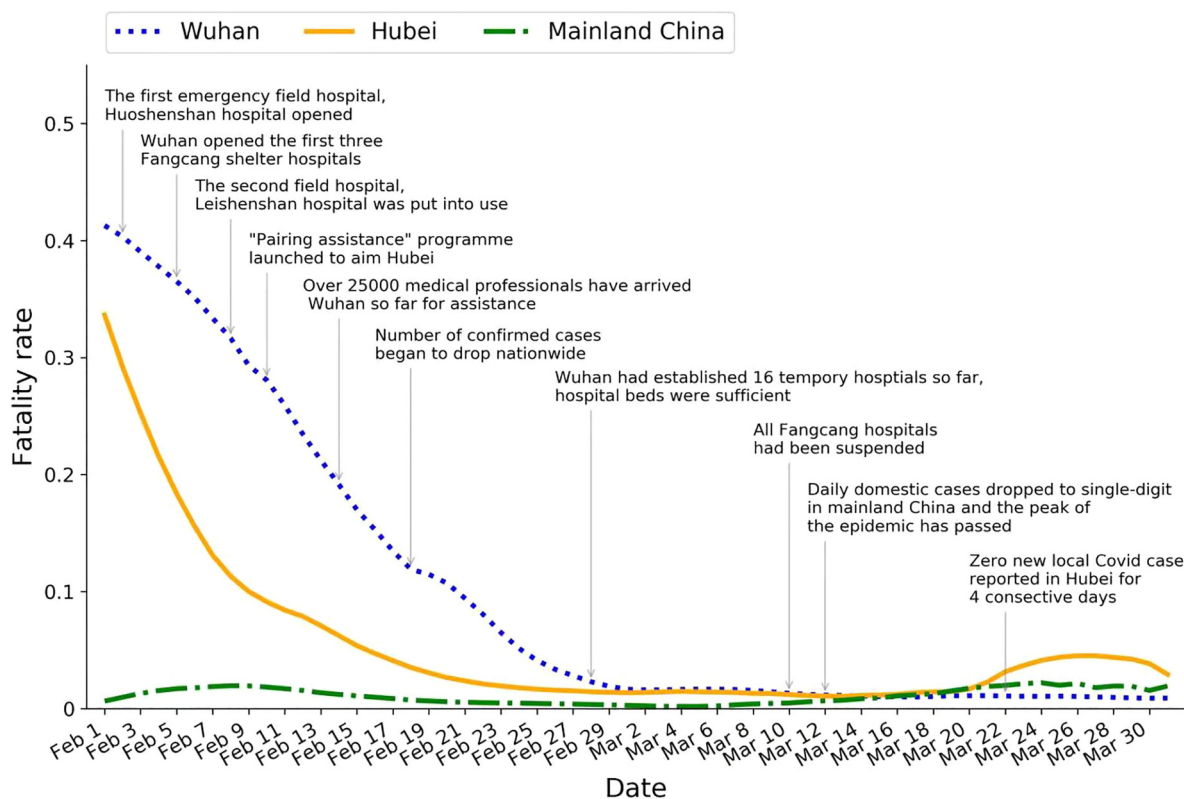


Figure 1. The estimated real-time fatality rate for the outbreak of COVID-19 in three separate clusters.

most of the cases were concentrated in Hubei province at the beginning of the outbreak, the hospitals and local health care systems were suddenly overwhelmed. Especially in Wuhan, many patients did not receive timely treatment, causing Wuhan to have the highest fatality rate, followed by the remaining cities in Hubei province. On the contrary, the lockdown measures implemented in Wuhan city delayed the epidemic growth in other provinces and provides valuable time for them to prepare. Therefore, the number of infections grew at a much slower rate compared with the supply of health care resources, contributing to a mitigating fatality rate in mainland China as a whole. To meet the shortage of medical resources in worst-hit areas, the Chinese government mobilized all the necessary resources nationwide to support virus control in Hubei and the city of Wuhan. Two new field hospitals, namely Huoshenshan and Leishenshan, were built in a few days and had been put into use in Wuhan in early February. Over 25,000 medical professionals from other provinces of China rushed to Wuhan for assistance as of 14 February.²⁵ The remaining 16 cities in Hubei province also received one-to-one paired assistance from other provinces.²⁶ In the meantime, a number of temporary hospitals, namely the Fangcang shelter hospitals, were constructed to provide enough beds to treat patients with mild to moderate symptoms. These temporary hospitals relieved the huge pressure on the health care system and allowed the designated hospitals to concentrate on treating patients with severe and critical conditions.²⁷ As of 28 February 2020, Wuhan had established 16 temporary hospitals, and the demand for hospital beds in Hubei was met. While the RTFR of mainland China stabilized at a low level, the RTFRs of Wuhan and Hubei declined continuously owing to increasing hospital beds and sufficient medical resources. Eventually, the RTFRs for the three clusters reach a similar level by the end of February and remained relatively low throughout March.

The proposed method is applied to examine the difference in RTFRs among areas over time. By treating day as the unit, there are $H = 60$ intervals in total with day 0 being 1 February 2020. We conduct the two-sample test for $H_1 : \pi_{\text{Wuhan}}(h) > \pi_{\text{Hubei}}(h)$, and the three-sample test for $H_1 : \pi_{\text{Wuhan}}(h) > \pi_{\text{Hubei}}(h) > \pi_{\text{China}}(h)$ for some $h < H = 60$, respectively. The typical weight function $w(h) = 1$ is used and the overall significance level is set to be 0.05. Specifically, on day h ($h = 1, 2, \dots, H$), we compared the test statistics $Z_2^\dagger(h)$ and $Z_3^\dagger(h)$ with their corresponding critical values b_h , respectively. The sequential test is terminated at $0 \leq h^* \leq H$ if the test statistic at h^* exceeds the critical value b_{h^*} . In line with the considerable gap in fatality rates among areas shown in Figure 1, the null hypothesis of equal RTFRs is rejected quickly on the seventh day (7 February) and fourth day (4 February) based on the two- and three-sample tests, respectively. We then conduct the same pair of tests for the period from 1 March to 1 April with $H = 32$, during which the medical resources availability is more or less the same across different clusters. As expected, both the two- and three-sample tests fail to reject the null hypothesis, which is further supported by the similar fatality rates in March 2020 among the three areas as displayed in Figure 1. This example shows that our proposed tests are sensitive in picking up changes in RTFRs, and it is useful to provide real-time signals at time $t \in [0, \tau]$ to the health authority on whether the existing measures or medical resources are adequate in some areas to contain the epidemic.

6 Discussion

A statistical test is proposed in this paper to compare the RTFRs among independent groups during the course of an ongoing epidemic. As the implementation of an effective control measure can reduce disease severity and save more lives, our method can provide an evidence-based assessment of the effectiveness of the implemented intervention and inform the policy-making process during an emerging epidemic. The asymptotic Brownian motion of the test statistic under the null hypothesis allows one to adopt a sequential design naturally. Therefore, the null hypothesis of no difference in severity among subgroups can be rejected as soon as sufficient information has accumulated over time. This property is particularly useful during the emerging epidemic as the government officials can identify the effective control measures at the earliest time and issue the recommendation for disease control promptly. A large-scale simulation study shows the good performance of our proposed test in two- and three-sample cases in terms of unbiasedness and the sensitivity in picking up the difference in severity among groups.

The proposed statistical test is applied to the COVID-19 data in mainland China to examine the difference in severity among three separate clusters. The results suggest that the severity of COVID-19 in mainland China is possibly associated with the accessibility of local health care resources. This emphasizes that medical supplies and resources play an important role in lowering the RTFR. As many countries are now struggling with the COVID-19 outbreak, these findings may suggest on disease prevention and control worldwide. Especially for the resource-limited countries, they should at least slow down the surge of infections to avoid the local medical system being overwhelmed. The illustrated example demonstrates that our method is simple to use and is widely applicable to all emerging infectious diseases.

We have shown that the proposed two-sample test can be easily generalized to accommodate the K -sample situation. Essentially, this enables us to deal with more clinical questions in practice. For example, investigating the discrepancy in RTFRs between multiple age groups could help to minimize the confounding effect and help us gain an in-depth

understanding of the other factors that affect fatality. Most importantly, noting the discrepancy of fatality rates between different treatment arms in clinical trials helps the clinicians to identify the most effective treatment for curbing the disease. Take COVID-19 as an example, over hundreds of clinical trials have been registered worldwide on clinical trials registries so far aiming to evaluate the performance of some possible treatments.^{28–30} Our proposed method can be one of the essential tools to evaluate the efficacy of different potential treatments, where superiority over other candidate treatments is indicated by a relatively improved fatality rate along the timeline.

The proposed tests have the advantage of using minimal information to gain timely assessment on the effectiveness of potential treatments or implemented measures based on a quantitative approach. During an outbreak of the emerging epidemic, the surveillance data are always incomplete, and individual data such as the time-at-infection, time-to-recovery, and time-to-death, are difficult to obtain. This is true especially for those areas with low public health awareness and with a poor health care system. For the ongoing COVID-19, the epidemiological data are hard to obtain, and, for most of the countries, only the cumulative counts on cases, death, and recovery are recorded. It is important for public health officials to make use of this simplest data structure to gain more insight into the disease so that prompt actions can be taken to suppress the disease fatality at the earliest possible time.

Declaration of conflicting interests


The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

KF Lam  <https://orcid.org/0000-0001-5453-994X>

Chun Yin Lee  <https://orcid.org/0000-0002-7207-2519>

References

1. Jones KE, Patel NG, Levy MA et al. Global trends in emerging infectious diseases. *Nature* 2008; **451**: 990–993.
2. World Health Organisation (WHO). Coronavirus disease (COVID-19) weekly epidemiological update and weekly operational update, 2020. <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports> (2020, accessed 10 December 2020).
3. World Health Organisation (WHO). Estimating mortality from COVID-19, 2020. <https://www.who.int/publications/i/item/WHO-2019-nCoV-Sci-Brief-Mortality-2020.1> (2020, accessed 18 November 2020).
4. Ghani AC, Donnelly CA, Cox DR et al. Methods for estimating the case fatality ratio for a novel, emerging infectious disease. *Am J Epidemiol* 2005; **162**: 479–486.
5. World Health Organisation (WHO). Estimating mortality from COVID-19: Scientific brief, 4 August 2020. Technical report, 2020.
6. Kucharski AJ and Edmunds WJ. Case fatality rate for Ebola virus disease in west Africa. *The Lancet* 2014; **384**: 1260.
7. Rajgor DD, Lee MH, Archuleta S et al. The many estimates of the COVID-19 case fatality rate. *Lancet Infect Dis* 2020; **395**: 1569–1578.
8. Mizumoto K, Saitoh M, Chowell G et al. Estimating the risk of Middle East respiratory syndrome (MERS) death during the course of the outbreak in the Republic of Korea, 2015. *Int J Infect Dis* 2015; **39**: 7–9.
9. Yip PSF, Lam KF, Lau EHY et al. A comparison study of realtime fatality rates: severe acute respiratory syndrome in Hong Kong, Singapore, Taiwan, Toronto and Beijing, China. *J R Stat Soc A Stat* 2005; **168**: 233–243.
10. Lam KF, Deshpande JV, Lau EHY et al. A test for constant fatality rate of an emerging epidemic: with applications to severe acute respiratory syndrome in Hong Kong and Beijing. *Biometrics* 2008; **64**: 869–876.
11. Reich NG, Lessler J, Cummings DA et al. Estimating absolute and relative case fatality ratios from infectious disease surveillance data. *Biometrics* 2012; **68**: 598–606.
12. Chen Z, Akazawa K and Nakamura T. Estimating the case fatality rate using a constant cure-death hazard ratio. *Lifetime Data Anal* 2009; **15**: 316–329.
13. Grein J, Ohmagari N, Shin D et al. Compassionate use of remdesivir for patients with severe COVID-19. *New Engl J Med* 2020; **382**: 2327–2336.
14. Wang Y, Zhang D, Du G et al. Remdesivir in adults with severe COVID-19: A randomised, double-blind, placebo-controlled, multi-centre trial. *The Lancet* 2020; **395**: 1569–1578.
15. Beigel JH, Tomashek KM, Dodd LE et al. Remdesivir for the treatment of COVID-19 preliminary report. *New Engl J Med* 2020; **383**: 1813–1826.
16. Andersen PK, Borgan O, Gill RD et al. *Statistical models based on counting processes*. New York: Springer Science & Business Media, 2012.

17. Yip PSF, Lau EHY, Lam KF et al. A chain multinomial model for estimating the real-time fatality rate of a disease, with an application to severe acute respiratory syndrome. *Am J Epidemiol* 2005; **161**: 700–706.
18. Oehlert GW. A note on the delta method. *Am Stat* 1992; **46**: 27–29.
19. Gordon Lan KK and DeMets DL. Discrete sequential boundaries for clinical trials. *Biometrika* 1983; **70**: 659–663.
20. Jennison C and Turnbull BW. *Group sequential methods with applications to clinical trials*. London: Chapman & Hall, 2020.
21. Lau H, Khosrawipour V, Kocbach P et al. The positive impact of lockdown in Wuhan on containing the COVID-19 outbreak in China. *J Travel Med* 2020; : –.
22. Wu JT, Leung K, Bushman M et al. Estimating clinical severity of COVID-19 from the transmission dynamics in Wuhan, China. *Nat Med* 2020; **26**: 506–510.
23. Ji Y, Ma Z, Peppelenbosch MP et al. Potential association between COVID-19 mortality and health-care resource availability. *Lancet Glob Health* 2020; **8**: e480.
24. Tencent. Tencent News, 2020. <https://news.qq.com/zt2020/page/feiyan.htm#> (2019, accessed 12 January 2021).
25. Jia J, Ding J, Liu S et al. Modeling the control of COVID-19: Impact of policy interventions and meteorological factors. *Electron J Differ Equ* 2020; **2020**: 1–24.
26. Chen T, Wang Y and Hua L. Pairing assistance the effective way to solve the breakdown of health services system caused by COVID-19 pandemic. *Int J Equity Health* 2020; **19**: 1–4.
27. Chen S, Zhang Z, Yang J et al. Fangcang shelter hospitals: a novel concept for responding to public health emergencies. *The Lancet* 2020; **395**: 1305–1314.
28. Yao X, Ye F, Zhang M et al. In vitro antiviral activity and projection of optimized dosing design of hydroxychloroquine for the treatment of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). *Clin Infect Dis* 2020; **71**: 732–739.
29. Cao B, Wang Y, Wen D et al. A trial of lopinavir–ritonavir in adults hospitalized with severe COVID-19. *New Engl J Med* 2020; : –.
30. Gautret P, Lagier JC, Parola P et al. Hydroxychloroquine and azithromycin as a treatment of COVID-19: results of an open-label non-randomized clinical trial. *Int J Antimicrob Agents* 2020; : .