RESEARCH PAPER

Risk Identification and Prediction for COVID-19 Mortality

Hanh Nguyen [a] **Qin Shao** [a]

Corresponding author(s): qin.shao@utoledo.edu

[a] Department of Mathematics and Statistics College of Natural Sciences and Mathematics, Toledo, Ohio 43614,

This paper studies several key metrics for COVID-19 using a public surveillance system data set. It compares the difference between two case fatality rates: the naive case fatality rate, which has been frequently mentioned in media outlets, and one which is the sample estimate for the mortality rate. A logistic regression model is applied to modeling the daily mortality rate. The conclusion is that time, gender, age and some of their interactions, appear to have a significant impact on the mortality rate; the daily mortality rate has been decreasing since the outbreak; males older than 60 has been the most vulnerable group. The receiver operating characteristics curve and the curve under the area show that the proposed logistic model is capable of predicting the outcome of a reported case with accuracy as high as 89%. These findings are helpful in assessing the magnitude of the risk posed by the COVID-19 virus to certain groups, predicting outcome severity, and optimally allocating medical resources such as intensive care units and ventilators.

COVID-19 | fatality rate | mortality rate | Logistic regression | receiver operating characteristics curve

Since the outbreak of the coronavirus (COVID-19) pandemic in December 2019 in China, researchers all over the world have been working on understanding the transmission mechanism (6, 7, 11, 29), estimating key metrics for assessing the magnitude of the risk posed by this virus (2, 13, 24, 27), and obtaining information for policy making (5, 8, 17, 26). Case fatality rate (CFR) is one of the key indicators of the severity of an infectious disease. However, it is challenging to obtain an accurate CFR, as both case and death counts of an infectious disease are in general unknown.

The simplest approach uses the daily naive CFR, which is the death count divided by the case count on day t. The daily naive CFR, denoted by r_t, is one of the statistics that numerous organizations and media have been updating based on the latest COVID-19 data. An advantage of r_t, for example, is that it is computationally straightforward, whereas the major disadvantage is that it is not accurate as a measure of disease severity, and sometimes is even misleading. As Ritchie and Roser (21) pointed out, it ignores deaths in cases with time lags. Since the deaths in the numerator are not a subset of the cases in the denominator, the naive CFR does not accurately reflect the severity. Another daily CFR, denoted by π_t, is the ratio of the death count to the case count on day t. Both r_t and π_t are relative frequencies of deaths and share the same denominator or case count on day t, but they have different numerators — the numerator of r_t is the death count on the same day, while that of n_t is the death count among the cases in the denominator. The deaths in the numerator of π_t consist of a subset of the denominator, although they can happen any time after the case onset dates. This fundamental disparity which will be examined and elaborated, distinguishes pi_t from r_t as a better description of the disease severity (22).

A daily mortality rate (MR), denoted by p_t, is the probability of death from a disease and is another measure of severity. However, the true probability is not observable and usually estimated by the CFR π_t. The relationship between daily COVID-19 mortality rate and several factors will be modeled using reported death and case counts as well as other relevant information provided by the public surveillance system of the State of Ohio. Shao et al. (23) considered how much the mortality rate can be explained by gender and age using the same reported system, but it treated MR as constant over time and did not take the change of MR into account. Several public policy measures could have had some impact on the daily counts since the outbreak of COVID-19. For example, how infectiousness of COVID-19 has been changing due to interventions (15), such as social distancing and curfew; it is possible that more and more easily accessible tests have led to large case counts recently; more and more effective treatments could have been contributing to the reduc-

Submitted: 03/23/2021, published: 08/31/2021.

tion of death counts (9). Thus, in the model development for daily MR, time is considered as one of the covariates for the purpose of identifying statistically significant factors based on statistical modeling. In this paper, the model proposed will be utilized to predict the likelihood of mortality for a reported case.

There are three goals of this paper: comparing the daily naive CFR r_t with CFR π_t, identifying risk factors that impact daily MR p_t using statistical inference, and making a prediction about the probability of mortality for a COVID-19 patient based on the risk factors. All the data analysis is conducted using the State of Ohio COVID-19 surveillance data, which includes information about each reported patient, in particular, gender, age, onset date, death date, and outcome. The paper is organized as follows: details about the data and statistical descriptions of several major characteristics are presented; r_t and π_t are examined and compared; the statistical inference based on logistic regression is provided in the findings about the relationship between r_t and π_t, statistical inference results about p_t, and the application of the model in prediction of death likelihood of a case based on age, gender, and time are elaborated; finally in the paper concludes with a discussion.

Materials and Methods

Data

The raw daily data in the study period, March 10, 2020 to January 31, 2021 inclusive, were downloaded from the State of Ohio COVID-19 dashboard (18). The rows of the raw data set are the records of patients, and the final data set is obtained by deleting all the rows that contain "unknown". Figure 1 shows that the case count increased dramatically until November and then dropped off in the last two months, while the death count did not change much throughout. Table 1 lists the monthly summary for the daily counts. The maximum case count suddenly jumped from 4,094 in October to 13,523 in November.

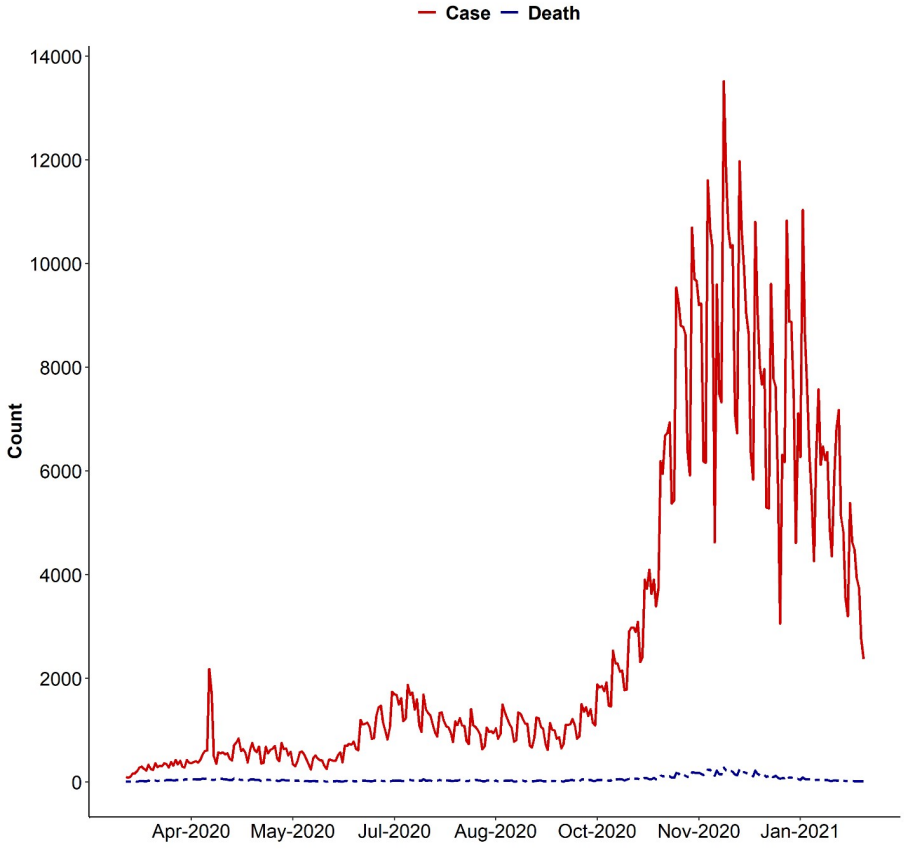

Figure 1. Daily Death and Case Counts, March 10, 2020 - January 31, 2021

Table 1. Summary of Daily Count Data by Month

Month	Daily Max		Daily Median		Daily Mean		Daily Min	
	Case	Death	Case	Death	Case	Death	Case	Death
March - 2020	383	39	280.5	18.0	256.0	18.9	81	4
April - 2020	2181	76	475.0	49.0	583.7	50.1	281	28
May - 2020	754	57	572.0	29.0	531.2	32.2	237	10
June - 2020	1437	29	622.5	15.0	694.8	16.0	249	7
July -2020	1877	55	1329.0	27.0	1324.4	27.0	820	12
August - 2020	1493	41	1034.0	23.0	1018.5	23.7	628	5
September - 2020	1501	48	1078.0	19.5	1025.0	21.9	621	7
October - 2020	4094	82	2286.0	44.0	2393.5	47.1	1089	15
November - 2020	13523	280	8064.0	144.0	8009.8	148.3	3729	66
December - 2020	11976	242	8028.0	136.0	8255.0	139.7	3057	62
January - 2021	11038	87	5539.0	28.0	5597.4	31.2	2370	8

A threshold of 21 days is chosen as the cutoff for survival for two reasons: according to the State of Ohio dashboard, a positive case is considered as "presumed recovered" after the symptom onset date larger than 21 days; according to Figure 2, all the monthly medians of days for death are less than 21 days. In other words, if a patient has not died of COVID-19 by February 21, 2021, he or she is considered to have survived.

For each reported positive case whose onset date is in the study period, define the dichotomous dependent variable y, which is the outcome indicator, as either 1 if the patient has died of COVID-19 by February 21, 2021 or 0 otherwise. Age is divided into two groups: the older group (at least 60 years old) and the younger group (from 0 to 59 years old).

Time t is introduced for the number of days between the beginning of the study and the onset date on the record of a case. For example, $t = 1$ for a case whose onset date was on March 10. The final data set to be analyzed contains $n = 898{,}228$ rows, with each row being the record of a positive case, and four columns being the outcome indicator y and the covariates. The column information is summarized in Table 2.

Table 2. Columns of Data Set

Column	Type	Values
Sex	factor with 2 levels	female, male
Age	factor with 2 levels	0, 1
Time	integer	1, 2, . . .
Outcome	factor	0, 1

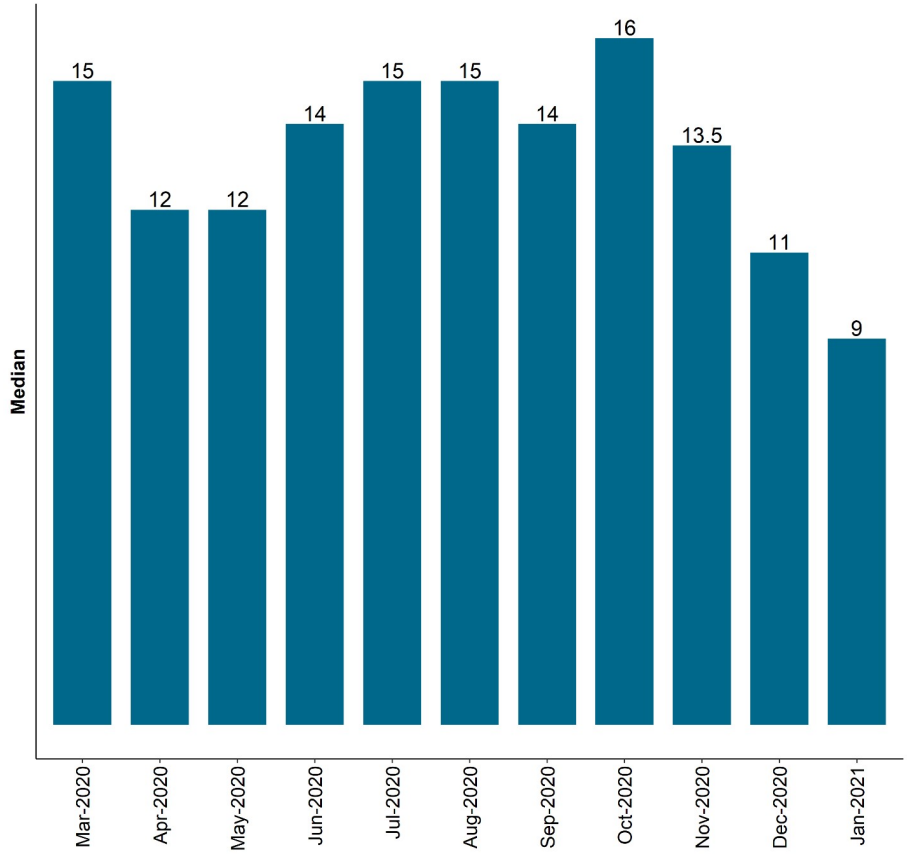

Figure 2. Monthly Medians of Days for Deaths

Case Fatality Rates

Table 3 summarizes the total case count, death count and overall naive CFR of each gender-by-age category up to and including January 31, 2021. The overall naive CFR is 0.0187, and these four gender-by-age groups have very different CFR's: the male older group has the largest CFR, which is 0.0089 and the female younger group has the smallest CFR, which is 0.0005. The odds ratio of these two groups is as large as 17.951. This naive CFR uses a possibly smaller numerator, as the outcomes of the most recent cases are ignored. Moreover, these CFR's are snapshots and do not take time into account.

Table 3. Case Fatality Rates (CFR=Death Count in Each Category/Total Case Count)

Age	Gender						Total		
	Female			Male			Case	Death	MR
	Case	Death	MR	Case	Death	MR			
<60	364536	461	0.0005	315177	693	0.0008	679713	1154	0.0013
⩾ 60	118749	7590	0.0085	99766	8037	0.0089	218515	15627	0.0174
Total	483285	8051	0.009	414943	8730	0.0097	898228	16781	0.0187

Given many factors could have impacted the counts, it is reasonable to take *t* into consideration. The daily CFR's r_t and π_t are respectively calculated as follows:

$$r_t = \frac{\text{Death Count on Day t}}{\text{Case Count on Day t}}$$

$$\pi_t = \frac{Death\ Count\ among\ Case\ Count\ on\ Day\ t}{Case\ Count\ on\ Day\ t}$$

Logistic Regression for Mortality Rate

Define $p_t = P\ (y_t = 1)$ which is the probability of death of a reported case or reported case mortality rate at time *t*. Hereafter p_t and $p_t\ (x)$ will be used interchangeably with the latter emphasizing covariates x. The daily cases are separated into four groups according to age and gender. The reference group includes all the cases who are younger females or females younger than 60, and three dummy variables are introduced for the other groups: $x_1 = 1$ for a female case whose age is older than 60 and 0 otherwise; $x_2 = 1$ for a male case whose age is younger than 60 and 0 otherwise; $x_3 = 1$ for a male case whose age is older than 60 and 0 otherwise. Logistic regression, which is typically implemented to model the relationship between a dichotomous dependent variable and covariates, is applied to *y*, age, gender and time. Interested readers can refer to (1) and (16) for comprehensive discussions about the theory and applications of logistic regression.

The full model that includes the covariates and all the interactions between time, age, gender is considered. Data analysis for logistic regression is carried out using the package glm in R (19) which is a free software environment for statistical computing and graphics. According to the Akaike information criterion, the following model is a good compromise between simplicity and adequacy:

$$\log\frac{p_t(x)}{1-p_t(x)} = \beta_0 + \sum_{i=1}^{3}\beta_i x_i + \beta_4 t + \beta_5 x_1 t + \beta_6 x_3 t, \quad [1]$$

where $x = (x_1, x_2, x_3, t)$. It is obvious that for the female younger positive cases which constitutes the reference group, model [1] becomes:

$$\log\frac{p_t(x)}{1-p_t(x)} = \beta_0 + \beta_4 t. \qquad [2]$$

It is straightforward to obtain the models for the other age-by-gender groups. For example, for the older female group, model [1] is rewritten as:

$$log\,\frac{p_t(x)}{1-p_t(x)} = \beta_0 + \beta_1 + (\beta_4 + \beta_5)t. \qquad [3]$$

From [2] and [3], $\beta_1 + \beta_5 t$ indicates the log odds ratio of older and younger groups of female cases. Similarly, it can be concluded that $\beta_3 + \beta_6 t$ is the log odds ratio between older and younger groups of male cases.

Results

The mathematical difference between r_t and π_t is the numerator. Unless a death from COVID-19 in the numerator of r_t happens on the same day when it is reported as a case, it is not among the case counts in the denominator. Thus, it is obvious that r_t mismatches these counts, which introduces bias, and on the other hand, π_t pairs the deaths with the cases and is a more reliable indicator for the severity or the death likelihood of a COVID-19 patient. Figure 3 illustrates the difference between relative frequencies of r_t and π_t.

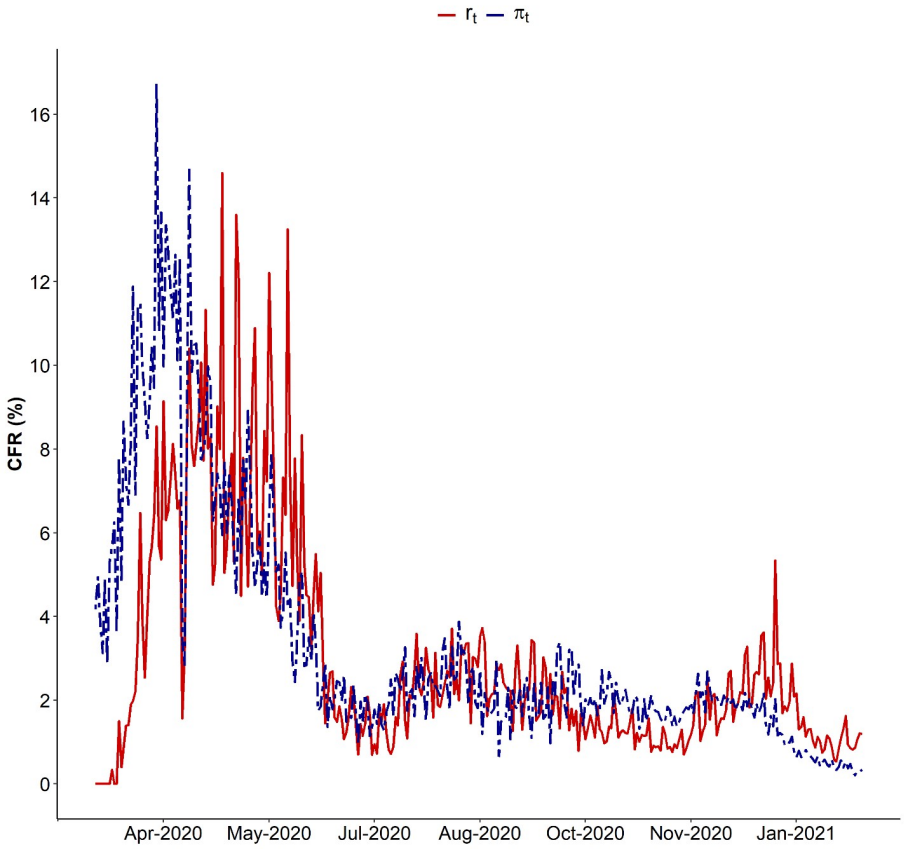

Figure 3. Case Fatality Rates r_t and π_t

The distinction was more manifest in the first 100 days, and π_t reached a peak sooner than r_t. Not only is there a time lag between r_t and π_t, but they display different patterns. In particular, the surge of r_t in December did not occur in π_t. The peak of π_t implies that the early cases were more likely to result in death.

Table 4 is the information for the estimates $\widehat{\beta} = (\widehat{\beta}_0 \ldots, \widehat{\beta}_6)$ of the parameters in model [1].

Table 4. Generalized Linear Model Coefficient Estimates

β	$\widehat{\beta}$	**Standard Error**	**95% Confident Interval**	***p* - value**
β_0	-4.288	0.098	(-4.480,-4.097)	<0.001
β_1	3.530	0.106	(3.323,3.737)	<0.001
β_2	0.523	0.060	(0.405,0.641)	<0.001
β_3	3.633	0.105	(3.427,3.840)	<0.001
β_4	-0.008	0.000	(-0.009,-0.007)	<0.001
β_5	0.002	0.000	(0.001,0.002)	<0.001
β_6	0.002	0.000	(0.001,0.003)	<0.001

There are several interesting observations from Table 4: first, a negative $\widehat{\beta}_4$ entails that the death probabilities of two younger groups of both genders are decreasing functions of time *t*; secondly, $\widehat{\beta}_4 + \widehat{\beta}_5 = \widehat{\beta}_4 + \widehat{\beta}_6 = -0.006$ suggests that the death probabilities of the older groups of both genders are also decreasing functions of time *t*, but that the change is slower than that of the younger groups; the MR of the male older group is the largest and that of the female younger group is the smallest; for females, log odds ratio of MR between older and younger is $3.530 + 0.002t$, and for the males the log odds ratio of MR between older and younger is $3.110 + 0.002t$, which implies that the differences become larger and larger; the odds ratio between the largest MR of the male older group and the smallest MR of the female younger group is an increasing function of t, which is $\exp(3.633 + 0.002t)$, and changes, for example, from 37.902 at $t = 1$ to 68.924 at $t = 300$. The model [1] is applied to predict the mortality risk of a case based on age and gender at time *t*. A large value of $\widehat{p}_t$ (x) is associated with greater risk. From model [1], it is straightforward to show that the estimate $\widehat{p}_t$ (x) can be calculated by:

$$\widehat{p_t}(x) = \frac{\exp\big(w_t(x)\big)}{1 + \exp\big(w_t(x)\big)}, \qquad [4]$$

where:

$$w_t(x) = \widehat{\beta_0} + \sum_{i=1}^{3} \widehat{\beta_i}\, x_i + \widehat{\beta_4} t + \widehat{\beta_5} x_1 t + \widehat{\beta_6} x_3 t$$

with $\widehat{\beta}$ being the estimates in Table 4. The observed daily CRF π_t and predicted values $\widehat{p}_t$ in Figure 4 match each other well, and the male older group has been having the greatest risk since the outbreak.

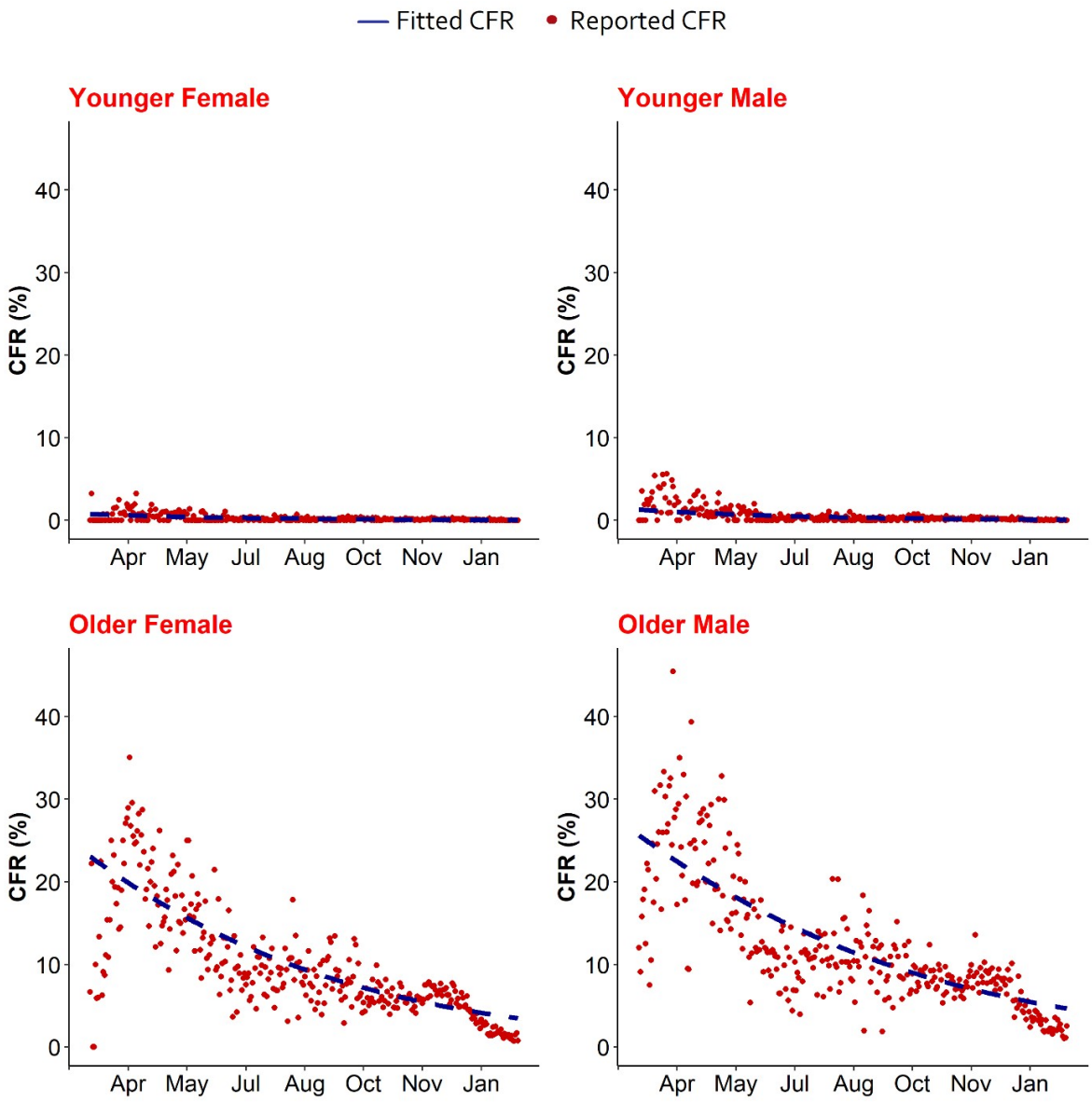

Figure 4. Case Fatality Rates π_t and Model Predicted Mortality Rates $\hat{p}_t$

A receiver operating characteristics (ROC) curve measures the accuracy of prediction. The higher a ROC curve is above the reference line y = x, the larger power it has. In other words, the closer to (0,1) the middle of the curve is, the more accurate the prediction using the model is.

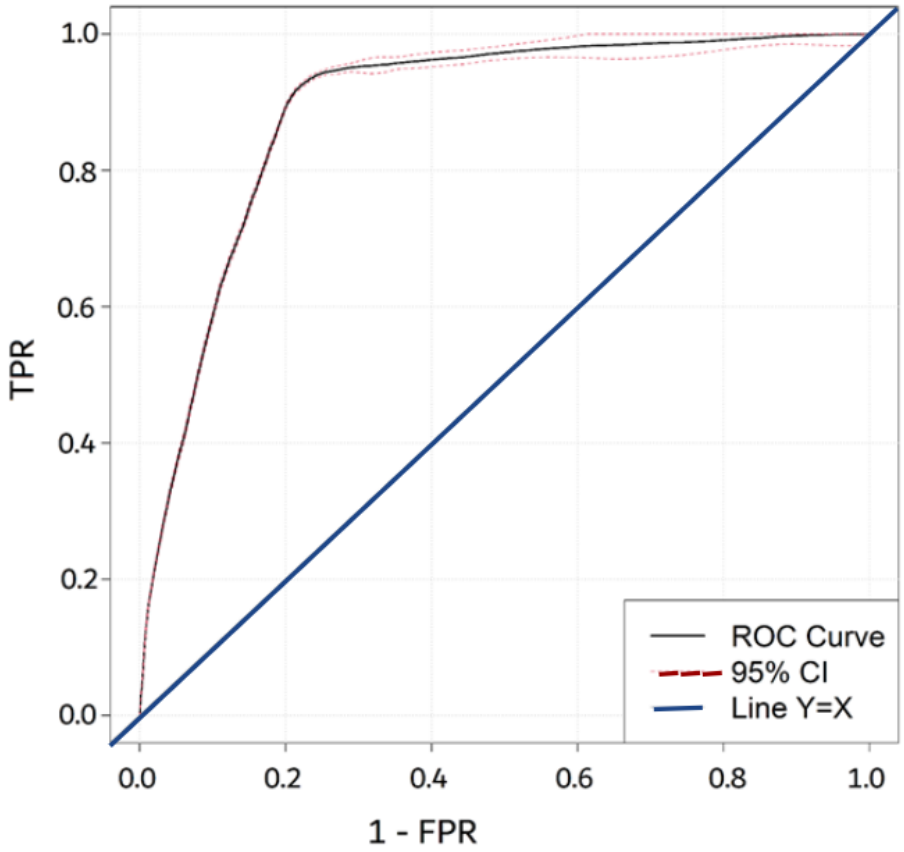

Figure 5. Case Fatality Rates π_t and Model Predicted Mortality Rates $\hat{p}_t$

Figure 5 is the receiver operating characteristics curve based on equation [4]. As early as 1966, Green and Swets (10) systematically introduced ROC curves and their applications. Figure 5 shows the prediction power of model [1]: it is within a 95% confidence interval and is high above the reference line y = x.

Another measure of prediction power is given by the areas under the curve (AUC). The AUC of model [1] is 89% close to one which is the largest possible value of AUC, and the 95% confidence interval is (88.67%,89.34%). Thus, both the ROC curve and the AUC indicate that the logistic regression model [1] is a powerful tool for prediction.

Discussion

The case fatality rate is one of the metrics that assess the severity of an infectious disease. The daily naive CFR r_t is constantly updated despite the fact that it is biased. According to the comparison based on the State of Ohio COVID-19 surveillance data, although r_t and π_t are different at the beginning of the COVID-19 outbreak, they share a common declining overall trend, and indicate the same most and least vulnerable groups. Therefore, r_t is informative despite its biasedness.

In the study, age, gender and time appear to be statistically significant in determining the likelihood of death for a case. In particular, the group of males older than 60 has been most vulnerable, which confirms a CDC recommendation. Moreover, the model that includes time, age, and gender provides a relatively high prediction accuracy as measured by the ROC curve and AUC. These findings are helpful in predicting outcome severity of certain groups and optimally allocating medical resources such as ICU's and ventilators.

This study has several limitations. First, our study relies on the Ohio surveillance data, and thus ignores unreported counts, such as asymptomatic patients. Secondly, outbreaks in clusters could have exaggerated the contagiousness. For example, many reported cases in nursing homes, could have resulted in an inflated total of reported case and death counts, given deaths in nursing homes in Ohio were about 32% of the total deaths by January 28, 2021 (25). Some research has been conducted for the purpose of estimating the society CFR. For example, Reich et al. (20) estimated death counts using log linear models by taking an incomplete reporting system into account; the work of Bendavid et al. (3) and Havers et al. (12) attempted to estimate society CFR in particular for COVID-19, by

sampling the population in certain geographical regions. Thirdly, although the logistic model [1] can explain the data reasonably well and shows strong power for prediction, pre-existing health conditions or comorbidities may be linked to the mortality rate and could improve model performance if such information was included. For example, Xu et al. (28) and Li et al. (14) studied how comorbidity contributed to the severity of COVID-19 patients' outcomes in China. Lastly, the prediction power could be enhanced if the record of some typical symptoms of each patient were accessible (4).

Conclusion

The proposed analysis procedure can be applied to similar COVID-19 data. For example, the national counterpart of r_t in Figure 6 exhibits the same changing pattern as that of the State of Ohio in Figure 3, and it is reasonable to conjecture that the proposed logistic regression is useful to modeling the national counterpart of π_t and p_t, which could be a future research project if such information was available.

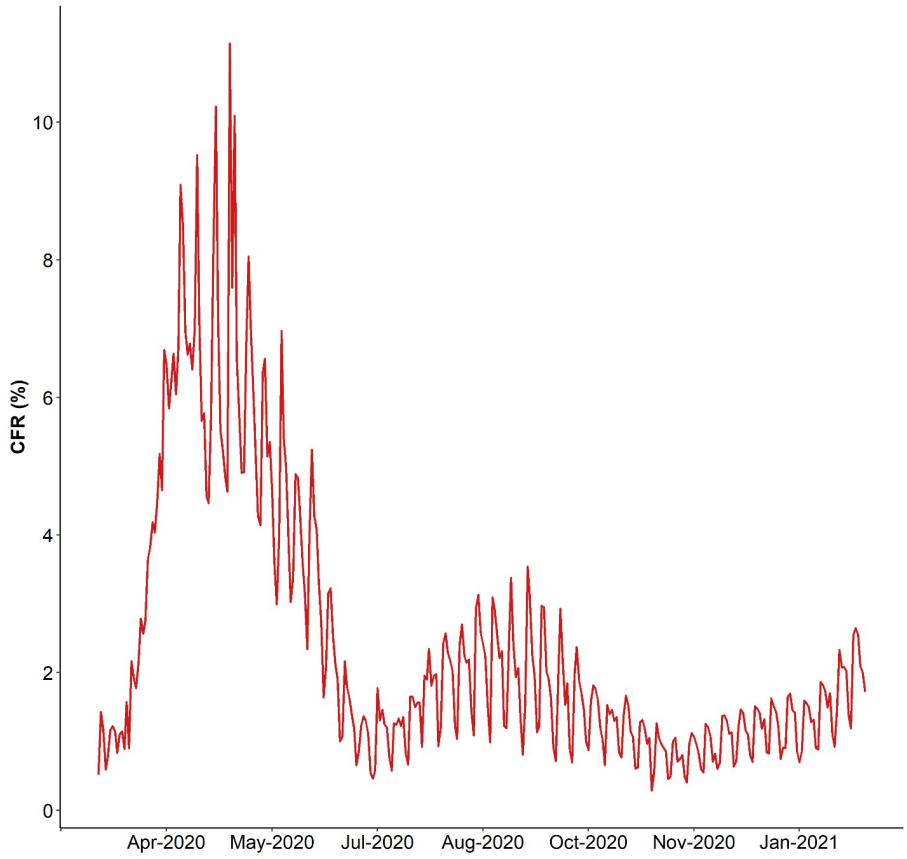

Figure 6: Case Fatality Rates rt of the United States, March 10, 2020 - January 31, 2021

Conflict of interest

Authors declare no conflict of interest.

Authors' contributions

HN performed data processioning and data analysis; QS reviewed literature and provided the significance of the research from the public health perspective. Both authors participated in revision of the manuscript, read and approved the final document.

1. Agresti A. Categorical Data Analysis, 2nd ed. New Jersey: Wiley-Interscience, 2002.
2. Angelopoulos AN, Pathak R, Varma R, Jordan, MI (2020) On identifying and mitigating bias in the estimation of the COVID-19 case fatality rate. Harvard Data Science Review [Internet]. 2020 Jul 16; Available from https://hdsr.mitpress.mit.edu/pub/y9vc2u36.
3. Bendavid E, Mulaney B, Sood N, et al. (2020). COVID-19 Antibody Seroprevalence in Santa Clara County, California. medRxiv, 2020.04.14.20062463.
4. Bertsimas D, Boussioux L, Cory-Wright R, et al. From predictions to prescriptions: A data-driven response to COVID-19. Health Care Management Science 2021 Jun 24(2): 253-272.
5. Bundgaard H, Bundgaard JS, Raaschou-Pedersen DET, et al. (2021) Effectiveness of Adding a Mask Recommendation to Other Public Health Measures to Prevent SARS-CoV-2 Infection in Danish Mask Wearers: A Randomized Controlled Trial. Annals of Internal Medicine 174: 335-343.
6. Chen N, Zhou M, Dong X, et al. (2020) Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. The Lancet 395: 507-513.
7. Chen Y, Liu Q, Guo D (2020) Emerging coronaviruses: genome structure, replication, andpathogenesis. Journal of Virology 92: 418-423.
8. Chiu WA, Fischer R, Ndeffo-Mbah ML (2020) State-level needs for social distancing and contact tracing to contain COVID-19 in the United States. Nature Human Behavior 4: 1080-1090.
9. Cortegiani A, Ingoglia G, Ippolito M, et al. (2020) A systematic review on the efficacy and safety of chloroquine for the treatment of COVID-19. Journal of Critical Care 57: 279-283.
10. Green DM, Swets JA. Signal Detection Theory and Psychophysics, New York: John Wiley and Sons, 1966.
11. Guan W, Ni Z, Hu Y, et al. (2019) Clinical characteristics of coronavirus disease 2019 in China. The New England Journal of Medicine 382: 1708-1720.
12. Havers FP, Reed C, Lim T, et al. (2020) Seroprevalence of antibodies to SARS-CoV-2 in 10 sites in the United States, March 23-May 12, 2020. JAMA Intern Med 180: 1576-86.
13. Kobayashi T, Jung S, Linton MN, et al. (2020) Communicating the risk of death from novel coronavirus disease (COVID-19). Journal of Clinical Medicine 9: 580.
14. Li X, Xu S, Yu M, et al. (2020) Risk factors for severity and mortality in adult COVID-19 inpatients in Wuhan. The Journal of Allergy and Clinical Immunology 146: 110-118.
15. Liu, Y., Gayle, A. A., Wilder-Smith, A., et al. (2020) The reproductive number of COVID-19 is higher compared to SARS coronavirus. Journal of Travel Medicine 27: 1-4.
16. McCullagh P, Nelder JA. Generalized Linear Models 2nd ed. Boca Raton: CRC, 1989.
17. Mizumoto K, Chowell G (2020) Estimating risk for death from coronavirus disease, China, January-February 2020. Emerging Infectious Diseases 26: 1251-1256.
18. 1Ohio Department of Health COVID-19 Dashboard https://coronavirus.ohio.gov/wps/portal/gov/covid-19/dashboards
19. 1R Core Team (2020) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
20. Reich NG, Lessler J, Cummings DAT, et al. (2012) Estimating absolute and relative case fatality ratios from infectious disease surveillance data. Biometrics 68: 598-606.
21. Ritchie H, Roser M (2020) What do we know about the risk of dying from COVID-19? https://ourworldindata.org/covid-mortality-risk
22. Ritchie H, Ortiz-Ospina E, Beltekian D, et al. (2020) Mortality risk of COVID-19. https://ourworldindata.org/mortality-risk-covid
23. Shao Q, Thompson G, Thompson A (2020) COVID-19 risk factor identification based on Ohio data. Translation: The University of Toledo Journal of Medical Sciences 8: 6-14.
24. Shen CY (2020) Logistic growth modelling of COVID-19 proliferation in China and its international implications. International Journal of Infectious Diseases 96: 582-589.
25. The COVID-19 Tracking Project. https://covidtracking.com/data/state/ohio/long-term-care
26. Zhou Y, Wang L, Zhang L, et al. (2020) A spatiotemporal epidemiological prediction model to inform county-level COVID-19 risk in the United States. Harvard Data Science Review [Internet] 2020 Aug 6; Available from: https://hdsr.mitpress.mit.edu/pub/qqg19a0r
27. Xu C, Dong Y, Yu X, et al. (2020) Estimation of reproduction numbers of COVID-19 in typical countries and epidemic trends under different prevention and control scenarios. Frontiers of Medicine 14: 613-622.
28. Xu K, Zhou M, Yang D, et al. (2020) Application of ordinal logistic regression analysis to identify the determinants of illness severity of COVID-19 in China. Epidemiology and Infection 148 e146: 1-11.
29. Zhang Q, Bastard P, Liu Z, et al. (2020) Inborn errors of type I IFN immunity in patients with life-threatening COVID-19. Science 370. 2020 Oct 23; 370(6515):eabd4570.