Innovationinfo

Journal of
Health Science and Development

ISSN: 2581-7310
Volume 4: 1
J Health Sci Dev 2021

Towards Understanding the COVID-19 Case Fatality Rate

Donghui Yan*[1]

Aiyou Chen[2]

Buqing Yang[3]

[1] *Department of Mathematics and Data Science, University of Massachusetts, Dartmouth, USA*

[2] *Independent Researcher*

[3] *Department of Actuarial Science, Shanghai University of Finance and Economics, Shanghai, China*

Abstract

An important parameter for COVID-19 is the case fatality rate (CFR). It has been applied to wide applications, such as measuring the severity of the infection, estimating the number of infected cases, risk assessment etc. However, there remains a lack of understanding on CFR, including relevant important population factors, the apparent discrepancy of CFRs across different countries, and how the age effect rolls in. We analyze CFRs at two different time snapshots, July 6 and Dec 28, 2020, during the first and second wave of the COVID-19 pandemic with the later just before the wide adoption of COVID-19 vaccines. Two important population covariates, age and GDP—as a proxy for the quality and abundance of public health—are considered. Our exploratory data analysis leads to interesting findings. There is a clear exponential age effect among different age groups, and, strikingly, the exponential index is almost invariant (0.0715 Vs 0.0704) across countries and over time during the pandemic. Meanwhile, the roles played by the age and GDP are a little surprising: during the first wave, age is a more significant factor than GDP, while their roles have switched during the second wave of the pandemic, which we attribute to the delay in time for the quality of public health to factor in.

Keywords: COVID-19, Case fatality rate, Age, GDP, Public health

Introduction

The COVID-19 pandemic has quickly reached a global scale, with the total confirmed cases at 96.24 million and death toll at 2.06 million as of Jan 18, 2021. An important parameter for COVID-19 is the case fatality rate (CFR), which is *defined as the ratio of the death toll and the number of infected cases*. The primary use of CFR is as a quantitative metric for the severity or lethality of the COVID-19 infection. It can be used as a reference in comparison to known infectious diseases such as the severe acute respiratory syndrome (SARS) or Ebola etc. An important application of CFR is to estimate the number of infected cases [1,2] through the death tolls, as it is commonly believed that the death toll is a relatively reliable quantity. It is also used as a proxy for risk assessment [3]. In order to apply the CFR properly, it is important to understand factors contributing to CFR. While it is clear that the mortality of COVID-19 is closely related to the health status or pre-existing conditions of an individual, these are not suitable to understand CFR at the population level, for example at the scale of a country. COVID-19 death is often mixed with various other

Article Information

Article Type: Research Article
Article Number: JHSD-133
Received Date: 15 April, 2021
Accepted Date: 23 June, 2021
Published Date: 30 June, 2021

***Corresponding author:** Donghui Yan, Department of Mathematics and Data Science, University of Massachusetts Dartmouth, 285 Westport Road, Dartmouth, MA 02747, USA. Tel: +(1)508-999-8746; Email: dyan@umassd.edu

Citation: Yan D, Chen A, Yang B (2021) Towards Understanding the COVID-19 Case Fatality Rate. J Health Sci Dev Vol: 4, Issu: 1 (30-38).

Copyright: © 2021 Yan D et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

diseases related to the lung or cardiovascular diseases etc. for an individual, which makes it challenging to characterize CFR at the population level. We need to understand CFR in terms of population parameters or covariates if we wish to understand the difference in CFRs across different countries.

The population parameter we are primarily interested in is the age. It has been acknowledged there is a strong age effect in the mortality among COVID-19 cases—while the CFR for the seniors is high, it would be very low for young people especially those below 30 years old. Such a sharp disparity is illustrated in figure 1 which shows the CFR by age groups for a number of countries; the countries are selected primarily due to the availability of the data and turn out to distribute fairly evenly over the world.

It can be seen that, the CFR for people younger than 30 is almost 0 while increasing very rapidly among those older than 60. Though differing in details, this pattern is fairly consistent for all countries shown in the figure. However, as a matter of fact, countries in the world differ significantly in terms of their age profile. For example, many countries in Africa have a median age of around 20, while a significant portion of European countries have a median age over 40. We expect that the CFR for a young population be smaller than a population where senior people dominate. If one can clarify the age effect in CFR, that will help understand potential discrepancy caused by different age structures across countries in comparing their CFRs, or to assess how well a particular country or region (termed broadly as country from now on for simplicity of description) is doing in *controlling* the CFR, or statistical inference on COVID-19 in one country using CFR related information from another.

Other relevant population parameters include the quality and abundance of medical service or public health, public policies, etc. The mortality of COVID-19 has been observed to be related to factors on the quality and abundance of health care and medical facilities, such as the number and capacity of hospitals and patient beds, testing coverage and accuracy, the quantity and quality of personal protection equipments, the experience of health workers and level of medical research on infectious disease etc. It is often challenging to quantify these or to access related data in many countries, and we will use the gross domestic production (GDP) per capita as a proxy for simplicity. Although there are limitations, the use of GDP per capita has often been used or mentioned in the literature [4-6] as a measure or an important factor for the well-being or healthcare quality of a population. A related work is [7], which considers the among-country variation of CFR in terms of age, GDP, and a number of other indicators for public health. However, this work only uses data up to Jun 11, 2020, and limits to European countries, with different findings.

We will carry out exploratory data analysis to investigate the role by age and GDP in CFR at the country level. We will start by considering the age effect, and then extend the analysis by including GDP. The remainder of this paper is organized as follows. In Section 2, we will describe the methods. This is followed by a presentation of data collection in Section 3 and the results in Section 4. Section 5 concludes the paper.

Methods

The observed CFR for a given population can be very noisy. For example, the death toll may be affected by the use of potentially different definitions in counting mortality, the difficulty in determining the exact cause of death when COVID-19 is mixed with other chronic diseases, as well as missing counts or inflation in the reported case mortality etc [2]. Meanwhile, the number of infected cases may be under-counted since it is limited to patients who have access to COVID-19 testing. Thus, the observed CFRs may be either over- or underestimated. We analyze observed CFRs by fitting regression models which would absorb all the noises into the error term. Note that this is a simplistic way to handle the noise or uncertainty with the reported infection or death counts. A more thorough approach would model the

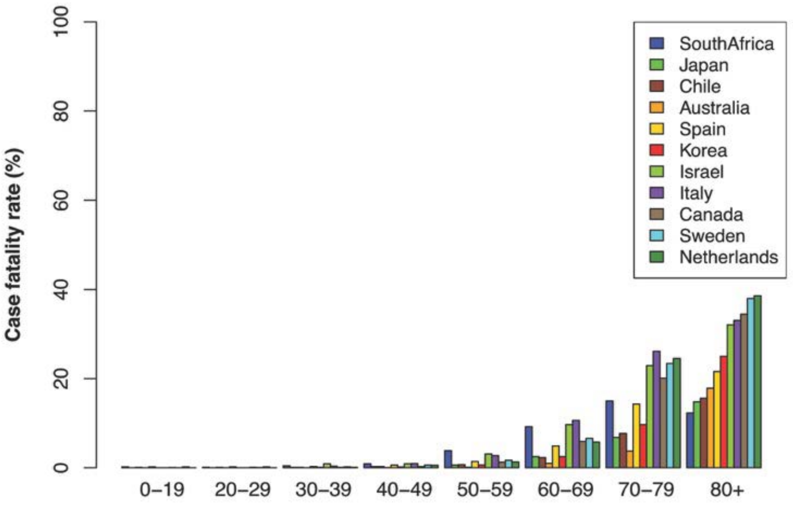

Figure 1: CFR by age groups for selected countries (as of July 6, 2020).

associated noise or uncertainty with relevant data, which are unfortunately not available for many countries in the world. Indeed, as our regression diagnosis reveals (c.f. Section 4.1), the error terms follow quite closely a normal distribution. Moreover, our goal is not to recover the underlying true CFR, but to unravel how age and GDP attribute to CFR across countries and over time. Our method is partially motivated by the observation made in figure 1, which shows that at crude level and in terms of the overall age trend, COVID-19 acts roughly similarly across different populations. The major population covariates under consideration are age and GDP.

The regression models can be expressed as

$$Y_i = f\left(X_i,\ \theta\right)\ +\ \varepsilon_i$$

where Xi and Yi stand for the population covariates and the observed CFR for the ith population, for i = 1, ..., n, θ is the parameter shared by all countries under consideration, and εi is used to model the noise in the observed CFR. Assume that Yi's are independent conditional on Xi. To be specific, we consider simple linear regression with f(X, θ) = θT X, which is powerful to discover strong main effects especially when the sample size is small.

Instead of using the CFR directly, we use the log-scale since the CFR appears to increase exponentially with the age as evident from figure 1. More directly, by visualizing CFRs in the log-scale as in figure 2, we see an almost linear increase (except for the age groups below 30) of the log-scaled CFR with the age. To better appreciate the magnitude of actual values of CFR for different age groups, we show as an example in table 1 the CFR by age groups in Canada.

Alternatively, one may consider the *Logit* transform, that is, convert CFR to log(CFR/(1 − CFR)). As the CFR's are typically quite small, it is similar to the log transform. Though different in details, the overall linear pattern is fairly consistent across different countries.

Data

The data we use in the analysis includes the following. The number of reported cases and the death toll are retrieved from the Worldometer [8], which we use to calculate the observed CFR for individual countries in the world. The median age is taken from Wikipedia [9]. The detailed age profile, i.e., percent by age groups, for countries is obtained from the United Nations web [10]. The GDP per capita data is also taken from the Worldometer [8]. Our initial analysis was carried out in the summer of 2020 using COVID-19 case data as of July 6, 2020. However, the pandemic had continued and deteriorated during the second half of the year. We were curious how that might impact our results. So we collected another snapshot of data, i.e., data sets as of Dec 28, 2020, also from the Worldometer. Note that Dec 28, 2020 is *also the time just before the wide adoption of COVID-19 vaccines*.

Results

In this section, we report results from the analysis on data collected on July 6, 2020 and Dec 28, 2020, respectively. We then make a comparison on these analyses, and report some interesting, maybe a little surprising, findings.

Analysis on data as of July 6, 2020

As of July 6, 2020, the observed CFR w.r.t. the median age for different countries is shown in figure 3. There appears to be an overall increasing trend of CFR with the median age in the population. We start by considering the following simple linear model

$$log\left(CFR\right)\ =\ \beta_0 + \beta_1 \cdot X\ +\ \varepsilon, \qquad (1)$$

where X is the median age of a population, and we term this as model I. In carrying out linear regression model fitting, we exclude countries with less than 3000 reported cases as the CFR for such populations would be very noisy. This leaves us a total of 99 observations (i.e.,countries) for linear regression; their total number of reported cases

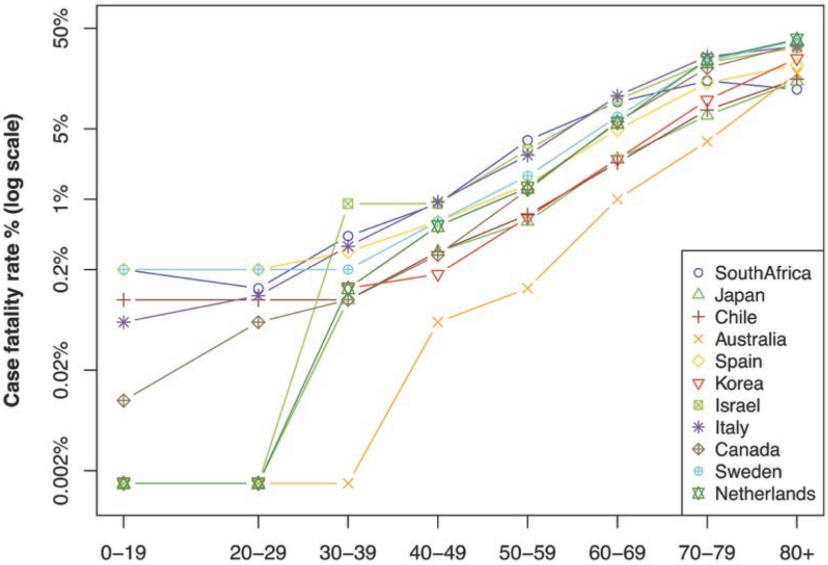

Figure 2: Log-scaled CFR by age groups for selected countries as of July 6, 2020.

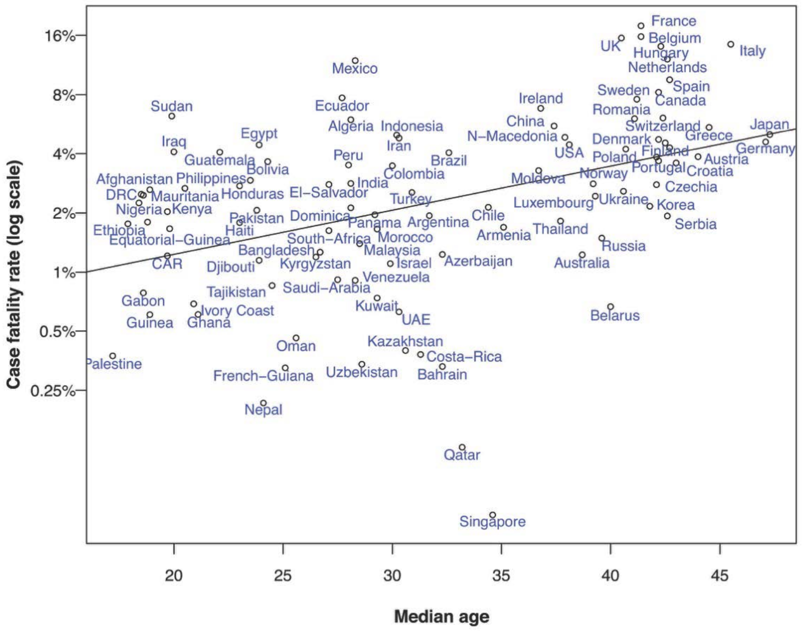

Figure 3: Scatter plot of CFR by median ages for individual countries as of Jul 6, 2020. The solid line is the regression line.

Age	0-19	20-29	30-29	40-49	50-59	60-69	70-79	80+
CFR	0.01%	0.06%	0.10%	0.28%	1.24%	5.59%	20.10%	34.42%

Table 1: CFR by age groups in Canada as of July 6, 2020.

Model		**2020/07/06**	**2020/12/28**
Model I	Age	0.0516 (1.9100e-5)***	-1.6300e-3 (0.8020)
	R^2	0.1726 (0.1640)	4.1500e-4 (6.1600e-3)
	F-stat	20.2300 (1.9100e-5)***	6.3400e-2 (0.8019)
Model II	GDP	7.2900e-6 (0.1460)	-0.3309 (6.0600e-3)**
	R^2	0.0217 (0.0116)	0.0491 (0.0428)
	F-stat	2.1520 (0.1457)	8.7520 (6.0600e-3)**
Model II	Age	7.1400e-2 (2.8100e-6)***	1.7800e-2 (0.0325)*
	GDP	-0.5537 (0.0284)*	-0.5453 (5.2500e-4)***
	R^2	0.2132 (0.1968)	0.0780 (0.0656)
	F-stat	13.0000 (1.0100e-5)***	6.2990 (2.3700e-3)**

Table 2: Regression coefficients and p-values (in the parentheses) under different models for data during the first wave and second wave of pandemic.

is 11,471,724 with a total death toll of 534,347. The fitted model parameters are

$$\beta_0 = -5.4288, \beta_1 = 0.0516,$$

with a reported R2 at 0.1726 (adjusted 0.1640), and a p-value of 1.9100 × 10−5 on F-test. All the coefficients are statistically significant with a p-value less than 1.9100 × 10−5. The fitted regression line is added as the solid line in figure 3. As expected, the estimated CFR increases with the age of a population. Observed CFR in many countries indeed follow this trend.

With model (1), we can estimate CFR for individual countries. For example, the CFR for USA, India, China and Korea are estimated as 3.13%, 1.87%, 3.02% and 3.19%, close to estimates at 2.85% given by [11], 2.20% by [12], 2.30% by [13], 2.36% by [14], respectively. The worldwide CFR is estimated to be 2.76%, close to the WHO published 3.40% as of Mar 2020; in contrast, a direct calculation from the reported cases and death toll would give 4.66%. A country that stands out is Singapore which has an extremely low observed CFR, given its above average median population age. We attribute this to the small size of this country and the painstaking efforts dedicated by its government in combating the pandemic.

In the linear regression analysis, we make two assumptions. These include the assumption of normality and of the constant variance. To validate these assumptions, we carry out some regression diagnostic analysis [12]. Figure 4 visualizes our results. The QQ-norm plot shows that, approximately, the regression residuals follow a normal distribution. We further perform a Kolmogorov-Smirnov test [15] of the regression residuals against a standard normal, which supports normality at a p-value of 0.3740. Next, we look at the constant variance assumption. The residual plot shows that the regression residuals have a roughly constant spreadout over the range of median ages. The Cook-Weisberg's constant variance test [16] gives p-value 0.8909, which suggests the compatibility of the data to homoscedasticity.

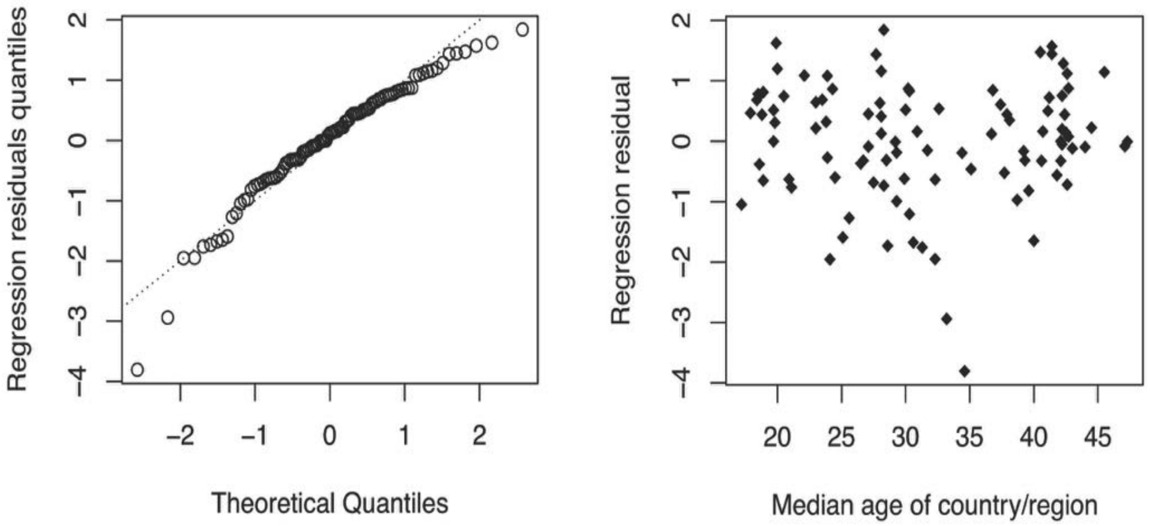

Figure 4: Regression diagnostics plots under Model I. The left and right panel are the QQ-norm plot of regression residuals and the residual plot, respectively. The dashed line in the QQ-plot is the qqline.

Next, we extend the above analysis by adding the GDP covariate, that is

$$log(CFR) = \beta_0 + \beta_1 \cdot X + \beta_2 \cdot GDP + \varepsilon, \qquad (2)$$

where X is the median age of a population. We term (2) as Model III. The GDP is coded as 1 if it is smaller than $10,000 per capita and 2 otherwise; the cutoff value of $10,000 is close to that (i.e., $12,000) used in determining if a country is a developing or developed country by the United Nation (indeed a cutoff value anywhere between $8,000 and $15,000 makes very little difference in our analysis). This is consistent with our understanding that, as long as the GDP per capita is above a certain threshold, it no longer has a major impact to the quality of healthcare. The fitted model parameters are

$$\beta_0 = -5.2550,\ \beta_1 = 0.0714,\ \beta_2 = -0.5537,$$

with a reported R2 at 0.2132 (adjusted 0.1968), and a p-value of 1.0060 × 10−5 on F-test. Using the original GDP value would lead to a slightly inferior model fit (with R2 at 0.1851). The coefficient for the age is statistically significant with a p-value less than 2.8100 × 10−6, but that for the GDP is not as significant with a p-value of 0.0284. It should be noted that the use of relative size of p-values in linear regression for factor significance is not necessarily a perfect measure; it is just a simple metric that is easy to operate on while having an apparent statistical interpretation.

Analysis on data as of Dec 28, 2020

Similar to the analysis on data as of July 6, 2020 in Section 4.2, we carry out analysis on data as of Dec 28, 2020, for which the total number of reported cases is 81, 597, 946 (more than 7 times of the July data) with a total death toll of 1,779,448 (slightly more than 3 times of the July data). An overall observation is that most countries have a reduced observed CFR than that by the July 6 data. This is consistent with a widely acknowledged view that the CFR gradually drops with the on-going of the pandemic after certain stage. For example, the observed CFR for the US is 5.56%, 5.43%, 4.14%, 3.09%, 2.87%, 2.70%, 2.35%, 1.87% as of May 6, June 6, through Dec 6, 2020, respectively. This could be due to various reasons: the population handles COVID-19 better and better after learning from early lessons, further mutations of the COVID-19 virus may have caused it to be less lethal over time, or simply because of the lack of enough testings in earlier stages (which in the analysis is assumed to be uniformly distributed across the age groups, but not over time).

We start by considering the effect of age on the CFR, using model (1). The result was a little surprising, and the median age of the population barely plays a role in the linear regression which finishes with an R2 almost 0, i.e., 4.1520e-4, and the p-value associated with the F-test at 0.8020. To get sense on why this is the case, we plot the observed CFR for individual countries in figure 5.

To facilitate easy comparison, we also include the observed CFR for data as of July 6, 2020. Figure 5 is quite revealing, and we see that most of the countries with a high CFR as of July 6, 2020 have seen a sharp decrease in their CFRs by Dec 28, 2020, while the decrease is marginal (or even increase a little) for those countries with a previously low CFR. The decreasing trend is most significant for countries with a relatively high median age.

We then consider model (2), and model fitting on Dec 28 data leads to a reported R2 at 0.0780 (adjusted 0.0656), and a p-value of 2.3660 × 10−3 on F-test. The fitted model parameters are

$$\beta_0 = -3.9261,\ \beta_1 = 0.0178,\ \beta_2 = -0.5453.$$

The GDP is statistically significant with a p-value 5.2500×10−4, but the age is not as significant with a p-value of 0.0325. Similarly, we have produced the diagnostics as before which suggest that the regression residuals have a roughly constant variance over the range of fitted values except with a moderate departure from normality. Linear

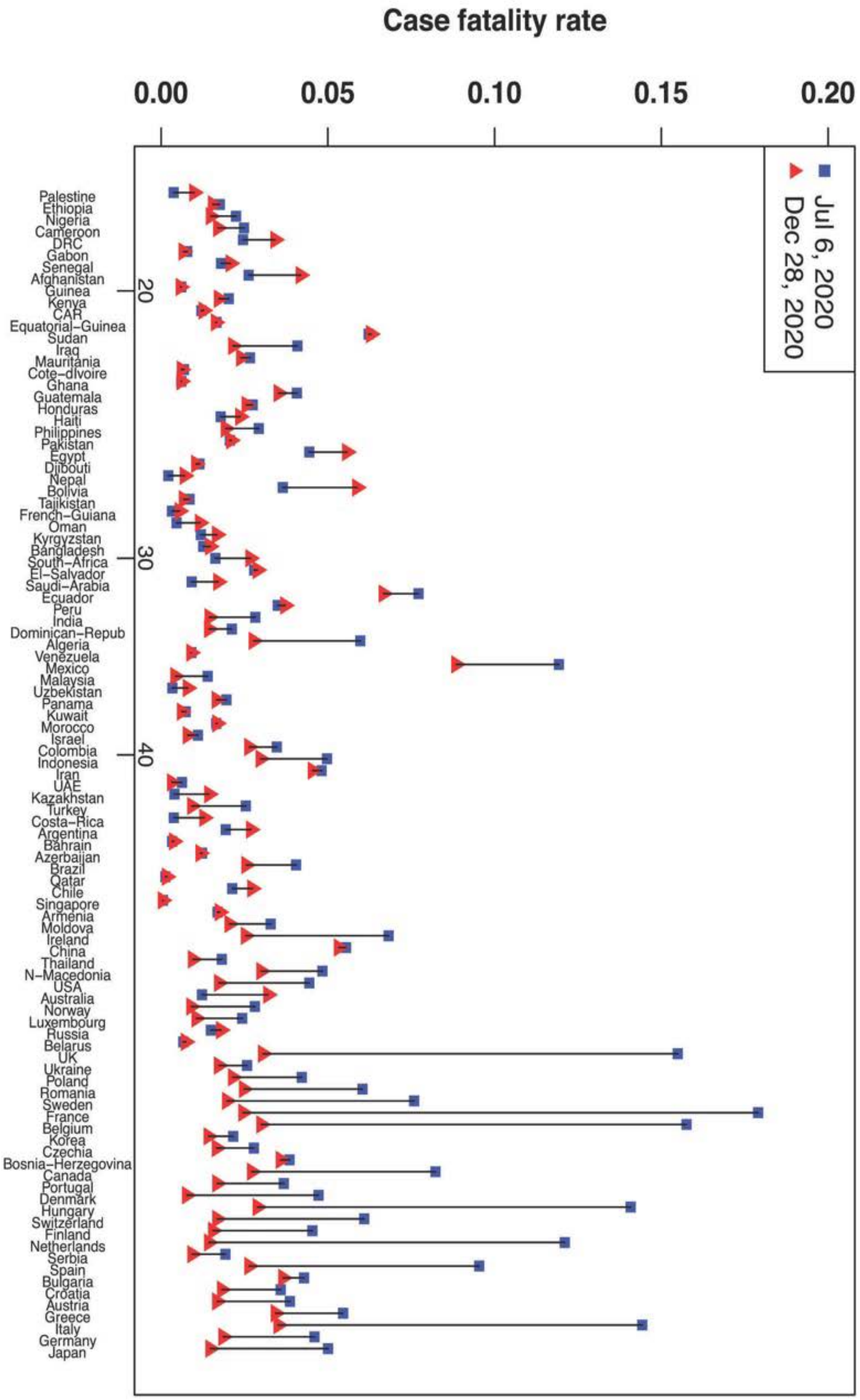

Figure 5: Changes in CFR from July 6, 2020 to Dec 28, 2020. The countries are sorted by median ages in an increasing order from left to right in the figure. The numbers on the x-axis are the median age.

regression using the original GDP leads to slighter lower R2. The effect of GDP on CFR can be visualized from figure 6, and higher GDP leads to a lower CFR. This is consistent with our understanding, as higher GDP typically implies better public health and medical facilities.

Findings in comparing the two analysis

We have carried out analysis of the CFR with the same models for COVID-19 data taken at two different time snapshots. Much has happened during the time, with a fast increasing and then slowing down pattern of the pandemic in different countries during the summer, followed by the general upward trend into the winter. It will be interesting to compare the results we have obtained. To facilitate our comparison, we summarize our results in table 2.

One particularly interesting observation is the reversing roles played by the two population covariates—age and GDP. Age is a significant covariate in the July 6 data, but no longer as important in the Dec 28 data; GDP is not an important covariate in the July 6 data but becomes significant in the Dec 28 data. What causes this? Our interpretation is that, by July 6, 2020, most of the countries are still trying to understand the mechanism of COVID-19 and exploring and learning how to effectively deal with COVID-19, so the quality of public health and abundance of medical facilities have not yet been reflected in the CFR; rather the more fundamental factor—the age—played a major role at this stage. As time goes by, both the public and health workers are gaining experiences in the handling and treating of COVID-19, so the quality of medical care has picked up and becomes a major factor in the CFR of a country; by this time, the age effect starts to shrink. Note that such a statement applies when we attempt to compare CFRs of many countries simultaneously. Our finding is consistent with analysis in [17] which claims no evidence of age-specific CFR changes up to mid Aug 2020. As from the beginning of the COVID-19 pandemic till mid Aug 2020, the age, at the country scale, stays the same thus our analysis would not imply major changes to CFRs (different story once the effect of public health sets in). On the other hand, there are a number of countries with marginal changes in CFRs till Dec 2020 as shown in figure 5 and such countries are typically those with less developed public health and thus the age plays a major role; if the age does not change much—which is true—then the CFRs will not change much either by our analysis, which is consistent with findings in [17,18].

Can we claim that the age effect is mostly disappearing after nearly a year since the start of the pandemic? This motivates our analysis in Section 4.4.

Invariance of the age effect in CFR

To answer the question posed in Section 4.3, we will look at CFR by age groups and by countries. This will help get rid of the country effect in CFR due to the difference in their population age structures, and also to standardize many other factors caused by differences among countries. For simplicity and constrained by the availability of the data (unfortunately, for most of the countries in the world, such statistics breakdown by age groups are not available), we will use the same 11 countries that we use to produce figure 1 and figure 2 based on the July 6 data. We will additionally analyze the CFR by age groups for these 11 countries using data around Dec 28, 2020.

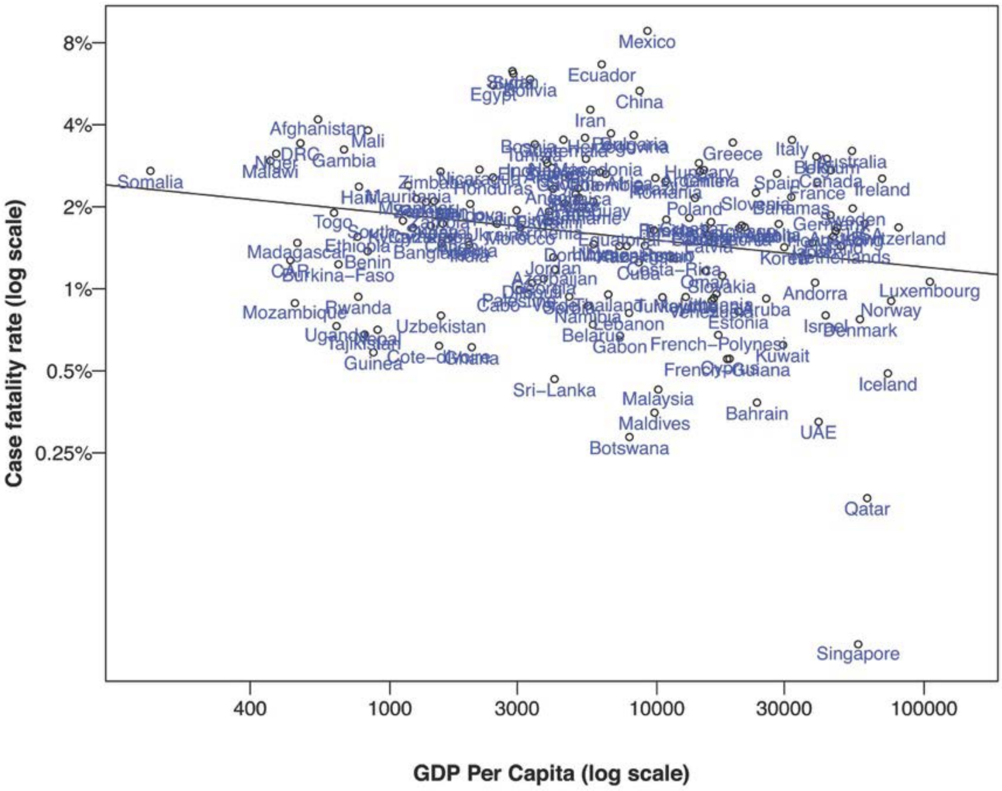

Figure 6: Scatter plot of CFR by GDP per capita for individual countries as of Dec 28, 2020. The solid line is the regression line.

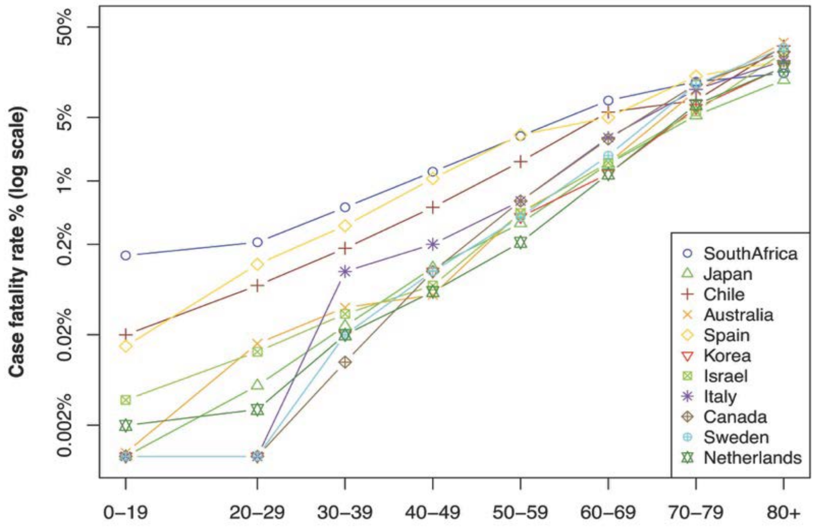

Figure 7: Log-scaled CFR by age groups for selected countries as of Dec 28, 2020.

We first carry out a simple linear regression on CFR (in log scale) versus age groups for the 11 countries involved similar as Model (1), except that we now treat each age group in a country as an instance of data. As the ages are given as a range, we take the middle of the age groups, i.e., 10, 25, 35, ..., 75, and 85, in linear regression. This leads to a fairly good fit to the linear model on the July 6 data, with the estimated coefficients as the following

$$\beta_0 = -3.006,\ \beta_1 = 0.0715,$$

and a reported R2 at 0.9102 (adjusted 0.8952) and p-value less than 2.3400e-4 for the F-test. So the age effect is significant, and in particular, there is an exponential increase in CFR with the moving up through the age groups.

A similar regression analysis is carried out using data as of Dec 28, 2020, from the same 11 countries. The model fits the data well, with a reported R2 at 0.9730 (adjusted 0.9685), and a p-value of 6.2000e-6 on the F-test. The fitted intercept and slope are as follows

$$\beta_0 = -2.9076,\ \beta_1 = 0.0704,$$

which are surprisingly close to that on the July 6 data.

Though the comparison between figure 2 and figure 7 suggests that the exponential age effect appeared more homogeneous across countries for the Jul 6 data and less so for the Dec 28 data, we see the same exponential age effect with almost the same exponential factor between age groups despite that the data are separated about a half year apart. This suggests that the exponential age effect is mostly invariant regardless of countries and time. Given that the 11 countries have a wide spectrum of median ages, ranging from 27.1 to 47.3, and GDP per capita, ranging from $6,120 to $54,075 per year. We expect such an invariance to widely hold across countries.

Conclusion

We have analyzed the CFR for countries in the world by including population covariates such as age, and GDP as proxy for the quality and abundance of healthcare. This allows us to understand the roles played by age and GDP in the apparently discrepant CFRs across countries despite the limitation of data accuracy. By analysis of data collected at two separate time snapshots, July 6 and Dec 28, 2020, we have arrived at some interesting findings. During the initial stage of pandemic, age is a significant factor in explaining discrepancy in CFR across countries while GDP plays a lesser role, and then as the pandemic continues with the public and health workers gradually gaining experience in handling and treating COVID- 19, GDP becomes a more significant factor than age. However, the exponential age effect is largely invariant across countries which are clearly exhibited on both data with nearly identical estimated exponent.

Acknowledgement

We thank all those who contributed to the collection, curation and sharing of data related to COVID-19.

References

1. Gupta S, Shankar R (2020) Estimating the number of COVID-19 infections in Indian hot-spots using fatality data. arXiv.
2. Jagodnik KM, Ray F, Giorgi FM, Lachmann A (2020) Correcting under-reported COVID-19 case numbers: estimating the true scale of the pandemic. medRxiv.
3. Schr ¨oder I (2020) COVID-19: A risk assessment perspective. ACS Chemical Health and Safety 27: 160-169.
4. Becker G, Philipson T, Soares R (2005) The quantity and quality of life and the evolution of world inequality. The American Economic Review 95: 277-291.
5. Oulton N (2012) Hooray for GDP. Occasional Paper 30, Centre for Economic Performance, London School of Economics and Political Science.
6. Raghupathi V, Raghupathi W (2020) Healthcare expenditure and economic performance: Insights from the United States Data. Front Public Health 8: 156.
7. Sorci G, Faivre B, Morand S (2020) Explaining among-country variation in COVID-19 case fatality rate. Scientific Reports 10: 18909.
8. Worldometer (2020) COVID-19 Coronavirous Pandemic.

9. Wikipedia (2020) List of countries by median age.

10. United Nations (2019) World Population Prospects.

11. Yan D, Xu Y, Wang P (2021) Estimating the number of infected cases in COVID-19 pandemic. Journal of Data Science 19: 348-364.

12. Philip M, Ray D, Subramanian S (2020) Decoding India's Low Covid-19 Case Fatality rate. Working Paper 27696, Natioanal Bureau of Economic Research.

13. von K¨ugelgen J, Gresele L, Sch¨olkopf B (2016) Simpson's paradox in Covid-19 case fatality rates: a mediation analysis of age-related causal effects. arXiv.

14. Rice JA (1995) Mathematical Statistics and Data Analysis. Duxbury Press.

15. Chakravarti I, Laha R, Roy J (1967) Handbook of Methods of Applied Statistics-Volume I. John Wiley and Sons.

16. Cook JD, Weisberg S (1983) Diagnostics for heteroscedasticity in regression. Biometrika 70: 1-10.

17. Signorelli C, Odone A (2020) Age-specific COVID-19 case-fatality rate: no evidence of changes over time. International Journal of Public Health 25: 1-2.

18. Shim E, Mizumoto K, Choi W, Chowell G (2020) Estimating the risk of COVID-19 death during the course of the outbreak in Korea, February-May, 2020. Journal of Clinical Medicine 9: 1641.