Clustering of age standardised COVID-19 infection fatality ratios and death trajectories

Thu-Lan Kelly¹,²*, Greer Humphrey¹, Caroline Miller¹,³, Jacqueline A Bowden³,¹, Joanne Dono¹, Paddy A Phillips⁴

¹South Australian Health and Medical Research Institute, South Australia, Australia
²University of South Australia, Clinical and Health Sciences, South Australia, Australia
³University of Adelaide, School of Public Health, South Australia, Australia
⁴SA Commission on Excellence and Innovation in Health, South Australia, Australia

*Corresponding author
E-mail: Lan.Kelly@sahmri.com (TK)

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.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

Background

An accurate measure of the impact of COVID-19 is the infection fatality ratio, or the proportion of deaths among those infected, which does not depend on variable testing rates between nations. The risk of mortality from COVID-19 depends strongly on age and current estimates of the infection fatality ratio do not account for differences in national age profiles. Comparisons of cumulative death trajectories allow the effect and timing of public health interventions to be assessed.

Our purpose is to (1) determine whether countries are clustered according to infection fatality ratios and (2) compare interventions to slow the spread of the disease by clustering death trajectories.

Methods

National age standardised infection fatality ratios were derived from age stratified estimates from China and population estimates from the World Health Organisation. The IFRs were clustered into groups using Gaussian mixture models. Trajectory analysis clustered cumulative death rates in two time windows, 50 and 100 days after the first reported death.

Findings

Infection fatality ratios from 201 nations were clustered into three groups: young, medium and older, with corresponding means (SD) of 0.20% (0.03%), 0.38% (0.11%) and 0.93% (0.21%).
At 50 and 100 days after the first reported death, there were two clusters of cumulative death trajectories from 113 nations with at least 25 deaths reported at 100 days. The first group had slowly increasing or stable cumulative death rates, while the second group had accelerating rates at the end of the time window. Fifty-two nations changed group membership between the time windows.

**Conclusion**

A cluster of younger nations have a lower estimated infection fatality ratio than older nations. The effect and timing of public health interventions in preventing the spread of the disease can be tracked by clustering death rate trajectories into stable or accelerating and comparing changes over time.
Introduction

The disease COVID-19 caused by the coronavirus SARS-CoV-2 was first described in Wuhan, China in December 2019 [1-2]. The impact the disease will have on the global population has been estimated through the case fatality rate (CFR), or the proportion of deaths among confirmed cases. Since the number of cases depends on testing rates which may include only symptomatic or severe cases, the CFR may be an over-estimate of the impact of the disease.

In contrast, the infection fatality ratio (IFR) is the proportion of deaths among infected individuals and is a more accurate estimate of disease mortality. However, it is difficult to determine the true number of infections in a population. Recent antibody prevalence studies have attempted to establish infection rates in the USA, Spain and elsewhere [3-5]. A recent meta-analysis calculated an IFR of 0.68 % (95% CI 0.53-0.82%), but with significant heterogeneity [6]. The authors concluded that different regions may experience different IFRs due to age structure and underlying co-morbidities and called for more research on age stratified IFRs. IFRs for China [7] and Italy [8] have been estimated both by age group and overall. Although the age stratified IFRs show a large increase in risk for older age groups in both countries, since Italy has an older population than China, overall IFRs vary from 0.66% (95% credible interval 0.39-1.33%) in China to 1.29 % (95% crI 0.89-2.01%) in Italy. An IFR of below 0.5% was ruled out in populations with more than 30% over 60 years old [8].

While age is a significant risk factor for mortality rates, the spread of infection through populations also depends on public health interventions such as physical distancing measures, mask wearing, testing, contact tracing, quarantine and border controls [9]. Comparing trajectories of death rates makes it possible to use real-world data to assess the impact of such
public health interventions on disease spread between nations. Clustering groups of countries with similar cumulative death rate trajectories over different time windows enables comparisons of the timing and effectiveness of such interventions.

Methods

Age standardisation of IFRs estimated from China by Verity et al [7] was performed using a weighting method (see Supporting Material). The overall estimate of the IFR from China (0.66% [7]) was multiplied by each nation’s weight to obtain a point estimate of the age adjusted IFR.

Model based clustering with Gaussian mixture models was used to cluster groups of nations with IFRs arising from the same normal distribution, using the ‘mclust’ R package version 5.4.6 [10]. Estimates of the mean IFR, SD and bootstrapped 95% confidence intervals of the mean were determined for each distribution.

To investigate factors related to death rates that are independent of national age profiles, such as public health interventions, the next stage was to analyse death rate trajectories. If infection rates are assumed to be equal across groups and age stratified IFRs relative to those in the 80+ age group are assumed to be the same for every nation, then the IFR weights can be used to age standardise the death rates per population. The age stratified IFRs relative to the 80+ age group in China are mostly in agreement with those estimated for Italy (S1 Table in the Supporting Material). Weighting reduced the effect of age on the trajectories so that other factors involved in the impact of the disease over time could be investigated. Trajectory analysis of cumulative death rates per population was used at two time points: 50 and 100 days after the first reported death from COVID-19. To ensure stable trajectories, only
countries with at least 25 deaths reported by 100 days after the first death were included in
the analysis. Rolling 14-day averages of weighted cumulative deaths rates were smoothed
using splines to reduce the effect of outlying data points. The R package ‘traj’ version 1.2,
which combines principal components of statistical measures of growth and cluster analysis,
was used to cluster cumulative death rate trajectories into groups, without requiring the
number of clusters to be determined a priori [11].

The R package ‘rworldmap’ version 1.3-6 [12] was used to visualise the IFR and trajectory
clusters on a global scale. R software version 4.0.2 (R foundation for Statistical Computing,
Vienna, Austria) was used for all analyses.

Data

Estimates of national populations in 2020 by five-year age groups from the World Health
Organisation were available from [13]. Daily cumulative death rates compiled by the
European Centre for Disease Control were obtained from the Our World in Data website
[14]. All datasets and R code used to produce the results are available from
https://github.com/lan-k/COVID19.

Results

Age adjusted IFRs were calculated for 201 countries. If the national IFRs were assumed to be
from one normal distribution, the mean IFR would be 0.54% (SD 0.34%); however, a
histogram of the IFRs showed the data were not drawn from a single normal distribution (Fig
Figure 1: Histogram of global Infection Fatality Ratios

Clustering of the IFRs produced three groups of nations with young, medium and older age profiles. Model fit diagnostics can be found in the Supporting Material (S1 and S2 Figs). Mean IFRs (SD) from the three distributions are 0.20% (0.03%), 0.38% (0.11%) and 0.93% (0.21%) (Table 1). Bootstrapped 95% CIs for the mean and SD of the three normal distributions and the minimum and maximum IFR in each cluster are also presented in Table 1. The countries included in each cluster are displayed in Fig 2. After excluding countries in the ‘Young’ cluster, the mean IFR (95% CI) from the remaining countries, assuming the data were from a single normal distribution, was 0.67% (0.62-0.72%), which is very close to the meta-analysis estimate of 0.68% (95% CI 0.53-0.82%) [6].

Table 1: Characteristics of the three IFR clusters.

| Cluster (N) | Age profile | Mean IFR (%) (95% CI) | SD of distribution (%) (95% CI) | IFR range (%) (min, max) |
|------------|-------------|-----------------------|---------------------------------|--------------------------|
| 1 (55)     | Young       | 0.20 (0.18-0.23)      | 0.03 (0.02-0.05)                | (0.14-0.25)              |
| 2 (75)     | Medium      | 0.38 (0.30-0.48)      | 0.11 (0.05-0.15)                | (0.25-0.61)              |
| 3 (71)     | Older       | 0.93 (0.78-1.03)      | 0.21 (0.13-0.28)                | (0.62-1.51)              |

Figure 2: National membership of three IFR clusters. Countries with missing data are shown in white.
For the trajectory analysis, cumulative death rates were available for 113 countries with at least 25 deaths 100 days after the first reported death, as of July 21, 2020. Trajectory analysis clustered countries based on the growth of weighted cumulative death rates over time and the groups are independent of the IFR clusters in Fig 2. Two groups were found; ‘Stable’ and ‘Accelerating’ (Table 2). Cluster group membership was based not on cumulative death rates at the end of the period, but on the shape of the trajectory. The first group, ‘Stable’, had cumulative death rates which had plateaued or slowly increased towards the end of the time window, while the second group, ‘Accelerating’, showed death rates which were rapidly increasing. Details of the statistical measures which defined the clusters at each time window and diagnostics can be found in S2 Table and S3 and S4 Figs in the Supporting Material. The median trajectories and interquartile ranges for each group at 50 and 100 days after the first reported death are shown in Fig 3. Large IQRs at 100 days compared with 50 days possibly reflect the timing of interventions between the time windows. Cluster group membership did not change in a sensitivity analysis using unweighted trajectories, since clusters were defined by the shape of the trajectories. Fig 4 shows national group membership of the two trajectory clusters at 50 and 100 days after the first reported death. Each trajectory cluster includes a mixture of young, medium and older nations, indicating that the spread of disease in a population over time depends on other factors besides age.

Table 2: Median cumulative death rate and interquartile range for clusters at 50 and 100 days after the first reported death.

| Cluster | 50 days (N) | 100 days (N) | Median cumulative death rate at 50 days (IQR) (per 100k) | Median cumulative death rate at 100 days (IQR) (per 100k) |
|---------|-------------|-------------|--------------------------------------------------------|--------------------------------------------------------|
| Stable    | 68 | 70 | 7.99 (3.72-15.8) | 30.9 (10.1-110.7) |
|-----------|----|----|-----------------|------------------|
| Accelerating | 45 | 43 | 18.6 (3.90-49.6) | 28.1 (11.8-83.1) |

**Figure 3: Median cluster trajectories.** Median trajectories (solid line) and interquartile range (dashed line) for the ‘Stable’ (blue) and ‘Accelerating’ (red) clusters at (A) 50 days and (B) 100 days after the first reported death.

**Figure 4: National trajectory cluster group membership.** Cluster group membership of the ‘Stable’ (blue) and ‘Accelerating’ (red) clusters at (A) 50 days and (B) 100 days after the first reported death. Countries excluded from the analysis are shown in white.

Between the two time windows, 52 out of 113 nations (46%) changed group membership (Fig 5). 25 countries moved from ‘Stable’ to ‘Accelerating’ (the ‘Worse’ group), while 27 changed from ‘Accelerating’ to ‘Stable’ (the ‘Improved’ group). Again, the change in group membership between the time points did not depend on the age profile of countries.

**Figure 5: Change in trajectory membership over time.** Change in group membership of trajectory clusters between 50 and 100 days after the first reported death. The ‘Stable’ (blue) and ‘Accelerating’ (red) groups remained unchanged between time points. Between 50 and 100 days, ‘Improved’ (purple) changed from ‘Accelerating’ to ‘Stable’, while ‘Worse’ (orange) changed from ‘Stable’ to ‘Accelerating’.
Discussion

The risk of death from COVID-19 is highly dependent on age. Current estimates of the IFR are between 0.5-1.3%, are difficult to calculate and assume a single value will describe the global impact of the disease. There have been few studies reporting IFRs for younger nations such as in Africa, possibly due to difficulties in testing, measuring the number of asymptomatic infections and reporting accurate death rates. It has been suggested the different regions will experience different IFRs due to age structure and co-morbidities [6, 15]. A single estimate of the IFR for all nations may not capture the true global distribution.

Our study has shown that national IFRs are not drawn from a single normal distribution, but from a mixture of three distributions with different means and standard deviations. This would explain some of the heterogeneity in the IFRs reported [6]. When countries from the ‘Young’ cluster were excluded, the mean of the remaining national IFRs, assuming a single normal distribution, is very close to the meta-analysis estimate. If data from younger countries become available, they may confirm our findings. While younger nations may have a lower IFRs from age alone, these countries may have less developed health systems and poorer health status, so the actual infection fatality ratio in these nations may be higher than that estimated due to age alone.

Mortality risk from COVID-19 does not depend on age alone. Apart from relatively immutable risk factors in a population, such as gender and co-morbidities, implementation of public health measures can make an important difference to the increase in deaths. Trajectory analysis showed the timing and effect (or lack) of implementation of public health interventions, independent of national age profiles. Each trajectory cluster included a mixture of nations from the IFR clusters, indicating that age alone does not explain the full impact of the disease on a population. At 50 days after the first reported death, North America, most of
Europe and Asia, parts of Central and South America and Australia were experiencing accelerating death rates. By 100 days, North America, China, Europe and Australia (the ‘Improved’ group in Fig 5) had stabilised their death trajectories through public health interventions such as lockdowns, increasing testing rates, contact tracing and border controls. In contrast, the number of countries with accelerating death rates in Central and South America, the Middle East and Africa (the ‘Worse’ group) had increased between 50 and 100 days, while countries such as India, Pakistan and Russia had not slowed their trajectories. Future analysis with longer time windows may indicate whether some nations are experiencing subsequent waves of infections due to the easing of restrictions or other factors, including those who had successfully suppressed the disease previously.

Our study has some limitations. The assumption that infection rates are equal are across age groups may be met only in nations with large outbreaks and high death rates or with high inter-generational mixing [16-17]. Age stratified IFRs relative to the 80+ age group may differ from those in China or Italy, particularly in countries where health system support is limited, overwhelmed or inequitable. However, cluster group membership in the trajectory analysis did not depend on weighting for age adjustment. COVID-19 mortality data may be under-reported and the calculated IFRs may be under-estimates. The IFR estimates were produced from data that was available in February, 2020, before large scale seroprevalence studies had been conducted [3-6]. If more up to date age stratified IFR estimates become available, the analysis can be updated.

Mortality also depends on gender, co-morbidities, ethnicity, obesity and other risk factors such as smoking [18-19], as well as access to health services. Increased risk of severe COVID-19 requiring hospitalisation due to underlying health conditions has been calculated at the national level [15]. As an extension to our analysis, the risk models developed for
infection hospitalisation ratios by Clark et al [15] could be adapted to the age standardised national IFR point estimates before clustering.

Conclusion

COVID-19 IFRs were clustered into three groups depending on national age profiles. A cluster of younger nations, predominantly in Africa, had a lower mean IFR than older nations. However, these countries may have less developed health systems and poorer overall health status, so the actual IFR in these nations may be higher than that estimated from age alone. Independently of risk due to age, clustering of death trajectories can track the effect of public health interventions in preventing the spread of the disease and changes over time.

References

1. Zhu N, Zhang D, Wang W, et al; China Novel Coronavirus Investigating and Research Team. A novel coronavirus from patients with pneumonia in China, 2019. N Engl J Med. 2020;382(8):727-733. doi:10.1056/NEJMoa2001017

2. Novel Coronavirus Pneumonia Emergency Response Epidemiology Team. Vital surveillances: the epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (COVID-19)—China, 2020. China CDC Weekly. 2020;2(8):113-122. Accessed April 16, 2020.

http://weekly.chinacdc.cn/en/article/id/e53946e2-c6c4-41e9-9a9b-fca8db1a8f51
3. Havers FP, Reed C, Lim T, et al. Seroprevalence of Antibodies to SARS-CoV-2 in 10 Sites in the United States, March 23-May 12, 2020. JAMA Intern Med. Published online July 21, 2020. doi:10.1001/jamainternmed.2020.4130

4. Pollán M, Pérez-Gómez B, Pastor-Barriuso R, Oteo J, Hernán MA, Pérez-Olmeda M, et al. Prevalence of SARS-CoV-2 in Spain (ENE-COVID): a nationwide, population-based seroepidemiological study. Lancet 2020; published online July 6. http://dx.doi.org/10.1016/S0140-6736(20)31483-5.

5. Stringhini S, Wisniak A, Piumatti G, Azman AS, Lauer SA, Bayssson H, et al. Seroprevalence of anti-SARS-CoV-2 IgG antibodies in Geneva, Switzerland (SEROCoV-POP): a population-based study. Lancet 2020; published online June 11. http://dx.doi.org/10.1016/S0140-6736(20)31304-0

6. Meyerowitz-Katz G, Merone L, A systematic review and meta-analysis of published research data on COVID-19 infection-fatality rates. Preprint published online July 07, 2020. medRxiv 2020.05.03.20089854; doi: https://doi.org/10.1101/2020.05.03.20089854

7. Verity R, Okell LC, Dorigatti I, et al. Estimates of the severity of coronavirus disease 2019: a model-based analysis. Lancet Infect Dis. Published online March 30, 2020. DOI:https://doi.org/10.1016/S1473-3099(20)30243-7

8. Rinaldi G, Paradisi M. An empirical estimate of the infection fatality rate of COVID-19 from the first Italian outbreak. Preprint published online May 18, 2020. https://www.ssrn.com/abstract=3582811

9. Ferguson NM, Laydon D, Nedjati-Gilani G, Imai N, Ainslie K, Baguelin M, et al. Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand. Imperial College London. Published online 16 March, 2020. doi: https://doi.org/10.25561/77482
10. Scrucca L, Fop M, Murphy TB and Raftery AE. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. The R Journal 2016; 8(1):289-317 https://doi.org/10.32614/RJ-2016-021

11. Sylvestre MP, McCusker J, Cole M, Regeasse. Classification of patterns of delirium severity scores over time in an elderly population. International Psychogeriatrics, 2006; 18(4):667-680. doi:10.1017/S1041610206003334.

12. South, A. rworldmap: A New R package for Mapping Global Data. The R Journal 2011; 3(1):35-43.

13. WHO World Population Prospects 2020 estimates. Accessed April 6, 2020. https://population.un.org/wpp/Download/Standard/Population/

14. Our World in Data, https://covid.ourworldindata.org/data/owid-covid-data.csv, extracted on July 21, 2020.

15. Clark A, Jit M, Warren-Gash C, Guthrie B, Wang HHX, Mercer SW, et al. Global, regional, and national estimates of the population at increased risk of severe COVID-19 due to underlying health conditions in 2020: a modelling study. Lancet Glob Health 2020; published online June 12. http://dx.doi.org/10.1016/S2214-109X(20)30264-3.

16. Dowd J, Andriano L, Brazel DM, Rotondi V, Block P, Ding X, et al. Demographic science aids in understanding the spread and fatality rates of COVID-19. Proceedings of the National Academy of Sciences. 2020; 117(18): 9696-9698; DOI: 10.1073/pnas.2004911117

17. Esteve A, Permanyer I, Boertien, Vaupel JW. National age and coresidence patterns shape COVID-19 vulnerability. Proceedings of the National Academy of Sciences. 2020; 117 (28) 16118-16120; DOI: 10.1073/pnas.2008764117
18. Yang J, Zheng Y, Gou X, Pu K, Chen Z, Guo Q, et al. Prevalence of comorbidities and its effects in coronavirus disease 2019 patients: A systematic review and meta-analysis. Int J Infect Dis. 2020;94:91–95. doi:10.1016/j.ijid.2020.03.017

19. Richardson S, Hirsch JS, Narasimhan M, Crawford JM, McGinn T, Davidson KW, et al. Presenting Characteristics, Comorbidities, and Outcomes Among 5700 Patients Hospitalized With COVID-19 in the New York City Area. JAMA. Published online April 22, 2020. doi:10.1001/jama.2020.6775
Supporting information captions

S1 Table: Relative death rates compared with the 80+ age group from China and Italy. Confidence limits for IFR/IFR 80+ were estimated using the upper and lower limits for IFR and dividing by the point estimate for IFR 80+ (IFR 81+ for Italy).

S1 Figure: Probability density function and histogram. Probability density function of fitted distributions from three clusters (solid line) and histogram of observed IFRs.

S2 Figure: Model diagnostics. Quantile-Quantile plots to test normality (left) and estimated and empirical cumulative density functions (CDF) (right).

S2 Table: Statistical measures of trajectory clusters. Statistical measures selected by factor analysis to describe the 50 and 100 day trajectories by cluster.

S3 Figure: Criteria used to determine optimal number of clusters for 50 day trajectories. Ccc criteria for 2-15 clusters (left) and within groups sum of squares for 1-15 clusters (right).

S4 Figure: Criteria used to determine optimal number of clusters for 100 day trajectories. Ccc criteria for 2-15 clusters (left) and within groups sum of squares for 1-15 clusters (right).
