The impact of population age structure for COVID-19 case fatality rates, a multicountry analysis

*Gibril Gaye, Jedidiah Awuku, Kayode Bamimore and Meron Kifle

*Names in alphabetical order

Summary:

Current evidence highlights that there is a substantial variation in the severity of the coronavirus disease 2019 (COVID-19) between countries. A significant amount of this variation may be accounted for by the country level age structure of individuals who are tested and diagnosed. This project aimed to assess the impact of age structure on covid-19 case fatality rates in Brazil, France, Mexico and the USA. Using demographic data linked with Covid-19 number of infections and deaths by age, we planned to analyse the open access country level data using Mann-Kendall test to assess the effect of age structure on CFR by country over time. Our analysis showed that the data was not suitable to do further processing to achieve our objectives. This is because of the unsatisfactory quality of the data structure and major violation of assumptions for our planned analysis plan. We therefore recommend that increasing data quality through harmonization of the evidence generation process from data collecting, curation, and maintenance should be given a priority by national and relevant international organizations.

Key words: Covid-19, Age structure, Case Fatality Rate

Introduction

The ongoing coronavirus disease 2019 (COVID-19) pandemic is causing unprecedented human and economic losses to countries with varying and often interconnected factors influencing disease transmission and death (1). Population age structure is usually one of the most visible factors associated with reported variances in disease transmission and mortality rates both across and between countries (2). Owing to the lack of availability of nuanced data and age being a strong confounder to many demographic, economic and health status of a population, it remains as one of the most preferred predictive tools to assess the burden of Covid-19 and subsequent planning in healthcare access, capacity deployment and other resources (3).

Typically, infectious disease transmission like Covid-19 are characterised by similar magnitude of transmission across a population but severity of the disease, quantified as case fatality rate (CFR), tends to be concentrated in a specific age group (4). This observation has been highlighted in the Covid-19 pandemic, particularly in its early stages where a higher proportion of Covid-19 related deaths in countries like Italy was seen among the older people whilst a relatively milder impact was felt in countries with a younger population pyramid (5). Hence, a better understanding of age specific stratification and interaction of population groups who may be at an increased risk of death would be important for two major reasons. First, the information is important to track inter country performance of Covid-19 treatment and care and to measure progress over time. Secondly, health policy makers can utilize this information to have realistic anticipation of the severity of the disease relative to their population structure and design public health policies and interventions that account for these variations.

Aim and Objectives

Within this context, the overall aim of this work was to investigate the impact of age structure on covid-19 case fatality rates across selected countries. Specifically, we aimed to investigate :

(a) the country level trends in case fatality rates (CFR) of Covid 19 in the selected countries

(b) how age distribution affects the case fatality rate (CFR) through time across the selected countries

Methods

Source data

The research is centred around the use of Covid-19 data, therefore, the project employed an open access "COVID-19 Data" (Available at: https://github.com/GoogleCloudPlatform/covid-19-open-data/blob/main/README.md). This source is one of the largest Covid-19 epidemiological databases that is open licensed and accessible to the public. It includes country and sub-country level datasets for Covid-19 cases, deaths, population demographics, economic indicators, geographical and climatic variables assembled from more than 20,000 global data sources.

Data processing

Four countries namely Brazil, France, Mexico and the USA were selected. These countries were selected because higher quality detailed age-disaggregated data were available at the time of the analysis. Four variables that can capture the objectives were identified for each country. These were time of data reporting, number of cases, number of deaths and age. Age categories were recategorized into three groups based on previous literature evidence on CFR levels by age across countries (6) and for the purposes of not dwelling on building unnecessarily or too many models but building parsimonious models. The case fatality rate was a composite variable computed by expressing the total number of cases by deaths as a percentage.

Environment setup for analysis

All data analysis was done using programming languages Python and R on Google Colab. R Studio, an open-source Integrated Development Environment for R, was used as the main working environment for the analysis. This environment and other relevant libraries were installed for processing of datasets and analysis. The final R Notebook is attached as an annex at the end of this paper. Jupyter Notebook, an open source web application was also used to generate a pandas-profiling report for the dataset to aid in exploratory analysis.

Data processing

Data processing is a crucial step for statistical modelling and analysis using a large dataset. The aim of pre-processing was to improve the quality of the data and create features that will be useful in modelling or providing inference.

The dataset being used was obtained from a stratification in terms of age of two large data sets within the Covid-19-Open-data namely Epidemiology and hospitalization. The data were heavily disaggregated with a lot of missing values and irregular format of entry, which made it very difficult to merge and link the datasets together.

The analysis required total confirmed cases and total deceased patients stratified by age for Brazil, France, US and Mexico. The original dataset was filtered to contain only these countries. Within each of these countries, there were regions that provided distinct counts of cases and deceased daily. Total daily cases and deaths were obtained by summing the counts from the individual regions. Variables that were highly sparse and not essential to the study were dropped. To simplify the analysis and facilitate comparisons between countries, we combined the initial nine age bands in the dataset to three large age bins/bands; 0-29, 30-69 and 70+. Furthermore, case fatality rate (CFR) was calculated using the total cases and deaths for each country and age band. The resulting data set was a time-series data of total cases, total deaths and CFR for each age band.

Data analysis

We started with exploratory data analysis with initial data profiling in order to gain insight into the dataset. Summary statistics on the total number of Covid-19 cases, deaths with crude CFRs and age adjusted CFRs by country were planned to be displayed in tables. Line charts of the CFRs showed nonlinear trends over time. We tested the hypothesis of an existing trend using a Mann-Kendall trend test. Sen’s slope was planned for the age bands and countries that showed a significant result in the hypothesis test. Sen’s slope allows the comparison of the rates of change of CFR over time.

H_0: The CFRs are all independent and identically distributed

H_1: There exist a monotonic trend

Results and discussion

This project aimed to explore trends of age specific CFRs at country level. However, we were unable to proceed further for two major challenges; quality of the data structure and major violation of assumptions for our planned analysis plan.

Data structure

The objectives of the study were one of the major factors that was seriously put into considerations while sourcing for data to be used for the analysis. Datasets that have required variables such as age, cases and confirmed death were considered. However, the majority of the datasets in various databases that were explored are not separated by country and sub country in a clear format and had highly variable patterns of recording.

The sub- country reports and figures could not be verified to be independent. It was not categorically stated whether the daily counts/observations followed a cumulative pattern or are independent. Because the original data was collated from different sources, it posed a serious challenge in the pre- processing phase. Merging and linking together of these various datasets caused hindrances during the data wrangling process.

The category of age band used by different countries also varies, which makes it difficult to have harmonic categorization. Reliability of the data could not be verified by external evidence, looking at researches that used the same data, no single record was found. All efforts made to opt for other sources was challenging because databases that had datasets recorded at country level were not found among the open access category.

Implications of the data structure on data analysis plan

Initially, the Mann-Kendall test was planned to test the effect of age structure on CFR by country over time. However, the assumption of independent realizations of the data required by the test was violated due to the data being cumulative for some data points and uncertain in others. Another analysis plan considered was to fit a Poisson regression model on the daily cases/deaths. This was considered to forecast future cases or deaths and provide more inference. Similarly this was not possible due to the same reason.

Conclusion and reflections on the exercise

Although the Covid-19 pandemic is an ongoing challenge on multiple ends, harmonization of the evidence generation process from data collecting, curation, and maintenance should be given a priority by countries and relevant international organizations. The initiative of open access data is certainly useful for democratizing data access and efficient evidence use by policy makers and researchers. However, further work is warranted to address the challenges discussed above because without ensuring high data quality standards, the initiative might render less than what it is intended.

References

1. Vivian Thangaraj, J., Murhekar, M., Mehta, Y., Kataria, S., Brijwal, M., Gupta, N., … Bhargava, B. (2020). A cluster of SARS-CoV-2 infection among Italian tourists visiting India, March 2020. Indian Journal of Medical Research, 151(5), 438–443. https://doi.org/10.4103/ijmr.IJMR_1722_20

2. Sudharsanan, N., Didzun, O., Bärnighausen, T., & Geldsetzer, P. (2020). The Contribution of the Age Distribution of Cases to COVID-19 Case Fatality Across Countries : A Nine-Country Demographic Study. Annals of Internal Medicine, 173(9), 714–720. https://doi.org/10.7326/M20-2973

3. United Nations, World Population Prospects 2019. https://population.un.org/wpp/DataQuery/. Accessed 13 March 2020

4. Mossong, J., Hens, N., Jit, M., Beutels, P., Auranen, K., Mikolajczyk, R., … Edmunds, W. J. (2008). Social contacts and mixing patterns relevant to the spread of infectious diseases. PLoS Medicine, 5(3), 0381–0391. https://doi.org/10.1371/journal.pmed.0050074

5. Onder G , Rezza G , Brusaferro S . Case-fatality rate and characteristics of patients dying in relation to COVID-19 in Italy. JAMA. 2020. [PMID: 32203977] doi:10.1001/jama.2020.4683

6. Jason Oke, Carl Heneghan. Global Covid-19 Case Fatality Rates.7th Oct 2020, available at : https://www.cebm.net/covid-19/global-covid-19-case-fatality-rates/