COVID-19 Mortality In Selected Countries of The Americas

Data Report

Johannes Garstenauer
Friedrich-Alexander-Universität Erlangen-Nürnberg
Erlangen, Germany
johannes.garstenauer@fau.de

***Abstract*—The COVID-19 pandemic has significantly impacted global economic, political, and societal structures, claiming many lives. This project analyzes mortality in the Americas, utilizing open data from various North and South American countries. It focuses on differences in mortality outcomes, both absolute and relative to population size, particularly comparing COVID-19 mortality between Hispanic nations and the Hispanic population in the U.S. The analysis is confined to 2020-2023, the period with the highest mortality and most available data.**

This research aims to provide insights into the pandemic's impact across different regions and communities, with a specific focus on the Hispanic community. The findings can serve as a basis for further research into correlations between societal factors and pandemic mortality.

I. Questions

1) How does COVID-19 mortality compare in the selected nations in South- and North America in absolute and in relative terms from 2020 to 2023?
2) How did COVID-19 mortality develop in the selected nations from 2020 to 2023?
3) How does COVID-19 mortality in Hispanic countries compare to COVID-19 mortality in the Hispanic population in the United States from 2020 to 2023?

II. Data Sources

For all datasets, an emphasis is laid upon those columns most relevant to answering the research questions. When it comes to data quality, we assess the properties of *Accuracy*, *Completeness*, *Consistency*, *Timeliness* and *Relevancy*, in no particular order.

A. United States of America: Center for Disease Control

1) Data Structure: This dataset contains data on COVID-19 deaths in the United States of America [1]. Double stratification is used, meaning the data is stacked based on two different criteria, leading to more accurate and reliable statistical inferences. In this case the data is split by categorical groups (column `group` with respective values in `subgroup1` and if required `subgroup2`) age, race/ethnicity, sex, and region, with race/ethnicity by age group and age group by race/ethnicity double stratification. Thus, in practice, mortality information can be easily extracted by each of the groups (race, age, ...) individually, as well as for race within age groups and vice versa.

Additionally, each row represents the data of a month-long period, denoted by the two continuous text columns (`data_period_start` and `data_period_end`), indicating the period start and end date, in the American date format (mm/dd/yyyy). The time interval of the dataset ranges from January the 1st, 2020 to September the 30th, 2024.

In addition to these categorical text columns, further columns contain continuous numerical data on the number of COVID-19 deaths, crude mortality rate per 100,000 people and the lower and upper bound of the 95% confidence interval of the crude mortality rate (`covid_deaths`, `crude_rate`...). Lastly a categorical text column `jurisdiction_residence` denotes whether the given numbers apply for an administrative region or for the entire country.

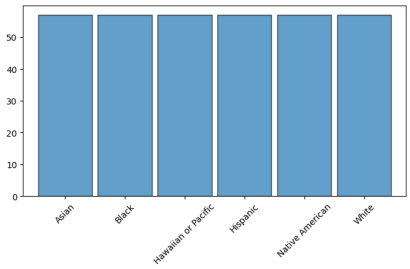

Fig. 1. Distribution of racial groups in US dataset.

2) Data Quality: Given that the data is sourced from an official government portal, it is reasonable to assume that rigorous data collection standards have been adhered to. Nonetheless, the COVID-19 pandemic has highlighted significant challenges associated with data collection in the context of emerging crises. This is reflected in the ongoing discourse within academic and media circles regarding the accuracy of mortality statistics. For example, in elderly patients, it can be challenging to ascertain whether a death was due to underlying health conditions or directly attributable to the virus. Resultingly, it is beyond the scope of this report to clearly determine the data accuracy, however the number of deaths in the dataset matches the widely accepted number of deaths in the US [3]. The data is timely as it is still being updated (at the time of writing) and encompasses the

whole analysis time interval. Data analysis has shown the data to be consistent in format and data types and obviously relevant to the topic of this project. Fig. 1 shows, that 57 months (1.1.2020-30.9.2024) of data are included for all racial subgroups[1]. Thus we confirm that, the dataset is complete and no time intervals are missing. However, the columns containing continuous values on mortality numbers and rates do have a significant number (ca. 27%) of missing values. As these are contained to the double stratified *Race and Age* and *Age and Race* categories, which are not required for our analysis, this does not impact its quality.

B. Republic of Colombia: Instituto Nacional de Salud

1) Data Structure: This dataset contains individual COVID-19 death cases in Colombia [2] ranging from March the 15th, 2020 to January the 12th, 2024. It is a community created view of an official dataset containing all COVID-19 cases in the country. In total, there are 23 columns, where a row represents an individual COVID-19 fatality. There are various location columns denoting the location and administrative region in categorical text or numerical data (`Nombre departamento`, `Estado`, ...), personal data in categorical text or numerical types (`Edad`, `Sexo`, ...), time-related columns denoting date of death, diagnosis, symptom onset and more in ISO 8601 format[2] (`Fecha de recuperación`, `Fecha de diagnóstico`, `Fecha de muerte`, ...). A categorical text column `Recuperado` indicates recovery status, which for this dataset view is always *Fallecido* (eng: "Deceased").

2) Data Quality: Similarly to above, we avoid making a claim about the real-world accuracy of data, can however note that the number of data points match widely accepted death counts for Colombia [3]. Data analysis has shown, that the entire analysis interval is covered. While there is a significant number of missing values in columns relating to personal information, location or diagnosis, none of the columns required for our analysis are affected. Again while there are noticeable inconsistencies, like the `Recuperado` column containing two unique values, which are identical apart from one being lower-cased, those columns most interesting for answering the research questions come in consistent and usable formats, such as dates in ISO 8601. The data covers the entire analysis interval.

C. Republic of Chile: Ministerio de Salud

1) Data Structure: Similarly to above dataset, the Chilean Ministry of Health provides individual COVID-19 related fatality cases beginning in March 2020 up to the present month [4]. There are a total of 27 columns, describing various attributes of a case. In terms of personal data there are numerical continuous columns related to age (`EDAD_CANT`) and a categorial text column related to gender (`SEXO_NOMBRE`). A few categorical text columns denote the location of the case (`COMUNA`, `NOMBRE_REGION`, ...), while another continous column denotes the time of death in ISO 8601 format[3] (`FECHA_DEF`). The remaining columns contain categorical text information about the diagnosis, with the possibility of having two different diagnoses.

2) Data Quality: Similarly, to before we avoid making a claim about the real-world accuracy of data, can however note that the number of data points does not match widely accepted death counts for Chile [3]. In fact, there is a discrepancy of about six thousand cases that are unaccounted for in the given dataset. Since the dataset time interval starts when noticeable numbers of COVID-19 cases started emerging in the country and ends at the present day, there is no time interval missing. Thus, without further information, we conclude, that completeness is significantly reduced. Furthermore there are missing value rates of about 100% for all rows related to a second diagnosis (next to COVID-19, being the first diagnosis for each case, in this dataset). As a result, these columns can be discarded, as they are also not required for answering our research questions. The remaining columns are complete and in consistent, usable formats.

D. World Bank: World Population Prospects (Population Total)

1) Data Structure: This dataset contains the total world population numbers from 1960 to 2023 for all officially recognized countries and their dependent territories [5]. The categorical text columns `Country Name` and `Country Code` denote the nation in question, and two more categorical text columns (`Indicator Name`, `Indicator Code`) describe what indicator is measured, in this case the total population. Thus they contain only one unique value each. The remaining columns contain continuous numerical data in floating point precision (`1960`, ..., `2023`), denoting the year for which the total population of a given country is provided.

2) Data Quality: The dataset contains a column `Unnamed: 68` filled with only missing values. It serves no purpose and can be discarded. Furthermore missing values exist for the population of stateless persons and *West Bank and Gaza* territory before 1990. Apart from that, the dataset is complete, in consistent format, spans the required time range and receives yearly updates. The data is sourced from the United Nations Population Division and national statistical offices, suggesting a high level of accuracy consistent with standard population data collection practices.

III. Licenses

Most datasets are licensed under various Creative Commons licenses or are under Public Domain (U.S. Govermnent). In the following we will summarize the permissions and obligations associated with these licenses and how we plan on following them.

- (World Bank [5]) *Creative Commons Attribution 4.0* (CC-BY 4.0) [8]. Free to: Share and Adapt. Obligated to: Give

[1]The same holds true for all the other groups. Proof is omitted for the sake of brevity.

[2]ISO 8601: (YYYY-MM-DD HH:MM:SS) is conveniently also used by Python's `datetime` library

[3]see also Colombia II-B)

attribution, Provide a link to license, Indicate if changes were made.

- (Chile [4]) *Creative Commons Non-Commercial* (CC BY-NC 4.0) [6]. See CC-BY 4.0 with the additional obligation of non-commercial use only.
- (Colombia [2]) *Attribution-ShareAlike 4.0 International* (CC BY-SA 4.0) [7]. See CC-BY 4.0 with the additional obligation of distributing under an identical license.
- (US [1]) *U.S. Public Domain Government* [9]. Free to: Share and Adapt. Obligated to: Avoid misrepresentation of government work, verify that data originates from government entity.

By giving attribution to dataset providers, indicating what licenses were provided and what changes to the data were made, we follow the most common licensing obligations. Additionally, we will not use the data commercially in any way, and publish the resulting project under the CC BY-SA 4.0 (*ShareAlike*) license [7]. Thus we follow all given obligations and are free to share and adapt the datasets.

IV. Data Pipeline

For this project we used Python to build an automated and resilient data pipeline following the ETL (Extract, Transform, Load) standard. For all datasets an HTTP download extracts the data into readable format, after which datasets are transformed, meaning unnecessary columns are dropped[4], missing values can be imputed using a variety of strategies[5] (mean, deletion, interpolation, ...) if a defined threshold of missing values has been passed for a given column. Additionally, rows, where certain conditions are true can be dropped. This comes in handy, when data is duplicated or stratified. Lastly, values of columns in various date formats are transformed into Python's `datetime` type[6] and the datasets are saved as local copies in `csv` format, which offers a lightweight and easy-to-use solution for our storage requirements. As a result, for the analysis there are uniform and consistent datasets of high quality available.

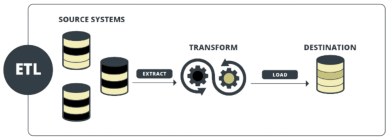

Fig. 2. The ETL Pipeline Architecture [11]

As we work with a variety of homogenous datasets, all pipeline steps described above are implemented as entirely modular and configurable functions with extensive documentation. We took particular inspiration from the *Pipes and Filters pattern* as proposed in [10], where functions are conceptualized as filters or modular parts of a pipeline. These filters inputs and outputs follow a predefined schema, such that filters can be arranged in any order and thus are highly parallelizable and scalable as long as the involved schemas match. Therefore, we deploy a global schema[7]

We emphasize robust error handling to ensure that the pipeline can continue processing remaining datasets even if one fails, while systematically logging exceptions. Additionally, each step in the pipeline is meticulously logged and designed to be resilient to failures by incorporating retry mechanisms wherever feasible.[8].

V. Results and Limitations

A. Resuling Datasets

Resulting from the pipeline, there are two datasets (*Chile*, *Colombia*) containing rows for each individual case of COVID-mortality, with two columns each. The first column denotes the time of passing in `datetime` format, whereas the second column contains categorial text data on the diagnosis. The third dataset (*USA*) contains two columns denoting a monthly period in `datetime` format, categorical text data on group and subgroup membership (f.e. race, or age) as well as the absolute numbers of deaths for that (sub-)group in the specified time frame and a relative mortality rate. The last dataset contains the population numbers for the above mentioned countries, with one column denoting the country name in categorical text data, and four columns denoting the population number for the years 2020-2023 in continuous numerical types. As such, all data sets contain a minimum necessary amount of rows and columns, have consistent data formats, no missing values and are complete and timely in terms of the time range of analysis.

B. Challenges and Limitations

Working with this data however also introduces a number of limitations and concerns. Firstly, the accuracy of the reported death tolls are controversial for reasons of difficulty of data collection under pandemic circumstances. Furthermore one dataset (see II-C) does not contain the expected number of COVID-19 fatalities. Additionally, the datasets still contain a degree of heterogeneity, where some contain indivudual cases and others aggregated case numbers for certain time periods. Lastly, even though the datasets are sourced from official and government sources, during development it has been observed, that they can go offline.

In summary, while there certain challenges, for the most part the data fulfills the properties of completeness, consistency, accuracy, timeliness and relevancy and is thus well suited to answer the research questions.

[4]In order to follow common best practices, desired columns are whitelisted, while the rest is automatically dropped

[5]Thus implementing the *Strategy Pattern*. For more see, Gamma, Erich. "Design patterns: elements of reusable object-oriented software." (1995).

[6]See: https://docs.python.org/3/library/datetime.html

[7]All inputs and outputs of transformation filters are in the `dataframe` type https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.html, thus satisfying this property.

[8]See: https://pypi.org/project/retry/

REFERENCES

[1] Center for Disease Control, Coronavirus and Other Respiratory Viruses Division (CORVD) (U.S.A.), "Monthly COVID-19 Death Rates per 100,000 Population by Age Group, Race and Ethnicity, Sex, and Region with Double Stratification" https://data.cdc.gov/Public-Health-Surveillance/Monthly-COVID-19-Death-Rates-per-100-000-Populatio/exs3-hbne/about_data (Last Access: 14.11.2024), October 21, 2024.

[2] Instituto Nacional de Salud: Salud y Protección Social (Rep. of Colombia), "Fallecidos COVID en Colombia" - Community created View, https://www.datos.gov.co/en/Salud-y-Protecci-n-Social/Fallecidos-COVID-en-Colombia/jp5m-e7yr/about_data (Last Access: 21.11.2024), January 18, 2024.

[3] worldometer.info, Worldometer Coronavirus Counter https://www.worldometers.info/coronavirus/ (Last Access: 21.11.2024).

[4] Ministerio de Salud (Rep. of Chile), "Defunciones por COVID19" https://datos.gob.cl/dataset/defunciones-por-covid19 (Last Access: 22.11.2024), November 19, 2024.

[5] World Bank Group,"World Development Indicators: Population Total", https://data.worldbank.org/indicator/SP.POP.TOTL?most_recent_year_desc=true (Last Access: 22.11.2024), 2023.

[6] Creative Commons: CC BY-NC 4.0, https://creativecommons.org/licenses/by-nc/4.0/ (Last Access: 22.11.2024), 2024.

[7] Creative Commons: CC BY-SA 4.0, https://creativecommons.org/licenses/by-sa/4.0/ (Last Access: 22.11.2024), 2024.

[8] Creative Commons: CC BY 4.0, https://creativecommons.org/licenses/by/4.0/ (Last Access: 22.11.2024), 2024.

[9] Goverment of the United States: U.S. Public Domain Government, https://www.usa.gov/government-copyright (Last Access: 22.11.2024), 2024.

[10] Microsoft Azure Architecture Center: Pipes and Filters pattern, https://learn.microsoft.com/en-us/azure/architecture/patterns/pipes-and-filters (Last Access: 26.11.2024), 2024.

[11] Software AG: A deep dive into data pipeline architecture, https://www.softwareag.com/en_corporate/blog/streamsets/data-pipeline-architecture-deep-dive.html (Last Access: 26.11.2024), 2024.