Analysis and Visualization of COVID-19 Mortality In Selected Countries of The Americas

Final Report

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

***Introduction*—The COVID-19 pandemic has profoundly affected global economic, political, and societal structures, with substantial loss of life. Understanding the impact of the pandemic in terms of mortality in different regions and demographics is crucial to developing effective public health strategies and policies. This project aims to analyze mortality related to COVID-19 in the Americas, leveraging open data mainly from government sources. The focus is on exploring differences in mortality outcomes, both in absolute terms and relative to population size, with a particular emphasis on comparing COVID-19 mortality between Hispanic nations and the Hispanic population within the United States.**

The analysis is confined to the period from 2020 to 2024 (excluded), a time frame characterized by the highest mortality rates and the most comprehensive availability of data.

By concentrating on the Hispanic community, this research seeks to uncover potential national and ethnic disparities in COVID-19 mortality. The insights gained from this study aim to inform future research in guiding more equitable health responses in future public health crises.

1) **How does COVID-19 mortality compare in the selected nations in South and North America in absolute and relative terms from 2020 to 2024?**
2) **How did COVID-19 mortality develop in the selected nations from 2020 to 2024?**
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 2024?**

These questions aim to show the impact of COVID-19 on mortality across diverse populations and regions, with a specific focus on the Hispanic demographic, thereby contributing to a deeper understanding of the pandemic's effects on different societal groups.

I. Data

For the purpose of answering these research questions, various open data datasets have been sourced from either official government portals or reputable alternative sources, as detailed below. These data sets are processed through an automated data pipeline of the **Extract**, **Transform**, and **Load** architecture, for the purpose of improving data quality and omitting redundant features or data points.

A. Datasets

1) United States of America: Center for Disease Control: This dataset contains data on COVID-19 deaths in the United States of America and originates from the *Center for Disease Control* government agency [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`) age, race/ethnicity, sex, and region, with race/ethnicity by age group and age group by race/ethnicity double stratification. Making use of this enabled us to extract mortality numbers for certain ethnic groups, as well as for the whole population. 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 ISO 8601 format[1]. The time interval of the dataset ranges from January the 1st, 2020 to September the 30th, 2024. Further columns contain continuous numerical data on the number of COVID-19 deaths and the crude mortality rate per 100,000 people.

2) Republic of Colombia: Instituto Nacional de Salud: This dataset contains individual COVID-19 fatalities in Colombia ranging from March the 15th, 2020 to January the 12th, 2024 and is a community created view of an official dataset from the health ministry [2]. All COVID-19 mortalities in the country are included. After processing, it contains two categorical columns, `Fecha de muerte` (or "Date of Death") denoting the day of passing in ISO 8601 format and a column `Recuperado` indicating recovery status, which in this case is always *Fallecido* (or "Deceased").

3) Republic of Chile: Ministerio de Salud: Similarly the dataset provided by the Chilean ministry of health, contains, after processing, a continuous date column in ISO 8601 denoting the day of passing `FECHA_DEF`) as well as a categorical text column `DIAG1` denoting the primary diagnosis which in this case is always one of three distinct COVID-19 related diagnoses.

4) World Bank: World Population Prospects (Population Total): This dataset contains the total world population numbers from 1960 to 2023 for all officially recognized countries and their dependent territories and is sourced from the World Bank organization [4]. The categorical text columns `Country Name` denotes the nation in question and the remaining

[1]ISO 8601: YYYY-MM-DD

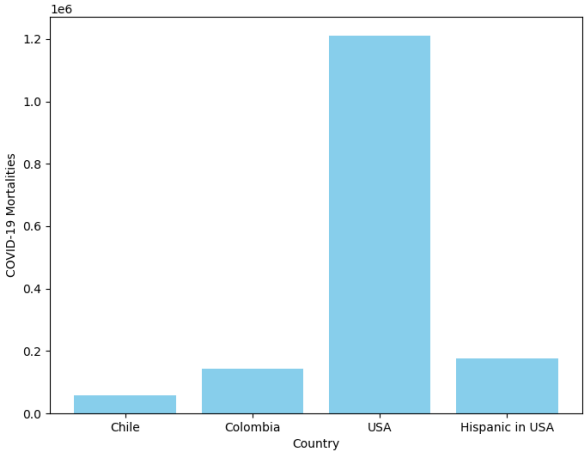

Fig. 1. Total COVID-19 Mortality in Millions

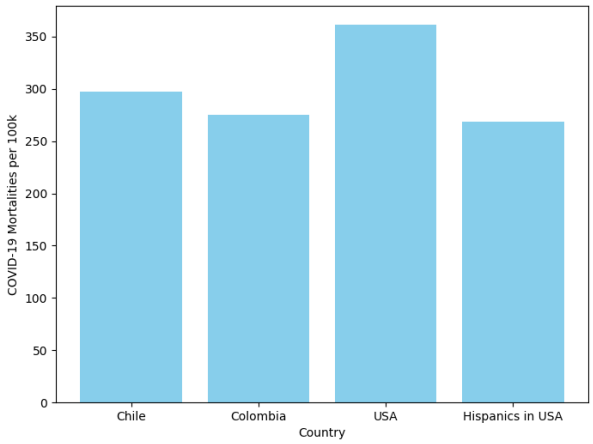

Fig. 2. COVID-19 Mortality Rate Relative to Population Size per 100k Citizens

columns (`2020`, ..., `2023`) contain continuous numerical data in floating point precision, denoting the total population of a country fro a given year.

B. Licensing

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 [4]) *Creative Commons Attribution 4.0* (CC-BY 4.0) [7]. Free to: Share and Adapt. Obligated to: Give attribution, Provide a link to license, Indicate if changes were made.
- (Chile [3]) *Creative Commons Non-Commercial* (CC BY-NC 4.0) [5]. 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) [6]. See CC-BY 4.0 with the additional obligation of distributing under an identical license.
- (US [1]) *U.S. Public Domain Government* [8]. 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-NC-SA 4.0 (*NonCommercial-ShareAlike*) license [6]. Thus we follow all given obligations and are free to share and adapt the datasets.

II. Analysis

The analysis section of this report aims to thoroughly examine the COVID-19 mortality data for selected nations in the Americas, by leveraging the well-structured datasets originating from the automated data pipeline in various visualizations. We focus on absolute and relative mortality rates, as well as the temporal development of these rates from 2020 to 2024. This section seeks to provide insights into how mortality outcomes differ across countries and demographic groups, with particular attention to the Hispanic population both in Latin American countries and within the United States.

A. Method

To address these questions, a methodical approach was employed, involving the pre-processing and aggregation of data to facilitate visualization and interpretation. Monthly accumulations of mortalities were created for each country, allowing for a detailed temporal analysis. The world population data enabled calculating COVID-19 death rates relative to population size.

We visualized the processed data using various plotting techniques. Histograms serve to visualize the distribution of mortality outcomes, helping to identify patterns and outliers. Area charts illustrate continuous data trends over time, enabling the comparison of mortality patterns between countries. The visualizations focused on both total and relative mortality rates to provide a complete understanding of the impact relative to population sizes.

B. Results

1) Total And Relative COVID-19 Mortality: When looking at total mortality rates in Fig. 1, the United States had the highest number of COVID-19 deaths, with Colombia and Chile reporting significantly lower numbers. However, total figures alone do not offer meaningful insights without considering population size. When adjusted for population numbers in Fig. 2, the United States still exhibited the highest relative mortality rates, followed by Chile and Colombia, with the Hispanic population in the United States showing even lower

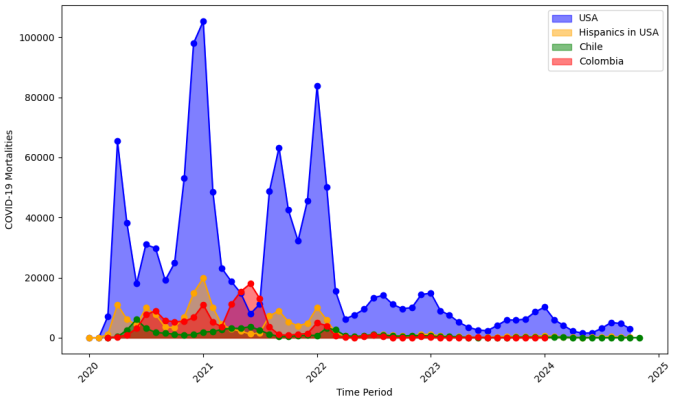

Fig. 3. Total Monthly COVID-19 Mortality During Pandemic

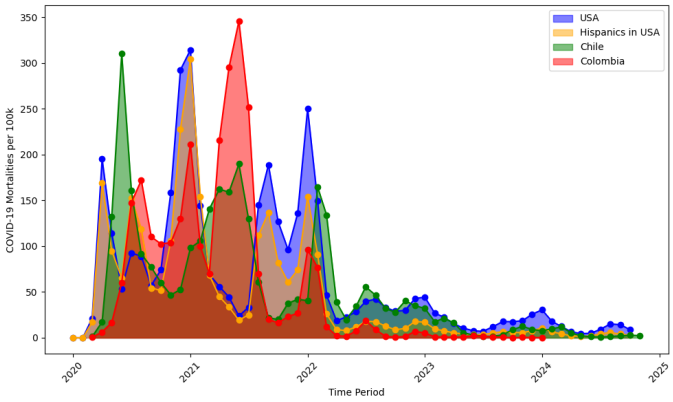

Fig. 4. Monthly COVID-19 Mortality Rate Relative to Population Size per 100k Citizens during the Pandemic

rates. This suggests that the Hispanic death rates in the U.S. are more aligned with those in the South American countries.

2) Development Of Total And Relative COVID-19 Mortality Over Time: Analyzing the temporal development of mortality rates in Fig. 3 and Fig. 4 revealed distinct patterns. Initial spikes, relating to the first wave of cases, in mortality rates were observed across all countries, but the timing and magnitude of these spikes varied. The United States and its Hispanic population exhibited a synchronized initial spike, whereas Chile and Colombia experienced theirs a few months later. While the U.S. also experienced a small second spike at that time, this second initial spike was relatively much more pronounced for the Hispanic population in the U.S., thus displaying a shared pattern to their South American counterparts.

Generally though, the United States and its Hispanic population exhibit remarkably synchronized mortality patterns, being the highest during the colder months. Columbian and Chilean mortality patterns are more similar to each other than to the U.S. and U.S.-Hispanic patterns. For instance, they exhibit a significant spike in Early to Mid 2021, that is not present in U.S. data. However apart from that, they display only a limited degree of synchronicity, beyond what is shared between all countries and populations in the analysis.

C. Interpretation

Consequently we showed, that South American nations were slightly more similar to each other than to the U.S. in terms of relative pandemic death rates as well as mortality trends, wich may be attributed to sharing some geographic, climactic, institutional and cultural factors. Furthermore, we pointed out, how Hispanic populations in the U.S. do show some similarities to Hispanic countries when it comes to death rates as well as their development during the pandemic years. This may be due to cultural reasons, like familial ties that contribute to spreading spikes in infections over borders by means of relative visitations or settling preferences that resemble home climates, and therefore produce similar pandemic outcomes. However we found, that Hispanics in the U.S. mostly shared mortality trends with the general population, leading to the conclusion, that the reasons mentioned above may only be applicable to a small degree.

Limitations of this report are multifaceted, stemming from the inherent challenges of data collection during a pandemic. Moreover, identifying the specific reasons behind the observed patterns is beyond the scope of this report, as it would require an in-depth examination of the various political, institutional, environmental, and cultural factors that impacted pandemic outcomes. Potential underreporting or misclassification of deaths presents a significant concern, as does the varying degree of data accuracy across countries. For instance, discrepancies in data reporting have been noted, such as 8,000 unaccounted cases in Chile that appear in other statistics [9] but are missing from the official dataset. Political factors and differences in public health infrastructure may have also influenced the reported mortality rates. In some cases, political interference might lead to under reporting of cases to present a more favorable image of governmental handling of the pandemic[2]. Differences in trust in official numbers and the capacity for accurate case documentation further complicate comparisons. Generally, more developed countries with robust health infrastructure may report higher case numbers due to better case detection and documentation.

Given that the data is sourced from official government portals, it is reasonable to assume adherence to rigorous data collection standards. However, the COVID-19 pandemic has underscored the challenges of data collection during emerging crises, reflected in ongoing debates within academic and media circles about the accuracy of mortality statistics. For instance, distinguishing whether deaths in elderly patients were due to underlying health conditions or directly attributable to the virus remains a critical issue.

The small sample size, comprising only two South Amer-

[2]In June 2020, during a campaign rally in Tulsa, Oklahoma, the President of the U.S. mentioned that he had asked his administration to ”slow the testing down” because increased testing was leading to more reported cases. This statement sparked significant controversy and concern among public health officials and was later redacted. https://edition.cnn.com/2020/06/20/politics/tulsa-rally-trump/index.html

ican countries and one North American country, limits the generalizability of the findings. This restricted geographic scope may not fully capture the diverse COVID-19 experiences across the Americas, potentially leading to an incomplete understanding of regional variations and their broader implications for public health policy.

Consequently, we must acknowledge significant limitations in the interpretability of our findings.

III. Conclusion

Overall, the analysis suggests that Chile and Colombia, both South American countries, exhibit more similar mortality patterns to each other than to the United States, a North American country. While the Hispanic population in the U.S. shows some parallels with Latin American countries, these similarities are dominated by their predominant alignment with broader U.S. temporal COVID-19 mortality trends. However, serious limitations to the interpretability must be noted.

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: 06.01.2025), January 18, 2024.

[3] Ministerio de Salud (Rep. of Chile), "Defunciones por COVID19" https://datos.gob.cl/dataset/defunciones-por-covid19 (Last Access: 06.01.2025), November 19, 2024.

[4] 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.

[5] Creative Commons: CC BY-NC 4.0, https://creativecommons.org/licenses/by-nc/4.0/ (Last Access: 06.01.2025), 2024.

[6] Creative Commons: CC BY-NC-SA 4.0, https://creativecommons.org/licenses/by-nc-sa/4.0/deed.de (Last Access: 06.01.2025), 2024.

[7] Creative Commons: CC BY 4.0, https://creativecommons.org/licenses/by/4.0/ (Last Access: 06.01.2025), 2024.

[8] Goverment of the United States: U.S. Public Domain Government, https://www.usa.gov/government-copyright (Last Access: 06.01.2025), 2024.

[9] Worldometer: COVID-19 cases and deaths, https://www.worldometers.info/coronavirus/country/chile/ (Last Access: 09.01.2025), April 13, 2024.