COVerAGE-DB: A database of age-structured COVID-19 cases and deaths
COVerAGE-DB
A database of age-structured COVID-19 cases and deaths

Key Features

- COVerAGE-DB is an open access database including cumulative counts of confirmed COVID-19 cases, deaths, and tests by age and sex. Original data and sources are provided alongside data and measures in age-harmonized formats.

- The database is still in development. It includes data since January 2020 and at this writing, it includes 93 countries, and 266 subnational areas.

- The database also documents international variations in definitions of all input data and indicators of reporting completeness across sources and over time.

- An international team, composed of more than 60 researchers, contributed to the collection of data and metadata in COVerAGE-DB from governmental institutions, as well as to the design and implementation of the data processing and validation pipeline. We encourage researchers interested in supporting this project to send a message to the email: coverage-db@demogr.mpg.de

Introduction

Information about pandemic dynamics is crucial to understand the potential impacts on populations, design mitigation strategies, and evaluate the efficacy of their implementation. Centralization, standardization, and harmonization of data is critical to enable comparisons of the demographic impact of COVID-19 that take into account differences in the age and sex-compositions of confirmed infections and deaths. The international data landscape must keep pace with the global march of the pandemic, and researchers must work to triangulate the available data to create comparable measures to monitor and predict its demographic impacts.

The COVID Age Database (COVerAGE-DB) aims to provide global coverage of key demographic aspects of the COVID-19 pandemic as it unfolds in an up-to-date, transparent, and open-access format. COVerAGE-DB offers data with standardized count measures and harmonized age groups, which is a necessary but not sufficient condition to allow comparisons between populations at national and subnational scales.
The database is currently under expansion through both the increase in coverage of national and subnational populations and the inclusion of more recent periods as the pandemic continues to develop. At this writing, the database contains daily counts of COVID-19 cases, deaths, and tests performed by age and sex for 93 national and 266 subnational populations around the world, depending on the available data for each source. The date range available for each country or subpopulation varies. In several country series, the database includes the earliest confirmed cases in January 2020. For most populations, the database includes daily time series, beginning from an initial starting date when the data were first released or collected by our team. Fig. 1 displays a map of countries included in the database, indicating at least one subnational population from 12 countries. A detailed overview of data availability is given in a searchable table [external link](https://bit.ly/3kVDrLD).

**Figure 1.** Availability of national and subnational information on COVID-19 cases, deaths, and tests in the countries included in the database as of 16 November, 2020.

**Data Collected:**

**Collection:** Official counts of COVID-19 cases, deaths, and tests are extracted from reports published by official governmental institutions, such as health ministries and statistical offices. Depending on the source, data are collected in a variety of formats, including machine-readable files, pdf tables, html tables, interactive dashboards, press releases, official announcements via Twitter, and in a few instances, from digitized graphics. A full list of data sources is available in a dashboard view (https://bit.ly/2Qg1MxL).

Generally, COVID-19 cases, deaths and tests in age groups are reported as counts, but some sources report data in other metrics (fractions, percents,
COVerAGE-DB: A database of age-structured COVID-19 cases and deaths

ratios) or as summary indicators such as case fatality ratios by age. Reported age intervals vary by source, ranging from single ages to 30-year or greater age bands, and sometimes reported age intervals change over time within sources. Usually data are reported as cross-sectional snapshots of cumulative counts, but some sources give full time series of new cases or deaths, in which case we cumulate counts over time. We also collect standard metadata on each of the sources to capture various characteristics of the collected data, such as the primary collection channels, definitions used, and notes on major disruptions or events. An overview of key fields from this metadata is shared as a spreadsheet (https://bit.ly/2FAmKFn).

Production: All source data is entered into standard spreadsheet templates hosted in a central folder on Google Drive. Data entry into the templates is either manual or automatic depending on the source.

R programs collect data from the source templates and compile the merged input database. The merged input file is then subject to a series of automatic validity checks. Initial checks are carried out by the individual responsible for data collection and entry using an interactive application (https://mpidr.shinyapps.io/cleaning_tracker/). Data is then harmonized to standard metrics (counts), measures (cases, deaths, tests), and age bands (5- and 10-year age intervals). Harmonization procedures include rescaling to ensure coherence between age distributions and reported total counts. Age group harmonization is done using the penalized composite link model for ungrouping, which was designed for splitting histograms of count data. Output data also includes a file containing selected diagnostics of data quality, such as completeness of age reporting, for each source and date.

The complete details on all steps of production are available in the COVerAGE-DB Method Protocol, which is publicly available on the web (https://osf.io/jcnw3/). A table listing which adjustments are applied to each population is available on the project website (https://bit.ly/2E61BSV). The merged input database, the harmonized output, and data quality files are uploaded daily as zipped csv files to an Open Science Framework repository (OSF) (https://osf.io/mpwjq/). A GitHub repository (https://bit.ly/2YbtPCJ), which is linked to OSF, contains all R scripts used in the complete production pipeline, including compilation, diagnostics, and harmonization.

Data Resource Use:

Since collection efforts began for COVerAGE-DB in late March 2020, we are aware of 13 studies using the data, many of which provide R code online and are fully reproducible. Broadly, these studies aim to measure the influence of demographic factors on mortality from COVID-19; assess the pandemic
impact on health and mortality within\textsuperscript{4,5} and across populations;\textsuperscript{6–11} analyze COVID-19 data availability and quality;\textsuperscript{12} and propose methodological innovations that allow comparisons of case fatality ratios,\textsuperscript{13} and the indirect estimation of infections in the population.\textsuperscript{14} The database is also used to monitor COVID-19 impacts in particular age ranges. For instance, UNICEF has used the database for monitoring the burden of the pandemic on the infant, child, and juvenile mortality around the world.

As an example of the analyses that COVerAGE-DB enables, Fig. 2 displays changes in the relation between age-specific deaths and cases rates in Colombia, inspired by Fig. 1 of Dudel et al.\textsuperscript{13} We divide both cases and deaths in each age band by the respective population sizes. Diagonal lines indicate implied age-specific case fatality rates (CFRs). The graphic illustrates a sharp increase in CFR over age for each sex, and displays considerable sex differences. For instance, men aged 60–69 in Colombia have almost the same case fatality rates (approximately 12% risk of death after COVID-19 disease is diagnosed) as women aged 70–79.

**Figure 2.** Relation between deaths and cases per 100,000 population by age group and sex in Colombia, until 7 November 2020. Diagonal lines indicate the case fatality rate.

We repeat this exercise to compare Colombia with Mexico (see Fig. 3), where standardizing by population size is more justified. Case fatality rates and death
rates are much higher in Mexico than in Colombia in each age band – around 2-fold –, except for ages 80+, which show a substantial reduction in the case fatality rate difference, and much higher death rates for Colombia.

This comparison between Colombia and Mexico allows us to illustrate several issues in data quality to be considered when comparing COVID-19 outcomes between populations. Besides the economic and sanitary conditions that make Latin American countries more vulnerable to the pandemic, the lack of unambiguous definitions of COVID-19 cases and deaths, and the limited testing capacity represent major challenges for data quality assessment.\textsuperscript{15–17} We focus here on definitions and testing strategies.

With respect to COVID-19 case and death definitions, criteria have varied since records started. At the time of data retrieval, both countries use laboratory, clinical, and epidemiological criteria to confirm SARS-CoV-2 infections.\textsuperscript{18,19} However, the vast majority of COVID-19 cases and deaths are confirmed with RT-PCR tests results in both populations (99.6% and 91.6% in Colombia and Mexico, respectively).\textsuperscript{20,21} Regarding the definition of tests, whereas in Colombia it refers to laboratory samples tested (4.5M as of 7 November), in Mexico it alludes to persons (2.3M). Because individuals may be tested more than once, comparison between these two units is not straightforward. Testing performance measures, such as positive rates (e.g., 30% in Colombia and 45% in Mexico\textsuperscript{22}), are essential for interpreting differences in cases and deaths across populations, because they help to assess the extent of infection underreporting.\textsuperscript{23} However, differences in test definitions pose serious challenges for direct comparisons. As to reported dates, these variables are directly comparable in both populations. Dates in both sources correspond to the occurrence of events and, as information in both sources rely on individual-level databases, delays in register of diagnosis and deaths are retrospectively adjusted.

A second factor to be considered is the difference in testing capacity and strategy between countries, which are key determinants for infection diagnosis. Given both the magnitude of contagion and limited resources in the region, Latin American countries have struggled to increase testing capacity proportionally to the spread of the infection.\textsuperscript{24,25} Although with very limited capacity, the testing approach of Colombia has been to test as many suspected cases as possible. In contrast, an important part of the test strategy in Mexico has focused on inferring the extent of contagion in the population by using nationally representative samples (which represent 36.5% of all confirmed infections at the date under observation), and it has gradually included a narrow proportion of suspected infections outside the surveillance system.\textsuperscript{19} On 7 November, Colombia performed 5-times more tests per capita than Mexico.\textsuperscript{22}
These differences in testing regimes between both countries may drive a substantial part of the CFR discrepancies observed in Figure 3.

The differences in definitions and testing strategies between populations highlight challenges in making comparisons, but also the need to produce data with sufficient detail to adjust for biases. For this reason, alongside data on cases, deaths, and tests, COVerAGE-DB offers additional information on metadata and quality diagnoses that are crucial for a cautious interpretation of the data and its limitations. It is our view that researchers should triangulate creatively from all available data rather than avoiding difficult comparisons.

**Figure 3.** Relation between deaths and cases per 100,000 population by age group in Mexico and Colombia, until 7 November 2020. Diagonal lines indicate the case fatality rate.

**Strengths and Weaknesses**

Since the beginning of the pandemic, it has been evident that population characteristics are key to understanding the prevalence, spread and fatality of COVID-19 across countries. However, data on cases, deaths, and tests disaggregated by age and sex are not easily comparable across countries, and sometimes not even accessible. The main strength of COVerAGE-DB is to provide a centralized, open-access, and fully reproducible repository of age- and sex-specific case, death, and test counts from COVID-19, collected
from official sources, and harmonized to standard output formats. The data harmonization process is transparent, following a strict protocol. The initial input data is provided alongside the harmonized counts, as well as the code used to harmonize the different input measures, metrics, and age groups into comparable granular output metrics. All scripts are written in the open-source R programming language. The data sources and limitations are documented for each country in a standard metadata framework.

A limitation of the COVerAGE-DB is the heterogeneous and difficult-to-evaluate quality of the underlying data. No single data source can currently claim accurate estimates of COVID-19 incidence or fatalities. Age-specific case counts are highly dependent upon the testing capacity, testing strategy, and differences in the definition of cases across sources and over time. Cases are underestimated everywhere, with underestimation expected to vary by age, given the relationship between age and case severity. The accuracy of diagnostic RT-PCR tests used to confirm infections is also known to vary. Furthermore, at any given date, cumulative counts are underestimated because of the lag between infection and a positive test result.

Death counts from COVID-19 are also likely underestimated for similar reasons, but also due to various kinds of delays in death registration. Media reports have circulated about intentional data manipulation in some of the official data covered in the database. Excess all-cause mortality has been observed across many regions. Although some of these deaths likely are from postponing or foregoing treatment from non-COVID-19 related causes, the magnitude of this excess is suggestive that numerous COVID-19 related deaths are classified under different causes. Populations also differ in whether deaths to suspected COVID-19 cases are included in official statistics and in post-mortem practices when an infection is suspected. Some populations only report deaths occurring in hospitals, neglecting a potentially sizable proportion of deaths occurring in institutional settings and at home. While most populations currently report all deaths to confirmed SARS-CoV-2 infections as COVID-19 deaths for this database, the underlying cause of death eventually reported on death certificates may differ in patients with severe comorbidities. To mitigate biases and misinterpretations due to different practices and definitions, such information is constantly updated and documented in the metadata of the database, which is freely accessible to users. Further, a supplementary data quality metrics file contains a suite of data quality indicators that is easily merged with the main output data. Quality metrics include age reporting completeness, some indicators on how aggressive age harmonization is, and two positivity measures from Our World in Data database on COVID-19 testing.
All of these issues compromise the comparability of the data contained within the COVerAGE-DB, both across populations at any given time and within populations over time. For these reasons, infection fatality ratios, which include in the denominator detected and undetected SARS-CoV-2 infections, should not be estimated from COVerAGE-DB data alone. Proper estimation of incidence and fatality will likely require triangulating data across numerous sources as these become available. To this end, the COVerAGE-DB was designed to be easily merged with other databases such as the Our World In Data testing or excess mortality data, the COVID-19 dashboard of Johns Hopkins, the World Population Prospects database, and the Short Term Mortality Fluctuations database. Moreover, given that we have near-complete time series capturing the whole pandemic curve in some places, careful modeling of the lag structures might allow some of these data-driven biases to be estimated.

Data Resource Access

Both merged input and harmonized output files can be downloaded directly from the OSF site (https://osf.io/mpwjq, DOI: 10.17605/OSF.IO/MPWJQ), which contains a folder called Data with four files of primary data. Fig 4 shows where to find the files in the OSF repository.

**Figure 4.** View of the Open Science Framework (OSF) repository, File section (https://osf.io/mpwjq/files/). To download data files, click on Data, and select one of the files.

<table>
<thead>
<tr>
<th>Name</th>
<th>Size</th>
<th>Version</th>
<th>Downloaded</th>
<th>Modified</th>
</tr>
</thead>
<tbody>
<tr>
<td>COVerAGE-DB: A database of COVID-19 cases and deaths by age</td>
<td>16.7 MB</td>
<td>80</td>
<td>253</td>
<td>2020-11-11 07:31 AM</td>
</tr>
<tr>
<td>GitHub: timriffe/covid_age (master)</td>
<td>1.5 MB</td>
<td>3</td>
<td>374</td>
<td>2020-10-27 03:05 PM</td>
</tr>
<tr>
<td>OSF Storage (Germany - Frankfurt)</td>
<td>8.0 MB</td>
<td>81</td>
<td>292</td>
<td>2020-11-11 07:32 AM</td>
</tr>
<tr>
<td>Data</td>
<td>13.6 MB</td>
<td>81</td>
<td>374</td>
<td>2020-11-11 07:32 AM</td>
</tr>
<tr>
<td>qualityMetrics.zip</td>
<td>1.7 MB</td>
<td>2</td>
<td>0</td>
<td>2020-11-07 07:32 AM</td>
</tr>
</tbody>
</table>

Each of the main data files has a stable link (see Tab. 1), which always points to the most recent version. Each file is a zipped csv file by the same name.
Table 1. The main data files, a description of their content, and their stable URLs.

<table>
<thead>
<tr>
<th>Filename</th>
<th>Description</th>
<th>Stable URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. inputDB.zip</td>
<td>Data in original metrics, measures, and age intervals</td>
<td><a href="https://osf.io/9dsfk/">https://osf.io/9dsfk/</a></td>
</tr>
<tr>
<td>2. Output_5.zip</td>
<td>Data with standardized metrics and measures, and harmonized age groups in 5-year intervals</td>
<td><a href="https://osf.io/7tnfh/">https://osf.io/7tnfh/</a></td>
</tr>
<tr>
<td>3. Output_10.zip</td>
<td>Data with standardized metrics and measures, and harmonized age groups in 10-year intervals</td>
<td><a href="https://osf.io/43ucn">https://osf.io/43ucn</a></td>
</tr>
<tr>
<td>4. qualityMetrics.zip</td>
<td>Selected data quality indicators by location and date.</td>
<td><a href="https://osf.io/qpfw5">https://osf.io/qpfw5</a></td>
</tr>
</tbody>
</table>

For stable links to download particular versions, click on the version number in the Version column seen in Fig. 4. Users can note versions either by referring to timestamps provided in the headers of data files or by referring to OSF file version numbers, which increment with each daily update.

A data dictionary is given in both the OSF wiki (https://osf.io/mpwjq/wiki/home/) and the Method Protocol. Files are shared in csv format to be as universally accessible as possible. A guide to getting started using the data in R is also provided (https://bit.ly/3g8nIVU), and tips for other statistical packages may also be added. Users are encouraged to reach out for further information or advice on using the database, or to express interest in the project: coverage-db@demogr.mpg.de.

Acknowledgments

We gratefully acknowledge the hard work of health ministries and statistical offices around the world in preparing and disseminating the data included in COVerAGE-DB.

References

Available from: http://www.who.int/bulletin/volumes/98/5/20–257600/en/
COVerAGE-DB: A database of age-structured COVID-19 cases and deaths


34. Wu J, McCann A, Katz J, Peltier E. 161,000 Missing Deaths: Tracking the True


