Datasets and future research suggestions concerning SARS-CoV-2

Tomasz Szmuda1, Shan Ali2, Cathrine Özdemir2, Mohammad Talha Syed2, Akshita Singh2, Tarjei Vevang Hetzger2, Philip Rosvall2, Karolina Fedorow2, Ahmed Alkhater2, Anders Majlöf2, Mohammed Albrahim2, Eyad Alquraya2, Rakan Al Dunquwah2, Zahraa Al-hakeem2, Emad Almohisin2, Mohammed Alradhi2, Weronica Magdalena Żydowicz2, Charlene Müller2, Ada Egeland2, Aurora Bergersen Kinstad2, Jessica Ngyuen2, Martin Bergersen Kinstad2, Ibraheem Al-Rubaye2, Yasmin Al-Khazragi2, Beatrice von Darde2, Bhakti Dave2, Paweł Słoniewski1, Justyna Fercho1, Sara Kierońska3

Tomasz Szmuda MD, PhD and Shan Ali contributed equally to the manuscript

1 Neurosurgery Department, Medical University of Gdańsk, Poland
2 Scientific Circle of Neurology and Neurosurgery, Neurosurgery Department, Medical University of Gdańsk, Poland
3 Neurology and Neurosurgery Department, University Hospital Collegium Medicum Nicolaus Copernicus University, Bydgoszcz, Poland

Abstract

We gathered publicly available online data and prepared a database of epidemiology, demographics, economics, Bacille Calmette-Guérin vaccination and online search trend statistics relevant to the coronavirus disease 2019 (COVID-19). Moreover, we provide several suggestions on the use of this bioresource and reference other relevant datasets to promote research on COVID-19.

Keywords: BCG · COVID-1 · SARS-CoV-2 · datasets · data · epidemiology · Google trends

Citation

Introduction

The first case of atypical pneumonia, which later was diagnosed as the Coronavirus disease 2019 (COVID-19), was reported to be on December 31st 2019 in China. At that time COVID-19 attracted relatively little public or scientific interest internationally [1]. However, by March of 2020 the situation evolved to COVID-19 pandemic and the novel SARS-CoV-2 virus became the subject of numerous research articles. It is noteworthy that several scientific journals stopped publishing articles not related to SARS-CoV-2. COVID-19 quickly prompted much scientific research and several scientific journals have called for articles concerning SARS-CoV-2. Several online databases regarding COVID-19 are available from well-known institutions such as the World Health Organisation (WHO), the European Centre for Disease Prevention (ECDC) and Johns Hopkins University (JHU). Several other COVID-19 datasets were made available regarding online conversations on Twitter, summaries of scholarly articles and epidemiology [2–5]. Our bioresource is novel, as it provides not only a concise dataset on epidemiology but also additional data about demographics, economics, tuberculosis (Bacille Calmette-Guérin, BCG) vaccination and online search trends. After proper statistical analysis, this data may be used to draw novel conclusions. We hope this dataset will be of use to researchers, particularly to those at the beginning of their career.

Data sources and initiatives

Repurposing of data for research is supported by the WHO and other medical organizations worldwide as this type of collaboration may lead to the discovery of new information concerning the COVID-19 threat. There are three major sources of daily-updated COVID-19 epidemiology (i.e. incidence and mortality) data: WHO, ECDC and JHU as seen on Table 1. Third-party aggregators, such as GitHub scrape data from the above repositories to make it simpler to view and analyse. In some cases, users may need to create an account for free to download the information.

Due to the limited capacity and accessibility of testing for SARS-CoV-2 in many countries worldwide, there may be a substantial difference between the confirmed number of COVID-19 cases and the total number of COVID-19 cases (See Table 1).

Dataset to repurpose

The data sources shown in Table 1 provide the information about the following variables: incidence,
mortality, country and population size. We added several new variables to the shared database:

- gross domestic product (GDP) and GDP per capita,
- the number of days since the first reported case in the country,
- the number of days since January 1, 2020 (the first report of the novel coronavirus in Wuhan, China),
- the number of days since January 25th 2020 (the first reported case in Europe),
- cumulative incidence/mortality in Europe and worldwide,
- incidence/mortality per 1000 citizens of a country,
- case fatality rate (CFR, the proportion of deaths from COVID-19 among all diagnosed individuals) was calculated. CFR in Europe/worldwide/country, indicated European countries and the European Union,
- tuberculosis (BCG) vaccination policies and practices [6].

Free statistical analysis software includes: R and Past [7,8]. In particular, the Past software supports a broad range of statistics such as Monte Carlo simulation, cross-correlation, analysis and removal of serial correlations in time series, principal coordination analysis, spherical data and Kernel densities. Moreover, the statistics derived from Past’s palaeontological science category may be applied in various clinical analyses [9]. MedCalc software (free 15-day trial) may also be useful [10]. The logarithmic increase of COVID-19 cases in the early phase of the pandemic may be analysed with geographical data. Whereas Our World in Data provides interesting and free to use/embed graphs [11].

**Potential uses**

We suggest several research questions to potentially explore in future studies:

- What is the influence of the COVID-19 pandemic on global mortality due to other illnesses? To what extent does the overall mortality due to COVID-19 differ from a country’s baseline mortality level? The EuroMOMO website [12] may be useful for this analysis as it provides information on all causes of mortality in 24 European countries.
- When, where and what kind of public policies significantly reduced the spread of COVID-19 and/or ended the epidemic? The effectiveness of public policies worldwide may be correlated with the graph shown in Figure 1.
- Are the incidence and/or CFR in a particular country correlated with its population density, social distancing policies and its society’s adherence to restrictions? The information collected by Our World in Data could help group countries and continents based on their CFR and incidence as seen on Figure 2.

**Daily confirmed COVID-19 deaths: are we bending the curve?**

Shown is the 7-day rolling average. Limited testing and challenges in the attribution of the cause of death means that the number of confirmed deaths may not be an accurate count of the true number of deaths from COVID-19.

![Figure 1. Daily confirmed COVID-19 deaths: are we bending the curve? [11]](source: European CDC - Situation Update Worldwide - Last updated 28th April, 11:30 (London time) - OurWorldinData.org/coronavirus • CC BY)
Datasets and future research suggestions concerning SARS-CoV-2

• Are the internet search trends correlated with the incidence and mortality of COVID-19 in a particular country? Or is this more due to media clamor? Google Trends may be helpful in this analysis [13].

• What is the educational quality of YouTube videos concerning COVID-19? Several studies were published on this topic [14–16]. The Google Chrome extension “vidIQ Vision for YouTube” may be used to access additional statistics that are normally not available on the YouTube website and provide the exact numbers of likes, dislikes and the like ratio, as seen on Figure 3.

• Are the internet search trends correlated with the incidence and mortality of COVID-19 in a particular country? Or is this more due to media clamor? Google Trends may be helpful in this analysis [13].

• How does the density and movement of people influence the incidence and mortality of COVID-19?

• What words are users worldwide searching for during the COVID-19 pandemic. The online software Keywords Explorer may be used for this study [17].

• When did people stop traveling and what influences them to maintain their social distancing? Recent data provided by Apple Inc. on their device travel patterns have been published online and may be useful [18].

Figure 2. Total confirmed COVID-19 deaths vs. cases [11]

Figure 3: A screenshot of the statistics obtained from “vidIQ Vision for YouTube.”
The data suggests that March 11, 2020 was the date that traffic in many European countries was officially restricted. This information may be correlated with media announcements.

- Online academic discussion forums like ResearchGate may offer additional new research ideas, links to datasets and an open discussion on various problems [19]. Data compiling initiatives such as Lens, provide an overview of information published on COVID-19 [20].
- The effectiveness of telemedicine in regards to COVID-19 treatment and treatment of other disease? [21,22].

**Summary**

Although several articles about COVID-19 are published daily, there are still several social, geographical and medical aspects which may be analysed. In this bioresource paper, we gathered relevant data concerning COVID-19 so that it is more convenient for researchers to analyse how the disease has developed over time. We hope that this bioresource paper and corresponding bioinformatics regarding the COVID-19 threat may encourage research, contribute to further understanding of epidemics, contribute to faster control of virus spread and help discover new scientific ideas.

**Contact information**

For those who are interested in scientific cooperation, who require an update to the database or who require help in realizing their scientific-oriented ideas may contact the corresponding author of this paper.

**References**