EUROPEAN JOURNAL OF TRANSLATIONAL
AND CLINICAL MEDICINE 2020;3(2):80-85

BIORESOURCE PAPER

Datasets and future research suggestions concerning SARS-CoV-2

Tomasz Szmuda[1], Shan Ali[2], Cathrine Özdemir[2], Mohammad Talha Syed[2], Akshita Singh[2], Tarjei Vevang Hetzger[2], Philip Rosvall[2], Karolina Fedorow[2], Ahmed Alkhater[2], Anders Majlöf[2], Mohammed Albrahim[2], Eyad Alquraya[2], Rakan Al Dunquwah[2], Zahraa Al-hakeem[2], Emad Almohisin[2], Mohammed Alradhi[2], Weronika Magdalena Żydowicz[2], Charlene Müller[2], Ada Egeland[2], Aurora Bergersen Kinstad[2], Jessica Ngyuen[2], Martin Bergersen Kinstad[2], Ibraheem Al-Rubaye[2], Yasmin Al-Khazragi[2], Beatrice von Dardel[2], Bhakti Dave[2], Paweł Słoniewski[1], Justyna Fercho[1], Sara Kierońska[3]

Tomasz Szmuda MD, PhD[1] and Shan Ali[2] 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

Szmuda T, ALi Sh, Özdemir C, Syed MT, Singh A, Hetzger TV, Rosvall P, Fedorow K, Alkhater A, Majlöf A, Albrahim M, Alquraya E, Al Dunquwah R, Al-hakeem Z, Almohisin E, Alradhi M, Żydowicz WM, Müller Ch, Egeland A, Kinstad AB, Ngyuen J, Kinstad MB, Al-Rubaye I, Al-Khazragi Y, von Dardel B, Dave B, Słoniewski P, Fercho J, Kierońska S. Eur J Transl Clin Med. 2020;3(2):80-85.
DOI: 10.31373/ejtcm/124734

Corresponding author:
Tomasz Szmuda, Neurosurgery Department, Medical University of Gdańsk, Gdańsk, Poland
e-mail: tszmuda@gumed.edu.pl
No external funds.
Available online: www.ejtcm.gumed.edu.pl

Download link to the database:
https://ejtcm.gumed.edu.pl/files/60

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,

Table 1. Main data sources comprising the national COVID-19 incidence, mortality, country and population size

Acronym	Organisation	Website	Raw Data or GitHub
WHO	World Health Organisation	https://covid19.who.int/table	Raw Data
ECDC	European Center for Disease Prevention	https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide	Raw Data
JHU	Johns Hopkins University	https://coronavirus.jhu.edu/map.html	Data could be scrubbed https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data
OWD	Our World in Data	https://ourworldindata.org/covid-testing	third-party data, downloadable spreadsheets, visualizations are licensed under Creative Commons Attribution 4.0 International (CC BY 4.0)

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 polices 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.

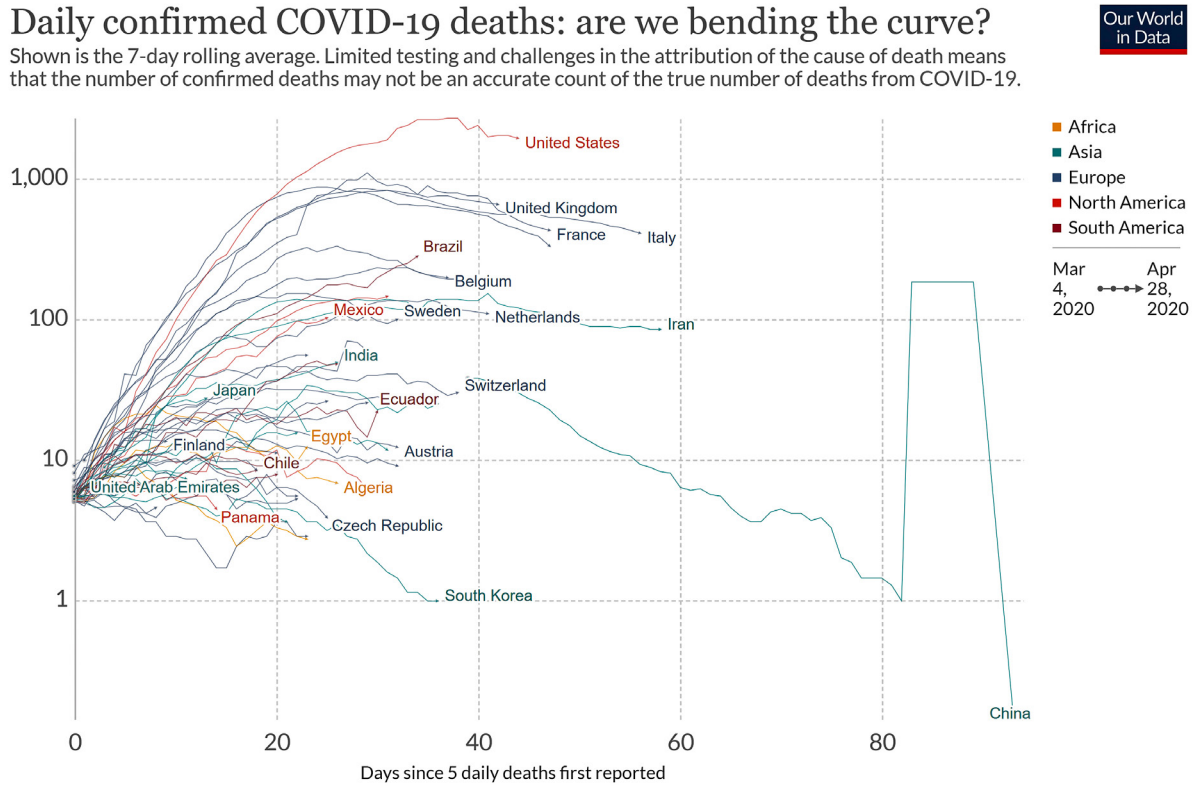

Figure 1. Daily confirmed COVID-19 deaths: are we bending the curve? [11]

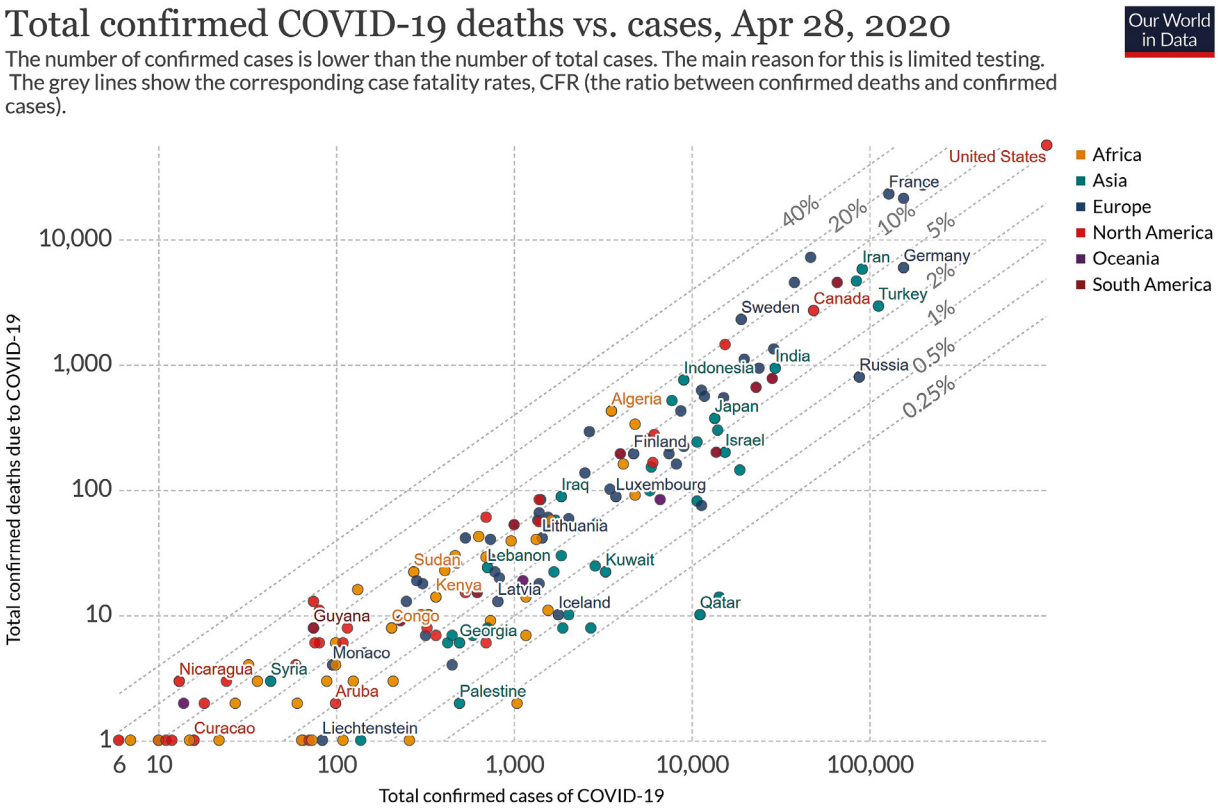

Figure 2. Total confirmed COVID-19 deaths vs. cases [11]

- 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].

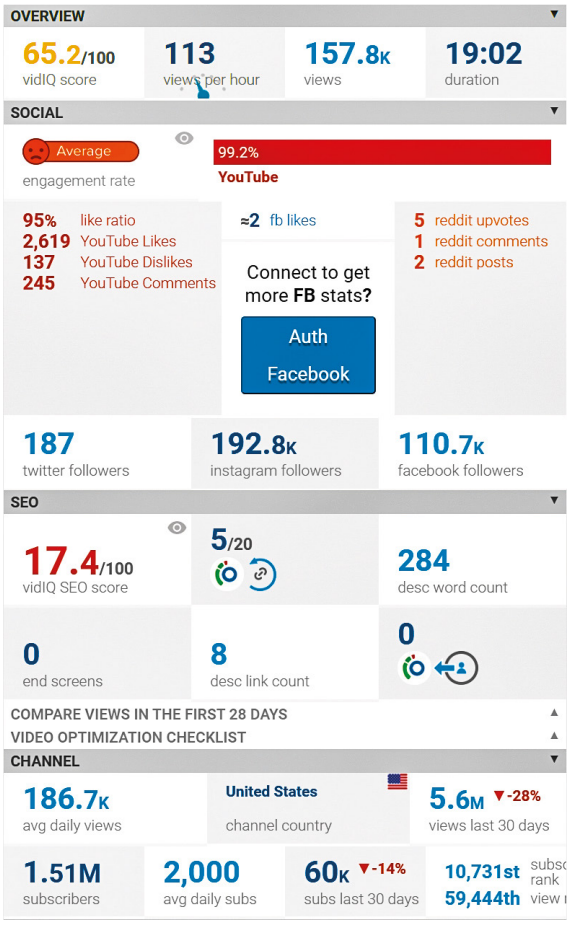

Figure 3: A screenshot of the statistics obtained from "vidIQ Vision for YouTube".

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

1. Smiatacz T. It didn't have to happen this way – what COVID-19 tells us about translational medicine. Eur J Transl Clin Med [Internet]. 2020 May 29;3(1):7–10. Available from: https://doi.org/10.31373/ejtcm/119455
2. Xu B, Gutierrez B, Mekaru S, Sewalk K, Goodwin L, Loskill A, et al. Epidemiological data from the COVID-19 outbreak, real-time case information. Sci Data [Internet]. 2020 Dec 24;7(1):106. Available from: http://www.nature.com/articles/s41597-020-0448-0
3. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis [Internet]. 2020 May;20(5):533–4. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1473309920301201
4. COVID-19 Open Research Dataset (CORD-19) [Internet]; [cited 2020 Jul 23]. Available from: https://www.semanticscholar.org/cord19
5. Chen E, Lerman K, Ferrara E. Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set. JMIR Public Heal Surveill [Internet]. 2020 May 29;6(2):e19273. Available from: http://publichealth.jmir.org/2020/2/e19273/
6. Zwerling A, Behr MA, Verma A, Brewer TF, Menzies D, Pai M. The BCG World Atlas: A Database of Global BCG Vaccination Policies and Practices. PLoS Med [Internet]. 2011 Mar 22;8(3):e1001012. Available from: https://dx.plos.org/10.1371/journal.pmed.1001012
7. R Core Team. The R Project for Statistical Computing [Internet]. The R Foundation; [cited 2020 Apr 26]. Available from: https://www.r-project.org/
8. University of Oslo. PAST [Internet]. Oslo: University of Oslo; [cited 2020 Apr 26]. Available from: https://past.en.lo4d.com/windows
9. Szmuda T, Słoniewski P, Ali S, Dzierżanowski J, Kamieniecki A, Siedlecki K. Can sectioning the posterior communicating artery be predicted with computed tomography angiography in the microsurgical clipping of basilar apex aneurysms? Acta Neurochir (Wien) [Internet]. 2020 Mar 20;162(3):567–79. Available from: https://doi.org/10.1007/s00701-019-04138-2
10. MedCalc. MedCalc statistical software [Internet]. MedCalc Software Ltd; [cited 2020 Apr 26]. Available from: https://www.medcalc.org/
11. Roser M, Ritchie H, Ortiz-Ospina E. Coronavirus Disease (COVID-19) – the data [Internet]. Available from: https://ourworldindata.org/coronavirus
12. European Centre for Disease Prevention and Control, World Health Organization. EuroMOMO [Internet]; [cited 2020 Apr 26]. Available from: https://www.euromomo.eu/
13. Google Inc. Google Trends [Internet]. [cited 2020 Apr 14]. Available from: https://trends.google.com/trends/?geo=UK
14. Szmuda T, Rosvall P, Hetzger TV, Ali S, Słoniewski P. YouTube as a Source of Patient Information for Hydrocephalus: A Content-Quality and Optimization Analysis. World Neurosurg [Internet]. 2020;138:e469–77. Available from: http://www.sciencedirect.com/science/article/pii/S1878875020304241

15. Khatri P, Singh SR, Belani NK, Yeong YL, Lohan R, Lim YW, et al. YouTube as source of information on 2019 novel coronavirus outbreak: a cross sectional study of English and Mandarin content. Travel Med Infect Dis [Internet]. 2020 May;35:101636. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1477893920301046
16. Szmuda T, Özdemir C, Fedorow K, Ali S, Słoniewski P. YouTube as a source of information for narcolepsy: A content-quality and optimization analysis. J Sleep Res [Internet]. 2020 Apr 21;e13053. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/jsr.13053
17. Ahrefs Pte. Ltd. Ahrefs Keywords Explorer [Internet]. Singapore: Ahrefs Pte. Ltd; [cited 2020 Apr 26]. Available from: https://ahrefs.com/keywords-explorer
18. Apple Inc. COVID-19 - Mobility Trends Reports [Internet]; [cited 2020 Apr 26]. Available from: https://www.apple.com/covid19/mobility
19. Madisch I, Hofmayer S, Fickenscher H. COVID-19 research community [Internet]. ResearchGate GmbH. 2020; [cited 2020 Apr 26]. Available from: https://www.researchgate.net/community/COVID-19/discussions
20. About The Lens: COVID-19 Datasets [Internet]; [cited 2020 Apr 26]. Available from: https://about.lens.org/covid-19/
21. Szmuda T, Ali S, Słoniewski P, Group NsW. Telemedicine in neurosurgery during the novel coronavirus (COVID-19) pandemic. Neurol Neurochir Pol [Internet]. 2020 Apr;54(2):207–8. Available from: https://journals.viamedica.pl/neurologia_neurochirurgia_polska/article/download/PJNNS.a2020.0038/50703
22. Szmuda T, Özdemir C, Ali S, Singh A, Syed MT, Słoniewski P. Readability of online patient education material for the novel coronavirus disease (COVID-19): a cross-sectional health literacy study. Public Health [Internet]. 2020 Aug;185:21–5. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0033350620302031