nCov2019: A R package for studying the COVID-19 coronavirus pandemic

Tianzhi Wu ¹, Erqiang Hu ¹, Xijin Ge ², Guangchuang Yu  Corresponding Author

¹ Department of Bioinformatics, School of Basic Medical Sciences, Southern Medical University, Guangzhou, China
² Department of Mathematics and Statistics, South Dakota State University, Brookings, United States

Corresponding Author: Guangchuang Yu
Email address: gcyu1@smu.edu.cn

Background. The global spreading of the COVID-19 coronavirus is still a serious public health challenge. Although there are a large number of public tools that provide statistics data, tools for retrospective historical data and convenient visualization are still valuable. To provide convenient access to epidemic statistics data and visualization on the virus spreading, we developed an R package, nCov2019 (https://github.com/YuLab-SMU/nCov2019).

Methods. We collect stable and reliable data of COVID-19 cases from multiple authoritative and up-to-date sources, and aggregate the most recent and historical data for each country or even province. Medical progress information, including global vaccine development and therapeutics candidates, were also collected and can be directly accessed in our package. The nCov2019 package provides an R language interfaces and designed functions for data operation and presentation, a set of interfaces to fetch data subset intuitively, visualization methods, and a dashboard with no extra coding requirement for data exploration and interactive analysis.

Results. As of January 14, the global epidemic crisis is still serious. The number of worldwide confirmed cases has reached 91,268,983. Following the US, India has reached 10 million level confirmed cases. The growth rate has shown that multiple peaks in many countries and some of them are now at the tail of the second epidemic wave. A third wave of COVID-19 remains a threat and is a cause for alarm. Under the efforts of researchers, 51 vaccines and 54 drugs are under development and 14 of these vaccines are already in Pre-clinical phase.

Discussion. The nCov2019 package provides detailed statistics data, visualization functions and Shiny web application, which allows researchers to keep abreast of the latest epidemic spread overview.
nCov2019: A R package for studying the COVID-19 coronavirus pandemic

Tianzhi Wu¹, Erqiang Hu¹, Xijin Ge²*, Guangchuang Yu¹*

¹ Department of Bioinformatics, School of Basic Medical Sciences, Southern Medical University, Guangzhou, China
² Department of Mathematics and Statistics, South Dakota State University, Brookings, SD 57007, USA.

Corresponding Author:
Guangchuang Yu¹
Xijin Ge²

¹Department of Bioinformatics, School of Basic Medical Sciences, Southern Medical University, Guangzhou, China
²Department of Mathematics and Statistics, South Dakota State University, Brookings, SD 57007, USA. correspondence

Email address: gcyu1@smu.edu.cn, Xijin.Ge@sdsate.edu

Abstract

Background. The global spreading of the COVID-19 coronavirus is still a serious public health challenge. Although there are a large number of public tools that provide statistics data, tools for retrospective historical data and convenient visualization are still valuable. To provide convenient access to epidemic statistics data and visualization on the virus spreading, we developed an R package, nCov2019 (https://cran.r-project.org/package=nCov2019).

Methods. We collect stable and reliable data of COVID-19 cases from multiple authoritative and up-to-date sources, and aggregate the most recent and historical data for each country or even province. Medical progress information, including global vaccine development and therapeutics candidates, were also collected and can be directly accessed in our package. The nCov2019 package provides an R language interfaces and designed functions for data operation and presentation, a set of interfaces to fetch data subset intuitively, visualization methods, and a dashboard with no extra coding requirement for data exploration and interactive analysis.
Results. As of January 14, the global epidemic crisis is still serious. The number of worldwide confirmed cases has reached 91,268,983. Following the US, India has reached 10 million level confirmed cases. The growth rate has shown that multiple peaks in many countries and some of them are now at the tail of the second epidemic wave. A third wave of COVID-19 remains a threat and is a cause for alarm. Under the efforts of researchers, 51 vaccines and 54 drugs are under development and 14 of these vaccines are already in Pre-clinical phase.

Discussion. The nCov2019 package provides detailed statistics data, visualization functions and Shiny web application, which allows researchers to keep abreast of the latest epidemic spread overview.

Introduction

The COVID-19 pandemic emerged at the end of 2019 from Wuhan, China\cite{1,2}. The virus has raged globally for more than 12 months and still in the process of accelerating spread. Currently it remains an extremely serious public health challenge, affecting more than 220 countries worldwide. After a year-long battle, the researchers have gained more insight into the transmission route, molecular structure, sequence information and pathogenesis of the novel COVID-19 virus\cite{3–6}. Experts in epidemiology and data science also play a great role in this; they collected data and developed many convenient tools to deliver information on the latest infection numbers, high-risk areas and so on. World Health Organization (WHO)\cite{7} and national Center for Disease Control and Prevention (CDCs) publish authoritative data but usually on a daily basis update frequency and the data often appear to lag behind. In contrast, statistics reported by aggregated news media, such as DXY\cite{8}, WorldoMeters\cite{9}, ourworldindata\cite{10} and BNO News\cite{11}, are usually updated more frequently and published before the CDCs. The Johns Hopkins Center for Systems Science and Engineering (CSSE) integrates statistics from these data source and provides an online visual dashboard\cite{12}, it is the most popular and frequently updated platform we could find. Their efforts have facilitated the public to get the latest information on the spread of the epidemic.

Data analysis and customized visualizations are essential to the medical researchers and epidemiology experts, for analyzing changes of spreading, monitoring new outbreak trends, assessing current health measures and so on. Obtaining data is a prerequisite for data analysis. As time goes on, there are many websites and resources that provide COVID data. Currently, several R packages are available to provide different types of data. For instance, covid19.analytics\cite{13} provides virus sequence queries, as well as dashboard and SIR models; COVID19\cite{14} and coronavirus\cite{15} packages provide detailed vaccine and case test data. Here, we provide an R package, nCov2019, which aims to access, visually explore and analyze epidemic related statistics data in R (Figure 1). The nCov2019 package is more comprehensive compare to other R packages. It provides more data types including real-time and historical infection statistics, therapeutic and vaccine data. In addition, our tool provides several convenient and practical visualization functions
and an interactive dashboard based on Shiny web application with no code requirements to help
users exploring the data visually. The nCov2019 package was developed since Jan, 2020 and is
one of the earliest R packages that designed to query COVID data to support epidemiology
modeling. It is available at the time there were few data resources available and was cited in several
academic articles\[16–19\].

Materials & Methods

The main purposes of developing nCov2019 are to reduce the barriers of data acquisition and
to provide essential visualization functions, including dynamic visualization to monitor the spread
of the virus in an easier and concise way. To achieve these goals, the nCov2019 package was
designed with four main parts: statistics data collection, statistics data query and operation,
geographic maps visualization and interactive dashboard.

Statistics data collection

The Statistics data usually contains the latest status and historical data. For the real-time latest
status, we chose WorldoMeters as our data source, which has the high update frequency. Our
historical data source is CSSE, which integrates data from multiple source centers and the data is
reliable and timely. In addition, after one year of continuous battle of the COVID-19 virus,
researchers have developed some vaccines, and at the same time, doctors have tried different
treatment options. These data are more important than simple summary of confirmed, recovered
and death cases. We also collected data on vaccine development and drug therapeutics progress
from two website of raps.org\[20,21\] and these data can also be accessed directly in our nCov2019
package.

Statistics data query and operation

For ease of use, we wrapped the data fetching process into a simple function, \( \text{query}() \). It usually
needs to be executed only once in a session, then users can obtain five datasets. The latest data
contains detailed statistics status at the time of query, and the historical data contains daily statistics
for each country, which is often useful to model epidemic growth. We also provide a global
summary to help understand the latest progress in the fight against the COVID-19 pandemic.

To facilitate downstream data analysis, we defined the \( \mathcal{I} \) operator which mimics the API of
data selection for data frame in R. So that data can be easily accessed by specific regions. For
example, let \( X \) be the historical data, then \( \text{Xfc("USA","UK")} \) will return the historical data table
for USA and UK only. These data are organized in long format and can be directly handed over to
ggplot2 for plotting in R.

The vaccine and drug therapeutics development status can be queried at the same time. The data contains the latest information about candidate medication class, mechanism of vaccine, trade name for drugs and their current trial phase. Both datasets have the details information such as background, develop aims and trial details. These data are helpful for users to understand the medicine progress of epidemic prevention and control.

Geographic maps visualization

Geographic visualization is an effective way to observe the spatial patterns of virus spread. We provided built-in and convenient geographic map visualization functions within the nCov2019 package. The visualization functions were wrapped into a simple and easy-to-use command as `plot()`, so that users can plot the distribution of cases on the maps of the worldwide or national scope. For example, let $X$ be the latest data in query result, then `plot(X)` will plot the world map contained global confirmed cases (Figure 2).

To review and analyze the spread of COVID-19 epidemic situation, a more informative application is to draw dynamic geographic maps at multiple time points. Users can easily get the dynamic map by just specifying the start and end time in our tool. For instance, by performing `plot(X, from="2020-03-01", to="2020-08-01")` function, an animation will be generated to reflect transmission and spread dynamic of the COVID-19 outbreak during these time points (Figure S1).

Interactive Dashboard

We also developed an interactive web dashboard to help users to access and explore these datasets by interactive mouse clicks. Built with the RStudio Shiny framework, the dashboard could be launched with the function, `dashboard()`. It enables users to select their regions of interest and check both the historical and real-time data. The statistics of confirmed, deaths, and recovered cases will be displayed in the Dashboard header, followed by a downloadable statistics data table, nearby their cumulative curve (Figure 3).

Multiple charts are designed on the bottom of dashboard. A chart of global statistics could display nine types of statistics on the world map, such as confirmed cases, active infected cases, number of detections, and population for each country and so on. Vaccine and therapeutics summary table is shown next to global statistics chart. For exploring the dataset easier, we provided an interaction plot; it can be used to explore the relationship between any two of the twenty different statistics, such as whether there is an association between confirmed cases number and the total detection number. Finally, we designed a curve plot to reflecting the growth intensity by daily increase cases for each country. This chart could be used to monitor the outbreak strength over time. One of its practical applications is to determine whether a second wave of outbreaks is occurring.
Results

At present, the global epidemic crisis is still serious. According to data calculations from the nCov2019 package, as of January 14, the number of worldwide confirmed cases is 91,268,983 (Figure 1), the cumulative number of deaths has reached to 1,951,790. Two country have more than 10 million confirmed cases, (23,616,515 confirmed cases, 236,631 per million population in the United States and 1,051,283 confirmed cases, 7578 per million population in India), which indicating that there is still great pressure for COVID-19 prevention and control.

Detection of infected persons and cutting off the transmission route are the key solutions to curb the virus spread. Detection status could be visualized by using the visual map function provided in our package (Figure 4). Intuitively, it can be seen that Asia and North America performed the highest number of detection tests for the virus, compared to Africa, which had the lowest number of detections. This may be due to a temporary lack or delay of reported data in Africa countries. But given the situation of general economic backwardness and large population, it is needed for other countries to support the fighting against the pandemic in Africa.

Over the past year, the epidemic has shown multiple peaks in many countries, such as Australia, Japan, Italy, Germany and China. The epidemic in some countries were under control by middle in 2020, but then the confirm cases has rebounded (Figure S2). While these countries are now at the tail end of the second wave of the epidemic, the possibility of a third wave in the future is a cause for alarm. Some countries, such as the United States, India, Russia, and South Africa, are still in the midst of a severe and rapid increase.

Researchers are playing an active role in the fight against the virus. There are currently 51 vaccines in development or in use worldwide. Due to the emergencies, most vaccines and regimens are using a simultaneous multi-clinical trial approach. Fourteen of these vaccines are already in pre-clinical phase. As for the therapeutics, there are currently 54 candidates, including HIV protease inhibitor, IL-6 receptor agonist, HIV-1 Rev protein inhibitor, Autologous adipose-derived stem cells, and Monoclonal antibody.

All of the above information is available directly in the R package either using command line or dashboard (Figure 3). Detailed vignette including the usages of data acquisition and visualization functions could be found in the supplemental file and the package is hosted on CRAN (https://cran.r-project.org/package=nCov2019).

Discussion

We provide practical and concise tools for searching outbreak data, as well as vaccine and treatment-related data. The data are sourced from reliable platforms. As a highlight, our nCov2019 package is designed not only to help clinicians and other researchers to obtain statistics table, but
also to help them reducing the difficulty of visualizing comprehensive data. Researchers could
utilize our package to track the latest statistic status, conduct retrospective analysis or establish
predictive models. Our package was collected in several resource list, such as COVID-19
Coronavirus Disease Resources[22] and Top 100 R resources on Novel COVID-19 Coronavirus[23],
and was served as the back end for COVID-19 Pandemic Analysis Platform[24].

The epidemic has been going on for over a year and there is now a lot of online dashboard[25] or
websites[26][27] with great appearance design. For the public who just want to know the latest status
of the epidemic, it is the most convenient way by simply browse from the internet. In contrast, our
R package is designed for both command line and dashboard interaction analysis, it is more
suitable for researchers and data analyzers who wish to explore the epidemic data by themselves.
At present, our package has provided data support for several academic research[16,18,19,28,29], and
we hope our tool will continue to facilitate similar studies.

Conclusions

Our nCov2019 package reduces the barrier for researchers and public health officials in
obtaining comprehensive, up-to-date data on this ongoing outbreak. With this package,
epidemiologists and other scientists can directly access data, produce convenient and rich
visualization images, facilitating mathematical modeling and forecasting of the COVID-19
outbreak. The interactive web dashboard is accessible to the general public and could also be easily
customized by researchers to produce other dashboards or track other countries. We hope our tool
be helpful to the COVID-19 relevant research work.

Acknowledgements

This work was supported by Startup fund from Southern Medical University (G618289088).

References

1. Zhu, N.; Zhang, D.; Wang, W.; Li, X.; Yang, B.; Song, J.; Zhao, X.; Huang, B.; Shi, W.; Lu, R.; Niu, P.; Zhan, F.; Ma,
   X.; Wang, D.; Xu, W.; Wu, G.; Gao, G.F.; Tan, W. A Novel Coronavirus from Patients with Pneumonia in China, 2019. New
   England Journal of Medicine Massachusetts Medical Society: 2020, 382(8), 727–733 10.1056/NEJMoa2001017.

2. Wang, C.; Horby, P.W.; Hayden, F.G.; Gao, G.F. A novel coronavirus outbreak of global health concern. Lancet
   (London, England) 2020, 395(10223), 470–473 10.1016/S0140-6736(20)30185-9.

3. Lippi, G.; Mattiuzzi, C.; Bovo, C.; Plebani, M. Current laboratory diagnostics of coronavirus disease 2019 (COVID-
   19). Acta Bio-Medica: Atenei Parmensis 2020, 91(2), 137–145 10.23750/abm.v91i2.9548.
245  24. COVID-19 Pandemic Analysis Platform. n.d., http://14.215.135.56:3838/COVID-19-EN/ [Oct 22 2020].
246  25. Kuwait Corona Virus Platform:COVID19. n.d., https://m-alenezi.shinyapps.io/CoronaKW3/ [Jan 15 2021].
247  26. WHO COVID-19 Explorer. n.d., https://worldhealthorg.shinyapps.io/covid/ [Jan 15 2021].
248  27. The 2019-2020 Coronavirus Pandemic - A Timeline. n.d., https://milano-r.github.io/erum2020-covidr-contest/hahn-covid-shinyline.html [Jan 15 2021].
249  28. Ran, J.; Zhao, S.; Han, L.; Chen, D.; Yang, Z.; Yang, L.; Wang, M.H.; He, D. The ambient ozone and COVID-19 transmissibility in China: A data-driven ecological study of 154 cities. The Journal of Infection 2020, 81(3), e9–e11 10.1016/j.jinf.2020.07.011.
250  29. Zhuang, Z.; Zhao, S.; Lin, Q.; Cao, P.; Lou, Y.; Yang, L.; Yang, S.; He, D.; Xiao, L. Preliminary estimates of the reproduction number of the coronavirus disease (COVID-19) outbreak in Republic of Korea and Italy by 5 March 2020. International Journal of Infectious Diseases Elsevier: 2020, 95, 308–310 10.1016/j.ijid.2020.04.044.

Figure 1: The Diagram of nCov2019 architecture. nCov2019 is designed for both command line (CLI part) and dashboard interaction analysis (GUI part), while `dashboard()` is the main entry for the GUI explore part, the `query()` is the main function used in CLI explore part. nCov2019 also provides a map visualization and an animation function, the latter can be used to draw dynamic maps.

Figure 2: The world map visualizes the global spread of the COVID-19. The map was filled by colors indicating confirm cases of each country and the global view shows that the COVID-19 pandemic is a serious public health challenge. The plot was generated on 2021-01-14 with build in function of the nCov2019 package.

Figure 3. The user interface of the dashboard provided in the nCOV2019 package. The panel includes statistical data query, map visualization, interactive analysis and summary of the therapeutics and vaccine development. User could launch it with the function `dashboard()`.

Figure 4: The map visualized by nCov2019, reflecting the amount of COVID-19 testing in different countries around the world. While Asia and North America performed the highest number of detection tests for the virus, Africa had the lowest number of detections, although this may be due to temporary lack or delay of reported data in Africa countries, it cause an alarm that more attentions should be given to African countries.
Figure S1: **Demo of the animated world-wide epidemic maps generation.** The set of maps in specify time range could be plotted by simple commands as show at the center.

Figure S2: **The epidemic has shown multiple peaks in many countries over the past year.** Similar chart can be explored interactively on the dashboard.
nCov2019 is designed for both command line (CLI part) and dashboard interaction analysis (GUI part), while `dashboard()` is the main entry for the GUI explore part, the `query()` is the main function used in CLI explore part. nCov2019 also provides a map visualization and an animation function, the latter can be used to draw dynamic maps.
Figure 2

Figure 2: The world map visualizes the global spread of the COVID-19. The map was filled by colors indicating confirm cases of each country and the global view shows that the COVID-19 pandemic is a serious public health challenge. The plot was generated on 2021-01-14 with build in function of the nCov2019 package.
Figure 3

Figure 3. The user interface of the dashboard provided in the nCOV2019 package.

The panel includes statistical data query, map visualization, interactive analysis and summary of the therapeutics and vaccine development. User could launch it with the function `dashboard()`.
Figure 4: The map visualized by nCov2019, reflecting the amount of COVID-19 testing in different countries around the world.

While Asia and North America performed the highest number of detection tests for the virus, Africa had the lowest number of detections, although this may be due to temporary lack or delay of reported data in Africa countries, it cause an alarm that more attentions should be given to African countries.