COVID19-Tracker: A shiny app to perform comprehensive data visualisation for SARS-CoV-2 epidemic in Spain

Aurelio Tobías[1], Joan Valls[2], Pau Satorra[3], Cristian Tebé[3]

(equal contribution)

1. Institute of Environmental Assessment and Water Research (IDEA), Spanish Council for Scientific Research (CSIC), Barcelona, Spain. (aurelio.tobias@idaea.csic.es)
2. Department of Mathematics, Universitat Autònoma de Barcelona (UAB), Bellaterra, Barcelona, Spain.
3. Biostatistics Unit, Institut d'Investigació Biomèdica de Bellvitge (IDIBELL), Barcelona, Spain.

(Version 3.0 – Date: April 27th, 2020)

Abstract

Data visualization is an essential tool for exploring and communicating findings in medical research, especially in epidemiological surveillance. The COVID19-Tracker web application systematically produces daily updated data visualization and analysis of the SARS-CoV-2 epidemic in Spain. It collects automatically daily data on COVID-19 diagnosed cases, and mortality from February 24th, 2020 onwards. Several analyses have been developed to visualize data trends and estimating short-term projections; to estimate the case fatality rate; to assess the effect of the lockdown measures on the trends of incident data; to estimate infection time and the basic reproduction number; and to analyse the excess of mortality. The application may help for a better understanding of the SARS-CoV-2 epidemic data in Spain.

1. INTRODUCTION

The first confirmed cases of SARS-CoV-2 in Spain were identified in late February 2020 (1). Since then, Spain became, by the April 27^{h}, the second most affected country worldwide (236,199 diagnosed cases) and recorded the third number of deaths (23,521 deaths) due to the SARS-CoV-2 pandemic (2). Since March 16th, lockdown measures oriented on flattening the epidemic curve were in place in Spain, restricting social contact, reducing public transport, and closing businesses, except for those essential to the country's supply chains (3). However, these measures, even they were in the right direction, did not show to be enough to change the rising trend of the epidemic. For this reason, a more restrictive lockdown was suggested (4), and eventually undertaken by the Spanish Government on March 30th (5). On April 13th, a partial opening of the economic activity was allowed by the Spanish Government, allowing non essential activities from businesses not open to public.

Data visualization and analysis is an essential tool for exploring and communicating findings in medical research, and especially in epidemiological surveillance. It can help researchers and policymakers to identify and understand trends that could be overlooked if the data were reviewed in tabular form. We have developed a Shiny app that allows users to evaluate daily time-series data from a statistical standpoint. The COVID19-Tracker app systematically produces daily updated data visualization and analysis of SARS-CoV-2 epidemic data in Spain. It is easy to use and fills a role in the tool space for visualization, analysis, and exploration of epidemiological data during this particular scenario.

2. SOFTWARE AVAILABILITY AND REQUIREMENTS

The COVID19-Tracker app has been developed in RStudio (6), version 1.2.5033, using the Shiny package, version 1.4.0. Shiny offers the ability to develop a graphical user interface (GUI) that can be run locally or deployed online. Last is particularly beneficial to show and communicate updated findings to a broad audience. All the analyses have been carried out using R, version 3.6.3.

The application has a friendly structure based on menus to shown data visualization for each of the analyses currently implemented: projections, fatality rates, intervention, infection time, reproducibility number and mortality register analysis (Figure 1).

- ***Projections*** and ***Projections by age*** display the trends for diagnosed cases and mortality since the epidemic began, and estimates a 3-day projection.
- ***Fatality*** and ***Fatality by age*** display the trends for the case fatality rates.
- **Intervention** calculates and displays the effect of the lockdown periods on the trend of incident daily diagnosed cases and mortality.
- **Infection time** estimates and displays the incubation period for COVID-19 between the interval of exposure to SARS-CoV-2 and the date of COVID-19 diagnosis.
- **Reproducibility number** estimates and displays the average number of secondary cases of disease caused by a single infected individual over his or her infectious period.
- **Mortality registry** displays the evolution of the observed all-cause mortality during the epidemic. It also compares the observed and expected deaths to assess the excess of mortality.

Figure 1. Home page of the COVID19-Tracker application, for visualization and analysis of data from the SARS-CoV-2 epidemic in Spain. Available at: https://ubidi.shinyapps.io/covid19/

Two additional menus to describe the ***Methodology***, to describe the statistical details on the analyses already implemented, and ***Other apps***, which collects applications also developed in Shiny by other users to follow the COVID19 epidemic in Spain and globally.

The app has an automated process to update data and all analyses every time a user connects to the app. It is available online at the following link: https://ubidi.shinyapps.io/covid19/ and shortly free available on github as an R package. The displayed graphs are mouse-sensitive, showing the observed and expected number of events through the plot. Likewise, when selecting any plot, the application allows the option of downloading it as a *portable network graphic* (* .png). All menus are available in English, Spanish, and Catalan.

3. DATA SOURCES

We collected daily data on COVID-19 diagnosed cases and mortality, from February 24th onwards. Data is collected automatically every day daily from the Datadista Github repository (7). This repository updates data according to the calendar and rate of publication of the Spanish Ministry of Health/Instituto de Salud Carlos III (8).

4. METHODS

4.1. Projections

To estimate the observed data trends for the number of events, we used a Poisson regression model (9), allowing for over-dispersion (10), fitting lineal, quadratic and cubic terms:

$$\log(E(c_t)) = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3$$

Where t = 1, 2, …, T, represents the time unit (from the first observed day until the last, T consecutive days in total), and c_t is the number of events. The estimated regression parameters and their standard errors are used to obtain the short-term projections, up to three days, and their 95% confidence interval (95% CI).

Results are available nationwide by default, and at the regional level accessing to the dropdown menu for this purpose (Figure 2a). Trends and projections are also calculated by age group (0-39, 40-49, 50-59, 60-69, 70-79, and 80 or more years) (Figure 2b).

We should note that in previous versions of the application, linear or quadratic models were considered. Based on the evolution of the epidemic, these models were compared using a likelihood ratio test based on their deviances. In the current version, the cubic model is the one showing the best goodness of fit. In any case, the models are regularly being evaluated, in case a model reformulation with a better fit is necessary during the course of the epidemic.

4.2. Case fatality rate

The case fatality rate is defined as the ratio between the number of deaths and the diagnosed cases (11). Thus, an offset is fitted into the Poisson regression model, also allowing for overdispension, as the logarithm of the diagnosed cases:

$$\log(E(m_t)) = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \log(c_t)$$

Where m_t is the daily number of deaths, and c_t is the daily number of diagnosed cases. Case fatality rates are also calculated for the same age groups (Figure 2c).

We should acknowledge that it is not possible to make an accurate estimate of the case fatality rates due to underreporting of cases diagnosed in official statistics (12). Nonetheless, the estimation and monitoring of the case fatality rates monitoring are of espeical interest in the current epidemic scenario.

4.3. Intervention analysis

To assess the effect of the lockdown on the trend of incident diagnosed cases and mortality, we used an interrupted time-series design (13). The data is analyzed with quasi-Poisson regression with an interaction model to estimate the change in trend:

$$\log(E(c_t)) = \beta_0 + \beta_1 t + \beta_2 \text{lockdown} + \beta_3 t * \text{lockdown}$$

Where *lockdown* is a variable that identifies the intervals before and during the lockdown periods imposed by the Spanish Government (3, 5) (0=before March 15th,

2020; 1=between March 16th and March 29th, 2020; 2= between Mach 30th and April 12th, 2020; 3=after April 13th, 2020).

We should acknowledge that this is a descriptive analysis without predictive purposes (Figure 2d). For an easy interpretation, and comparison of the effectiveness of lockdown measures between countries, a linear trend is assumed before and during each lockdown period (14). Although not accounted for residual autocorrelation, the estimates are unbiased but possibly inefficient. This analysis also shows the results nationwide in table reporting the daily percentage (%) mean increase, and its 95% CI.

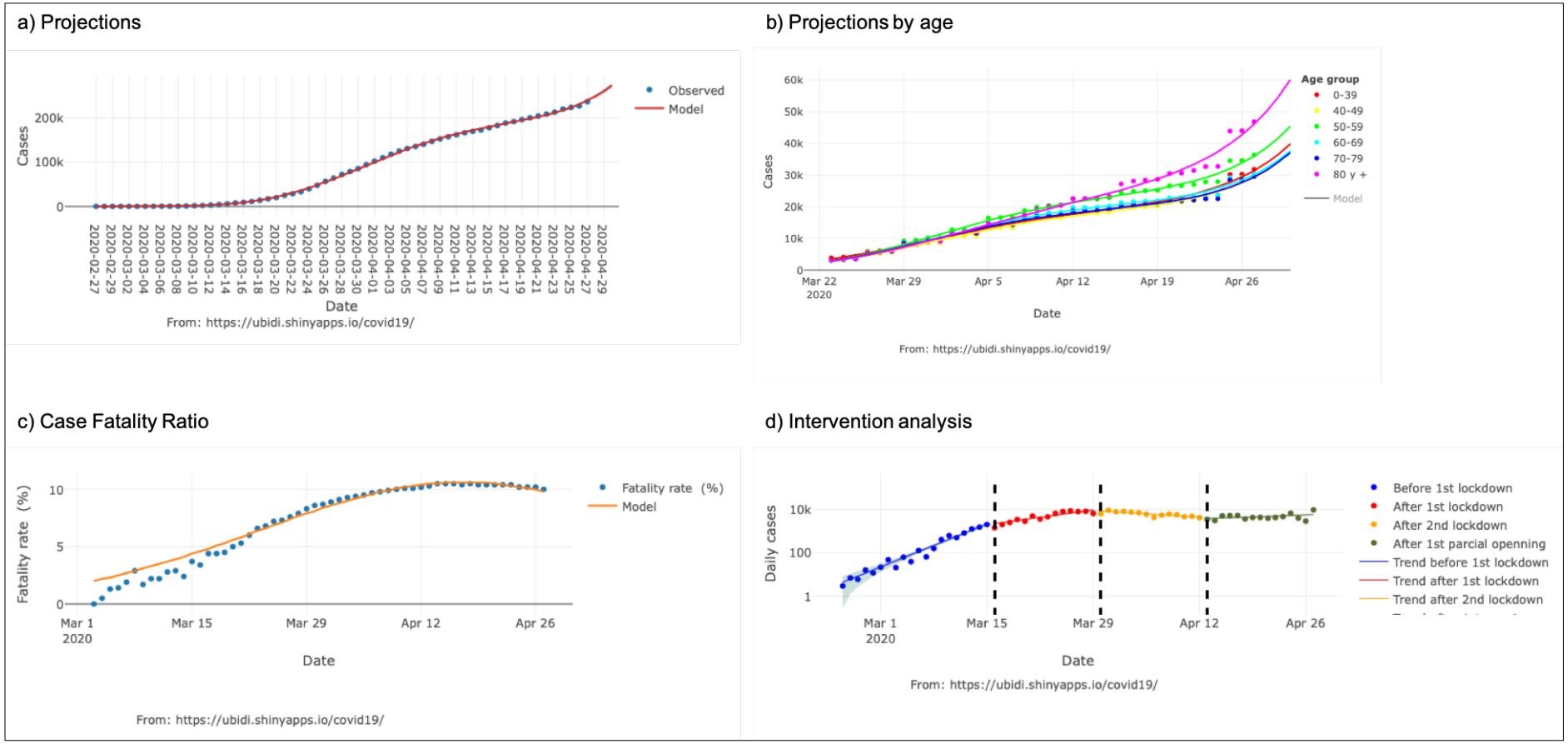

Figure 2. Standard output display of the COVID19-Tracker application (results updated to April 27th, 2020), for trend analysis and its 3-day projection at the national level (a) and by age group (b), of the fatality rate (c), and intervention analysis to evaluate the effect of alarm states on incident data (d).

4.4. Estimation of time to infection

Lauer et al. (15) have recently analyzed the incubation period for COVID-19 in a cohort of symptomatic patients. From each patient, they collected the interval of exposure to SARS-CoV-2 and the date of appearance of symptoms. They assumed that the incubation time would follow, as in other viral respiratory tract infections, a Lognormal distribution.

$$Lognormal(mu,sigma^2) = Lognormal(1.621\ ,\ 0.418)$$

In a naïve exercise with some limitations, we have replicated this distribution in the group of diagnosed cases to approximate the date of exposure to SARS-CoV-2 recursively:

$$q(i) = \sum_{j=1}^{14} P(j) \times c_{j+i}$$

Where p is the number of diagnosed cases on day i, q is the number of infected cases on day i-j, and j = 1, 2, …, 14 the maximum time it is expected that the disease can develop. P(j) is the probability of presenting symptoms on day j according to a Lognormal law with the parameters defined by Lauer et al. (15)

To estimate the last 14 days, since the information on the diagnosed cases was not available for the next 14 days, a quadratic model was used to project diagnosed cases. These latest estimates are displayed in the application with a different color.

Results are available nationwide by default (Figure 3a), and at the regional level accessing to the dropdown menu for this purpose .

4.5. Basic reproduction number

The basic reproduction number (R_0) is the average number of secondary cases of disease caused by a single infected individual over his or her infectious period (16). This statistic, which is time and situation specific, is commonly used to characterize pathogen transmissibility during an epidemic. The monitoring of R_0 over time provides feedback on the effectiveness of interventions and on the need to intensify control efforts, given that the goal of control efforts is to reduce R_0 below the threshold value of 1 and as close to 0 as possible, thus bringing an epidemic under control.

Here, we use the R package *EpiEstim* to estimate the basic reproduction number R_t using the Wallinga and Teunis method (16). This parametric method assumes a gamma distribution for the serial interval. The serial interval is the time between the onset of symptoms in a primary case and the onset of symptoms of secondary cases, which is needed to estimated R_t over the course of an epidemic.

The mean and standard deviation of the serial interval distribution can vary depending on the disease (16). Recently Nishiura et al. (17) estimated a mean and standard deviation for COVID-19 of 4.7 and 2.9 days, being theses the values we are using in our analysis application foer the gamma a priori distribution.

Results are available nationwide by default (Figure 3b), and at the regional level accessing to the dropdown menu for this purpose.

4.6. Mortality registry.

We show the evolution of the observed all-cause mortality during the epidemic, jointly with the expected mortality according to MoMo the model by the Instituto de Salud Carlos III (18).

Results are available nationwide by default, and at the regional level accessing to the dropdown menu for this purpose (Figure 3c). We also calculated the observed and expected mortality ratios, and their 95% CI, by age, gender and region, since March 4th, 2020, the date on which the first fatality from COVID-19 is official reported in Spain (7, 8) (Figure 3d)

Figure 3. Standard output display of the COVID19-Tracker application (results updated to April 27th, 2020), estimated infected time (a) and basic reporicibility number (b), mortality register (c), and observed versus expected mortality ratios by age, gender and region (d).

5. Further developing

So far, the COVID19-Tracker app has been very well received online, with a large number of connections generating an outsized memory usage on our server (Figure 4).

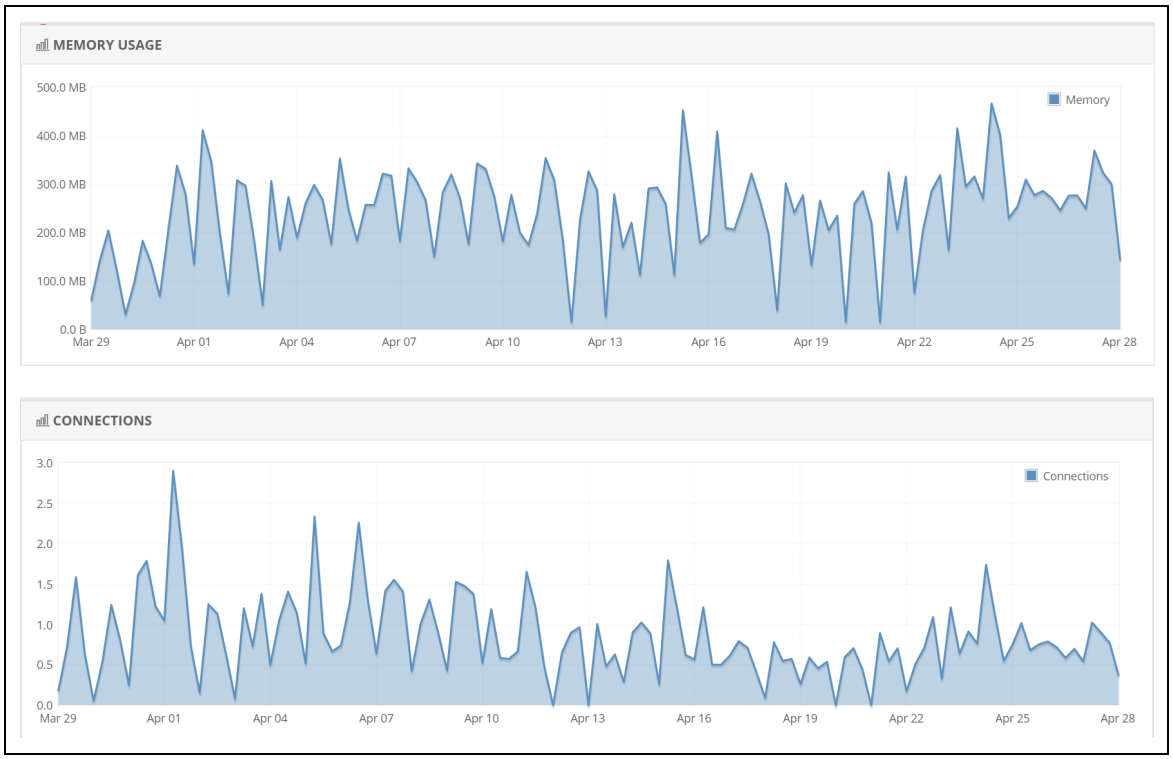

Figure 4. Number of connections and memory usage since March 27th to April 27th, 2020.

We keep improving the application by uploading new data visualizations, which may help for a better understanding of the SARS-CoV-2 epidemic data in Spain. Moreover, the COVID19-Tracker app could also be extensible to data visualizations across other countries and geographical regions.

Discussion

The COVID19-Tracker application presents a set of tools for updated analysis and graphic visualization that can be very useful for a better understanding of the evolution of the COVID-19 epidemic in Spain and its epidemiological surveillance.

As limitations, we should be note that the application does not take into account the changes in the definition of a case diagnosed by COVID-19, nor the population

exposed. So, the number of events is modeled directly instead of the incidence rate, assuming that the entire population is at risk, except for the case fatality rate. On the other hand, the analyzes are not free from the biases linked to the source of information provided by the Ministry of Health (8), being collected on a daily basis through the Datadista github (7).

We continue to plan improvements to the app to include new analytics and visualizations. Also, the application could be extensible for use in other countries or geographic areas. In summary, this application, easy to use, come to fill a gap in this particular scenario for the visualization of epidemiological data for the COVID-19 epidemic in Spain.

Funding

None.

Acknowledgements

None.

Conflict of interest

None.

References

1. Saglietto A, D'Ascenzo F, Zoccai GB, De Ferrari GM. COVID-19 in Europe: the Italian lesson. The Lancet. 2020. doi: 10.1016/s0140-6736(20)30690-5.
2. Our World in Data. Coronavirus Disease (COVID-19) Statistics and Research. Oxford Martin School, The University of Oxford, Global Change Data Lab; 2020. [Accessed April 27th, 2020]. Available from: https://ourworldindata.org/coronavirus/
3. Ministerio de la Presidencia, Relaciones con las Cortes y Memoria Democrática. Real Decreto 463/2020, de 14 de marzo de 2020, por el que se declara el estado de alarma para la gestión de la situación de crisis sanitaria ocasionada por el COVID-19.
4. Mitjà O, Arenas À, Rodó X, Tobias A, Brew J, Benlloch JM. Experts' request to the Spanish Government: move Spain towards complete lockdown. The Lancet. 2020. doi: 10.1016/s0140-6736(20)30753-4.
5. Ministerio de la Presidencia, Relaciones con las Cortes y Memoria Democrática. Real Decreto-ley 10/2020, de 29 de marzo de 2020, por el que se regula un permiso retribuido recuperable para las personas trabajadoras por cuenta ajena que no presten servicios esenciales, con el fin de reducir la movilidad de la población en el contexto de la lucha contra el COVID-19.
6. Team R. RStudio: Integrated Development for R. RStudio, Inc. Boston, MA; 2015.
7. Datadista. Extracción, limpieza y normalización de las tablas de la situación diaria acumulada de la enfermedad por el coronavirus SARS-CoV-2 (COVID-19) en España en un formato accesible y reutilizable. [Accessed April 27th, 2020]. Available from: https://github.com/datadista/datasets/tree/master/COVID%2019
8. Ministerio de Sanidad, Consumo y Bienestar Social. Situación de COVID-19 en España. [Accessed April 27th, 2020]. Available from: https://covid19.isciii.es/
9. Dyba T, Hakulinen T. Comparison of different approaches to incidence prediction based on simple interpolation techniques. Statistics in Medicine. 2000;19(13):1741-52.
10. Navarro A, Utzet F, Puig P, Caminal J, Martín M. La distribución binomial negativa frente a la de Poisson en el análisis de fenómenos recurrentes. Gac Sanit. 2001;15(5):447-52.
11. Rothman K, Greenland S. Modern epidemiology. Philadelphia, PA: Lippincott-Raven Publishers; 1998.
12. Battegay M, Kuehl R, Tschudin-Sutter S, Hirsch HH, Widmer AF, Neher RA. 2019-novel Coronavirus (2019-nCoV): estimating the case fatality rate - a word of caution. Swiss Med Wkly. 2020;150:w20203.
13. Bernal JL, Cummins S, Gasparrini A. Interrupted time series regression for the evaluation of public health interventions: a tutorial. Int J Epidemiol. 2017;46(1):348-55.

14. Tobías A. Evaluation of the lockdowns for the SARS-CoV-2 epidemic in Italy and Spain after one month follow up. Science of The Total Environment. 725: 138539.
15. Lauer SA, Grantz KH, Bi Q, Jones FK, Zheng Q, Meredith HR, et al. The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application. Ann Intern Med. 2020, doi: 10.7326/M20-0504.
16. Cori A, Ferguson NM, Fraser C, Cauchemez S. A new framework and software to estimate time-varying reproduction numbers during epidemics. Am J Epidemiol. 2013;178(9):1505-12.
17. Nishiura H, Linton NM, Akhmetzhanov AR. Serial interval of novel coronavirus (COVID-19) infections. Int J Infect Dis. 2020;93:284-6.
18. Leon-Gomez I, Delgado-Sanz C, Jimenez-Jorge S, Flores V, Simon F, Gomez-Barroso D, et al. Exceso de mortalidad relacionado con la gripe en España en el invierno de 2012. Gac Sanit. 2015;29(4):258-65.