PRE-PRINT – Not Peer Reviewed

22-02-2020

Epidemiological data from the COVID-19 outbreak, real-time case information

Bo Xu*[1,2], Bernardo Gutierrez*[2], Sumiko Mekaru*[3,4], Kara Sewalk*[3], Lauren Goodwin*[3], Alyssa Loskill*[3], Emily Cohn*[3],Yulin Hswen*[3], Sarah C. Hill*[2], Maria M Cobo*[2], Alexander Zarebski*[2], Sabrina Li*[2], Chieh-Hsi Wu*[5], Erin Hulland*[6], Julia Morgan*[6], Lin Wang*[7], Katelynn O'Brien*[3], Samuel V. Scarpino[8], John S. Brownstein[3,9], Oliver G. Pybus[2], David M. Pigott[$6], Moritz U. G. Kraemer[$2,3,9] , for the COVID-19 data working group

1. Tsinghua University, Beijing, China
2. Department of Zoology, University of Oxford, United Kingdom
3. Computational Epidemiology Group, Boston Children's Hospital, Boston, United States
4. Booz Allen Hamilton, Westborough Massachusetts, United States
5. Mathematical Sciences, University of Southampton, United Kingdom
6. Department of Health Metrics, University of Washington, Seattle, United States
7. Mathematical Modelling of Infectious Diseases Unit, Institut Pasteur, UMR2000, CNRS, Paris, France
8. Network Science Institute, Northeastern University, Boston, United States
9. Department of Pediatrics, Harvard Medical School, Boston, United States

* jointly contributed equally to this work

[$] correspondence should be addressed to pigottdm@uw.edu and moritz.kraemer@zoo.ox.ac.uk

Abstract

Cases of a novel coronavirus were first reported in Wuhan, Hubei province, China, in December 2019 and have since spread across the world. Epidemiological studies have indicated human-to-human transmission in China and elsewhere. To aid the analysis and tracking of the COVID-19 epidemic we collected and curated individual-level data from national, provincial, and municipal health reports, as well as additional information from online reports. All data are geo-coded and

where available include symptoms, key dates (date of onset, admission, and confirmation), and travel history. The generation of detailed, real-time, and robust data for emerging disease outbreaks is important can help to generate robust evidence that we support and inform public health decision making.

Background and summary

In December 2019 a number of novel coronavirus-infected penumonia (NCIP) cases were recorded in a large metropolitan City in China, Wuhan, caused by infection with a novel coronavirus named SARS-CoV-2[1]. The outbreak subsequently spread to other cities in Hubei province and across China. Increasingly, epidemiological studies are performed in real-time during an outbreak to understand key metrics such as the epidemic's mean reproduction number, serial interval distribution, incubation period, risk of international spread[2,3]. Geo-positioned records of case data can be important for risk communication and evaluation during outbreaks, especially when they are available in real-time [4].

Epidemiological data is needed during emerging epidemics to best monitor and anticipate spread of infection. In order to provide openly available, accurate and robust data during the COVID-19 outbreak, we collected and continue to curate a real-time database of the individual-level epidemiological data.

Methods

We use a range of different sources to update and curate our database. First, we use official government sources and peer-reviewed scientific papers that report primary data as the gold standard for data inclusion. Government sources include press releases on official websites for Ministries of Health or Provincial Public Health Commissions, as well as updates provided by the official social media accounts of governmental or public health institutions. Second, to find additional details for each case or patient we augment these data with online reports, mainly captured through news websites (e.g., 163.com) or via news aggregators (e.g., https://bnonews.com/). We recorded all data sources in our datasheet. Third, in some instances more detailed data are available, typically through peer-reviewed research articles[1,5], which were subsequently used to modify existing records in the database. We added a full list of sources and a summary of the frequency they were used in Supplementary Table 1.

We collected data on the following information: a) Key dates, which include the date of onset of disease, date of admission to hospital, date of confirmation of infection, and date of travel. b) Demographic information about the age and sex or patients/cases. c) Geographic information, at the highest resolution available down to the district level. We excluded information that were at the building level so that cases could not be identified. Geographic information was subdivided into administrative units (admin 0 = country, admin 1 = province, admin 2 = county, admin 3 = city, and where available, specific locations). d) Symptoms, e) any additional information such as exposure to the Huanan seafood market or record of exposure to infected individuals.

We created a best-practices document to reduce the risk of duplicate efforts or erroneous entries. In addition to the methods described above, we documented best practices specific to different locations. For example, some Chinese provinces reported new cases more than once a day, with each report providing only new data. Other provinces provided updates throughout the day and then provided a final update listing all new cases, inclusive of earlier reports. In the latter case, entry of all the newly reported cases would result in duplication of cases from earlier updates. Additionally, as countries began to report asymptomatic PCR-positive individuals, their referencing or indexing of patients sometimes changed. For example, Japan's Ministry of Health identified novel coronavirus pneumonia cases ordinally up to the country's eighth case. The next three cases were identified during testing of a Japanese national flown from Wuhan on a charter flight for repatriation. One of those cases became the Ministry of Health's ninth case while the other two were asymptomatic and not considered the 10th and 11th cases in their press release. As this distinction was not made in other countries, the practice was documented to avoid confusion of cases in the line-list.

Geo-positioning of data

The administrative unit name were used together with all contextual information provided about the site position to determine its latitudinal and longitudinal coordinates using Google Maps (https://www.maps.google.co.uk), Google Earth (http://www.google.co.uk/intl/en_uk/earth), or using simple Google searches as done in previous data extraction work[8]. Location names are often duplicated within a country, so contextual information was used to ensure the correct site was selected. When the site name was not found, information from the text was also used to scan sites in the approximate area to check for alternate spellings of the site name. We had curators with language skills: English, mandarin Chinese, Cantonese, Spanish, and Portuguese. The distribution of geographic locations where cases have been reported is shown in Figure 1. To provide real-time visualization we designed an interactive web application using Mapbox and automatically update the results using JavaScript. This visualisation is available at https://www.healthmap.org/ncov2019/.

Data records

The database has been made available publicly as of 20th January, 2020. It can be downloaded freely from (https://github.com/beoutbreakprepared/nCoV2019) and directly from a GoogleDrive Link: https://docs.google.com/spreadsheets/d/1itaohdPiAeniCXNlntNztZ_oRvjh0HsGuJXUJWET008/edit#gid=0 in .csv format and be imported it into a variety of software programmes. We also established a Github repository available https://github.com/beoutbreakprepared/nCoV2019 and provide code for importing the data into R statistical software. The epidemiological situation regarding the COVID-19 outbreak is continuously evolving. We therefore have made available an archive data folder through our Github repository updated daily. Each of the rows represents a single individual case and ID. A description of the fields in the database are shown below:

ID - unique identifier for reported case. Currently ID is run concurrently for cases reported from Hubei, China and cases reported outside of Hubei, China. ID order does not necessarily reflect epidemiological progression, or reporting date, and should not be used to order cases in temporal progression.

age - age of case. When not reported, N/A is used. Reported in years. Age ranges are recorded as start_age-end_age e.g. 50-59.

sex - sex of case. When not reported, N/A is used.

city - Generic geographic metadata is reported here. Intended to be cleaned and replaced.

province - Name of the first administrative division in which the case is reported. Intended to be cleaned and replaced by “admin1”.

country - Name of country in which case is reported. Note that imported cases will be assigned to the country in which confirmation occurred - this is typically in the arrival country, rather than the site of infection. “Travel_histroy_location” will describe other locations of travel for such instances.

wuhan(0)_not_wuhan(1) - binary flag to distinguish cases from Wuhan, Hubei, China, from all other cases. 0 denotes case is reported in Wuhan, 1 denotes case is reported elsewhere in the world.

latitude - the latitude of the specific location (denoted as “point” in “geo_resolution”) where the case was reported, or the latitude of a representative location (denoted as “admin” in “geo_resolution”) within the administrative unit the case is reported.

longitude - the longitude of the specific location (denoted as “point” in “geo_resolution”) where the case was reported, or the longitude of a representative location (denoted as “admin” in “geo_resolution”) within the administrative unit the case is reported.

geo_resolution - an indicative field in which the spatial representativeness of “latitude” and “longitude” are described. “point” indicates that a specific location is being represented by these coordinates. “admin” denotes that the coordinates are representative of the administrative unit in which coordinates lie. Subsequent “admin3”, “admin2”, “admin1” and corresponding “admin_id” and “shapefile” will allow for a more specific representation to be had.

date_onset_symptoms - date when the reported case was recorded to have become symptomatic. Specific dates are reported as DD.MM.YYYY. Ranges are recorded as DD.MM.YYYY - DD.MM.YYYY. Ranges with uncertain start of finish dates are recorded as - DD.MM.YYYY and DD.MM.YYYY - respectively.

date_admission_hospital - date when the reported case was recorded to have been hospitalized.Specific dates are reported as DD.MM.YYYY. Ranges are recorded as DD.MM.YYYY - DD.MM.YYYY. Ranges with uncertain start of finish dates are recorded as - DD.MM.YYYY and DD.MM.YYYY - respectively.

date_confirmation - date when the reported case was confirmed as having nCoV-2019 using rt-PCR. Confirmation accuracy is contingent on the data source used. Specific dates are reported as

DD.MM.YYYY. Ranges are recorded as DD.MM.YYYY - DD.MM.YYYY. Ranges with uncertain start of finish dates are recorded as - DD.MM.YYYY and DD.MM.YYYY - respectively.

symptoms - list of symptoms recorded in the description of the case.

lives_in_Wuhan - recorded relationship of patient with city of Wuhan, Hubei, China. "yes" indicates case was a resident of Wuhan. "no" indicates that the case is not a resident in Wuhan (residential). No information indicates that no data was available.

travel_history_dates - recorded travel dates to and from Wuhan. Specific dates are reported as DD.MM.YYYY and indicate date when individual left Wuhan. Ranges are recorded as DD.MM.YYYY - DD.MM.YYYY when both are available. Ranges with uncertain start of finish dates are recorded as - DD.MM.YYYY and DD.MM.YYYY - respectively.

travel_history_location - an open field describing the recent recorded travel history of the case.

reported_market_exposure - an open field indicating yes if there was reported market exposure and no if there was no. N/A in case no information is provided.

additional_information - any additional information that may be informative about the case, such as the occupation of the patient, the purpose of their travels, the hospital they were admitted to, etc,.

chronic_disease_binary - 0 represents if case was reported to have no chronic disease and 1 if there was a reported chronic disease

chronic_disease - reported chronic condition(s) of reported case.

source - URL identifying source of this information

sequence_available - if there was a genomic sequence available the accession number is inserted here.

outcome - patients outcome as either 'died' or 'discharged' from hospital

date_death_or_discharge - reported date of death or discharge in DD.MM.YYYY format.

At time of publication the database contained 26034 geopositioned records from December 1, 2019 to February, 22nd 2020. The mean age of all cases was 45 years (std. dev. = 17) and the overall distribution of age and sex was 56% male. A map of all records can be viewed in real-time here:https://www.healthmap.org/ncov2019/.

Technical validation

After initial data entry the database was checked using two complementary methodologies to identify possible duplicate records. One was a machine enabled one and the other was done manually by the data curators. The first algorithm proceeds in 5 steps. 1) columns with no variability across all records were removed, 2) the remaining data were hashed using a 32-bit variant of MurmurHash3 implemented in the R package *FeatureHashing* version 0.9.1.3[6,7], 3) a principle component analysis on the centered, scaled hashed feature matrix is performed for dimension reduction, with principle components having standard deviations greater than 0.5 retained, 4) pairwise, Euclidean distances are then calculated and are normalized based on the smallest observed distance between records, and 5) records that have pairwise distances less than the 0.5th percentile are flagged as duplicates. Duplicate are defined as cases that refer to the same case. Code for these methods are hosted on our GitHub repository (https://github.com/beoutbreakprepared/nCoV2019). Records identified as possible duplicates were communicated to data curators via Github and flagged in the database. Curators then discussed amongst themselves via an online chat system (www.slack.com) to reach a consensus on how to address the possible duplications.

Usage Notes

These data can be used to for investigate the epidemiological COVID-19 outbreak in China and elsewhere and . This includes descriptive mapping of occurrences through time and estimation of key epidemiological parameters using mathematical models. The data are openly available and we will continue to curate the database as new information is made available. However, if the epidemic continues to grow then public health agencies are unlikely to continue to report indvidual-level case data, and instead will switch to reporting only total numbers (or estimates thereof) of confirmed or suspected cases as done for previous large outbreaks such as pandemic flu H1N1[9]. When detailed data becomes increasingly less available as the epidemic is growing we may transition to an augmented database structure that only reports total new cases per location. Other groups have presented similar datasets which are complimentary to the one presented in this publication[10]. However, the dataset presented here includes fine grained geographic details and the most comprehensive list of cases.

There are possible changes of reporting during the first month of the outbreak. For example, we find that demographic information were reported initially as case numbers were small but detailed case information became less available after the 23rd of January. Initial cases from Wuhan are well described, mostly thanks to epidemiological studies published towards the end of January[5]. Even though we made the best attempt to report data as accurately as possible, given the dynamic nature of the outbreak we caution that the database cannot be guaranteed to be free from error, and we apologize in advance if there are missing entries that were not picked up using our standardised protocol. We encourage users of the database to contact us directly if potential errors or omissions have been found. This can be done by either emailing the corresponding authors or, preferably, by submitting a request via the Github repository (https://github.com/beoutbreakprepared/nCoV2019).

Additional information

References

1. Li, Q. *et al.* Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia. *N. Engl. J. Med.* NEJMoa2001316 (2020). doi:10.1056/NEJMoa2001316
2. Bogoch, I. I. *et al.* Pneumonia of Unknown Etiology in Wuhan, China: Potential for International Spread Via Commercial Air Travel. *J. Travel Med.* (2020). doi:10.1093/jtm/taaa008
3. Zhao, S. *et al.* Preliminary estimation of the basic reproduction number of novel coronavirus (COVID-19) in China, from 2019 to 2020: A data-driven analysis in the early phase of the outbreak. *Int. J. Infect. Dis.* (2020). doi:10.1016/j.ijid.2020.01.050
4. Brownstein, J. S., Freifeld, C. & Madoff, L. C. Digital Disease Detection - Harnessing the Web for Public Health Surveillance. *N. Engl. J. Med.* **360**, 2153–2157 (2009).
5. Chen, N. *et al.* Epidemiological and clinical characteristics of 99 cases of 2019 novel

coronavirus pneumonia in Wuhan, China: a descriptive study. *Lancet* **6736**, 1–7 (2020).

6. Weinberger, K., Dasgupta, A., Attenberg, J., Langford, J. & Smola, A. Feature Hashing for Large Scale Multitask Learning. *Proc. 26th Int. Conf. Mach. Learn. ICML 2009* 1113–1120 (2009).
7. R Core Team. R: A language and environment for computing. Vienna, Austria. *R Found. Stat. Comput.* (2019).
8. Kraemer, M. U. G. *et al.* The global compendium of Aedes aegypti and Ae. albopictus occurrence. *Sci. Data* **2**, 150035 (2015).
9. Fraser, C. *et al.* Pandemic potential of a strain of influenza A (H1N1): Early findings. *Science* **324**, 1557–1562 (2009).
10. Sun, K., Chen, J. & Viboud, C. Early epidemiological analysis of the COVID-19 outbreak based on a crowdsourced data. *medRxiv* (2020). doi:2020.01.31.20019935

Acknowledgements

We thank all those individuals and organizations across the world who have been willing and able to report data in as open and timely manner as possible. This work attempts to synthesize information from across a myriad set of data sources. We acknowledge support from the Oxford Martin School.

Contributions

Data curation: All authors contributed to curating the database. Technical validation: AZ, C-HW, DMP, SVS, BG. Oversaw project: JSB, SC, OGP, DMP, MUGK. All authors contributed to writing and editing the manuscript.

Competing Interests

SM declares employment at Booz Allen Hamilton while engaged in the research project. All other authors declare no competing financial interests.

Members of the COVID-19 data working group: Bo Xu, Bernardo Gutierrez, Sumiko Mekaru, Kara Sewalk, Lauren Goodwin, Alyssa Loskill, Emily Cohn, Sarah C. Hill, Maria M Cobo, Alexander Zabreski, Sabrina Li, Chieh-Hsi Wu, Erin Hulland, Julia Morgan, Lin Wang, Katelynn O'Brien, Samuel V. Scarpino, John S. Brownstein, Oliver G. Pybus, David M. Pigott, Moritz U. G. Kraemer