

# Coronavirus disease 2019 (COVID-19) outbreak in China, spatial temporal dataset

Wenyuan Liu<sup>1\*,+</sup>, Peter Tsung-Wen Yen<sup>1\*,+</sup>, Siew Ann Cheong<sup>1</sup>,

March 27, 2020

1. School of Physical and Mathematical Sciences, Nanyang Technological University, 21 Nanyang Link, Singapore 637371, Singapore

\*corresponding author(s): Wenyuan Liu (wenyuan.liu@ntu.edu.sg), Peter Tsung-Wen Yen (peteryen2017@gmail.com)

+ equal contribution

## Abstract

We present Coronavirus disease 2019 (COVID-19) statistics in China dataset: daily statistics of the COVID-19 outbreak in China at the city/county level. For each city/country, we include the six most important numbers for epidemic research: daily new infections, accumulated infections, daily new recoveries, accumulated recoveries, daily new deaths, and accumulated deaths. We hope this dataset can help researchers around the world better understand the spread dynamics of COVID-19 at a regional level, to inform intervention and mitigation strategies for policymakers.

## Background & Summary

Starting in East Asia at the beginning of 2020, COVID-19 is by now a global pandemic. At the time of writing there are more than 392,000 confirmed cases in more than 190 territories with no sign of slowing down. To meet this great healthcare challenge of our time, we need to combine efforts from the medical, pharmaceutical, epidemiological, transport, and even political realms. Thus far scientific efforts made to understand how COVID-19 spreads, are mainly based numbers from China at province level. As the situation improves in China, we can derive a more complete and detailed picture of COVID-19 spreading in the territory. This picture can help other countries develop their own strategies to combat the coronavirus. Unfortunately, at the time of writing, there is no high-quality publicly-available datasets, so this is a big obstacle to scientists all over the world. In reality, Chinese authorities announce the ongoing situation daily after 20 Jan 2020, although (1) cities and provinces put their daily reports only in their own homepages and in different formats, and (2) most of these reports are in Chinese. It's therefore difficult for scientists who cannot read Chinese to

do any research based on these reports. For our own research, and also to make the data more widely accessible, we collected all daily report available from the official websites, extracted the data and organized them in several .csv files. Researchers can then use their favorite tools to analyze the data, and it is our hope that this dataset can help people understand and fight COVID-19 better.

## Methods

At the end of 2019, the novel coronavirus was first discovered in Wuhan City, Hubei province, China. Since then, the viral infection spread out to nearby provinces and eventually to all China. Starting from 21 Jan 2020, provincial authorities have decided to release new and accumulated infected cases, newly recovered and accumulated recovered cases, death tolls, and other information to the public daily. This information is published on the official Health Commission websites of each province once or twice a day depending on whether the infection situation is changing rapidly. There are 22 provinces, 5 autonomous regions, 4 municipalities, two special administrative regions (HongKong and Macao), and also Taiwan. In these official COVID-19 reports, cases are reported down to the administrative region level (the equivalent of a county). For example, in Hubei, there are 17 administrative regions, such as Wuhan city, Huangshi city, Shiyan city, Yichang city, Xiangyang city, and so on.

Here, let us describe the procedures we used to extract essential information from the COVID-19 daily reports. First, to ensure the reliability and verifiability of our data, we used a browser tool called *Save Page WE* to download all the daily reports and save them locally as html files. For consistency, the html files are named in the format “Province\_dd-mm-yyyy.html”. We organized these source files into folders named after the provinces or regions. The whole dataset will be made accessible to all readers. Next, we describe what information the daily reports reveal. A typical daily report contains a duration of time, say for example 1600 hour on 23 Jan 2020 to 0900 hour on 24 Jan 2020. If this duration is within one calendar day, we treat new cases reported therein as for that calendar day, whereas if the duration straddles two calendar days, we treat new cases reported therein as belonging to the calendar day with the longer coverage; in cases where the duration is split evenly between the two calendar days, we count new cases reported therein towards the earlier day. In the current version, we only extract (1) new and accumulated infected cases, (2) new and accumulated recovered cases, and (3) new and accumulated death cases, so we end up with six types of data for each of the administrative regions. Because of their larger populations, municipalities report cases down to the district level. Since municipalities are similar in sizes to counties, we decided to collect aggregated data for the municipalities so that our data set is uniform geographically. Some provinces like Liaoning Province offer only aggregated data, and do not go down to the administrative region level. For these cases, we collect and show only aggregated data. Here we introduce three formulas that we used to count on day  $i$  the new infected ( $NI_i$ ), recovered ( $NR_i$ ), and

deaths ( $ND_i$ ) from the accumulated infected ( $TI_i$ ), recovered ( $TR_i$ ), and deaths ( $TD_i$ ):

$$NI_i = TI_i - TI_{i-1}, \quad (1)$$

$$NR_i = TR_i - TR_{i-1}, \quad (2)$$

$$ND_i = TD_i - TD_{i-1}. \quad (3)$$

After we extract the reported cases from the daily reports, we use the above formulas to deduce the number of new cases for our dataset.

As a side note, in the early stage of this project, we planned to do data extraction automatically. Unfortunately this was not successful because the report format for each province was different, making scripting approaches impractical and unreliable. We show in Figures 1, 2, and 3 to illustrate how different the formats can be and the level of difficulty to automate the collection process. Nonetheless, we will continue to explore ways to make the automation procedure feasible in the future..

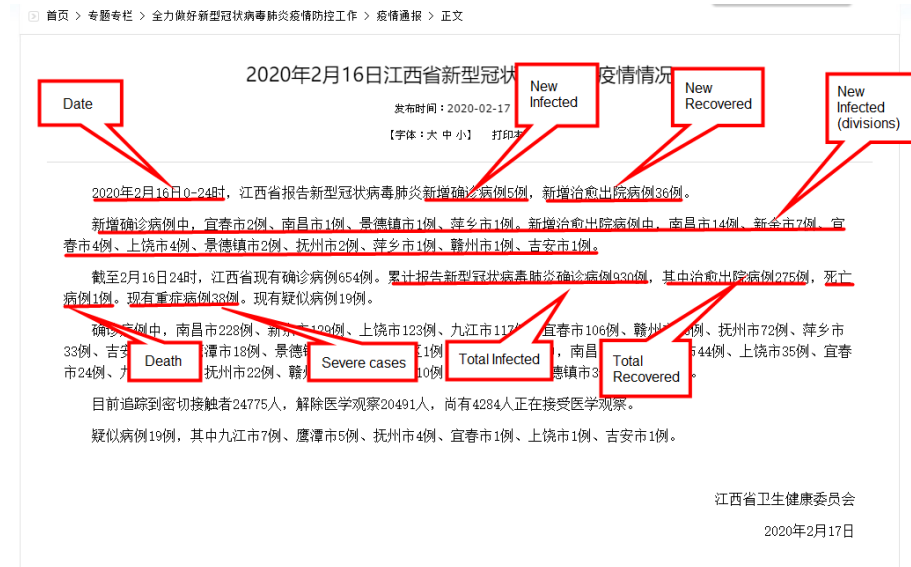


Figure 1: The COVID-19 raw data of Jiangxi Province. We used red boxes to indicate the data we extracted and stored in the dataset.

## Data Records

We made the dataset available through Github [1] under Creative Commons Zero v1.0 Universal (CC0-1.0) license. We provide six .csv files to cover basic

2020年2月17日12-24时,山东省报告新增新型冠状病毒肺炎确诊病例2例,累计确诊病例543例(含重症病例13例,危重症病例14例,治愈出院193例,死亡病例2例);新增疑似病例15例,现有疑似病例31例。截至目前,追踪到密切接触者15909人,已解除医学观察12650人,诊断为疑似或确诊病例289例,尚有2970人正在接受医学观察。

Date Recovered Death Newly infected Total infected

Severe cases

详见下表(单位:例):

地级市	当日12-24时新增确诊病例		累计确诊病例		重症病例	危重症病例	出院人数	死亡人数	当日12-24时疑似病例	
	总数	分布	总数	分布					新增	现有
济南市	0	--	47	市中区28, 历下区3, 长清区4, 历城区4, 天桥区4, 槐荫区1, 章丘区1	0	0	13	0	1	13
青岛市	1	黄岛区1	58	黄岛区13, 市南区11, 市北区8, 市南区7, 李沧区6, 崂山区4, 平度市3, 莱西市3, 城阳区和胶州市1	2	1	26	1	7	28
淄博市	0	--	29	博山区10, 张店区9, 淄川区8, 高村区1, 桓台县1	1	1	5	0	0	1
枣庄市	0	--	24	市中区5, 薛城区3, 滕州市5, 滕州区和山亭区4, 台儿庄区1	0	1	7	0	0	0
烟台市	0	--	47	芝罘区10, 莱山区10, 福山区9, 莱阳市6, 招远市6, 龙山区3, 牟平市3, 莱山区1	3	0	13	0	0	0
潍坊市	0	--	44	奎文区28, 昌乐县6, 诸城市4, 临朐县3, 高密市2, 寿光市1, 安丘市1, 安亭镇1	2	2	11	0	1	1
济宁市	0	--	62	任城区28, 曲阜市13, 兖州区6, 汶上县2, 邹城市1, 嘉祥县1, 金乡县1	2	0	20	0	0	3
泰安市	0	--	32	泰山区12, 新泰市4, 肥城市2, 东平县2, 宁阳县1, 岱岳区1	0	1	6	0	1	3
威海市	0	--	38	环翠区22, 乳山市8, 文登区7, 荣成市1	1	1	9	0	0	0
日照市	0	--	16	东港区12, 岚山区2, 五莲县2	0	0	4	0	0	0
临沂市	1	沂水区和1	48	兰山区12, 河东区11, 沂南县8, 沂水县4, 平邑县4, 费县3, 沂水县3, 郯城县3, 临沭县2, 兰陵县1	0	3	35	0	0	2
德州市	0	--	37	武城县13, 德城区9, 乐陵市2, 禹城市3, 宁津县3, 齐河县2	2	4	9	1	5	10
聊城市	0	--	38	东阿县27, 阳谷县4, 莘县3, 东阿县2, 临清市1, 茌平县1	0	0	12	0	0	0
滨州市	0	--	15	滨城区7, 惠民县3, 无棣县3, 沾化区1, 阳信县1	0	0	10	0	0	0
菏泽市	0	--	18	巨野县7, 牡丹区6, 曹县1, 单县1, 成武县1, 鄄城县1, 东明县1	0	0	13	0	0	0
合计	2	--	543		13	14	193	2	15	61

备注:根据要求,按病例确诊时医院所在县区统计。

Figure 2: The COVID-19 raw data for Shandong Province.

吉林省卫生健康委员会关于新型冠状病毒肺炎疫情情况通报 (2020年2月16日公布)

Date New Infected 02-16 来源:吉林省卫生健康委员会 字体显示:小 中 New Recovered

2月15日0-24时,全省新增确诊病例1例(长春市),系由疑似病例转为确诊病例,新增治愈出院3例(长春市1例,四平市2例)。

截至2月15日24时,全省累计报告确诊病例89例,累计治愈出院28例(长春市14例,吉林市3例,延边州2例,四平市4例,辽源市1例,松原市1例,公主岭市2例,梅河口市1例),死亡1例(四平市),现在院隔离治疗确诊病例60例,其中长春市31例,吉林市2例,延边州3例,四平市9例,通化市3例,白城市1例,辽源市6例,松原市1例,公主岭市4例。

在院治疗确诊病例中,54例为普通病例,3例为重症(长春市2例,松原市1例),3例为危重症(长春市1例,白城市1例,公主岭市1例)。上述确诊病例的密切接触者3659人,已解除医学观察211人,集中隔离或居家隔离医学观察1548人,其中主动开展核酸检测筛查出确诊患者13例。

2月15日0-24时,全省原有27例疑似病例中,转为确诊病例1例,排除5例;新增疑似病例16例;现有疑似病例37例,已全部隔离治疗,目前正在进一步明确诊断。

病例,女,1960年出生,系黑龙江省输入病例,与黑龙江省七台河市确诊病例有密切接触史,该患者于2月13日从黑龙江省七台河市乘自驾车到长春市。2月14日到医院就诊,被诊断为疑似病例。2月15日转为确诊病例,住址为长春市高新区融创上城。

近期正值复工复产高峰,人员流动性加大,广大群众应注意避免到人群密集场所,做好个人防护,勤洗手,戴口罩,少出门,勿聚集,保护好自己和家人的健康。

提醒广大群众,如您是从疫情发生省份返回人员,或是与以上确诊病例同住一个小区有密切接触的人员,应主动到当地社区做好筛查登记,配合专业人员开展医学观察,一旦出现发热、咳嗽等急性呼吸道症状,请到当地定点医院发热门诊就诊。

Total Death Total Infected cases

Figure 3: The COVID-19 raw data for Jilin Province.

information of COVID-19 pandemic in China, namely the number of new confirmed infections (`China_daily_new_infections.csv`), the number of accumulated confirmed infections (`China_accumulated_infections.csv`), the number of new recovered patients (`China_daily_new_recoveries.csv`), the number of accumulated recovered patients (`China_accumulated_recoveries.csv`), the number of new death case(s) (`China_daily_new_deaths.csv`), and the

number of accumulated death case(s) (**China\_accumulated\_deaths.csv**) on each day in each city.

Each file contains 368 lines and 44 columns: the first row is the header, the name for each column, while other rows are the data for all cities/counties. For each row, the first four columns are names for city/county: the first column is the name of city/county in English, the second column is the name of the provincial-level region this city/county belongs to in English, the third column is the name of city/county in Chinese, and the fourth column is the name of the provincial-level region this city/county belongs to in Chinese. The remaining columns are dates ranging from 20 January 2020 to February 29 2020 (in YYYY-MM-DD format). For example, in **China\_accumulated\_infections.csv**, for row 169, column 1 is 'Wuhan', whereas, column 19 (2020-02-04) is 8351. This tells us that there are 8351 confirmed cases reported in Wuhan up till 24:00 4 February 2020.

## References

- [1] Github repository, <https://github.com/cheongsa/Coronavirus-COVID-19-statistics-in-China>