COVID-19 Public Datasets & Data-Driven Tools

This resource list has been prepared in collaboration with the **MIT COVID-19 Challenge** (http://covid19challenge.mit.edu) and **HealthDSA: Health Data Science & Analytics Community** (www.HealthDSA.org) to assist healthcare and public health professionals, scientists, researchers, engineers, technologists, innovators, entrepreneurs, and policy experts with research and development of solutions for the COVID-19 Pandemic.

1) COVID-19 Open Research Dataset (CORD-19):
https://pages.semanticscholar.org/coronavirus-research

- 29,000 articles on Coronavirus for text mining
- Formatted as a series of compressed folders with json files
- Used in a current Kaggle competition

2) 2019 Novel Coronavirus COVID-19 (2019-nCoV) Epidemiological Data Repository by Johns Hopkins University Center for Systems Science and Engineering (JHU CCSE):
https://github.com/CSSEGISandData/COVID-19

- Provides aggregate information on daily reports from around the world
- Each day, data in each country about number of cases, mortality, and recoveries
- Actual data is here

3) European Center for Disease Control and Prevention (ECDC) - Epidemiological Data:
https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide

- Single dataset with daily reports on number of COVID-19 cases and mortality
- Updated daily into Excel file

4) U.S. Hospital Capacity Estimates (Harvard Global Health Institute):
https://globalepidemics.org/2020-03-17-caring-for-covid-19-patients/

- State by state information on the expected number of available and needed hospital beds and ICU beds in 6, 12, 18 months

5) Italy COVID-19 Data: https://github.com/pcm-dpc/COVID-19

- In Italian, but English descriptions available
- Regional statistics on number of hospitalized patients, ICU patients, recoveries, deaths, positive tests, total tests per region, number of people in home confinement

6) WHO Data - COVID-19 Cases & Deaths in China (by Province) and other countries: https://data.humdata.org/dataset/coronavirus-covid-19-cases-data-for-china-and-the-rest-of-the-world

- Daily reports of confirmed cases, deaths per country

7) ACAPS COVID-19: Government Measures Dataset: https://data.humdata.org/dataset/acaps-covid19-government-measures-dataset

- Rich dataset with information on protective measures taken by governments
- Includes country, date implemented, category, and actual measure
- Can be joined with other datasets to empirically measure the effect of some interventions

8) World Bank Indicators of Interest - regarding population health and healthcare systems worldwide, relevant to the COVID-19 Outbreak: https://data.humdata.org/dataset/world-bank-indicators-of-interest-to-the-covid-19-outbreak

- Data on a wide range of indicators related to healthcare
- Historic data goes as far back as 1961 and as recent as 2019

9) GeneBank COVID-19 Genetic Sequences: https://www.ncbi.nlm.nih.gov/genbank/sars-cov-2-seqs

- Sequencing data - lots of different sequencing runs
- Hundreds of different sequencing runs available between nucleotide sequencing and SRA sequencing

10) Next Strain - COVID-19 Genomics Database: https://nextstrain.org/ncov

- Genomics database with scripts and methods for analyzing and visualizing sequencing data

11) U.S. State COVID Testing Data: https://covidtracking.com

- Contains testing information per state, including positive, negative, and pending tests
- Can be accessed via API: https://covidtracking.com/api/
- AWS Version: https://aws.amazon.com/marketplace/pp/prodview-a2ev4blctqkwc?qid=1585087643935

12) U.S. State-Level and County-Level COVID-19 Count Data (cases and deaths): https://github.com/nytimes/covid-19-data

- Series of data files released by the New York Times with cumulative counts of coronavirus cases and deaths in the United States, at state/county level, over time.
- Time series data compiled from state and local governments and health departments

13) U.S. State-Specific Projections for Hospital Resource Utilization:
http://www.healthdata.org/covid/

- Model by Institute for Health Metrics and Evaluation / University of Washington
- Data and projections can be downloaded

14) COVID-19 Twitter Datasets

#1 - http://www.panacealab.org/covid19/
#2 - https://ieee-dataport.org/open-access/corona-virus-covid-19-tweets-dataset
#3 - https://github.com/echen102/COVID-19-TweetIDs

- Social media tweets relating to Coronavirus Pandemic
- Each dataset may be based on different methods and time period
- Can be used for NLP / text mining models

15) UnaCast Social Distancing Scoreboard:
https://www.unacast.com/covid19/social-distancing-scoreboard

- Compare community's social distancing activity to its activity prior to COVID-19 at state and county levels

16) Collection of COVID-19 Data APIs (variety of data sources):
https://covid-19-apis.postman.com

17) New York City COVID-19 Dataset: https://www1.nyc.gov/site/doh/covid/covid-19-data.page

18) Synthetic COVID-19 EHR Dataset from MITRE / Veterans Health Administration

NOTE: Synthetic data are most suitable for methods development & proof-of-concept analyses; avoid use for direct clinical & critical operational applications.

Datasets - CSV files (#1 Civilian Population; #2 Veteran Population):
https://www.dropbox.com/sh/3xcsz3bzb7rjjwy/AADPY4gmouSKDa8XaC200g8za?dl=0

Data Dictionary: https://github.com/synthetichealth/synthea/wiki/CSV-File-Data-Dictionary

Guide to Synthetic Dataset: https://github.com/synthetichealth/synthea/wiki/Getting-Started

Methods Article: https://doi.org/10.1093/jamia/ocx079

19) Definitive Healthcare U.S. hospital capacity data (number of beds, ICU beds, ventilator capacity by state/county).

Version 1 (GitHub): https://github.com/rsowers-dhc/covid19

Version 2 (AWS): https://aws.amazon.com/marketplace/pp/USA-Hospital-Beds-COVID-19-Definitive-Healthcare/prodview-yivxd2owkloha

20) Amazon Web Services (AWS) Data Lake with Public COVID-19 Datasets - includes several datasets on this list which are stored, updated, and ready-for-analysis on AWS: https://aws.amazon.com/blogs/big-data/a-public-data-lake-for-analysis-of-covid-19-data/