Epidemiology Datasets

CORD-19

- Description: 'In response to the COVID-19 pandemic, the White House and a coalition of leading research groups have prepared the COVID-19 Open Research Dataset (CORD-19). CORD-19 is a resource of over 500,000 scholarly articles, including over 200,000 with full text, about COVID-19, SARS-CoV-2, and related coronaviruses. This freely available dataset is provided to the global research community to apply recent advances in natural language processing and other AI techniques to generate new insights in support of the ongoing fight against this infectious disease. There is a growing urgency for these approaches because of the rapid acceleration in new coronavirus literature, making it difficult for the medical research community to keep up.'[1]
- How to access: Available for download at this Kaggle link or at `https://www.semanticscholar.org/cord19/download`.
- Access Process: For Kaggle link - Open website, scroll down to Data Explorer on the left-hand side of website, then download both CSV files containing data (No account sign up required)
- Other notes: See `https://www.kaggle.com/covid-19-contributions`

Health Indicators

- Description: 'The United Nations established, in September 2015, the Sustainable Development Goals (SDGs), which specify 17 universal goals, 169 targets, and 232 indicators leading up to 2030. Drawing from GBD 2016, this dataset provides estimates for 37 health-related SDG indicators for 188 countries from 1990 to 2016, as well as projections, based on past trends, from 2017 to 2030. These 37 SDG indicators were used to construct the health-related SDG index, a summary measure of overall performance across the health-related SDGs.'
- How to access: See `http://ghdx.healthdata.org/record/ihme-data/gbd-2016-health-related-sdgs-1990-2030`
- Access Process: Open website, select Files (2) tab located to the right of the General Info and Citation tabs, select on the appropriate files to download data (No account sign up required)
- Other notes:

Tobacco Use Prevalence

- Description: 'The Global Burden of Disease Study 2019 (GBD 2019), coordinated by the Institute for Health Metrics and Evaluation (IHME), estimated the burden of diseases, injuries, and risk factors for 204 countries and territories and selected subnational locations.

  Estimates of smoking tobacco use and the burden attributable to this risk factor were produced by sex, age group, and year for 204 countries and territories for 1990-2019. Files available in this record include estimates of the prevalence of smoking tobacco use, number of people that currently use smoked tobacco products, and supply-side tobacco availability and consumption. Estimates of disease burden attributable to smoking tobacco use are available through the GBD Results Tool.'
- How to access: For smoking tobacco use, see `http://ghdx.healthdata.org/record/ihme-data/gbd-2019-smoking-tobacco-use-prevalence-1990-2019`
  For chewing tobacco use, see

[1] https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge

`http://ghdx.healthdata.org/record/ihme-data/gbd-2019-chewing-tobacco-use-prevalence-1990-2019`

- Access Process: Open website, select Files tab located to the right of the General Info and Citation tabs, select on the appropriate files to download data (No account sign up required)
- Other notes:

MaskedFaceNet

- Description: 'An image editing approach and three types of masked face detection dataset; namely, the Correctly Masked Face Dataset (CMFD), the Incorrectly Masked Face Dataset (IMFD) and their combination for the global masked face detection (MaskedFace-Net). Realistic masked face datasets are proposed with a twofold objective: i) detecting people having their faces masked or not masked, ii) detecting faces having their masks correctly worn or incorrectly worn (e.g.; at airport portals or in crowds).'
- How to access: See `https://github.com/cabani/MaskedFace-Net`
- Access Process: Open website, scroll down to Dataset subsection, select the OneDrive links available to access the data (No account sign up required)
- Other notes:

Canadian Chronic Disease Surveillance System

- Description: 'The Canadian Chronic Disease Surveillance System (CCDSS) is the result of a collaborative network of provincial and territorial surveillance systems, supported by the Public Health Agency of Canada (PHAC). The system collects data on all residents who are eligible for provincial or territorial health insurance and can generate national estimates and trends over time for over 20 chronic diseases and other selected health outcomes. To identify people with chronic diseases, provincial and territorial health insurance registry records are linked using a unique personal identifier to the corresponding physician billing claims, hospital discharge abstract records and prescription drug records.'
- How to access: See `https://health-infobase.canada.ca/ccdss/Index`
- Access Process: To access the datasets used by the system, you may click on the TICK mark next any of the conditions in "Health outcomes related to chronic diseases and conditions included in the CCDSS" table on `https://health-infobase.canada.ca/ccdss/Index`. This will lead to one of the three types of data comparisons, each of these pages will have a "Download detailed table" button for accessing the data shown in the system generated tables.

  For further resources, refer to `https://open.canada.ca/data/en/dataset/9525c8c0-554a-461b-a763-f1657acb9c9d`

Canada Health Inequalities Data

- Description: 'The Health Inequalities Data Tool contains data on indicators of health status and health determinants, stratified by a range of social and economic characteristics (i.e. social stratifiers) meaningful to health equity. Indicators are grouped into twelve framework components.'
- How to access: See `https://health-infobase.canada.ca/health-inequalities/Index`

- Access Process: Click on the "Use the Health Inequalities Data tool" green button to get access to the system. This will redirect to "Data-Inequalities Measures" page of the system like `https://health-infobase.canada.ca/health-inequalities/data-tool/Index`, where after selecting the variables you may click on the "Download Detailed Table: Indicator" button. The same follows for the "Data-Rates by Province/Territory" page.

City of Toronto COVID-19 Pandemic Data

- Description: This is a set of official dataset provided by the City of Toronto for the COVID-19 Pandemic. The data includes case counts, epidemiological summaries, neighbourhood maps, ethno-racial identity and income, and vaccine data.
- How to access: See
  `https://drive.google.com/file/d/1-7j48S_KQY-I-4Qu3N3lsEOALXON2StG/view`,
  `https://drive.google.com/file/d/11KF1DuN5tntugNc10ogQDzFnW05ruzLH/view`,
  `https://drive.google.com/file/d/1euhrML0rkV_hHF1thiA0G5vSSeZCqxHY/view`,
  `https://drive.google.com/file/d/1jzH64LvFQ-UsDibXOOMOtvjbL2CvnV3N/view`,
  `https://drive.google.com/file/d/1hfYt1Wv_hL4U5JGzOH1-wbxqiqx1DAJp/view`,
  `https://www.toronto.ca/home/covid-19/covid-19-latest-city-of-toronto-news/covid-19-pandemic-data-covid-19-vaccine-data/`[2]
- Access Process: Click on Google Drive links and then download appropriate data (No account sign up required)

[2]There are multiple sheets available under "Download excel"

Clinical Practice Datasets

OASIS: Open Access Series Of Imaging Studies

- Description: Open Access Series of Imaging Studies (OASIS) is a project aimed at making neuroimaging datasets freely available to the scientific community. They provide three datasets that consist of cross-sectional and longitudinal neuroimaging data for aiding neuroimaging, clinical and cognitive research studies. Following are the three datasets:
  (i) OASIS-1: Cross-sectional MRI Data in Young, Middle Aged, Nondemented and Demented Older Adults
  (ii) OASIS-2: Longitudinal MRI Data in Nondemented and Demented Older Adults
  (iii) OASIS-3: Longitudinal Neuroimaging, Clinical, and Cognitive Dataset for Normal Aging and Alzheimer's Disease
- How to access: See `https://www.oasis-brains.org`
- Other notes: Only OASIS-1 and OASIS-2 datasets are accessible to students. To access the datasets, click on the "Apply to Access OASIS Data" button to complete the Data Use Agreement form.

OpenfMRI

- Description:The OpenfMRI database is a repository of human brain imaging data collected using MRI and EEG techniques. At present, 95 neuroimaging datasets are accessible for research studies.
- How to access: `https://openfmri.org/dataset/` (Main website) `https://openfmri.org/how-to-extract-data/` (Information for accessing data)
- Other notes: Data is available for download through the above link.

Action to Control Cardiovascular Risk in Diabetes (ACCORD)

- Description: The purpose of this study was to determine if intensive glycemic control, multiple lipid management and intensive blood pressure control could prevent major cardiovascular events (myocardial infarction, stroke or cardiovascular death) in adults with type 2 diabetes mellitus. Secondary hypotheses included treatment differences in other cardiovascular outcomes, total mortality, microvascular outcomes, health-related quality of life and cost-effectiveness.
- How to access: `https://biolincc.nhlbi.nih.gov/studies/accord/(Mainwebsite)` `https://clinicaltrials.gov/ct2/show/NCT00000620`
- Registration notes: To request access for this dataset, visit the main page and register to BioLINCC. Following this, you will be asked to select the type of request you wish to submit and an appropriate request form is can be filled. Further registration instructions can be found, **here***. You may also look at the **BioLINCC FAQs***.

Vivil - Center for Global Clinical Research Data

- Description: An independent general access electronic data repository and search engine through which individual participant-level data and metadata from clinical trials conducted by researchers in academic, industry, foundation, and non-profit entities can be identified, hosted, shared and analyzed.

- How to access: `https://vivli.org/`
- Other notes: For data request review process, follow instructions on `https://vivli.org/about/data-request-review-process/`

Medical Information Mart for Intensive Care (MIMIC)

- Description: MIMIC is a large, freely-available database comprising deidentified health-related data from patients who were admitted to the critical care units of the Beth Israel Deaconess Medical Center. This includes various versions of MIMIC released from 2001-2019.
- How to access: `https://mimic.mit.edu/docs/gettingstarted/`
- Registration notes: To access the data files, one must request to be a credential user `https://physionet.org/settings/credentialing/`. Once granted with a credential, the teams must sign a data use agreeement and follow the citation instructions for the specific MIMIC dataset version.
- Other notes: Access to the data is granted to individuals only, not to teams or groups. Each person (each member of each team) who wishes to work with the data must apply separately for access.

Deep Lesion

- Description: The National Institutes of Health's Clinical Center has made a large-scale dataset of over 32,000 annotated lesions identified on CT images that are publicly available to the scientific community. It covers images of over 4,400 anonymized unique patients who have consented for NIH research endeavours. Read more: `https://www.nih.gov/news-events/news-releases/nih-clinical-center-releases-dataset-32000-ct-images`
- How to access: All files found on `https://nihcc.app.box.com/v/DeepLesion`
- Other notes:

MURA: MSK Xrays

- Description: MURA (musculoskeletal radiographs) is a large dataset of bone X-rays. Algorithms are tasked with determining whether an X-ray study is normal or abnormal. Musculoskeletal conditions affect more than 1.7 billion people worldwide, and are the most common cause of severe, long-term pain and disability, with 30 million emergency department visits annually and increasing. We hope that our dataset can lead to significant advances in medical imaging technologies which can diagnose at the level of experts, towards improving healthcare access in parts of the world where access to skilled radiologists is limited.
- How to access: See `https://stanfordaimi.azurewebsites.net/datasets/3e00d84b-d86e-4fed-b2a4-bfe3effd661b`
- Other notes: You must login to the site to be able to export the dataset available on the above the link.

VOICED Database

- Description: This database includes 208 voice samples, from 150 pathological, and 58 healthy voices. The healthy voices or the presence of each vocal fold's disorders were clinically verified by the medical experts involved in the project. All diagnoses were made according to indications of the SIFEL protocol, a clinical protocol compiled by the Italian Society of Phoniatrics and Logopaedics. The database includes information such as gender, age, pathology, lifestyle habits (e.g. smoking, alcohol and coffee consummation), occupational status, and the results of two specific medical questionnaires: the Voice Handicap Index (VHI) and Reflux Symptom Index (RSI).
- How to access: See `https://physionet.org/content/voiced/1.0.0/`
- Other notes: Anyone can access the files, as long as they conform to the terms of the specified license.

OpenTrials

- Description: OpenTrials is a collaborative and open database for all available structured data and documents on all clinical trials, threaded together by individual trial. With a versatile and expandable data schema, it is initially designed to host and match the following documents and data for each trial: registry entries; links, abstracts, or texts of academic journal papers; portions of regulatory documents describing individual trials; structured data on methods and results extracted by systematic reviewers or other researchers; clinical study reports; and additional documents such as blank consent forms, blank case report forms, and protocols.
- How to access: Can download the data dumps at `https://explorer.opentrials.net/data`.
- Other notes: Read more about the dataset at `https://trialsjournal.biomedcentral.com/articles/10.1186/s13063-016-1290-8`

Diagnosis Data for International Electronic Health Record-Derived COVID-19 Clinical Course Profile: The 4CE Consortium

- Description: Diagnosis data collected for the 4CE manuscript "International Electronic Health Record-Derived COVID-19 Clinical Course Profile: The 4CE Consortium" (Preprint: `https://doi.org/10.1101/2020.04.13.20059691`) and the 4CE website about the same manuscript (`https://covidclinical.net/publications/paper-01.html`).
- How to use: Download data directly from the website `https://figshare.com/articles/dataset/Diagnosis_Data_for_International_Electronic_Health_Record-Derived_COVID-19_Clinical_Course_Profile_The_4CE_Consortium/12152967/1`
- Other notes:

Genetics Datasets

SARS-CoV-2 Genetics

- Description: Data collected on all of the strains of SARS-CoV-2 reported to GenBank
- How to access: `https://www.kaggle.com/rtwillett/sarscov2-genetics`
- Access Process: Open website, scroll down to Data Explorer on the left-hand side of website, then download both CSV files containing data (No account sign up required)
- Other notes: Working to sequence as many strains of the virus as possible to try and track the genetic progression of the evolution of viral strains

Gene Expression Omnibus(GEO) Datasets

- Description: 'Gene Expression Omnibus(GEO) is an international public repository that archives and freely distributes microarray, next-generation sequencing, and other forms of high-throughput functional genomics data submitted by the research community.'
- How to access: `https://www.ncbi.nlm.nih.gov/gds/`
- Access Process: Open website, search up name of dataset interested in, scroll down list to find and select appropriate dataset, open the tab related to selected dataset, scroll down to 'Download Family' section, then download files containing data (No account sign up required)
- Other notes:

Multiple Single Cell RNA Expressions ARCHS4

- Description: Multiple gene expression data from the GEO database
- How to access: `https://www.kaggle.com/alexandervc/multiple-single-cell-rna-expressions-archs4`
- Access Process: Open website, scroll down to Data Explorer on the left-hand side of website, then download both CSV files containing data (No account sign up required)
- Other notes:

Genetic Variant Classifications

- Description: Variants are classified by clinical laboratories on a categorical spectrum ranging from benign, likely benign, uncertain significance, likely pathogenic, and pathogenic. Variants that have conflicting classifications (from laboratory to laboratory) can cause confusion when clinicians or researchers try to interpret whether the variant has an impact on the disease of a given patient.
- How to access: `https://www.kaggle.com/kevinarvai/clinvar-conflicting`
- Access Process: Open website, scroll down to Data Explorer on the left-hand side of website, then download both CSV files containing data (No account sign up required)
- Other notes:

Breast Cancer Proteomes Datasets

- Description: This data set contains published iTRAQ proteome profiling of 77 breast cancer samples generated by the Clinical Proteomic Tumor Analysis Consortium (NCI/NIH).
- How to access: `https://www.kaggle.com/piotrgrabo/breastcancerproteomes`
- Access Process: Open website, scroll down to Data Explorer on the left-hand side of website, then download both CSV files containing data (No account sign up required)
- Other notes:

ALS Genes Datasets

- Description: Genetic factors are an important cause of ALS, with variants in more than 25 genes having strong evidence, and weaker evidence available for variants in more than 120 genes. With the increasing availability of next-generation sequencing data, non-specialists, including health care professionals and patients, are obtaining their genomic information without a corresponding ability to analyze and interpret it.
- How to access: `https://www.kaggle.com/mpwolke/cusersmarildownloadsgenescsv`
- Access Process: Open website, scroll down to Data Explorer on the left-hand side of website, then download both CSV files containing data (No account sign up required)
- Other notes:

Pathogen-Phenotype Datasets

- Description: This is a database containing pathogen-to-phenotype associations mined from the scientific literature.
- How to access: `https://zenodo.org/record/2592941#.Xqrx3KhKh`
- Access Process: Open website, scroll down to Files section, press download on the right hand side of the file name to access data (No account sign up required)
- Other notes:

Autistic Human Gene Datasets

- Description: The SFARI Gene web portal seamlessly integrates different kinds of genetic data that are being generated by research studies, and in so doing encourages the generation of new hypotheses. SFARI Gene utilizes a systems biology approach, linking information on autism candidate genes within its original "Human Gene" module to corresponding data from a diverse array of supplementary data modules.
- How to access: `https://gene.sfari.org/database/human-gene/`
- Access Process: Open website, select 'Download this dataset' on the bottom right-hand side of website to access data (No account sign up required)
- Other notes: