# COVID-19 Open Research Dataset Challenge (CORD-19)

An AI challenge with AI2, CZI, MSR, Georgetown, NIH & The White House

**Dataset 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 29,000 scholarly articles, including over 13,000 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.

**Call to Action**

We are issuing a call to action to the world's artificial intelligence experts to develop text and data mining tools that can help the medical community develop answers to high priority scientific questions. The CORD-19 dataset represents the most extensive machine-readable coronavirus literature collection available for data mining to date. This allows the worldwide AI research community the opportunity to apply text and data mining approaches to find answers to questions within, and connect insights across, this content in support of the ongoing COVID-19 response efforts worldwide. There is a growing urgency for these approaches because of the rapid increase in coronavirus literature, making it difficult for the medical community to keep up.

A list of our initial key questions can be found under the [Tasks](https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge/tasks) section of this dataset. These key scientific questions are drawn from the NASEM’s SCIED (National Academies of Sciences, Engineering, and Medicine’s Standing Committee on Emerging Infectious Diseases and 21st Century Health Threats) [research topics](https://www.nationalacademies.org/event/03-11-2020/standing-committee-on-emerging-infectious-diseases-and-21st-century-health-threats-virtual-meeting-1) and the World Health Organization’s [R&D Blueprint](https://www.who.int/blueprint/priority-diseases/key-action/Global_Research_Forum_FINAL_VERSION_for_web_14_feb_2020.pdf?ua=1) for COVID-19.

Many of these questions are suitable for text mining, and we encourage researchers to develop text mining tools to provide insights on these questions.

**Acknowledgements**

Une image contenant dessin

Description générée automatiquement

This dataset was created by the Allen Institute for AI in partnership with the Chan Zuckerberg Initiative, Georgetown University’s Center for Security and Emerging Technology, Microsoft Research, and the National Library of Medicine - National Institutes of Health, in coordination with The White House Office of Science and Technology Policy.

**Task**

##### **What do we know about COVID-19 risk factors?**

**Details**

What do we know about COVID-19 risk factors? What have we learned from epidemiological studies?

Specifically, we want to know what the literature reports about:

* Data on potential risks factors
  + Smoking, pre-existing pulmonary disease
  + Co-infections (determine whether co-existing respiratory/viral infections make the virus more transmissible or virulent) and other co-morbidities
  + Neonates and pregnant women
  + Socio-economic and behavioral factors to understand the economic impact of the virus and whether there were differences.
* Transmission dynamics of the virus, including the basic reproductive number, incubation period, serial interval, modes of transmission and environmental factors
* Severity of disease, including risk of fatality among symptomatic hospitalized patients, and high-risk patient groups
* Susceptibility of populations
* Public health mitigation measures that could be effective for control

**Expected Submission**

To be valid, a [submission](https://www.kaggle.com/product-feedback/121068) must be contained in a single notebook made public on or before the submission deadline. Participants are free to use additional datasets in addition to the official Kaggle dataset, but those datasets must also be publicly available on either Kaggle, Allen.ai, or Semantic Scholar in order for the submission to be valid. Participants must also accept the [competition rules](https://www.kaggle.com/cord-19-rules-accept-form" \t "_blank).

**Reflection**

**1. Pipeline to describe.**

**2. Identify potential shortcomings with the current pipeline**

**3. Suggest possible improvements to the pipeline**

A possible improvement would be to ...

Another potential improvement could be to ...