Leveraging Data Science To Combat COVID-19: A Comprehensive Review

Siddique Latif [1,1], Muhammad Usman [2], Sanaullah Manzoor [2], Waleed Iqbal [2], Junaid Qadir [2], Gareth Tyson [2], Ignacio Castro [2], Adeel Razi [2], Maged N. Kamel Boulos [2], Adrian Weller [2], and Jon Crowcroft [2]

[1]University of Southern Queensland
[2]Affiliation not available

November 8, 2023

Abstract

COVID-19, an infectious disease caused by the SARS-CoV-2 virus, was declared a pandemic by the World Health Organisation (WHO) in March 2020. At the time of writing, more than 2.8 million people have tested positive. Infections have been growing exponentially and tremendous efforts are being made to fight the disease. In this paper, we attempt to systematise ongoing data science activities in this area. As well as reviewing the rapidly growing body of recent research, we survey public datasets and repositories that can be used for further work to track COVID-19 spread and mitigation strategies.

As part of this, we present a bibliometric analysis of the papers produced in this short span of time. Finally, building on these insights, we highlight common challenges and pitfalls observed across the surveyed works.

Leveraging Data Science To Combat COVID-19: A Comprehensive Review

Siddique Latif[1,2], Muhammad Usman[3,4], Sanaullah Manzoor[5], Waleed Iqbal[6], Junaid Qadir[5], Gareth Tyson[6,11], Ignacio Castro[6], Adeel Razi[7,8], Maged N. Kamel Boulos[9], and Jon Crowcroft[10,11]

[1]University of Southern Queensland, Australia
[2]Distributed Sensing Systems Group, Data61, CSIRO Australia
[3]Seoul National University, Seoul, South Korea
[4]Center for Artificial Intelligence in Medicine and Imaging, HealthHub Co. Ltd., South Korea
[5]Information Technology University, Punjab, Pakistan
[6]Queen Mary University of London, United Kingdom
[7]Turner Institute for Brain and Mental Health & Monash Biomedical Imaging, Monash University, Australia
[8]Wellcome Centre for Human Neuroimaging, University College London, United Kingdom
[9]Sun Yat-sen University, Guangzhou, China
[10]University of Cambridge, United Kingdom
[11]Alan Turing Institute, United Kingdom

***Abstract*—COVID-19, an infectious disease caused by the SARS-CoV-2 virus, was declared a pandemic by the World Health Organisation (WHO) in March 2020. At the time of writing, more than 1.8 million people have tested positive. Infections have been growing exponentially and tremendous efforts are being made to fight the disease. In this paper, we attempt to systematise ongoing data science activities in the area. As well as reviewing the rapidly growing body of recent research, we survey public datasets and repositories that can be used for further work to track COVID-19 spread and mitigation strategies. As part of this, we present a bibliometric analysis of the papers produced in this short span of time. Finally, building on these insights, we highlight common challenges and pitfalls observed across the surveyed works.**

I. Introduction

The SARS-CoV-2 virus and the associated disease (designated COVID-19) was first identified in Wuhan city (China) in December 2019 [1]–[3], and was declared a pandemic by the World Health Organisation (WHO) on 11 March 2020.[1] At the time of writing,[2] the Centre for Systems Science and Engineering at Johns Hopkins University reports 1,846,963 confirmed cases, 114,185 deaths, and 421,728 recovered.

Since December 2019, over 24,000 research papers from peer-reviewed journals as well as sources like medRxiv are available online [4]. Understanding this rapidly moving research landscape is particularly challenging since much of this literature has not been vetted through a peer-review process yet. This paper tries to overcome this challenge by presenting a detailed overview and survey of data science research related to COVID-19. It is intended as an (evolving) community resource to facilitate accessibility to the large volume of data and papers published in recent months. We use the term 'data science' as an umbrella term that encompasses all techniques that use scientific methods, algorithms, and systems to learn from structured and unstructured data. In examining this growing landscape of data science research around COVID-19, we make the following five contributions. *First*, we present pressing research problems related to COVID-19, for which data scientists may be able to contribute. *Second*, we summarise publicly available COVID-19 datasets that are being used to drive research, and list how they could be utilised to address some of the aforementioned problems. *Third*, we survey some of the ongoing research in the area, highlighting the main topics covered. As our primary audience is computer scientists and engineers, we theme our discussion around types of data analysis. *Fourth*, we broaden our analysis and present a bibliometric study of the rapidly growing literature on COVID-19. Bringing together our observations, *fifth*, we highlight some of the common challenges in this fast-moving space. We intentionally cast a wide net, covering research from several technical areas surrounding data science.

This paper builds upon recent reviews and perspective papers [5], [6] to help systematise existing resources and support the research community in building solutions to the COVID-19 pandemic. We have attempted in this review to be comprehensive and provide an up to date literature review and highlight important applications, community resources, publication trends, and challenges of data science research in COVID-19. However, in a dynamic rapidly-evolving field such as this, it is not possible for us to aim for exhaustiveness. Nonetheless we hope that our work will provide a solid introduction to the field for all researchers interested in this area.

The rest of paper is organised as follows. In Section II, we present the possible use cases that can help address COVID-19 challenges. In Section III, we present the details of available

Email: siddique.latif@usq.edu.au

[1]https://tinyurl.com/WHOPandemicAnnouncement

[2]1:57 am Monday, 13 April 2020 Coordinated Universal Time (UTC)

datasets and resources. In Section IV, we review contributions made by data scientists including image analysis, textual data mining, audio analysis, and embedded sensing. In Section V we present a bibliometric analysis of the COVID-19 related papers. Next, we discuss common challenges facing this research in Section VI. Finally, Section VII concludes the paper.

II. Applications of Data Science for COVID-19

Data science is a broad term covering topics such as Machine Learning (ML), statistical learning, time-series modelling, data visualisation, expert systems and probabilistic reasoning. Van der Schaar et al. [6] identify five major healthcare challenges where such technologies could help address in the fight against COVID-19: (i) managing limited healthcare resources; (ii) developing personalised plans for patient treatment/management; (iii) informing policies and enabling effective collaboration; (iv) understanding and accounting for uncertainty; and (v) expediting clinical trials. Building on this, we start by summarising some of the key research use cases that data scientists may be able to contribute to.

A. Risk Assessment and Patient Prioritisation

Healthcare systems around the world are facing unprecedented pressures on their resources (e.g., availability of intensive care beds, respirators). This creates the need to rapidly assess and manage patient risk, while allocating resources appropriately. In periods of peak load, this must be done rapidly and accurately, creating a substantial challenge for healthcare professionals who may not even have access to historical patient data. Various studies have already proposed algorithmic risk assessments of diseases such as cancer [7], diabetes [8], and cardiac-related diseases [9] with Artificial Neural Networks (ANNs). Due to diverse symptoms and disease trajectories, researching technologies for data-driven risk assessment and management in individual COVID-19 patients would be useful. For instance, traits like age, gender, or health state can be utilised to provide an estimate of mortality risk. This is particularly important when resources are limited, e.g., for patient prioritisation when Intensive Care Unit (ICU) resources are insufficient.

B. Screening and Diagnosis

A major issue facing countries with growing COVID-19 infection rates is the lack of proper screening and diagnosis facilities. This further complicates capacity management as well as social distancing measures, since those with mild symptoms are often unaware they carry the disease. A key use case is to develop remote computational diagnosis tools. Some already exist, which could be expanded, e.g., Babylon is a mobile app that provides medical advice via questioning. More advanced solutions could also rely on audio, e.g., COVID-19 Sounds is a mobile app collecting audio of breathing symptoms to help perform diagnosis.[3] We posit that such research will be particularly useful in developing countries that have a shortage of healthcare facilities [10]. Automated tools can also be developed to facilitate screening in larger groups of people (e.g., at airports), e.g., using computer vision based thermal imaging to detect fever [11].

[3] http://www.covid-19-sounds.org/

C. Epidemic Modelling

Healthcare systems require accurate epidemic models to perform capacity management and public policy formulation. In epidemiology, *compartmental models* are the most widely used for this [12]. In these models, populations are divided into compartments and the flow of people among compartments is modelled using ordinary differential equations. For example, COVID-19's spread has recently been modelled using the SEIR model [13], which models the flow of people between four states (or compartments): susceptible (S), exposed (E), infected (I), and recovered (R). *Generative models* define another broad class of models that generate consequences (i.e., data) from causes (i.e., hidden states and parameters). An example generative model, developed at University College London, is based upon ensemble or population dynamics that generate outcomes (new cases of COVID-19 over time) [14]. This approach captures the effects of interventions (e.g., social distancing) and differences among populations (e.g., herd immunity) to predict what might happen in different circumstances. For interested readers, websites offering COVID-19 forecasting have emerged,[4] each using a different model (although they should be treated with caution due to the uncertainty of such predictions [15] [16]).

Parameterising the above models requires up-to-date information of the virus spread. Thus, an important use case is finding ways to better capture such data. For instance, this could be done by processing social media information to identify people who have been infected, or even analysing ambulance call out data [17]. Another beneficial use case would be to develop ways to more accurately evaluate "*what-if*" scenarios with these models [15]. As an example, the initial policy of the UK government (of adopting almost no social isolation measures) was later changed based on results from an extended SEIR model from Imperial College London [18]. This model projected that without interventions there would be up to half a million fatalities, highlighting the importance of accurate predictions. More generally, a comprehensive review focused on modelling infectious disease dynamics in the complex landscape of global health can be seen at [19].

D. Contact Tracing

Most countries reacted to the early stages of COVID-19 with containment measures. This typically involves rapidly identifying infected individuals, followed by quarantine and contact tracing. Countries, such as South Korea, conducted rigorous testing campaigns, which allowed other potentially infected contacts to be quickly quarantined. This approach has been seemingly successful in containing the outbreak [20]. A valuable use case can be therefore facilitating more rapid and comprehensive contact tracing at scale [21]. Smartphone contact

[4] For example: (1) *COVID-19 worldwide peak forecasting method* (https://www.people.vcu.edu/~tndinh/covid19/en/) and (2) *COVID-19 forecasting* (http://epidemicforecasting.org/)

Table I: Organisation of paper and summary of different sections.

Sections	Subsection
(§II) Applications of Data Science for COVID-19 This section highlights different use cases related to the application of data-driven methodologies for addressing COVID-19. It also discusses some examples of these use cases.	(§II-A) Risk Assessment and Patient Prioritisation
	(§II-B) Screening and Diagnosis
	(§II-C) Epidemic Modelling
	(§II-D) Contact Tracing
	(§II-E) Logistical Planning
	(§II-F) Automated Primary Care
	(§II-G) Supporting Drug Discovery and Treatment
	(§II-H) Understanding Social Interventions
	(§II-I) Supporting Economic Recovery
(§III) Datasets and Resources This section provides information about numerous datasets related to COVID-19. It also gives information about ongoing data science competitions, and online resources.	(§III-A) COVID-19 Case Data
	(§III-B) COVID-19 Textual Data
	(§III-C) COVID-19 Biomedical Data
	(§III-D) Other Supportive Datasets
	(§III-E) COVID-19 Competition Datasets
(§IV) Survey of Ongoing Research This section surveys ongoing work across several types of data. It also provides brief summaries of outcomes and methodologies.	(§IV-A) Image Data Analysis
	(§IV-B) Textual Data Analysis
	(§IV-C) Voice Sound Data Analysis
	(§IV-D) Embedded Data Analysis
	(§IV-E) Pharmaceutical Research
(§V) Bibliometric Analysis of COVID-19 Research This section presents a bibliometric analysis of COVID-19 research.	(§V-A) Bibliometric Data Collection
	(§V-B) Peer-reviewed vs. Non-peer-reviewed publications
	(§V-C) Research Topics
	(§V-D) COVID-19 vs. Earlier Epidemics
(§VI) Cross-Cutting Challenges This section highlights challenges that researchers may face when performing data-driven research related to COVID-19.	(§VI-A) Data Limitations
	(§VI-B) Correctness of Results vs. Urgency
	(§VI-C) Security, Privacy, and Ethics
	(§VI-D) The Need For Multidisciplinary Collaboration
	(§VI-E) New Data Modalities
	(§VI-F) Solutions for the Developing World
(§VII) Conclusions	

sensing, online surveys and automated diagnosis have all been proposed to rapidly identify exposure [22]. For example, there are ongoing efforts to survey general populations via social media to learn of symptoms within individuals' social networks [23]. Even prior to COVID-19, FluPhone [24] used Bluetooth communications to identify contacts between people, and BlueDot monitors outbreaks of infectious diseases to alert governments, hospitals, and businesses [25].

E. Logistical Planning

COVID-19 has created serious challenges for healthcare supply chains and provisioning. This includes personal protective equipment such as masks and gowns, alongside intensive care equipment like test kits, beds, and ventilators. There is a history of applying machine learning to logistical planning, e.g., by Amazon Fulfilment.[5] A simple use case would be to apply data science techniques to help supply chain management for healthcare provisioning. This can also be used to preemptively allocate resources, e.g., researchers from the University of Cambridge are using depersonalised data (like lab results and hospitalisation details) to predict the need for ventilation equipment.[6] This use case could be critical for ensuring appropriate equipment is available on time.

F. Automated Patient Care

The pandemic has triggered a shortage of healthcare workers (e.g., in primary care). To alleviate this, automated primary care tools, such as remote chatbots and expert systems, could be developed and/or improved. Such systems can help people in providing information about the outbreak, symptoms, precautionary measures, etc. For instance, an interactive chatbot by the WHO and Rakuten Viber aims to provide accurate information about COVID-19 to people in multiple languages [26].

Automated healthcare methods could also be utilised to help monitor the conditions of COVID-19 patients in emergency care [27]. Another use case would be to gather and collate observational data to monitor (and then generalise) the efficacy of treatments for certain patient types. Similarly, physical and psychological (self) recordings could be used to scalably generate personal plans. Due to the need to rapidly discharge patients from hospitals, further monitoring could continue remotely. For instance, AliveCor [28] has launched Kardia Mobile 6L, which allows healthcare professionals to measure QTc (heart rate corrected interval). Similarly, TeleICU has been used to identify respiratory deterioration [29].

G. Supporting Drug Discovery and Treatment

The international effort to discover or re-purpose drug treatments and vaccines can also benefit from extensive data science work predating COVID-19 [30]. For example, computational methods can reduce the time spent on examining data, predicting protein structures and genomes [31], [32]. It can also assist in identifying eligible patients for clinical trials [33], an often costly and time consuming part of drug development. There is also substantial scope for applying advanced methods to managing trials, such as applying Bayesian clinical trials to adapt treatments based on information that accrues during the

[5] https://services.amazon.co.uk/services/fulfilment-by-amazon/features-benefits.html

[6] https://tinyurl.com/CambridgeCenterAIMedicineCOVID

trial [34]. This may be critical in expediting the delivery of drug treatments, and we argue this is another area where data scientists can contribute.

H. Understanding Social Interventions

Governments have taken steps to manage social interactions as part of their response to COVID-19. We highlight two main use cases of relevance.

1) Monitoring of Social Distancing: Many governments have implemented social distancing strategies to mitigate the spread of COVID-19. This is a non-pharmaceutical intervention that reduces human contact within the population [35] and therefore constrains the spread of COVID-19 [36]. Data science can support contact tracing the monitoring of social distancing, for instance by extracting social media data and using language processing techniques [37], [38]. These analyses could also help in keeping record of interactions to be used as individuals develop symptoms. Furthermore, these could be used for general population tracking to understand compliance with social distancing. This could then be complemented with other datasets (e.g., cellular trace data or air pollution monitoring [39]) to better understand human mobility patterns in the context of social distancing.

2) Controlling Misinformation & Online Harms: The spread of misinformation can undermine public health strategies [40] and has potentially dangerous consequences [41], [42]. For example, online rumours accusing 5G deployments of causing COVID-19 led to mobile phone masts being attacked in the UK [43]. Wikipedia maintains an up-to-date list of misinformation surrounding COVID-19 [44]. This confirms the spread of a number of dangerous forms of misinformation, e.g., that vinegar is more effective than hand sanitiser against the coronavirus. Naturally, users who believe such misinformation could proceed to undermine public health. One important use case would therefore be to develop classifiers and techniques to stem this flow. For example, Pennycook et al. [45] are testing simple interventions to reduce the spread of COVID-19 misinformation. An infodemic observatory analysing digital response in online social media to COVID-19 has been created by CoMuNe lab at Fondazione Bruno Kessler (FBK) institute in Italy and is available online.[7] The observatory uses ML techniques based on Twitter data to quantify collective sentiment, social bot pollution, and news reliability and displays this visually.

I. Supporting Economic Recovery

Social distancing measures are having a major impact on the global economy [46], [47]. As organisations emerge from economic hibernation they will be challenged to return to normal levels of service and operation given disruptions to their workforce. Data scientists might be able to assist in identifying problems limiting recovery. For instance, governments can use data science techniques to identify optimal economic interventions at a high level of granularity and companies can use data science to detect unusual patterns of behaviour in the market or in their own customer base.

[7]COVID19 Infodemics Observatory: https://covid19obs.fbk.eu/

III. Datasets and Resources

To enable research by the community, it is vital that datasets are made available. Next, we survey public datasets, some of which we summarise in Table II.

A. COVID-19 Case Data

The number of COVID-19 cases along with their geo-locations can help to track the growth of the pandemic and the geographical distribution of patients. Many countries are collecting and sharing infection information. One of the most used datasets is collated by John Hopkins University, which contains the daily number of positive cases, the number of cured patients and the mortality rates at a country as well as state/province level [48]. A further source of daily COVID-19 case data is available at Kaggle [49]. This dataset is annotated with other attributes such as patient demographics, case reporting date and location. Another epidemiological dataset, nCOV2019 [65], contains national and municipal health reports of COVID-19 patients. The key attributes are geo-location, date of confirmation, symptoms, and travel history. Similarly, the New York Times is compiling a state-wise dataset consisting of the number of positive cases and death count [54]. Whereas the above datasets are mostly based on statistics compiled by governmental administrations, other datasets are being collated using community surveys, requesting people to report infection rates among their social networks [23]. Common data science applications used with such data in the literature include data visualisation and predictive analytics [70].

A key limitation in these datasets is the divergence of testing regimes,[8] which makes it challenging to compare results across countries. It is estimated in one study[9] that the average detection rate of SARS-CoV-2 infections is a meagre 6 percent worldwide. Similarly, variations in interventions, population densities and demographics have a major impact, as can be seen when contrasting, for example, Japan vs. USA.[10] As such, regional prediction tasks are non-trivial, and we posit that temporal models such as Auto Regressive Integrated Moving Average (ARIMA) [71] and Long Short Term Memory (LSTM) [72] neural networks may be effective here.

B. COVID-19 Textual Data

The availability of rich textual data from various online sources can be used to understand the growth, nature and spread of COVID-19.

One prominent source is *social media*, for which datasets are already available covering COVID-19 discussions. There are open Twitter datasets covering Tweet IDs [55] and tweet text data [56]. These were gathered using Twitter's Streaming API to record tweets containing a series of related keywords, including "Coronavirus", "COVID-19", "N95", "Pandemic", etc. Another dataset of 2.2 million tweets, alongside the code to collect more data is available [73]. This data could be

[8]https://tinyurl.com/theconversationCoronaVirus
[9]https://tinyurl.com/cov6percent
[10]http://nrg.cs.ucl.ac.uk/mjh/covid19/index.html

Table II: A List of Prominent COVID-19 Datasets

Dataset Name	Country/Region	Modality	Attributes	Ref.
JHU CSSE COVID-19 Data	All Countries	Case statistics	Number of infections, number of cured patients, total mortality count, location	[48]
Novel Corona-virus 2019 dataset	All Countries	Case statistics	Patient demographics, case reporting date, location, brief history	[49]
Coronavirus Source Data	All Countries	Case statistics	Time series of confirmed daily COVID-19 cases for countries around the world	[50]
CHIME	All Countries	Case statistics	Daily number of susceptible, infected and recovered patient	[51]
COVID-19 Korea Dataset	Korea	Case statistics	Patient routes, age, gender, diagnosed date	[52]
hCOV-19	All Countries	Genomic epidemiology	Genetic sequence and metadata	[53]
New York Times dataset	USA	State-wise cumulative cases	Date, state name, number of cases, death count	[54]
Public Corona-virus Twitter Dataset	All Countries	Tweet IDs	Twitter ID with location	[55]
Coronavirus COVID19 Tweets	All Countries	Public Tweets on COVID-19	UserID, location, hashtags, tweet text	[56]
COVID-19 Open Research Challenge	All Countries	Research articles dataset	Published date, author list, journal name, full text	[57]
LitCovid	All Countries	Dataset of research articles	Up-to-date research topics and geographic locations	[58], [59]
Global research on COVID-19	All Countries	Database of research articles	Date, location, authors and journal	[60]
COVID-19 Community Mobility Reports	131 Countries	Mobility statistics with textual reports	Presence of people at grocery stores, pharmacies, recreational spots, parks, transit stations, workplaces, and residences	[61]
COVID-19 DATABASE	Italy	Radiology data	Xrays and demographics	[62]
RKI COVID19	Germany	Cases data	Number of infection cases	[63]
BSTI Imaging Database	United Kingdom	CT scans data	Patient CT scans	[64]
nCoV2019 Dataset	China, Japan, South Korea, Hong Kong, Taiwan, Thailand, Singapore	Epidemiological data	Patient demographics, case reporting date, location, brief history	[65]
COVID Chestxray Dataset	Italy	Chest X-ray scans and reports	X-Ray Image, date, patient, demographics, findings, location and survival information	[66]
COVID-19 CT segmentation dataset	Italy	Lungs CT scans	JPG image scans with segmentation and label report	[67]
COVID-CT-Dataset	All Countries	Chest CT-scans	Scans with associated labels	[67]
RCSB Protein Data Bank	All Countries	Clinical and pathology	Gemonic sequences	[68]
Kinsa Smart Thermometer Weather Map	USA	U.S. Health Weather Map	Temperature readings from internet-connected thermometers made by Kinsa Health.	[69]

used to monitor the spread of COVID-19, as well as people's reactions (e.g., to social distancing measures) using existing natural language processing techniques [74]–[76]. Sharma et al. [77] also made a public dashboard[11] available summarising data across more than 5 million real-time tweets.

The wealth of *academic publications* in recent weeks is also creating a deluge of textual information. Information extraction from clinical studies is already being performed [78] using language processing models such as [79]. These bibliometric datasets can easily be collected from pre-print services such as arXiv, medRxiv, and biorXiv [80]–[82]. The White House has also released an open research articles dataset [57]. This dataset contains nearly $45,000$ articles related to COVID-19, SAR-CoV-2 and other coronaviruses. These activities are mirrored across other organisations. For instance, in the US, The National Center for Biotechnology Information (NCBI) is providing up-to-date COVID-19 scientific literature [58], and WHO is compiling a database of recent research publications [60]. Closely related is the wealth of activity on Wikipedia, a community-driven encylopedia, which already contains substantial information about COVID-19. The entirety of Wikipedia can be downloaded for offline analysis [83], and there are already pre-processed Wikipedia datasets focussing on COVID-19 available.[12]

C. COVID-19 Biomedical Data

Biomedical data can be used to support diagnosis, prognosis and treatment. A major source of data are physical medical reports (such as X-rays) or clinical pathology reports (genomic sequencing). As the current diagnosis and prognosis of COVID-19 often requires human interpretations, there is potential for applications of computer vision research, e.g., automated diagnosis from chest X-rays. Currently, there are some open-source COVID-19 X-ray scans such as the *COVIDx* dataset [84]. These can be used for training COVID-19 infection assessment and diagnosis models (exploiting known computer vision techniques [85]). Other X-ray datasets that are publicly available for research are [66], [86]. The latter contains date, patient, demographics, findings, location and survival information. However, there are some intrinsic challenges related to these X-ray datasets, such as the requirement of radiologists or clinicians for data labelling and annotation (before training models). As such, the datasets are still small, limiting the application of methods like convolutional neural networks.

Lung Computed Tomography (CT) scans can also be used for COVID-19 diagnosis and prognosis. Currently, there are datasets of lungs CT scans available. One of the datasets [67] covers 60 patients and comprises three class labels: ground glass, consolidation, and pleural effusion. The dataset is collected from 6, March to 13 March, 2020. A larger dataset of 288 CT scans collected from 19 January to 25 March, 2020 [87]. The dataset has 275 CT scans of COVID-19 patients, which to the best of our knowledge, is the largest publicly available.

Besides the above physical scans, there are important genomic sequencing datasets available. The study of drug impact, protein-protein interactions and RNA structure in genomic data is an essential part of diagnosis test evaluations.

[11] https://usc-melady.github.io/COVID-19-Tweet-Analysis/
[12] http://covid-data.wmflabs.org/

Available datasets related to epidemiological and clinical data include RCSBdata [68] and *GHDDI* [88]. However, as COVID-19 has emerged very recently, these datasets are mostly incomplete or too small. For example, the biomedical datasets (see [87]) range from just a few up to 300 patients.

D. Other Supportive Datasets

As part of monitoring secondary factors related to COVID-19 and the surrounding interventions, there are several other relevant datasets. For example, air quality index statistics can be used as an indirect measure of social distancing polices, i.e., if movements are restricted there will be fewer vehicles (and pollution). A recent study showed that the air quality of six populous world cities has improved between February and March 2020 due to the measures to combat COVID-19 [39]. The data is publicly available [89] as well as the related COVID-19 case data [48]. Mobility trace data [90] can also serve a similar purpose—a collection of such logs is available here [91]. Note that mobility datasets have already been re-purposed: Google has released community mobility reports for public health officials in 131 countries [61]. These reports are compiled using Google Maps and describe how busy places such as grocery stores, transit stations, and workplaces are.

E. COVID-19 Competition Datasets

To facilitate and promote research in this area, there are several recent open data science competitions established on Kaggle (summarised in Table III). These are mostly based on the previously discussed data. For instance, the White House in coalition with some leading research groups (e.g., Kaggle and SGS Digicomply) has opened a new challenge using the earlier mentioned dataset of $45,000$ research articles [57]. For this, there a few questions posed; for example, "*What do we know about virus genetics, origin, and evolution?*" For each task, there is an associated prize of $1000.

Table III: COVID-19 Related Kaggle Competitions

Challenges	Aims
Answer 9 key questions	This challenge asks data scientists to understand COVID-19 faster by exploring 47,000 scholarly articles about COVID-19 and related coronaviruses.
COVID19 Global Forecasting	This challenge asks data scientists to predict the number of cases and fatalities by city between April 9 and May 7.
UNCOVER COVID-19	This challenge asks data scientists to use exploratory analysis to answer research questions that help support frontline responders .

The Roche Data Science Coalition (RDSC) also established the challenge "Uncover COVID19" [92]. RDSC has rolled-out a multi-modal dataset collected from 20 sources and has posed questions prepared by front-line healthcare experts, medical staff, WHO and governmental policymakers. This dataset is mainly collected from John Hopkins, the WHO, New York Times and the World Bank. It includes local and national COVID-19 cases, geo-spatial data and social distancing polices. Participants are required to design solutions to address questions like "*Which populations are at risk of contracting COVID-19?*" and "*Which populations have contracted COVID-19 and require ventilators?*".

Finally, the White House Office of Science and Technology Policy (OSTP) has opened a weekly challenge to predict the number of COVID-19 cases and fatalities at particular locations around the world [93]. Competitors are also required to unveil the factors associated with COVID-19 transmission rate. At the time of writing, participants are required to forecast the number of COVID-19 cases and deaths between 1-April-20 to 07-May-20.

For those wishing to engage in these competitions, there are several helpful tools and guideline blogs available. These resources provide support for data preprocessing, visualisations, and the implementation of different frameworks. We provide a list in Table IV.

Table IV: Prominent COVID-19 Community Resources.

Resources	Details
AI against COVID-19	This webpage contains information related to recent papers, projects, and datasets for COVID-19.
AitsLab-Corona	This is an NLP toolbox and related text processing resources for SARS-CoV-2 and COVID-19 NLP research.
Amazon AWS	Amazon AWS public data lake for COVID-19 data analysis
CDC, USA	Centers for Disease Control and Prevention (CDC) COVID-19 research articles downloadable database.
COVID-19 Graphs	This repository provides the tools to visualise the different statistics of COVID-19 using case data.
MATLAB resource	MATLAB based tutorial on deep learning based techniques for detecting COVID-19 using chest radiographs (in MATLAB).
ChemML [94]	ChemML is a machine learning and informatics program that support and advance the data-driven research in the domain of chemical and materials.
JHU's CSSE	Coronavirus COVID-19 Resource Page by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU).
MONTREAL.AI	This contains the details of multiple open source codes and tools to model different aspects of COVID-19.
NIH NLM LitCovid	LitCovid is a curated literature hub for tracking up-to-date scientific information about COVID-19. It provides central access to more than 3558 relevant articles in PubMed.
Vector Institute	This is a webpage that provides information about various resources and research tools for COVID-19.
WHO resource	This is a webpage of the WHO which contains updated details on the global research on COVID-19.

IV. Survey of Ongoing Data Science Related COVID-19 Research

The above provides an overview of publicly available datasets that could be used by researchers wishing to contribute to the COVID-19 crisis. Next, we detail some of the ongoing research in this space. We theme this section around the above datasets and summarise key studies in Table V.

A. Image Data Analysis

Various studies ([95]–[97]) have used computer vision algorithms to speed up the process of disease detection across several imaging modalities with some studies demonstrating that image analysis techniques have the potential to outperform expert radiologists [98], [99]. To diagnose COVID-19, two medical imaging modalities (CT and X-ray) have been experimented with, which we discuss below.

1) Computed Tomography (CT) Scans: Recent studies have found that radiologists can diagnose COIVD-19 using Chest CT scans with lower false positive rates [100], [101] than other imaging modalities such as X-ray and Ultrasound scans. Thus, many deep learning (DL) techniques related to CT scans

have been proposed to expedite the diagnosis process. Wang et al. [102] utilise DL methods to detect radiographical changes in COVID-19 patients. They evaluate the proposed model on the CT scans of pathogen-confirmed COVID-19 cases and show that DL can extract radiological features suitable for COVID-19 diagnosis. Xiaowei et al. [103] present a method for the automatic screening of COVID-19 in pulmonary CT scans using a 3D DL model with location-attention. They achieve promising accuracy to identify COVID-19 infected patients scans from other well-known infections. Chen et al. [104] exploit the UNet++ architecture [105] to detect suspicious lesions on CT scans. They trained their model on 289 scans and test on 600 scans. They achieve 100% accuracy in identifying the suspicious areas in CT scans of COVID-19 patients, which suggests their techniques have potential for clinical utilisation. Ophir et al. [106] employ 2D and 3D convolutional neural networks (CNNs) to calculate the Corona score (which represents the evolution of the disease in the lungs). They estimate the presence of the virus in each slice of CT scan with a 2D CNN and detect other lung diseases (i.e., lung nodule) by using a 3D CNN. Similarly in [107], a neural network (COVNet), is developed to extract visual features from volumetric chest CT exams for the detection of COVID-19. The study suggests that DL-based models can accurately detect COVID-19 and differentiate it from community acquired pneumonia and other lung diseases.

2) X-ray Scans: Ongoing image processing work is not limited to CT scans, and there has been work on other modalities, namely X-rays. Ezz et al. [108] propose a DL-based framework (COVIDX-Net) to automatically diagnose COVID-19 in X-ray images. COVIDX-Net includes seven different CNN models, such as VGG19 [109] and Google MobileNet [110]. The models can classify the patient status as either COVID-19 negative or positive. However, due to a lack of data, the technique is validated on only 50 X-ray images, among which 25 were of confirmed corona patients. Linda et al. [84] introduce another DL-based solution tailored for the detection of COVID-19 cases from chest X-ray images. They also develop a dataset named COVIDx and leverage it to train a deep CNN. In [111], three different CNN-based models (i.e., ResNet-50, Inception and InceptionResNet) are employed to detect COVID-19 in X-rays of pneumonia infected patients. The results show that the pre-trained ResNet-50 model [112] performs well, achieving 98% accuracy. Similarly, Farooq et al. [113] provide the steps to fine-tune a pre-trained ResNet-50 [112] architecture to improve model performance for detecting COVID-19 related abnormalities (called COVID-ResNet). Prabira et al. [114] use DL to extract the meaningful features from chest X-rays, and then trained a support vector machine (using the extracted features) to detect infected patients.

We also briefly note that several companies have released commercial solutions, some of which are freely available, e.g., VUNO Med [115] for chest CT and X-ray scans, which help diagnosis.

B. Textual Data Analysis

Researchers are currently utilising text mining to explore different aspects of COVID-19, mainly from social media and bibliometric data. To assist in this, Kazemi et al. [148] have developed a toolbox for processing textual data related to COVID-19. This toolbox comprises English dictionaries related to the disease, virus, symptoms and protein/gene terms.

In terms of social media research, Lopez et al. [149] explore the discourse around the COVID-19 pandemic and government policies being implemented. They use Twitter data from different countries in multiple languages and identify the popular responses to the pandemic using text mining. Similarly, Saire and Navarro [150] use text mining on Twitter data to show the epidemiological impact of COVID-19 on press publications in Bogota, Colombia. Intuitively, they find that the number of tweets is positively correlated with the number of infected people in the city. Schild et al. [134] inspect Twitter and 4Chan data to measure how sinophobic behaviour, driven by the pandemic, has evolved. This includes studying the impact that real world events, such as regional containment measures, have on online hate. Cinelli et al. [151] analyse Twitter, Instagram, YouTube, Reddit and Gab data about COVID-19. They find different volumes of misinformation on each platform. Singh et al. [152] are also monitoring the flow of (mis)information flow across 2.7M tweets, and correlating it with infection rates to find that misinformation and myths are discussed, but at lower volume than other conversations. For those seeking easy acceess to this information, FBK institute is collecting COVID-19 related tweets to visualise the presence of bots and misinformation.[13]

In terms of bibliometric analysis, Li et al. [153] analyse research publications on other coronaviruses (e.g, SARS, MERS). This is used to build a network-based drug re-purposing platform to identify drugs for the treatment of COVID-2019. Using module detection and drug prioritisation algorithms, the authors identify 24 disease-related human pathways, five modules and suggest 78 drugs to re-purpose. The rapid growth in COVID-19 related literature further led Hossain et al. [154] to perform a bibliometric analysis of COVID-19 related studies published since the outbreak. They review relationships, citations and keywords, which could be useful to new researchers in the area.

Finally, there is work processing text data from patient records. Roquette et al. [155] train a deep neural network to forecast patient admission rates using the unstructured text data available for triage. There are also other studies that utilise text data mining techniques to explore the important aspect of current situation.

C. Voice Sound Data Analysis

The most common symptoms of COVID-19 are linked to pneumonia, and the main mortality risk is cardiovascular disease followed by chronic respiratory disease. Hence, audio analysis has been considered a potential means for lightweight diagnosis. There is work performing diagnosis with respiratory and lung sound analysis [156], which can work even with low-cost smartphones [157]. High mortality risk groups, including the elderly, can also be continuously monitored using speech analysis [158]. The patterns of coughs [159], [160] and sneezing

[13] https://covid19obs.fbk.eu/

Table V: Summary of data science work related to COVID-19. Papers are categorised based on the dataset used.

Authors	Area	Modality/Data Type	Technique	Methodology
Wang et al. [102]	Image Analysis	Chest CT scans	InceptionNet on random ROIs	InceptionNet is used to detect the anomalies related to COVID-19 infection in lungs CT scan.
Xu et al. [103]		Chest CT scans	3D CNNs	3-D CNN models used to classify the COVID-19 infected regions in CT scans
Chen et al. [104]		Chest CT scans	UNet++	UNet++ architecture has been used to identify the suspicious areas in CT scans
Gozes et al. [106]		Chest CT scans	2D + 3D CNNs	2D and 3D CNNs models have been simultaneously employed to quantify the infection in the lungs of COVID19 patients.
Lin et al. [107]		Chest CT scans	CNN	COVNet; CNN-based model is developed to detect COVID-19 in chest CT scans
Shan et al. [116]		Chest CT scans	DNN	DL-based segmentation system is developed to quantify infected ROIs lung CT scans
Zhang et al. [117]		Chest CT scans	DenseNet	Used DenseNet-like architecture and optimised it for classification task to detect COVID-19 infection.
Wang et al. [118]		Chest CT scans	Pre-training + DNN	Pre-trained DNN has been used to improve detection of COVID-19 in lungs scans.
Mucahid et al. [119]		Chest CT scans	Conventional Feature Extraction techniques + SVM	GLCM, LDP, GLRLM, GLSZM, and DWT algorithms are used as feature extraction and SVM for classification
Zhao et al. [87]		Chest CT scans	CNN	Developed a public dataset and employed CNN for COVID-19 detection on chest CT scans.
Gozes et al. [120]		Chest CT scans	U-Net + ResNet	Used UNet for lung segmentation, ResNet for 2D slice classification and fine grain localisation for detection of infected regions in lungs
Asnaui et al. [121]		Chest X-rays and CT images	Fine tuning + CNNs	Various CNN-based models used for binary classification in COVID-19 detection on pneumonia affected X-ray and CT images.
Ezz et al. [108]		Chest X-rays	CNN-based models	Introduced COVIDX-Net, which includes seven different CNN models for classification of COVID-19 infected X-rays
Linda et al. [84]		Chest X-rays	ResNet	Open source solution (COVID-Net), which detects COVID-19.
Narin et al. [111]		Chest X-rays	ResNet50, InceptionV3 and InceptionResNetV2	Different CNN-based models are used to detect COVID-19 pneumonia infected patients chest X-rays
Prabira et al. [114]		Chest X-rays	DNN + SVM	Used DNN to extract meaningful information from X-rays and SVM for classification of corona affected X-rays.
Farooq et al. [113]		Chest X-rays	Fine-tuning + ResNet	Devised multi-stage fine-tuning scheme to improve performance and training time.
Abbas et al. [122]		Chest X-rays	Transfer learning (TL) + CNN	Employed TL and used previously developed CNN, called Decompose, Transfer, and Compose (DeTraC)
Chowdhury et al. [123]		Chest X-rays	CNN + Image argumentation	An image argumentation technique has been proposed to create the chest X-ray images for training
Alqudah et al. [124]		Chest X-rays	CNN, SVM, and Random Forest (RF)	Applied various ML techniques for classification of COVID-19 infected X-rays.
Goshal et al. [125]		Chest X-rays	Bayesian Convolutional Neural Networks (BCNN)	Investigated the significance of dropping weights BCNN
Fatima et al. [126]		Chest X-rays	CNN	Trained CNN for COVID-10 detection in X-rays
Xin et al. [127]		Chest X-rays	DenseNet	Used DenseNet Architecture [128] for COVID-10 detection in X-rays
Karim et al. [129]		Chest X-rays	DNN	Used neural ensemble method for classification
Ioannis et al. [130]		Chest X-rays	TL + CNN	TL is used for extracting patterns from common bacterial pneumonia patients X-rays using CNN to detect COVID-19
Jahanbin et al. [131]	Text data Mining	Twitter data	Evolutionary algorithm	In this work, authors used the unstructured data from Twitter and used a fuzzy rule-based evolutionary algorithm to timely detect outbreaks of the COVID-19
Zhao et al. [132]		Sina Microblog hot search list	Content mining algorithms	This work investigates the public's response at the beginning (December 31, 2019, to February 20, 2020) of the COVID-19 epidemic in China.
Li et al. [133]		Weibo data	SVM, Naïve Bayes (NB), Random Forest (RF)	Authors performed a case study on Weibo data to characterise the propagation of situational information in social media during COVID-19.
Schild et al. [134]		Twitter & 4Chan data	word2vec	Authors look at rise of COVID-19 related sinophobic abuse on Twitter and 4Chan.
Prabhakar et al. [135]		Twitter data	Topic modelling	In this work, the information flow on twitter during COVID-19 pandemic was studies using topic modelling.
Stephany et al. [136]		Risk reports data	Multiple text mining algorithms	In this work, authors used a data mining approach to identify industry-specific risk assessments related to COVID-19 in real-time.
Zhavoronkov et al. [88]	Pharmaceutical Research	Crystal structure, homology modelling, and co-crystallised ligands	Generative models	The authors utilised generative models ([137], [138]) to generate the molecules for the 3C-like protease that can act as potential inhibitors for SARS-CoV-2.
Hofmarcher et al. [139]		Drug-discovery databases	DNNs	In this work, authors utilise ChemAI [140], [141], a DNN trained on million of data pints across 3.2 million of molecules, for screening favourable inhibitors from the ZINC database [142] for SARS-CoV-2.
Beck et al. [143]		SMILES strings, amino acid sequences	Deep learning model	Authors utilise a pre-trained drug-target interaction model to predict commercially available antiviral drugs for COVID-19.
Kim et al. [144]		SMILES strings, amino acid sequences	AI-based prediction platform	A binding affinity prediction platform is used to detect available FDA approved drugs that can block SARS-CoV-2 from entering cells.
Richardson et al. [145]		Biomedical data	AI-driven knowledge graph	Authors use BenevolentAI to search for approved drugs that can block the viral infection process.
Stebbing et al. [146]		Biomedical data	AI-driven knowledge graph	This study examines approved antiviral and anti-inflammatory treatments for COVID-19.
Vijil et al. [147]		SMILES	Generative models	Design drug candidates specific to a given target protein sequence. They release around 3000 novel COVID-19 drug candidates.

[159], throat clearing and swallowing sounds [161] can all be analysed using speech and sound processing. At present, COVID-19 related speech data has limited availability, although the potential benefits are highlighted in [156]. Thus, mobile apps like COVID-19 Sounds are attempting to collect large audio datasets. In [162], the authors present an app called AI4COVID-19 for the preliminary diagnosis of COVID-19. It requires a 2 second cough sample and provides the preliminary diagnosis within a minute. This work confirms the feasibility of COVID-19 detection using cough samples with promising results (90% detection rate).

D. Embedded Sensor Data Analysis

Embedded data (e.g., from smartphones and sensors) is being used for remote patient care and diagnosis [163]. This can include mobility data, physiological vital signs, blood glucose, body temperature, and various other movement-related signals. In [164], the authors develop a system utilising real-time information, including demographic data, mobility data, disease-related data, and user-generated information from social media. The proposed system, called α-Satellite, can provide hierarchical community-level risk assessment that can inform the development of strategies against the COVID-19 pandemic. Google has also been using location data from smartphones to show people's movement during the pandemic [165]. Another study [166] presents the design of a low-cost framework for the detection of COVID-19 using smartphone sensors. They propose the use of the mobile phones of radiologists for virus detection. They highlight that the proposed framework is more reliable as it uses multi-readings from different sensing devices that can capture symptoms related to the disease.

Another recent study [21] concluded that COVID-19's "spread is too fast to be contained by manual contact tracing". To address this, disease tracking apps [24] use contact/location sensor data. The simplest ones aim to understand the spread of the disease, particularly mild cases that are not routinely lab tested. For example, the COVID Symptom Tracker app[14] and COVID Near You[15] service. Others, like Hong Kong's StayHomeSafe and Poland's Home Quarantine app [167], try to monitor if people obey quarantine rules (via geofencing). More advanced solutions can notify users if they have come into contact with somebody infected. Examples include China's Close Contact Detector app [168], China's complementary QR health code system [169]), Singapore's TraceTogether [170] app, and Israel's HaMagen [171] app. At the time of writing in early April 2020, the UK is also planning to launch a similar app [172].

We note that one critical challenge in the above apps is protecting user privacy [173], [174]. For instance, uploading contact data for server-side computation could create a nation-wide database of social relationships, particularly in countries where usage is mandatory. Recently, Decentralised Privacy-Preserving Proximity Tracing (DP-3T) [175] was proposed by the the Pan-European Privacy-Preserving Proximity Tracing (PEPP-PT) [176] consortium.[16] This is a mobile app that offers privacy-preserving alerts for people who may have recently been in contact with an infected person. TraceSecure [177] supports similar features based on homomorhpic encryption, whereas [178] offers privacy guarantees via private set intersection. Apple and Google have announced a partnership to develop their own privacy-preserving contact tracing specifications based on Bluetooth.[17]

E. Pharmaceutical Research

There is extensive ongoing work in using new experimental technologies to support the search for COVID-19 pharmaceuticals. This has received substantial attention in recent months in an attempt to build models to explore the 3D structure of SARS-CoV-2 (the virus that causes COVID-19). In [179], the authors use the AlphaFold model to predict the structures of six proteins related to SARS-CoV-2. AlphaFold [180] is a DL model based on a dilated ResNet architecture [112], which predicts the distance and the distribution of angles between amino acid residing on protein structure. In [181], the authors use a DNN-based model for de novo design of new small molecules capable of inhibiting the chymotrypsin-like (3CL) protease—the protein targets for corona-viruses. Based on the results they were able to identify 31 potential compounds as ideal candidates for testing and synthesis against SARS-CoV-2. Studies also attempt to improve the RT-PCR test by utilising ML and novel genome technologies. Metsky et al. [182] employ CRISPR[18] to develop assay designs for the detection of 67 respiratory viruses including SARS-CoV-2.

As well as the above, studies have utilised ML models to speed up drug development. Hu et al. [183] exploit a multi-task DNN for the prediction of potential inhibitors against SARS-CoV-2. They aim to urgently identify existing drugs that can be re-purposed. Based on the results, they list 10 potential inhibitors for SARS-CoV-2. Zhang et al. [184] perform DL-based drug screening against 4 chemical compound databases and tripeptides for SARS-CoV-2. Based on the results, they provide a list of potential inhibitors that can help facilitate drug development for COVID-19. Tang et al. [185] propose the use of reinforcement learning (RL) models to predict potential lead compounds targeting SARS-CoV-2. Similarly, in [186] the authors propose a collaborative and open antiviral discovery approach using deep RL technique to discover new molecules to fight COVID-19.

Finally, pharmaceutical interventions must go through clinical trials before being deployed. Accelerated clearance pathways for COVID-19 studies have been established by several regulators including the WHO, the European Medicines Agency, the UK Medicines and Healthcare products Regulatory Agency and the US Food and Drug Administration [187]. As of March 24, 2020, 536 relevant clinical trials were registered. A major barrier though is recruiting suitable patients. Data-driven solutions are available to rapidly identify eligible trial

[14] https://covid.joinzoe.com/

[15] https://www.covidnearyou.org/

[16] See also the European Commission's recommendation at https://ec.europa.eu/info/files/recommendation-apps-contact-tracing_en

[17] https://www.apple.com/covid19/contacttracing/

[18] A tool that uses an enzyme to edit genomes by cleaving specific strands of genetic code

participants [33], [188], and data collection platforms already exist for monitoring symptoms remotely [189]

V. Bibliometric Analysis of COVID-19 Research

We next augment the survey in the previous section with a brief bibliometric analysis of the research literature related to COVID-19. This gives a broader understanding of how publications have evolved across the short lifespan of the pandemic.

A. Bibliometric Data Collection

There are many data repositories which contain COVID-19 research articles, both peer-reviewed [58], [190], [191] and non-peer reviewed [80]–[82]. We use Scopus to crawl peer reviewed articles, and arXiv, medRxiv, and biorXiv for non-peer reviewed articles. Peer-reviewed articles in our dataset are from top venues in science, including Nature [192], Science [193], the Lancet [194], and the British Medical Journal (BMJ) [195]. See Table VI for a complete list of peer-reviewed articles and the number of articles in our dataset of COVID-19 publications. We developed scripts to gather this data from pre-print archives and database queries to fetch data from the Scopus database. Each entry includes title, authors, journal, publication date, etc. Our dataset covers papers on COVID-19 from all of the mentioned sources till April 9th, 2020. We extracted these papers from the corpus of papers using keyword matching on titles and abstract of the paper. We use "COVID-19", "COVID", "CoronaVirus", "Corona Virus", "Pandemic", "Epidemic", and "SARS-CoV-2" as candidate keywords. Finally we did a manual check to confirm that extracted papers do not include any unrelated papers. In total, the dataset covers 2752 publications, of which 1469 are pre-prints and 1283 are from peer reviewed journals.

Table VI: Peer-reviewed journals and the number of COVID-19 articles in our dataset.

Journal Name	**Article Count**
The Lancet	228
Nature	204
BMJ	183
Science	113
Journal of Medical Virology	102
New England Journal of Medicine	58
JAMA	55
Clinical Infectious Diseases	49
Journal of Infection	39
Travel Medicine Infectious Disease	35
International Journal of Infectious Diseases	26
Eurosurveillance	25
Emerging Infectious Diseases	20
Radiology	19
Viruses	19
Infection Control Hospital Epidemiology	18
Emerging Microbes Infections	17
Journal of Hospital Infection	16
Annals of Internal Medicine	13
International Journal of Antimicrobial Agents	9
Journal of Clinical Medicine	8
Journal of the American Academy of Dermatology	4
European Respiratory Journal	4
Journal of Travel Medicine	4
Journal of Virology	4
Methods in Molecular Biology	2
Circulation	2
Canadian Journal of Anaesthesia	1
American Journal of Transplantation	1
Anaesthesia	1
Chinese Journal of Laboratory Medicine	1
American Journal of Roentgenology	1
European Review for Medical Pharmacological Sciences	1
Indian Journal of Medical Research	1
Chinese Journal of Hospital Administration	1
Anesthesia Analgesia	1

B. Peer-reviewed vs. Non-peer-reviewed publications

The pandemic has resulted in the rapid production of academic material, much of which is yet to go through the peer review process due to the urgency of dissemination. Due to the need for rapid information, COVID-19 researchers are also looking towards preprint articles. For example, Wynants et al. [196] have presented a systematic review and critical appraisal of prediction models for diagnosis and prognosis of COVID-19 infection that considered both preprint and peer reviewed articles.

Figure 1 presents the cumulative number of COVID-19 related papers published since December, 2020 including non-peer-reviewed COVID-19 literature. We see that the number of papers has increased dramatically since the beginning of January. To date, non-peer-reviewed articles are the most numerous (bioRxiv, medRxiv and arXiv combined), whereas peer reviewed articles are substantially fewer (although growing). By far the most active outlet is medRxiv, which has published 61% of all non-peer reviewed papers in our dataset. Table VI presents the number of publications from each journal covered. We see a highly skewed distribution: The Lancet, Nature, the BMJ, Science and Journal of Medical Virology contain more publications than all other journals combined. Of course, we anticipate this will change in the long-term, as more pre-prints move into peer reviewed journals.

Figure 2 complements the above analysis by presenting the geo-distribution of both groups of publications. As the initial epicentre of COVID-19 pandemic, a major part of COVID-19 research has been contributed by China. The USA holds the second position in terms of research contributions. Both hold roughly the same ratio of peer reviewed vs. non peer reviewed articles (2/3 are pre-print), e.g., China has 368 peer reviewed article and 735 non peer reviewed articles on COVID-19 in our dataset.

C. Research Topics

We next use topic modelling to identify core sub-topics within the publications. For this, we use Latent Dirichlet Allocation (LDA) [197]. This algorithm extracts and clusters abstract topics that exist within the papers. We divide our dataset into two groups: (1) all papers, and (2) data science related paper. We have tagged these papers manually based on their title and abstract.

Table VII shows the latent clusters of topics discussed in all papers in our dataset. Note that we split the results into peer

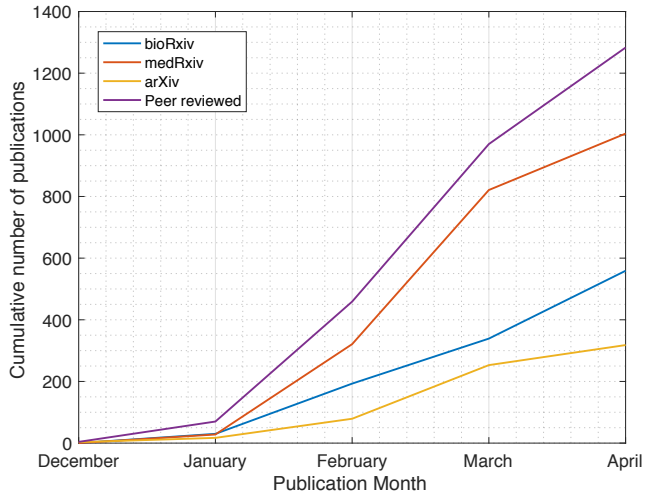
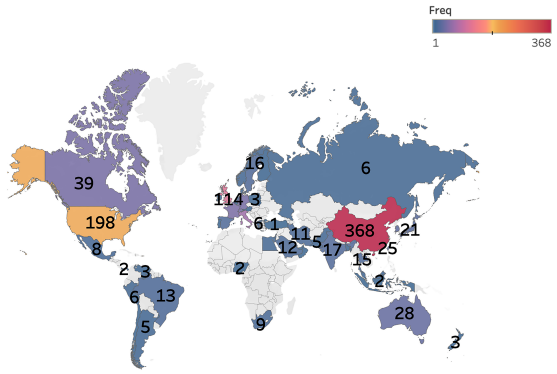
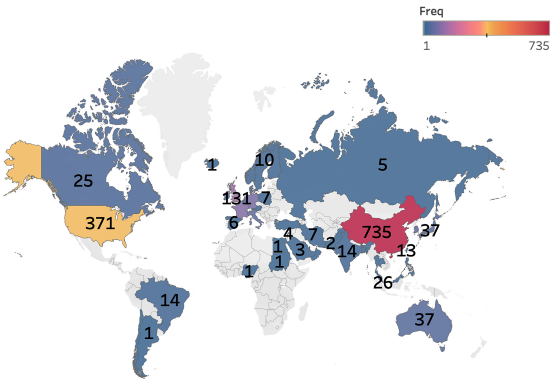


Figure 1: Cumulative distribution of publications per month on COVID-19 (data gathered till April 9th, 2020).



(a) Peer reviewed papers



(b) Pre-print (non-peer-reviewed) papers

Figure 2: Publication count of different countries on COVID-19.

reviewed vs. pre-print publications. The results are intuitive, covering many important aspects of COVID-19 research, e.g., disease cure, transmission of COVID-19, the role of different animals, social distancing, the impact of COVID-19 on crime rates, and the impact of age on COVID-19 positive patients. In contrast, Table VIII, shows the list of topics observed in data science related COVID-19 papers. These topics show that data science research on COVID-19 is being carried out using many important techniques and algorithms. Noteworthy algorithms and techniques include multidimensional kernel estimation, Bayesian learning, and deep learning based epidemic forecasting with synthetic information (TDFESI). We hope that these results will be useful to the community in identifying key topics receiving coverage.

D. COVID-19 vs. Earlier Epidemics

We conclude our bibliometric analysis by briefly comparing the rate of publication for COVID-19 research vs. prior epidemics. For this, we select Ebola and SARS-CoV-1. Figure 3 presents a time series for the first 3 years of publications. Note that the X-range differs and, naturally, we only have data since December 2019 for COVID-19.

We see that COVID-19 literature is growing faster than any prior epidemic. There have been more peer-reviewed publications ($\sim$1000) in around 3 months for COVID-19 than there were in 3 years for SARS-COV-1 and Ebola. Furthermore, as noted in the earlier subsection, there are even more pre-prints being released which means that COVID-19 has rapidly overtaken other epidemics in terms of academic attention. Of course, this is driven in-part by the wider geographic coverage of COVID-19, impacting numerous highly research active countries (e.g., China, USA, UK, Germany)

VI. Challenges in Data Science Related COVID-19 Research

In this section, we highlight some of the most important data science challenges. We specifically focus on cross-cutting challenges that impact all of the previously discussed use cases.

A. Data Limitations

Machine learning models are demanding in terms of data. Ideally, the data should be of high fidelity and voluminous. For many of the above use cases, extensive labelled datasets are not yet available, e.g., for speech analysis. Although there are few publicly available datasets for medical images and textual analysis, these datasets are small compared to the requirements of deep learning models. For example, in the case of biomedical data, sample sizes range from a few up to 60 patients (see [67]). The scarcity of measured data is frequently due to the distributed nature of many data sources. For example, electronic healthcare records are often segregated on a national, regional, or even per-hospital level. A key challenge is therefore federating these sources, and overcoming practical differences across each source, e.g., in terms of schemas. Thus, better and more automated approaches to data munging, data wrangling etc. may be critical in attaining fast, reliable and robust outcomes.

Beyond these challenges regarding availability of data, there are also major challenges within the data itself. The time-critical nature of this research is causing hurdles in developing certain types of high-quality dataset. For instance, by the time social media data is collected, curated and annotated it can become

Table VII: Top topics discussed in *overall* COVID-19 research literature

(a) Top topics discussed in *non-peer-reviewed papers* on COVID-19

Topic No.	Extracted Topics
1	(risk, grade, shed, infect, effect, strong, parturients, viral, favipiravir, arbidol, monkey, glucose)
2	(coronavirus, sequence, genome, virus, infect, viral, bat, human, novel, outbreak)
3	(antibodies, patient, sample, infect, detect, serology, assay, negative, serum, swab)
4	(estimate, transmission, countries, outbreak, number, infect, epidemic, spread, travel, reproduction)
5	(specimen, sputum, influenza, throat, nasal, heat, primer-prob, respiration, inactive, psychiatry)
6	(virus, vaccine, epitope, genome, coronavirus, mutate, sequence, strain, differ, region, viral, human, immune, high, peptide, develop)
7	(intervention, health, public, social, populate, reduce, quarantine, epidemic, outbreak, distance)
8	(temperature, treatment, hypoxemia, meteorology, coronavirus, effect, degree, receive, progress, screen)
9	(drug, effect, virus, antivirus, potential, inhibitor, protease, coronavirus, model, compound)
10	(positive, age, data, negative, differ, rate, older, associate, variant, fatal)

(b) Top topics discussed in *peer-reviewed* papers on COVID-19

Topic No.	Extracted Topics
1	(patient, pneumonia, severe, infect, treatment, lung, symptom, coronavirus, chest, hospital)
2	(coronavirus, sequence, virus, human,bat, genome, respiration, gene, animal, origin)
3	(health, medic, care, patient, staff, pandemic, social, service, score, protect)
4	(infect, coronavirus, disease, outbreak, case, spread, virus, health, transmission, respiratories)
5	(infect, effect, strong, viral, glucose, genome, virus, antibodies, patient, sample)
6	(viral, bat, human, novel, outbreak, nasal, sequence, strain, inhibitor, infect)
7	(patient, sample, infect, detect, serum, swab, antibodies, influenza, test, sputum)
8	(countries, outbreak, number, infect, epidemic, spread, travel, reproduction)
9	(effect, virus, protease, coronavirus, model, compound, degree, receive, potential, inhibitor)
10	(intervention, health, public, social, data, negative, differ, rate, outbreak, distance)

Table VIII: Top topics discussed in COVID-19 *data science based* research papers

(a) Top topics in *non-peer-reviewed data science based* COVID-19 papers

Topic No.	Extracted Topics
1	(model, number, use, countries, passengers, access, china, reduction, outbreak, result)
2	(dimensions, kernel, complex, structure, spectral, time, network, distance, base, infection-link)
3	(learn, image, covid-19, detect, dataset, feature, patient, predict, risk, death)
4	(sample, network, estimate, image, detrace, mean, transfer, covid-19, x-ray, medics)
5	(epidemic, risk, data, detect, method, health, influence, outbreak, measure, covid-19)
6	(model, graph, number, mixture, rate, predict, infect, algorithm, china, covid-19)
7	(sepsis, learn, feature, clinic, severe, treatment, differ, disease, auroc, automate)
8	(forecast, data, epidemic, high-resolute, tdefsi, method, ili, disease, mds, perform)
9	(data, world, period, trend, death, register, pandemic, epidemic, model, covid-19)
10	(crime, virus, sars-cov-2, genotype, isolate, mutate, global, genome, public, policies)

(b) Top topics in *peer-reviewed data science based* COVID-19 papers

Topic No.	Extracted Topics
1	(estimate, number, outbreak, model, epidemic, method, data, rate, dynamics, coronavirus)
2	(infect, estimate, death, risk, disease, quarantine, asymptomatic, coronavirus, intervene, individual)
3	(number, case, infect, model, data, epidemic, patient, control, peak, forecast)
4	(report, forecast, use, cumulative, predict, growth, data, outbreak, transmission, improve)
5	(coronavirus, quarantine, countries, data, suspect, measure, effect, ratio, intervention, transmission)
6	(outbreak, coronavirus, period, transmission, peak, predict, reproduction, mean, intervention)
7	(case, cities, model, number, outbreak, fit, dynamics, prevent, trend, predict)
8	(outside, travel, cause, viral, range, detect, phase, pneumonia, incubate, quarantine)
9	(case, estimate, epidemic, global, export, forecast, risk, incident, reproduction, severe)
10	(control, outbreak, trace, isolate, transmission, symptom, prevent, model, onset, strategies)

out-of-date. Due to this, real-time datasets often contain poorly quantified biases. For example, daily infection rates in Japan exhibit few similarities to those in the Italy. Training models on unrepresentative datasets will lead to poor (and even dangerous) outcomes. Whereas techniques such as transfer learning could allow models to be specialised with regional characteristics, the fast-moving nature of the problem can make it difficult to perform informed model selection and parameterisation. A key challenge is devising analytical approaches that can work with these data limitations.

B. Correctness of Results vs. Urgency

There is a clear need for rapid results, yet the methods surveyed in this paper are largely based on statistical learning on (quickly produced) datasets. In a recent systematic review of prediction models for diagnosis and prognosis of COVID-19 infection, Wynants et al. [196] have reported that all 31 reviewed prediction models had a high risk of bias (due to non-representative selection of control patients and model overfitting) and presently lacked external validation (which would require time). The reported models are therefore susceptible to errors. This is an inherent risk in all scientific work but, given the fast-moving nature of the situation, errors can have severe consequences. It should be remembered that the outcomes of research may impact healthcare policy. For example, predictions may be used by governments to decide social distancing policies. Yet political actors are often less well placed to understand the nuance of scientific studies. We therefore posit that a key challenge is balancing exigency vs. the need for well-evidenced and reproducible results that can inform policy.

A clear challenge is finding ways to capture the uncertainty of conclusions produced within this flurry of research. Bayesian methods can be used to capture uncertainty, although we have seen limited quantification of uncertainty in studies so far [198]. To ensure the correctness of data analysis, researchers must also facilitate reproducible conclusions, e.g., sharing code, data and documentation. This, again, can create challenges as such requirements are balanced against the need for urgency. Another promising avenue is 'Explainable AI' [199], which can be used to provide context to results. That said, it is not clear if this will protect against problems such as unintentional bias [200] or even adversarial scenarios [201].

C. Security, Privacy, and Ethics

Most works discussed imply the *sharing* and/or *use* of potentially personal and sensitive data. Devising solutions that exhibit good results but also protect privacy and adhere

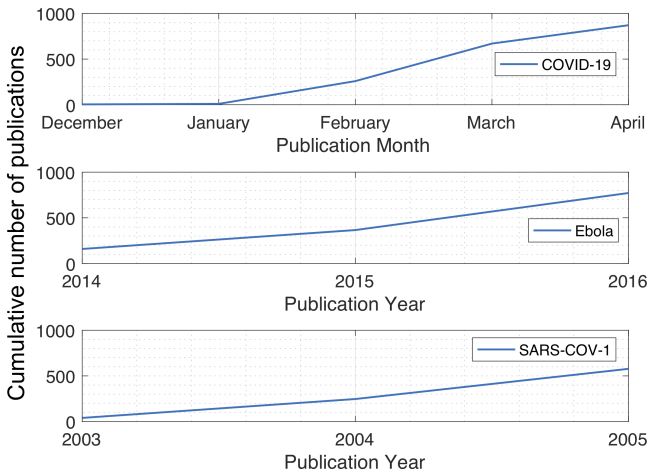


Figure 3: Cumulative publication rates for peer-reviewed publications in COVID-19, SARS-CoV-1, Ebola. Note the different X-ranges.

to high ethical standards is a key challenge. We argue that this could be vital for encouraging uptake among populations, particularly as infrastructure setup may persist beyond the pandemic [202]. There are already substantial efforts to build privacy-preserving medical analytics. For example, MedCo [203] uses homomorphic encryption to allow sites to federate datasets with privacy guarantees. Drynx [204] supports privacy-conscious statistical analysis on distributed datasets. This links closely into the quality of data (see §VI-A), as often data can only be shared when robust privacy guarantees are in place.

Broadly speaking, there is some consensus as outlined in Floridi et al. [205] on the five main "AI ethics principles": (1) *beneficence*, (2) *non-maleficence*, (3) *autonomy*, (4) *justice*, and (5) *explicability*. However, in the situation imposed by COVID-19, decisions may balance between these AI ethics virtues [206]. For example, to what extent does the current situation warrant the prioritisation of "public health" and "beneficence" over "individual privacy" and "autonomy". And even if this is warranted in the short-term, how can we ensure that these compromises do not become permanent and it is possible to roll back these tradeoffs in the future as the situation changes. Other difficult questions include the question of allocation of scarce resources and the tradeoffs involved therein. As highlighted in the Call of Action presented in March 2020 by a coalition of experts on data governance [207], there is also a needed for data sharing between public and private sectors to ensure that data is used for "beneficence" where it is needed. In effect, the failure to share data in such contexts may be considered maleficence since withholding critical data may block an opportunity that data science models can leverage to bring potential benefit. That said, good governance mechanisms with suitable regulations should be in place to oversee ethical use of data as much as possible.

Privacy may become particularly challenging when considering the roll-out of interventions (e.g., targeted social distancing measures) as the intervention itself may expose sensitive information. This, for example, may apply to contact tracing apps, which strive to notify users when they have been in contact with an infected person. Although privacy-preserving implementations exist (e.g., DP-3T, TraceSecure), notification may still allow users to guess who the infected person is (see [208] for a discussion of security issues in tracing apps). To move ahead, simple measures can be adopted to help ensure ethical data science research. For example, data collected should be transparent (the users should be informed about what data is being collected) and stewarded with a limited purpose (even when it is anonymised) and governed with ethical oversight and appropriate safeguards (e.g., with time limits and sunset provisions). Interested readers are referred to comprehensive resources on data ethics [205], [209]–[213], and to a recent report from the TUM Institute for Ethics in Artificial Intelligence on the ethical challenges involved in using AI for managing the COVID-19 outbreak [206].

D. The Need For Multidisciplinary Collaboration

Our understanding of COVID-19's long-term impact is still limited. Contributing serious insights will require a mix of domain expertise from multiple fields, and there is already a push for better international collaboration and tracking of COVID-19 [214]. For example, the use of black-box models might yield a superficially practical solution, but could be useless without the involvement of (international) medical and biotechnology expert interpretations. This will further have implications for licensing technologies and engendering uptake (as healthcare professionals are unlikely to engage with technologies developed without medical expertise). Rapidly bringing together cohorts of complementary expertise is therefore important. This also brings many further challenges, e.g., ensuring a team's interpretation of things like ethics, benefits and risks are coherent.

E. New Data Modalities

The data science community has limited exposure to certain modalities of data that may prove critical in combating COVID-19. A natural challenge is rapidly adapting existing techniques to reflect these new data types. For example, whereas the community has substantial expertise in computer vision tasks, there is less experience in processing ultrasound scans. Yet these have shown good results that are similar to chest CT scans and superior to standard chest radiography for the evaluation of pneumonia and/or acute respiratory distress syndrome in corona patients [215], [216]. They also have the benefit of greater ease of use, absence of radiation, and low cost. Despite these advantages, to the best of our knowledge, no study has yet explored the potential of automatically detecting COVID-19 infections via ultrasound scans. Similarly, magnetic resonance imaging (MRI) is considered the safest imaging modality as it is a non-invasive and non-ionising technique, which provides a high resolution image and excellent soft tissue contrast [96]. Some studies like [217] have described the significance of MRI in fighting against COVID-19 infections. Yet the modality remained under-explored by the computer vision community due to a lack of sufficient training data. Thus, a challenge is

to rapidly develop a well-annotated dataset of such medical imaging modalities.

F. Solutions for the Developing World

The COVID-19 pandemic poses unique challenges to populations that have limited access to healthcare (e.g. in developing countries), particularly as such people are disproportionately affected by limited access to information [218]. A key challenge is therefore developing technologies that are designed so that they are globally inclusive. These should expressly consider how such technologies could be deployed in both rural and economically deprived regions [219]–[221], as well as how they might be misused in certain contexts. This subsumes several practical challenges that naturally vary based on the specific use case. For example, if building a mobile app for contact tracing, it should be low cost and require limited resources; it should be designed with limited network connectivity in-mind; it should also support multiple languages and be accessible to illiterate users or those with disabilities. We emphasise that ensuring wide accessibility of technological solutions is critical for addressing this *global* pandemic.

VII. Conclusions

Data scientists have been active in addressing the emerging challenges related to COVID-19. This paper has been written to rapidly make available a summary of ongoing work for the wider community. We have attempted to make five broad contributions. We first presented relevant use cases of data science, which have the potential to help in the pandemic. This is by no means a comprehensive list and we expect the set to expand in the coming months. We then focused on summarising publicly available datasets for use by researchers. Again, this is intended as a community resource to shorten the time taken to discover relevant data. Following this, we surveyed some of the ongoing research in this area. As the paper is mainly intended for a computer science and engineering audience, we again themed our analysis around the different types of datasets available. Following this, we broadened our analysis and presented a bibliometric study of thousands of publications in recent months. Finally, we highlighted some of the common challenges we observed as part of our systematic review, e.g., availability of data and privacy concerns. We also note that many of the systems discussed in this paper are not operational yet. In view of this, we intend to update the paper repeatedly with new information.

References

[1] C. Huang, Y. Wang, X. Li, L. Ren, J. Zhao, Y. Hu, L. Zhang, G. Fan, J. Xu, X. Gu *et al.*, "Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China," *The Lancet*, vol. 395, no. 10223, pp. 497–506, 2020.

[2] J. T. Wu, K. Leung, and G. M. Leung, "Nowcasting and forecasting the potential domestic and international spread of the 2019-ncov outbreak originating in Wuhan, China: a modelling study," *The Lancet*, vol. 395, no. 10225, pp. 689 – 697, 2020.

[3] N. Zhu, D. Zhang, W. Wang, X. Li, B. Yang, J. Song, X. Zhao, B. Huang, W. Shi, R. Lu *et al.*, "A novel coronavirus from patients with pneumonia in China, 2019," *New England Journal of Medicine*, 2020.

[4] K. Hao, "*Over 24,000 coronavirus research papers are now available in one place*, published on 16-March-2020; accessed on: 6-April-2020." [Online]. Available: https://tinyurl.com/MITTECHREV24000papers

[5] J. Bullock, K. H. Pham, C. S. N. Lam, M. Luengo-Oroz *et al.*, "Mapping the landscape of artificial intelligence applications against COVID-19," *arXiv preprint arXiv:2003.11336*, 2020.

[6] M. van der Schaar, A. Alaa, A. Floto, A. Gimson, S. Scholtes, A. Wood, E. McKinney, D. Jarrett, P. Lio, and A. Ercole, "*How artificial intelligence and machine learning can help healthcare systems respond to COVID-19*, published on 27-March-2020; accessed on: 1-April-2020."

[7] A. A. Khorana, "Artificial intelligence for cancer-associated thrombosis risk assessment–author's reply," *The Lancet Haematology*, vol. 5, no. 9, pp. e391–e392, 2018.

[8] J. Dave, V. N. Dubey, D. Coppini, and J. Beavis, "Predicting diabetic neuropathy risk level using artificial neural network based on clinical characteristics of subjects with diabetes," *Diabetic Medicine*, pp. 144–144, 2019.

[9] K. Wattanakit, G. Harshavardhan, S. Mungee, and M. Imtiaz, "Artificial intelligence based clinical risk assessment in predicting cardiac related chest pain in patients presenting to emergency room," *Circulation*, vol. 140, no. Suppl_1, pp. A14 100–A14 100, 2019.

[10] S. Latif, M. Y. Khan, A. Qayyum, J. Qadir, M. Usman, S. M. Ali, Q. H. Abbasi, and M. A. Imran, "Mobile technologies for managing non-communicable diseases in developing countries," in *Mobile applications and solutions for social inclusion*. IGI Global, 2018, pp. 261–287.

[11] "*Startup Uses Fever Detection Technology To Stop Spread of Coronavirus*, accessed on: 4-April-2020," 2020. [Online]. Available: https://tinyurl.com/feverdetectiontechnology

[12] W. O. Kermack, A. G. McKendrick, and G. T. Walker, "A contribution to the mathematical theory of epidemics," *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, vol. 115, no. 772, pp. 700–721, 1927.

[13] A. J. Kucharski, T. W. Russell, C. Diamond, Y. Liu, J. Edmunds, S. Funk, R. M. Eggo, F. Sun, M. Jit, J. D. Munday *et al.*, "Early dynamics of transmission and control of COVID-19: a mathematical modelling study," *The Lancet Infectious Diseases*, 2020.

[14] K. J. Friston, T. Parr, P. Zeidman, A. Razi, G. Flandin, J. Daunizeau, O. Hulme, A. J. Billig, V. Litvak, R. J. Moran, C. J. Price, and C. Lambert, "Dynamic causal modelling of COVID-19," *arXiv preprint arXiv:2004.04463*, 2020.

[15] Z. Tufekci, "*Don't Believe the COVID-19 Models: That's not what they're for*, The Atlantic, Data Published: April 2, 2020," 2020. [Online]. Available: https://www.theatlantic.com/technology/archive/2020/04/coronavirus-models-arent-supposed-be-right/609271/

[16] M. Koerth, L. Bronner, and J. Mithani, "*Why It's So Freaking Hard To Make A Good COVID-19 Model*," https://fivethirtyeight.com/features/why-its-so-freaking-hard-to-make-a-good-covid-19-model/, 2020.

[17] A. Noulas, C. Moffatt, D. Hristova, and B. Gonçalves, "Foursquare to the rescue: Predicting ambulance calls across geographies," in *Proceedings of the 2018 International Conference on Digital Health*, 2018, pp. 100–109.

[18] N. Ferguson, D. Laydon, G. Nedjati Gilani, N. Imai, K. Ainslie, M. Baguelin, S. Bhatia, A. Boonyasiri, Z. Cucunuba Perez, G. Cuomo-Dannenburg *et al.*, "Report 9: Impact of non-pharmaceutical interventions (npis) to reduce covid19 mortality and healthcare demand," 2020.

[19] H. Heesterbeek, R. M. Anderson, V. Andreasen, S. Bansal, D. De Angelis, C. Dye, K. T. Eames, W. J. Edmunds, S. D. Frost, S. Funk *et al.*, "Modeling infectious disease dynamics in the complex landscape of global health," *Science*, vol. 347, no. 6227, p. aaa4339, 2015.

[20] J. Hellewell, S. Abbott, A. Gimma, N. I. Bosse, C. I. Jarvis, T. W. Russell, J. D. Munday, A. J. Kucharski, W. J. Edmunds, F. Sun *et al.*, "Feasibility of controlling covid-19 outbreaks by isolation of cases and contacts," *The Lancet Global Health*, 2020.

[21] L. Ferretti, C. Wymant, M. Kendall, L. Zhao, A. Nurtay, L. Abeler-Dörner, M. Parker, D. Bonsall, and C. Fraser, "Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing," *Science*, 2020.

[22] A. S. S. Rao and J. A. Vazquez, "Identification of COVID-19 can be quicker through artificial intelligence framework using a mobile phone-based survey in the populations when cities/towns are under quarantine," *Infection Control & Hospital Epidemiology*, pp. 1–18, 2020.

[23] C. B. et al., "Coronasurveys: Monitoring COVID-19 incidence via open polls," 2020. [Online]. Available: http://coronasurveys.com/

[24] E. Yoneki and J. Crowcroft, "Epimap: Towards quantifying contact networks for understanding epidemiology in developing countries," *Ad Hoc Networks*, vol. 13, pp. 83–93, 2014.

[25] "*AI could help with the next pandemic—but not with this one*, MIT Technology Review, accessed on: 1-April-2020." [Online]. Available: https://www.technologyreview.com/s/615351/ai-could-help-with-the-next-pandemicbut-not-with-this-one/

[26] "*WHO and Rakuten Viber fight COVID-19 misinformation with interactive chatbot*, Accessed on: 4-April-2020." [Online]. Available: https://tinyurl.com/WHORatukenChatBot

[27] CPR News, "*Health Care Workers' Stress Compounded By Long Days And Concerns About People Not Taking COVID-19 Seriously*, accessed on: 1-April-2020." [Online]. Available: https://www.cpr.org/2020/03/23/colorado-coronavirus-stress-healthcare-workers-covid-19-spread/

[28] "AliveCor, accessed on: 1-April-2020." [Online]. Available: https://www.alivecor.com/

[29] "CLEW Medical, accessed on: 1-April-2020." [Online]. Available: https://clewmed.com/

[30] J. B. Mitchell, "Artificial intelligence in pharmaceutical research and development," 2018.

[31] K.-K. Mak and M. R. Pichika, "Artificial intelligence in drug development: present status and future prospects," *Drug discovery today*, vol. 24, no. 3, pp. 773–780, 2019.

[32] A. Zhavoronkov, "Artificial intelligence for drug discovery, biomarker development, and generation of novel chemistry," 2018.

[33] G. Tyson, A. Taweel, S. Miles, M. Luck, T. Van Staa, and B. Delaney, "An agent-based approach to real-time patient identification for clinical trials," in *International Conference on Electronic Healthcare*. Springer, 2011, pp. 138–145.

[34] D. A. Berry, "Bayesian clinical trials," *Nature reviews Drug discovery*, vol. 5, no. 1, pp. 27–36, 2006.

[35] A. Wilder-Smith and D. Freedman, "Isolation, quarantine, social distancing and community containment: pivotal role for old-style public health measures in the novel Coronavirus (2019-ncov) outbreak," *Journal of Travel Medicine*, vol. 27, no. 2, 2020.

[36] F.-J. Schmitt, "A simplified model for expected development of the SARS-CoV-2 (corona) spread in germany and us after social distancing," *arXiv preprint arXiv:2003.10891*, 2020.

[37] C. St Louis and G. Zorlu, "Can twitter predict disease outbreaks?" *Bmj*, vol. 344, p. e2353, 2012.

[38] A. Signorini, A. M. Segre, and P. M. Polgreen, "The use of Twitter to track levels of disease activity and public concern in the US during the influenza A H1N1 pandemic," *PloS one*, vol. 6, no. 5, 2011.

[39] M. Cadotte, "Early evidence that COVID-19 government policies reduce urban air pollution," Mar 2020. [Online]. Available: eartharxiv.org/nhgj3

[40] J. Zarocostas, "How to fight an infodemic," *The Lancet*, vol. 395, no. 10225, p. 676, 2020.

[41] L. Bode and E. K. Vraga, "See something, say something: Correction of global health misinformation on social media," *Health communication*, vol. 33, no. 9, pp. 1131–1140, 2018.

[42] P. M. Waszak, W. Kasprzycka-Waszak, and A. Kubanek, "The spread of medical fake news in social media–the pilot quantitative study," *Health policy and technology*, vol. 7, no. 2, pp. 115–118, 2018.

[43] N. Parveen and J. Waterson, "Uk phone masts attacked amid 5g-coronavirus conspiracy theory," https://www.theguardian.com/uk-news/2020/apr/04/uk-phone-masts-attacked-amid-5g-coronavirus-conspiracy-theory, 2020.

[44] "Misinformation related to the 2019–20 coronavirus pandemic," https://en.wikipedia.org/wiki/Misinformation_related_to_the_2019%E2%80%9320_coronavirus_pandemic, 2020.

[45] G. Pennycook, J. McPhetres, Y. Zhang, and D. Rand, "Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy nudge intervention," 2020.

[46] W. J. McKibbin and R. Fernando, "The global macroeconomic impacts of COVID-19: Seven scenarios," 2020.

[47] A. Atkeson, "What will be the economic impact of COVID-19 in the US? rough estimates of disease scenarios," National Bureau of Economic Research, Tech. Rep., 2020.

[48] CSSEGISandData, "CSSEGISandData/COVID-19," Mar 2020. [Online]. Available: https://github.com/CSSEGISandData/COVID-19

[49] sudalairajkumar Data, "Novel corona-virus dataset," Mar 2020. [Online]. Available: https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset

[50] ECDC, "European Centre for Disease Prevention and Control (ECDC)," 2020. [Online]. Available: https://ourworldindata.org/coronavirus-source-data

[51] CHIME, "COVID-19 Hospital Impact Model for Epidemics," 2020. [Online]. Available: https://github.com/CodeForPhilly/chime

[52] "COVID-19 Korea Dataset with Patient Routes," 2020. [Online]. Available: https://github.com/ThisIsIsaac/Data-Science-for-COVID-19

[53] GISAID, "Genomic epidemiology of hCoV-19," 2020. [Online]. Available: https://www.gisaid.org/epiflu-applications/next-hcov-19-app/

[54] NewYork-Times, "New york times dataset," Mar 2020. [Online]. Available: https://github.com/nytimes/covid-19-data

[55] E. Chen, K. Lerman, and E. Ferrara, "COVID-19: The first public Coronavirus Twitter dataset," *arXiv preprint arXiv:2003.07372*, 2020.

[56] Smith, "Coronavirus (covid19) tweets," Mar 2020. [Online]. Available: www.kaggle.com/smid80/coronavirus-covid19-tweets

[57] Allen-Institute, "CORD-19 research challenge," Mar 2020. [Online]. Available: https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge

[58] NCBI, "LitCovid," 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/research/coronavirus/

[59] Q. Chen, A. Allot, and Z. Lu, "Keep up with the latest coronavirus research," *Nature*, vol. 579, no. 7798, p. 193, 2020.

[60] WHO, "Global research on novel coronavirus-2019," 2020. [Online]. Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/global-research-on-novel-coronavirus-2019-ncov

[61] G. Inc., "COVID-19 Community Mobility Reports," Mar 2020. [Online]. Available: https://www.google.com/covid19/mobility/

[62] SIRM, "COVID-19 DATABASE," 2020. [Online]. Available: https://www.sirm.org/category/senza-categoria/covid-19/

[63] NPGEO, "Dataset of infections in germany," 2020. [Online]. Available: https://npgeo-corona-npgeo-de.hub.arcgis.com/datasets/dd4580c810204019a7b8eb3e0b329dd6_0/data

[64] SIRM, "Covid-19-BSTI Imaging Database," 2020. [Online]. Available: https://www.bsti.org.uk/training-and-education/covid-19-bsti-imaging-database/

[65] nCoV2019Data, "ncov2019 epidemiological data," Mar 2020. [Online]. Available: https://github.com/beoutbreakprepared/nCoV2019

[66] J. P. Cohen, P. Morrison, and L. Dao, "COVID-19 image data collection," *arXiv 2003.11597*, 2020. [Online]. Available: https://github.com/ieee8023/covid-chestxray-dataset

[67] MegSeg, "COVID-19 CT segmentation dataset," Mar 2020. [Online]. Available: http://medicalsegmentation.com/covid19/

[68] D. S. Goodsell, C. Zardecki, L. Di Costanzo, J. M. Duarte, B. P. Hudson, I. Persikova, J. Segura, C. Shao, M. Voigt, J. D. Westbrook *et al.*, "RCSB protein data bank: Enabling biomedical research and drug discovery," *Protein Science*, vol. 29, no. 1, pp. 52–65, 2020.

[69] Kinsa Health, "U.S. Health Weather Map," Mar 2020. [Online]. Available: https://healthweather.us/?mode=Atypical

[70] A. Koubaa, "Understanding the covid19 outbreak: A comparative data analytics and study," *arXiv preprint arXiv:2003.14150*, 2020.

[71] G. P. Zhang, "Time series forecasting using a hybrid ARIMA and neural network model," *Neurocomputing*, vol. 50, pp. 159–175, 2003.

[72] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[73] C. Jacobs, "Coronada," 2020. [Online]. Available: https://github.com/BayesForDays/coronada

[74] E. Aramaki, S. Maskawa, and M. Morita, "Twitter catches the flu: detecting influenza epidemics using Twitter," in *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 2011, pp. 1568–1576.

[75] A. Culotta, "Towards detecting influenza epidemics by analyzing Twitter messages," in *Proceedings of the first workshop on social media analytics*, 2010, pp. 115–122.

[76] V. Lampos, T. De Bie, and N. Cristianini, "Flu detector-tracking epidemics on Twitter," in *Joint European conference on machine learning and knowledge discovery in databases*. Springer, 2010, pp. 599–602.

[77] K. Sharma, S. Seo, C. Meng, S. Rambhatla, A. Dua, and Y. Liu, "Coronavirus on social media: Analyzing misinformation in Twitter conversations," *arXiv preprint arXiv:2003.12309*, 2020.

[78] D. Zhao, F. Yao, L. Wang, L. Zheng, Y. Gao, J. Ye, F. Guo, H. Zhao, and R. Gao, "A comparative study on the clinical features of COVID-19 pneumonia to other pneumonias," *Clinical Infectious Diseases*, 2020.

[79] C. Manning, "Understanding human language: Can NLP and deep learning help?" in *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, 2016, pp. 1–1.

[80] "arXiv, accessed on: 12-April-2020." [Online]. Available: https://arxiv.org/

[81] "medRxiv, accessed on: 12-April-2020." [Online]. Available: https://www.medrxiv.org/

[82] "bioRXiv, accessed on: 12-April-2020." [Online]. Available: https://www.biorxiv.org/

[83] "Wikipedia database downloaod," https://en.wikipedia.org/wiki/Wikipedia:Database_download, 2020.

[84] L. Wang and A. Wong, "COVID-Net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest radiography images," 2020.

[85] D. Mery, "Computer vision for X-Ray testing," *Switzerland: Springer International Publishing.–2015*, vol. 10, pp. 978–3, 2015.

[86] "COVID Chest-Xray Dataset," Mar 2020. [Online]. Available: https://github.com/ieee8023/covid-chestxray-dataset

[87] J. Zhao, Y. Zhang, X. He, and P. Xie, "COVID-CT-Dataset: a CT scan dataset about COVID-19," *arXiv preprint arXiv:2003.13865*, 2020.

[88] A. Zhavoronkov, V. Aladinskiy, A. Zhebrak, B. Zagribelnyy, V. Terentiev, D. S. Bezrukov, D. Polykovskiy, R. Shayakhmetov, A. Filimonov, P. Orekhov *et al.*, "Potential COVID-2019 3c-like protease inhibitors designed using generative deep learning approaches," *Insilico Medicine Hong Kong Ltd A*, vol. 307, p. E1, 2020.

[89] W. Inc., "World's air pollution: Real-time air quality index," 2020. [Online]. Available: https://waqi.info

[90] N. Oliver, E. Letouzé, H. Sterly, S. Delataille, M. De Nadai, B. Lepri, R. Lambiotte, R. Benjamins, C. Cattuto, V. Colizza *et al.*, "Mobile phone data and COVID-19: Missing an opportunity?" *arXiv preprint arXiv:2003.12347*, 2020.

[91] "Data world," 2020. [Online]. Available: https://data.world/datasets/mobile

[92] R. data science coalition, "Uncover covid19 challenge," Mar 2020. [Online]. Available: https://www.kaggle.com/roche-data-science-coalition/uncover

[93] "Covid19 Global Forecasting Challenge, The White House Office of Science and Technology," Mar 2020. [Online]. Available: https://www.kaggle.com/c/covid19-global-forecasting-week-2/overview

[94] M. Haghighatlari, G. Vishwakarma, D. Altarawy, R. Subramanian, B. U. Kota, A. Sonpal, S. Setlur, and J. Hachmann, "Chemml: A machine learning and informatics program package for the analysis, mining, and modeling of chemical and materials data," *Wiley Interdisciplinary Reviews: Computational Molecular Science*, p. e1458, 2019.

[95] S. Robertson, H. Azizpour, K. Smith, and J. Hartman, "Digital image analysis in breast pathology—from image processing techniques to artificial intelligence," *Translational Research*, vol. 194, pp. 19–35, 2018.

[96] M. Usman, S. Latif, M. Asim, B.-D. Lee, and J. Qadir, "Retrospective motion correction in multishot MRI using generative adversarial network," *Scientific Reports*, vol. 10, no. 1, pp. 1–11, 2020.

[97] M. Usman, B.-D. Lee, S. S. Byon, S. H. Kim, and B. IlLee, "Volumetric lung nodule segmentation using adaptive roi with multi-view residual learning," *arXiv preprint arXiv:1912.13335*, 2019.

[98] Z. Z. Qin, M. S. Sander, B. Rai, C. N. Titahong, S. Sudrungrot, S. N. Laah, L. M. Adhikari, E. J. Carter, L. Puri, A. J. Codlin *et al.*, "Using artificial intelligence to read chest radiographs for tuberculosis detection: A multi-site evaluation of the diagnostic accuracy of three deep learning systems," *Scientific reports*, vol. 9, no. 1, pp. 1–10, 2019.

[99] R. Singh, M. K. Kalra, C. Nitiwarangkul, J. A. Patti, F. Homayounieh, A. Padole, P. Rao, P. Putha, V. V. Muse, A. Sharma *et al.*, "Deep learning in chest radiography: detection of findings and presence of change," *PloS one*, vol. 13, no. 10, 2018.

[100] T. Ai, Z. Yang, H. Hou, C. Zhan, C. Chen, W. Lv, Q. Tao, Z. Sun, and L. Xia, "Correlation of chest CT and RT-PCR testing in Coronavirus disease 2019 (COVID-19) in China: a report of 1014 cases," *Radiology*, p. 200642, 2020.

[101] Y. Li and L. Xia, "Coronavirus disease 2019 (covid-19): Role of chest ct in diagnosis and management," *American Journal of Roentgenology*, pp. 1–7, 2020.

[102] S. Wang, B. Kang, J. Ma, X. Zeng, M. Xiao, J. Guo, M. Cai, J. Yang, Y. Li, X. Meng *et al.*, "A deep learning algorithm using CT images to screen for corona virus disease (COVID-19)," *medRxiv*, 2020.

[103] X. Xu, X. Jiang, C. Ma, P. Du, X. Li, S. Lv, L. Yu, Y. Chen, J. Su, G. Lang *et al.*, "Deep learning system to screen Coronavirus disease 2019 pneumonia," *arXiv preprint arXiv:2002.09334*, 2020.

[104] J. Chen, L. Wu, J. Zhang, L. Zhang, D. Gong, Y. Zhao, S. Hu, Y. Wang, X. Hu, B. Zheng *et al.*, "Deep learning-based model for detecting 2019 novel Coronavirus pneumonia on high-resolution computed tomography: a prospective study," *medRxiv*, 2020.

[105] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, 2018, pp. 3–11.

[106] O. Gozes, M. Frid-Adar, H. Greenspan, P. D. Browning, H. Zhang, W. Ji, A. Bernheim, and E. Siegel, "Rapid AI development cycle for the Coronavirus (COVID-19) pandemic: Initial results for automated detection & patient monitoring using deep learning CT image analysis," *arXiv preprint arXiv:2003.05037*, 2020.

[107] L. Li, L. Qin, Z. Xu, Y. Yin, X. Wang, B. Kong, J. Bai, Y. Lu, Z. Fang, Q. Song *et al.*, "Artificial intelligence distinguishes COVID-19 from community acquired pneumonia on chest CT," *Radiology*, p. 200905, 2020.

[108] E. El-Din Hemdan, M. A. Shouman, and M. E. Karar, "COVIDX-Net: A framework of deep learning classifiers to diagnose COVID-19 in X-Ray images," *arXiv*, pp. arXiv–2003, 2020.

[109] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[110] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[111] A. Narin, C. Kaya, and Z. Pamuk, "Automatic detection of Coronavirus disease (COVID-19) using x-ray images and deep convolutional neural networks," *arXiv preprint arXiv:2003.10849*, 2020.

[112] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[113] M. Farooq and A. Hafeez, "COVID-ResNet: A deep learning framework for screening of COVID19 from radiographs," *arXiv preprint arXiv:2003.14395*, 2020.

[114] P. K. Sethy and S. K. Behera, "Detection of coronavirus disease (covid-19) based on deep features," *Preprints*, 2020.

[115] V. Inc., "VUNO Med-LungQuant & Chest X-ray solutions," 2020. [Online]. Available: https://covid19.vunomed.com/

[116] F. Shan+, Y. Gao+, J. Wang, W. Shi, N. Shi, M. Han, Z. Xue, D. Shen, and Y. Shi, "Lung infection quantification of COVID-19 in ct images with deep learning," *arXiv preprint arXiv:2003.04655*, 2020.

[117] J. Zhang, Y. Xie, Y. Li, C. Shen, and Y. Xia, "Covid-19 screening on chest x-ray images using deep learning based anomaly detection," *arXiv preprint arXiv:2003.12338*, 2020.

[118] S. Wang, Y. Zha, W. Li, Q. Wu, X. Li, M. Niu, M. Wang, X. Qiu, H. Li, H. Yu *et al.*, "A fully automatic deep learning system for COVID-19 diagnostic and prognostic analysis," *medRxiv*, 2020.

[119] M. Barstugan, U. Ozkaya, and S. Ozturk, "Coronavirus (covid-19) classification using ct images by machine learning methods," *arXiv preprint arXiv:2003.09424*, 2020.

[120] O. Gozes, M. Frid-Adar, N. Sagie, H. Zhang, W. Ji, and H. Greenspan, "Coronavirus detection and analysis on chest ct with deep learning," *arXiv preprint arXiv:2004.02640*, 2020.

[121] K. E. Asnaoui, Y. Chawki, and A. Idri, "Automated methods for detection and classification pneumonia based on x-ray images using deep learning," *arXiv preprint arXiv:2003.14363*, 2020.

[122] A. Abbas, M. M. Abdelsamea, and M. M. Gaber, "Classification of COVID-19 in chest x-ray images using DeTraC deep convolutional neural network," *arXiv preprint arXiv:2003.13815*, 2020.

[123] M. E. Chowdhury, T. Rahman, A. Khandakar, R. Mazhar, M. A. Kadir, Z. B. Mahbub, K. R. Islam, M. S. Khan, A. Iqbal, N. Al-Emadi *et al.*, "Can AI help in screening viral and COVID-19 pneumonia?" *arXiv preprint arXiv:2003.13145*, 2020.

[124] A. M. Alqudah, S. Qazan, H. Alquran, I. A. Qasmieh, and A. Alqudah, "COVID-2019 detection using x-ray images and artificial intelligence hybrid systems."

[125] B. Ghoshal and A. Tucker, "Estimating uncertainty and interpretability in deep learning for Coronavirus (COVID-19) detection," *arXiv preprint arXiv:2003.10769*, 2020.

[126] F. M. Salman, S. S. Abu-Naser, E. Alajrami, B. S. Abu-Nasser, and B. A. Ashqar, "Covid-19 detection using artificial intelligence," 2020.

[127] X. Li and D. Zhu, "Covid-xpert: An ai powered population screening of covid-19 cases using chest radiography images," *arXiv preprint arXiv:2004.03042*, 2020.

[128] Y. Zhu and S. Newsam, "Densenet for dense flow," in *2017 IEEE international conference on image processing (ICIP)*. IEEE, 2017, pp. 790–794.

[129] M. Karim, T. Döhmen, D. Rebholz-Schuhmann, S. Decker, M. Cochez, O. Beyan *et al.*, "Deepcovidexplainer: Explainable covid-19 predictions based on chest x-ray images," *arXiv preprint arXiv:2004.04582*, 2020.

[130] I. D. Apostolopoulos and T. A. Mpesiana, "Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks," *Physical and Engineering Sciences in Medicine*, p. 1, 2020.

[131] K. Jahanbin and V. Rahmanian, "Using Twitter and web news mining to predict COVID-19 outbreak," 2020.

[132] Y. Zhao and H. Xu, "Chinese public attention to COVID-19 epidemic: Based on social media," *medRxiv*, 2020.

[133] L. Li, Q. Zhang, X. Wang, J. Zhang, T. Wang, T.-L. Gao, W. Duan, K. K.-f. Tsoi, and F.-Y. Wang, "Characterizing the propagation of situational information in social media during COVID-19 epidemic: A case study on weibo," *IEEE Transactions on Computational Social Systems*, 2020.

[134] L. Schild, C. Ling, J. Blackburn, G. Stringhini, Y. Zhang, and S. Zannettou, ""go eat a bat, chang!": An early look on the emergence of sinophobic behavior on web communities in the face of COVID-19," 2020.

[135] D. Prabhakar Kaila, D. A. Prasad *et al.*, "Informational flow on Twitter–Corona virus outbreak–topic modelling approach," *International Journal of Advanced Research in Engineering and Technology (IJARET)*, vol. 11, no. 3, 2020.

[136] F. Stephany, N. Stoehr, P. Darius, L. Neuhäuser, O. Teutloff, and F. Braesemann, "The CoRisk-index: A data-mining approach to identify industry-specific risk assessments related to COVID-19 in real-time," *arXiv preprint arXiv:2003.12432*, 2020.

[137] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey, "Adversarial autoencoders," *arXiv preprint arXiv:1511.05644*, 2015.

[138] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[139] M. Hofmarcher, A. Mayr, E. Rumetshofer, P. Ruch, P. Renz, J. Schimunek, P. Seidl, A. Vall, M. Widrich, S. Hochreiter *et al.*, "Large-scale ligand-based virtual screening for SARS-CoV-2 inhibitors using deep neural networks," *Available at SSRN 3561442*, 2020.

[140] A. Mayr, G. Klambauer, T. Unterthiner, M. Steijaert, J. K. Wegner, H. Ceulemans, D.-A. Clevert, and S. Hochreiter, "Large-scale comparison of machine learning methods for drug target prediction on ChEMBL," *Chemical science*, vol. 9, no. 24, pp. 5441–5451, 2018.

[141] K. Preuer, G. Klambauer, F. Rippmann, S. Hochreiter, and T. Unterthiner, "Interpretable deep learning in drug discovery," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer, 2019, pp. 331–345.

[142] T. Sterling and J. J. Irwin, "ZINC 15–ligand discovery for everyone," *Journal of chemical information and modeling*, vol. 55, no. 11, pp. 2324–2337, 2015.

[143] B. R. Beck, B. Shin, Y. Choi, S. Park, and K. Kang, "Predicting commercially available antiviral drugs that may act on the novel coronavirus (SARS-CoV-2) through a drug-target interaction deep learning model," *Computational and Structural Biotechnology Journal*, 2020.

[144] J. Kim, Y. Cha, S. Kolitz, J. Funt, R. Escalante Chong, S. Barrett, B. Zeskind, R. Kusko, H. Kaufman *et al.*, "Advanced bioinformatics rapidly identifies existing therapeutics for patients with coronavirus disease-2019 (COVID-19)," 2020.

[145] P. Richardson, I. Griffin, C. Tucker, D. Smith, O. Oechsle, A. Phelan, and J. Stebbing, "Baricitinib as potential treatment for 2019-ncov acute respiratory disease," *The lancet*, vol. 395, no. 10223, pp. e30–e31, 2020.

[146] J. Stebbing, A. Phelan, I. Griffin, C. Tucker, O. Oechsle, D. Smith, and P. Richardson, "Covid-19: combining antiviral and anti-inflammatory treatments," *The Lancet Infectious Diseases*, 2020.

[147] V. Chenthamarakshan, P. Das, I. Padhi, H. Strobelt, K. W. Lim, B. Hoover, S. C. Hoffman, and A. Mojsilovic, "Target-specific and selective drug design for covid-19 using deep generative models," *arXiv preprint arXiv:2004.01215*, 2020.

[148] S. Kazemi Rashed, J. Frid, and S. Aits, "English dictionaries, gold and silver standard corpora for biomedical natural language processing related to SARS-CoV-2 and COVID-19," *arXiv*, pp. arXiv–2003, 2020.

[149] C. E. Lopez, M. Vasu, and C. Gallemore, "Understanding the perception of COVID-19 policies by mining a multilanguage Twitter dataset," *arXiv preprint arXiv:2003.10359*, 2020.

[150] J. E. C. Saire and R. C. Navarro, "What is the people posting about symptoms related to Coronavirus in Bogota, Colombia?" *arXiv preprint arXiv:2003.11159*, 2020.

[151] M. Cinelli, W. Quattrociocchi, A. Galeazzi, C. M. Valensise, E. Brugnoli, A. L. Schmidt, P. Zola, F. Zollo, and A. Scala, "The COVID-19 social media infodemic," *arXiv preprint arXiv:2003.05004*, 2020.

[152] L. Singh, S. Bansal, L. Bode, C. Budak, G. Chi, K. Kawintiranon, C. Padden, R. Vanarsdall, E. Vraga, and Y. Wang, "A first look at COVID-19 information and misinformation sharing on Twitter," *arXiv preprint arXiv:2003.13907*, 2020.

[153] X. Li, J. Yu, Z. Zhang, J. Ren, A. E. Peluffo, W. Zhang, Y. Zhao, K. Yan, D. Cohen, and W. Wang, "Network bioinformatics analysis provides insight into drug repurposing for COVID-2019," 2020.

[154] M. M. Hossain, "Current status of global research on novel Coronavirus disease (COVID-19) : A bibliometric analysis and knowledge mapping," *Available at SSRN 3547824*, 2020.

[155] B. P. Roquette, H. Nagano, E. C. Marujo, and A. C. Maiorano, "Prediction of admission in pediatric emergency department with deep neural networks and triage textual data," *Neural Networks*, 2020.

[156] B. W. Schuller, D. M. Schuller, K. Qian, J. Liu, H. Zheng, and X. Li, "Covid-19 and computer audition: An overview on what speech & sound analysis could contribute in the SARS-CoV-2 Corona crisis," *arXiv preprint arXiv:2003.11117*, 2020.

[157] I. Song, "Diagnosis of pneumonia from sounds collected using low cost cell phones," in *2015 International Joint Conference on Neural Networks (IJCNN)*, 2015, pp. 1–8.

[158] R. Rana, S. Latif, R. Gururajan, A. Gray, G. Mackenzie, G. Humphris, and J. Dunn, "Automated screening for distress: A perspective for the future," *European journal of cancer care*, vol. 28, no. 4, p. e13033, 2019.

[159] S. Amiriparian, S. Pugachevskiy, N. Cummins, S. Hantke, J. Pohjalainen, G. Keren, and B. Schuller, "Cast a database: Rapid targeted large-scale big data acquisition via small-world modelling of social media platforms," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2017, pp. 340–345.

[160] P. Moradshahi, H. Chatrzarrin, and R. Goubran, "Improving the performance of cough sound discriminator in reverberant environments using microphone array," in *2012 IEEE International Instrumentation and Measurement Technology Conference Proceedings*, 2012, pp. 20–23.

[161] T. Olubanjo and M. Ghovanloo, "Tracheal activity recognition based on acoustic signals," in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2014, pp. 1436–1439.

[162] A. Imran, I. Posokhova, H. N. Qureshi, U. Masood, S. Riaz, K. Ali, C. N. John, and M. Nabeel, "AI4COVID-19: AI Enabled Preliminary Diagnosis for COVID-19 from Cough Samples via an App," *arXiv preprint arXiv:2004.01275*, 2020.

[163] S. Latif, J. Qadir, S. Farooq, and M. A. Imran, "How 5g wireless (and concomitant technologies) will revolutionize healthcare?" *Future Internet*, vol. 9, no. 4, p. 93, 2017.

[164] Y. Ye, S. Hou, Y. Fan, Y. Qian, Y. Zhang, S. Sun, Q. Peng, and K. Laparo, "α-satellite: An AI-driven system and benchmark datasets for hierarchical community-level risk assessment to help combat COVID-19," *arXiv preprint arXiv:2003.12232*, 2020.

[165] "Google uses location data to show which places are complying with stay-at-home orders — and which aren't, *The Verge*, accessed on: 4-April-2020." [Online]. Available: https://www.theverge.com/2020/4/3/21206318/google-location-data-mobility-reports-covid-19-privacy

[166] H. S. Maghdid, K. Z. Ghafoor, A. S. Sadiq, K. Curran, and K. Rabie, "A novel AI-enabled framework to diagnose Coronavirus COVID 19 using smartphone embedded sensors: Design study," *arXiv preprint arXiv:2003.07434*, 2020.

[167] "*Poland: App helps police monitor home quarantine*, accessed on: 1-April-2020." [Online]. Available: https://privacyinternational.org/examples/3473/poland-app-helps-police-monitor-home-quarantine/

[168] M. N. Kamel Boulos and E. M. Geraghty, "Geographical tracking and mapping of coronavirus disease COVID-19/severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) epidemic and associated events around the world: how 21st century GIS technologies are supporting the global fight against outbreaks and epidemics," *International Journal of Health Geographics*, vol. 19, no. 8, 2020.

[169] J. Ye, "*China's QR health code*, published on 19-Feb-2020; accessed on: 6-April-2020." [Online]. Available: https://www.abacusnews.com/tech/chinas-qr-health-code-system-brings-relief-some-and-new-problems/article/3051020

[170] GovTech, "*TraceTogether - behind the scenes look at its development process*, published on 25-March-2020; accessed on: 6-April-2020."

[171] T. Cohen, "*Israelis using voluntary coronavirus monitoring app*, published on 01-April-2020; accessed on: 6-April-2020." [Online]. Available: https://tinyurl.com/reutersCoronaIsraelApps

[172] L. Kelion, "*Coronavirus: UK considers virus-tracing app to ease lockdown*, published on 31-March-2020; accessed on: 6-April-2020." [Online]. Available: https://www.bbc.com/news/technology-52095331

[173] H. Cho, D. Ippolito, and Y. W. Yu, "Contact tracing mobile apps for COVID-19: Privacy considerations and related trade-offs," *arXiv preprint arXiv:2003.11511*, 2020.

[174] R. A. Calvo, S. Deterding, and R. M. Ryan, "Health surveillance during COVID-19 pandemic," *BMJ*, vol. 369, 2020. [Online]. Available: https://www.bmj.com/content/369/bmj.m1373

[175] "D3tp documentation," 2020. [Online]. Available: https://github.com/DP-3T/documents

[176] Pan-European, "*Pan-European Privacy-Preserving Proximity Tracing*, accessed on: 6-April-2020." [Online]. Available: https://www.pepp-pt.org/

[177] J. Bell, D. Butler, C. Hicks, and J. Crowcroft, "Tracesecure: Towards privacy preserving contact tracing," 2020.

[178] A. Berke, M. Bakker, P. Vepakomma, R. Raskar, K. Larson, and A. Pentland, "Assessing disease exposure risk with location histories and protecting privacy: A cryptographic approach in response to a global pandemic," *arXiv preprint arXiv:2003.14412*, 2020.

[179] J. John, T. Kathryn, K. Pushmeet, H. Demis, and A. Team, "Computational predictions of protein structures associated with COVID-19," March 2020.

[180] A. W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Žídek, A. W. Nelson, A. Bridgland *et al.*, "Protein structure prediction using multiple deep neural networks in the 13th critical assessment of protein structure prediction (CASP13)," *Proteins: Structure, Function, and Bioinformatics*, vol. 87, no. 12, pp. 1141–1148, 2019.

[181] N. Bung, S. R. Krishnan, G. Bulusu, and A. Roy, "De novo design of new chemical entities (NCEs) for SARS-CoV-2 using artificial intelligence," 2020.

[182] H. C. Metsky, C. A. Freije, T.-S. F. Kosoko-Thoroddsen, P. C. Sabeti, and C. Myhrvold, "Crispr-based surveillance for COVID-19 using genomically-comprehensive machine learning design," *bioRxiv*, 2020.

[183] F. Hu, J. Jiang, and P. Yin, "Prediction of potential commercially inhibitors against SARS-CoV-2 by multi-task deep model," *arXiv preprint arXiv:2003.00728*, 2020.

[184] H. Zhang, K. M. Saravanan, Y. Yang, M. T. Hossain, J. Li, X. Ren, and Y. Wei, "Deep learning based drug screening for novel coronavirus 2019-ncov," 2020.

[185] B. Tang, F. He, D. Liu, M. Fang, Z. Wu, and D. Xu, "AI-aided design of novel targeted covalent inhibitors against SARS-CoV-2," *bioRxiv*, 2020.

[186] V. Boucher, "*Open and Collaborative De Novo Discovery of Antiviral Agents for COVID-19 with Deep Reinforcement Learning and OpenAI Gym*, MONTREAL.AI, accessed on: 7-April-2020." [Online]. Available: https://montrealartificialintelligence.com/covid19/

[187] "Global coalition to accelerate COVID-19 clinical research in resource-limited settings," *The Lancet*, 2020. [Online]. Available: https://doi.org/10.1016/S0140-6736(20)30798-4

[188] S. Mahmoud, G. Tyson, S. Miles, A. Taweel, T. Vanstaa, M. Luck, and B. Delaney, "Multi-agent system for recruiting patients for clinical trials," Kings College London, Tech. Rep., 2014.

[189] J. A. Anguera, J. T. Jordan, D. Castaneda, A. Gazzaley, and P. A. Areán, "Conducting a fully mobile and randomised clinical trial for depression: access, engagement and expense," *BMJ innovations*, vol. 2, no. 1, pp. 14–21, 2016.

[190] "Scopus, accessed on: 12-April-2020." [Online]. Available: https://www.scopus.com/home.uri

[191] CDC, "CDC research Repositories, accessed on: 12-April-2020." [Online]. Available: https://www.cdc.gov/library/researchguides/2019novelcoronavirus/databasesjournals.html

[192] "Nature, accessed on: 12-April-2020." [Online]. Available: https://nature.com/

[193] "Science, accessed on: 12-April-2020." [Online]. Available: https://sciencemag.org/

[194] "The Lancet, accessed on: 12-April-2020." [Online]. Available: https://thelancet.com/

[195] "BMJ, accessed on: 12-April-2020." [Online]. Available: https://bmj.com/

[196] L. Wynants, B. Van Calster, M. M. Bonten, G. S. Collins, T. P. Debray, M. De Vos, M. C. Haller, G. Heinze, K. G. Moons, R. D. Riley *et al.*, "Prediction models for diagnosis and prognosis of COVID-19 infection: systematic review and critical appraisal," *BMJ*, vol. 369, 2020.

[197] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[198] N. Fenton and M. Neil, "The use of Bayes and causal modelling in decision making, uncertainty and risk," *CEPIS Upgrade*, vol. 12, no. 5, pp. 10–21, 2011.

[199] D. Gunning, "Explainable artificial intelligence (XAI)," *Defense Advanced Research Projects Agency (DARPA), nd Web*, vol. 2, 2017.

[200] S. Latif, A. Qayyum, M. Usama, J. Qadir, A. Zwitter, and M. Shahzad, "Caveat emptor: the risks of using big data for human development," *IEEE Technology and Society Magazine*, vol. 38, no. 3, pp. 82–90, 2019.

[201] A. Qayyum, J. Qadir, M. Bilal, and A. Al-Fuqaha, "Secure and robust machine learning for healthcare: A survey," *arXiv preprint arXiv:2001.08103*, 2020.

[202] Y. N. Harari, "The world after coronavirus, financial times, accessed on: 1-April-2020." [Online]. Available: https://www.ft.com/content/19d90308-6858-11ea-a3c9-1fe6fedcca75

[203] J. L. Raisaro, J. R. Troncoso-Pastoriza, M. Misbach, J. S. Sousa, S. Pradervand, E. Missiaglia, O. Michielin, B. Ford, and J.-P. Hubaux, "Medco: Enabling secure and privacy-preserving exploration of distributed clinical and genomic data," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 16, no. 4, pp. 1328–1341, 2018.

[204] D. Froelicher, J. R. Troncoso-Pastoriza, J. S. Sousa, and J.-P. Hubaux, "Drynx: Decentralized, secure, verifiable system for statistical queries and machine learning on distributed datasets," *IEEE Transactions on Information Forensics and Security*, 2020.

[205] L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi *et al.*, "AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations," *Minds and Machines*, vol. 28, no. 4, pp. 689–707, 2018.

[206] "Ethical Implications of the Use of AI to Manage the COVID-19 Outbreak," April 2020. [Online]. Available: https://ieai.mcts.tum.de/wp-content/uploads/2020/04/April-2020-IEAI-Research-Brief_Covid-19-FINAL.pdf

[207] "*Call for Action: Toward Building the Data Infrastructure and Ecosystem We Need to Tackle Pandemics and Other Dynamic Society and Environmental Threats*. The Gov Lab, New York University," Mar 2020. [Online]. Available: http://www.thegovlab.org/static/files/publications/ACallForActionCOVID19.pdf

[208] R. Anderson, "Contact tracing in the real world," https://www.lightbluetouchpaper.org/2020/04/12/contact-tracing-in-the-real-world/, 2020.

[209] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of AI ethics guidelines," *Nature Machine Intelligence*, vol. 1, no. 9, pp. 389–399, 2019.

[210] T. Hagendorff, "The ethics of AI ethics: An evaluation of guidelines," *Minds and Machines*, pp. 1–22, 2020.

[211] N. Bostrom and E. Yudkowsky, "The ethics of artificial intelligence," *The Cambridge handbook of artificial intelligence*, vol. 1, pp. 316–334, 2014.

[212] C. Tucker, A. Agrawal, J. Gans, and A. Goldfarb, "Privacy, algorithms, and artificial intelligence," *The Economics of Artificial Intelligence: An Agenda*, pp. 423–437, 2018.

[213] R. Shokri and V. Shmatikov, "Privacy-preserving deep learning," in *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, 2015, pp. 1310–1321.

[214] E. Segal, F. Zhang, X. Lin, G. King, O. Shalem, S. Shilo, W. E. Allen, Y. H. Grad, C. S. Greene, F. Alquaddoomi *et al.*, "Building an international consortium for tracking coronavirus health status," *medRxiv*, 2020.

[215] Q.-Y. Peng, X.-T. Wang, L.-N. Zhang, C. C. C. U. S. Group *et al.*, "Findings of lung ultrasonography of novel corona virus pneumonia during the 2019–2020 epidemic," *Intensive Care Medicine*, p. 1, 2020.

[216] E. Poggiali, A. Dacrema, D. Bastoni, V. Tinelli, E. Demichele, P. Mateo Ramos, T. Marcianò, M. Silva, A. Vercelli, and A. Magnacavallo, "Can lung us help critical care clinicians in the early diagnosis of novel coronavirus (COVID-19) pneumonia?" *Radiology*, p. 200847, 2020.

[217] N. Poyiadji, G. Shahin, D. Noujaim, M. Stone, S. Patel, and B. Griffith, "COVID-19–associated acute hemorrhagic necrotizing encephalopathy: CT and MRI features," *Radiology*, p. 201187, 2020.

[218] F. Ahmed, N. Ahmed, C. Pissarides, and J. Stiglitz, "Why inequality could spread COVID-19," *The Lancet Public Health*, Apr. 2020. [Online]. Available: https://doi.org/10.1016/s2468-2667(20)30085-2

[219] J. Qadir, M. Mujeeb-U-Rahman, M. H. Rehmani, A.-S. K. Pathan, M. A. Imran, A. Hussain, R. Rana, and B. Luo, "IEEE access special section editorial: health informatics for the developing world," *IEEE Access*, vol. 5, pp. 27 818–27 823, 2017.

[220] S. Latif, R. Rana, J. Qadir, A. Ali, M. A. Imran, and M. S. Younis, "Mobile health in the developing world: Review of literature and lessons from a case study," *IEEE Access*, vol. 5, pp. 11 540–11 556, 2017.

[221] J. Quinn, V. Frias-Martinez, and L. Subramanian, "Computational sustainability and artificial intelligence in the developing world," *AI Magazine*, vol. 35, no. 3, p. 36, 2014.