Journal of the American Medical Informatics Association, 28(9), 2021, 2050–2067
doi: 10.1093/jamia/ocab098
Advance Access Publication Date: 21 June 2021
Review

Review

The application of artificial intelligence and data integration in COVID-19 studies: a scoping review

Yi Guo [1,2] Yahan Zhang,[3] Tianchen Lyu,[1,2] Mattia Prosperi [4] Fei Wang,[5] Hua Xu [6] and Jiang Bian [1,2]

[1]Department of Health Outcomes and Biomedical Informatics, College of Medicine, University of Florida, Gainesville, Florida, USA, [2]Cancer Informatics Shared Resource, University of Florida Health Cancer Center, Gainesville, Florida, USA, [3]Department of Pharmaceutical Outcomes and Policy, College of Pharmacy, University of Florida, Gainesville, Florida, USA, [4]Department of Epidemiology, College of Public Health and Health Professions & College of Medicine, University of Florida, Gainesville, Florida, USA, [5]Department of Population Health Sciences, Weill Cornell Medicine, New York, New York, USA, and [6]School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, Texas, USA

Corresponding Author: Jiang Bian, PhD, 2197 Mowry Road, Suite 122, Gainesville, FL 32610, USA (bianjiang@ufl.edu)

Received 7 January 2021; Revised 3 May 2021; Editorial Decision 5 May 2021; Accepted 6 May 2021

ABSTRACT

Objective: To summarize how artificial intelligence (AI) is being applied in COVID-19 research and determine whether these AI applications integrated heterogenous data from different sources for modeling.

Materials and Methods: We searched 2 major COVID-19 literature databases, the National Institutes of Health's LitCovid and the World Health Organization's COVID-19 database on March 9, 2021. Following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guideline, 2 reviewers independently reviewed all the articles in 2 rounds of screening.

Results: In the 794 studies included in the final qualitative analysis, we identified 7 key COVID-19 research areas in which AI was applied, including disease forecasting, medical imaging-based diagnosis and prognosis, early detection and prognosis (non-imaging), drug repurposing and early drug discovery, social media data analysis, genomic, transcriptomic, and proteomic data analysis, and other COVID-19 research topics. We also found that there was a lack of heterogenous data integration in these AI applications.

Discussion: Risk factors relevant to COVID-19 outcomes exist in heterogeneous data sources, including electronic health records, surveillance systems, sociodemographic datasets, and many more. However, most AI applications in COVID-19 research adopted a single-sourced approach that could omit important risk factors and thus lead to biased algorithms. Integrating heterogeneous data for modeling will help realize the full potential of AI algorithms, improve precision, and reduce bias.

Conclusion: There is a lack of data integration in the AI applications in COVID-19 research and a need for a multilevel AI framework that supports the analysis of heterogeneous data from different sources.

Key words: machine learning, deep learning, neural networks, natural language processing, coronavirus

INTRODUCTION

In just a few months, the 2019 novel coronavirus disease (COVID-19), severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has rapidly spread around the globe, and at the time of this writing, there are over 100 million confirmed COVID-19 cases and a few million confirmed deaths from COVID-19 worldwide.[1] The COVID-19

pandemic is now the second deadliest pandemic in over 100 years, behind only the 1918 influenza pandemic (ie, Spanish Flu).[2] While the COVID-19 pandemic is still raging, and the number of cases are growing exponentially, the scientific communities around the world have reacted promptly by directing effects and resources to research studies on the etiology, transmission, detection, treatment, and prevention and control of COVID-19. In about a year, an outstanding number of over 100 000 research articles on COVID-19-related topics have been published according to PubMed.[3]

Recent advances in artificial intelligence (AI) have provided novel methods and tools for combating global pandemics, such as COVID-19. In classic computer science textbooks, AI is broadly defined as the study of intelligent agents, machines or devices that can imitate human cognitive functions to learn the environment and take actions.[4] The learning process is often implemented through mathematical or statistical models in computer programs. Machine learning, of which deep learning is a subset, is a branch of AI that trains algorithms that allow computer programs to automatically (ie, without explicit programming) improve through data.[5] In the fields of public health and medicine, AI techniques—especially machine learning and, more recently, deep learning methods—have been widely used for disease surveillance, health risks and outcomes prediction, medical diagnostics and therapeutics, clinical decision-making, and many more.[6–8]

With surveillance tools, patient reporting systems, and clinical studies emerging quickly, large amounts of novel data have been rapidly accumulated during the COVID-19 pandemic. There is growing interest in leveraging these data to develop AI solutions for COVID-19 challenges. However, developing AI models in the era of precision health is not a trivial task. Precision health adopts a unified approach to understanding the full range of determinants of health for health promotion, prevention, diagnosis, and treatment.[9,10] The vision of precision health can only be realized through the integration and examination of a comprehensive list of determinants of health that include genetic, biological, environmental, as well as social and behavioral factors. On the other hand, these determinants of health exist in various data sources that are heterogeneous in syntax (eg, file formats), schema (eg, data models and structures), and semantics (eg, meanings or interpretations of the variables). One of the first and most important challenges in building precision health AI models is integrating relevant data that contain determinants of health from the heterogeneous sources.

In this study, we conducted a scoping review of AI applications in COVID-19 research with a focus on heterogeneous data integration. Our goal was to summarize the COVID-19 research areas in which AI is being applied, the AI models being used in these research applications, and the data sources being used to build the AI models. We were particularly interested in examining whether these AI applications integrated heterogenous data from different sources for building the models and treated missing data in the variables of interest. Although a few published reviews have summarized the applications of AI or machine learning methods in COVID-19 research,[11–15] none of them examined data integration, and many focused on a specific area of COVID-19 research (eg, medical imaging[15]). Note that we focused on the use of AI methods for data analysis and excluded other AI fields, such as robotics.

MATERIALS AND METHODS

Search strategy

We searched 2 major COVID-19 literature databases, the National Institutes of Health (NIH) LitCovid (part of PubMed)[3] and the World Health Organization (WHO) COVID-19 database[16] for articles published through March 9, 2021. LitCovid is an open-resource literature hub developed by the NIH for tracking up-to-date scientific information about COVID-19. It provides a central access to all COVID-19-related articles in PubMed.[3] The WHO COVID-19 database contains global literatures of scientific findings and knowledge on COVID-19 gathered by the WHO.[16] Both databases are updated daily with newly published articles. The following query and keywords were used to search the databases: "artificial Intelligence" or "machine learning" or "supervised learning" or "unsupervised learning" or "deep learning" or "neural networks" or "natural language processing."

Literature screening

Following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guideline,[17] we screened the articles retrieved from the databases in 2 rounds. First, we screened the titles and abstracts of the identified articles and excluded those that: (1) did not use any AI methods for data analysis, (2) were unrelated to COVID-19, (3) were reviews, editorials, opinions, letters to editor, or case reports, or (4) were not written in English. Second, we screened the full texts of the remaining articles to further exclude articles that met our exclusion criteria. Two reviewers (YZ and TL) independently reviewed all the articles in the 2 rounds of screening. Any conflicts between the 2 reviewers were reviewed and solved by a third reviewer (YG). We extracted and summarized COVID-19- and AI-related information from the retained articles.

RESULTS

Summary

We summarized our review procedure in a PRISMA flow diagram in Figure 1. We identified 1311 and 1218 studies in the LitCovid and WHO COVID-19 databases, respectively. After removing duplicated studies, we included 1338 studies in the first round of screening. In the first round of screening of titles and abstracts, 492 studies were excluded according to our exclusion criteria, while 846 studies were included in the full-text review. In the second round of screening, another 52 studies were excluded based on full-text review and eventually, 794 studies were included in the final qualitative analysis.

The AI applications covered in these 794 studies can be categorized into the following areas of COVID-19 research: Disease forecasting (n = 161), Medical imaging-based diagnosis and prognosis (n = 322); Early detection and prognosis (non-imaging) (n = 152); Drug repurposing and early drug discovery (n = 53); Social media data analysis (n = 44); Genomic, transcriptomic, and proteomic data analysis (n = 24); and Other COVID-19 research topics (survey studies, literature mining, surveillance, clinical trials, miscellaneous topics) (n = 38). We listed the full citations of all 794 studies by research area in the Supplementary Table S1. In the following sections, we summarized what and how AI techniques were applied in these areas. In particular, we determined whether the studies integrated heterogeneous data to expand the list of inputs (or predictors) for building the AI models. In line with Lenzerini 2002,[18] we defined data integration as the action of combining data that are heterogeneous in syntax, schema, and semantics and extracting predictors from these data for modeling. The total number of studies and the number of studies with data integration in each research area were summarized in Figure 2.

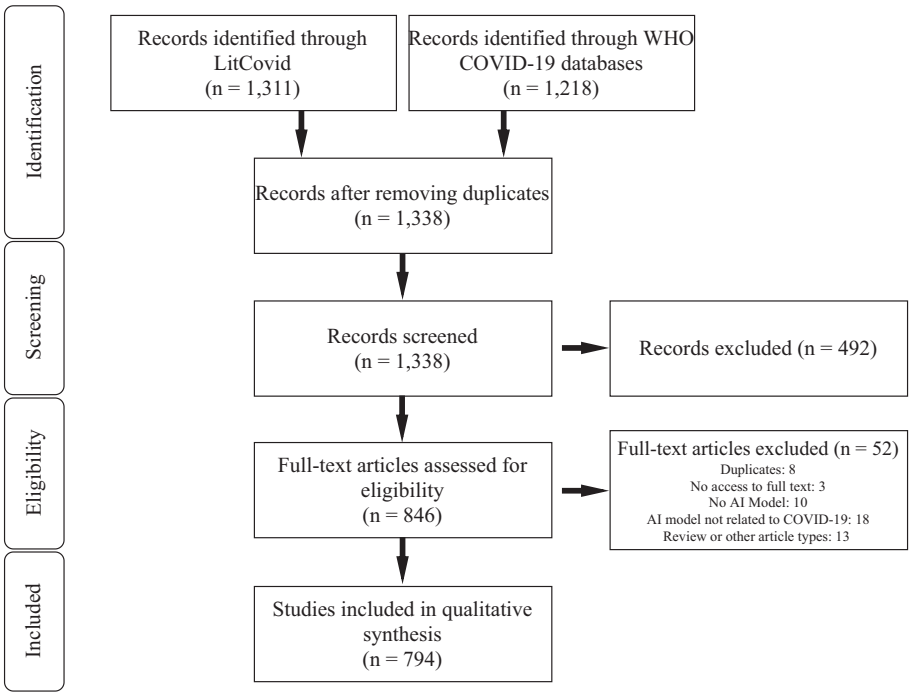

Figure 1. Search and review procedure.

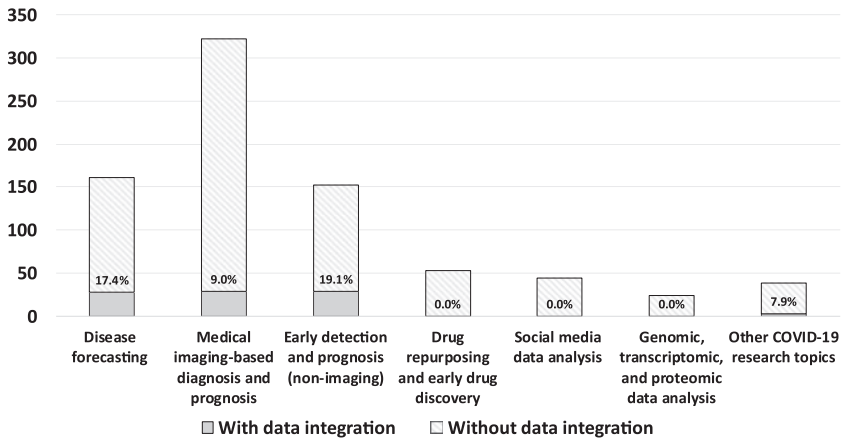

Figure 2. Number and percentage of studies with data integration in each research area.

Disease forecasting

A total of 161 studies described the use of AI for COVID-19 forecasting (Supplementary Table S1). In these studies, 106 predicted future COVID-19 incidence or mortality using historical data only, 43 predicted future or confirmed COVID-19 cases using potential risk factors as inputs, 8 characterized country-level differences in COVID-19 outcomes worldwide (clustering studies), and 4 predicted future demands for hospital resources or medical consumables.

The majority of the 106 studies on predicting future COVID-19 incidence or mortality used COVID-19 data from the Johns Hopkins University Center for Systems Science and Engineering,[19] or local health authorities. In these studies, the long short-term memory (LSTM), a class of recurrent neural networks (RNN), was the most commonly used deep learning model. Other popular models included other types of artificial neural networks (ANN); machine learning models, such as random forest, support vector machines (SVM), and gradient boosting machine (GBM); statistical time series models, such as the autoregressive integrated moving average (ARIMA) model; and epidemiological models, such as the Susceptible-Infectious-Recovered and Susceptible-Exposed-Infected-Removed models. None of the 106 studies integrated heterogeneous data for modeling since only historical COVID-19 data were used as inputs.

In the 43 studies on COVID-19 risk factors, 27 examined environmental exposures, while the remaining 16 examined a range of other risk factors, such as population characteristics, socioeconomic status, or other health-related factors. Most of these studies used machine learning models, among which random forest and GBM were the most popular algorithms. A small portion of these studies used ANN, among which the multilayer perceptron (MLP) was the most popular. Among these 43 studies, slightly over half (n = 24, 55.8%) integrated heterogeneous data on predictors for modeling (Table 1). Three of these studies imputed missing data. Two studies

Table 1. Studies on COVID-19 forecasting that integrated heterogeneous data

Study	Region	Outcome	Data source	Model	Heterogeneous data[a]	Missing data imputation
Environmental factors						
Brooks et al[20]	Worldwide	COVID-19 mortality rate	World Bank, Worldometer, Index Mundi, Wikipedia, Our World in Data, JHU, BCG Atlas, WHO, Oxford, GHS Index	k-means, linear regression	Socioeconomic, health system readiness, environmental, existing disease burden, demographics, vaccination programs, and response to the pandemic	Imputed with mean values
Cao et al[21]	China	COVID-19 incidence and growth rate	Chinese NHC, Baidu Qianxi, China Health & Family Planning Statistical Yearbook, China City Statistical Yearbook, CMA, CNIC	XGBoost	Travel-related, medical, socioeconomic, environmental, and influenza-like illness factors	No
Cazzolla-Gatti et al[22]	Italy	SARS-CoV-2 mortality and infectivity	Italian Civil Protection, ARPA, I.Stat, EpiCentro, Italian MoH, ENAC, ACI.it	RF	Environmental, health, socioeconomic factors	No
Chakraborti et al[23]	Worldwide	COVID-19 incidence and deaths	ECDC, World Bank, Google	RF, GB	Natural (climatic, environmental) and human (socioeconomic, demographic) factors	No
Gujral et al[24]	USA	COVID-19 incidence	JHU, US EPA,	EDEM	Air pollution, meteorological data, county-level demographics	No
Haghshenas et al[25]	Italy	COVID-19 incidence	Unspecified	ANN (PSO, DE)	Historical data, climate and urban factors	No
Kasilingam et al[26]	Worldwide	COVID-19 incidence	WHO, World Bank, Weather Underground	LR, DT, RF, SVM	Infrastructure, environment, policies, and infection-related factors	No
Khan et al[27]	China	COVID-19 incidence	Chinese NHC, IDIS, NBS, NCEP/NCAR	K-means, SIR	Temperature, population density, and demographic information	No
Kuo et al[28]	USA	COVID-19 incidence	NYT, USDA ERA, gridMET, Google, Federal Reserve Bank of Dallas	EN, PCR, PLSR, k-NN, RT, RF, GB, 2-layer ANN	County-level demographic, environmental, and mobility data	Imputed with median values
Li et al[29]	Worldwide	COVID-19 incidence and deaths	JHU, NOAA, KG system, CIA, Wikipedia, ESPN, CIES, Hupu, BBC, UN, WEO, World Bank, WHO, Knoema, FAO, OICA,	LASSO	Factors on politics, economy, culture, demographics, geography, education, medical resources, scientific development, environment, diseases, diet, and nutrition	No
Mollalo et al[30]	USA	COVID-19 incidence	USAFacts (CDC, JHU CSSE), US Census, GHDx	ANN (MLP)	Historical data, sociodemographic and environmental factors, disease mortality	No
Nikolopoulos et al[31]	USA, India, UK, Germany, Singapore	COVID-19 incidence and growth rate	WHO, JHU, Beihan University, Mayer Brown, WPR, WHR, World Bank, OECD, Google	52 statistical, epidemiological, machine- and deep-learning models	Climate, travel restrictions and curfews, population density, disease rates (lung, heart, diabetes), GDP spent on healthcare, air pollution, import data, Google trends	No
Pourghasemi et al[32]	Iran	COVID-19 incidence and deaths	Iranian MOHME, Open Street Map, WorldClim	RF	Historical data, anthropogenic and climatic factors	No

(continued)

Table 1. continued

Study	Region	Outcome	Data source	Model	Heterogeneous data[a]	Missing data imputation
Torrats-Espinosa[33]	USA	COVID-19 incidence and death rate	Unspecified[b]	Double-Lasso Regression	County-level demographics, density and potential for public interaction, social capital, health risk factors, capacity of the healthcare system, air pollution, employment in essential businesses, and political views	No
Zawbaa et al[34]	Italy, USA, China, Japan, Iran, Egypt, Alegria, Kenya, Cote d'Ivoire	COVID-19 incidence and death rate	JHU, ECDC	ANN (MLP)	Average age, average weather temperature, BCG vaccination, malaria treatment	No
Other factors						
Cobb et al[35]	USA	COVID-19 incidence	US local health departments, US Census	RF	SIP orders, county metrics	No
Galvan et al[36]	Brazil	COVID-19 incidence and deaths	Brazil MoH, IBGE, SUS, BCB, ADHB	ANN (SOM)	Socioeconomic, health, and safety data	No
Hasan et al[37]	Bangladesh	COVID-19 incidence	WHO, IEDCR, survey	LSTM, ANFIS, ANN (MLP)	Governing authorities, compliance, probability of infection and test positivity	No
Liu et al[38]	China	COVID-19 incidence	China CDC, Baidu Search data, Media Cloud, GLEAM	Complete linkage hierarchical clustering, LASSO	Official health reports, COVID-19-related internet search activity, news media activity, daily forecasts of COVID-19 activity	No
Mehta et al[39]	USA	COVID-19 incidence	NYT, CDC, GHDx	XGBoost	County-level population statistics, county-level disease rate and mortality	No
Pandit et al[40]	Worldwide	COVID-19 mortality rate	WHO, GSAID	LogitBoost, AdaboostM1	Age, SARS-CoV-2 clade information	No
Roy et al[41]	USA	COVID-19 incidence and deaths	WPR, Wikipedia, KFF, AHRQ, Hud Exchange, Kaggle, Worldometer, Census Bureau, CDC, NYCOpenData	SVM, SGD, NC, DTs, Gaussian NB	Social, economic, environmental, demographic, ethnic, cultural and health factors	No
Sun et al[42]	USA	COVID-19 incidence	Local DOH, CMS, LTCF, NICSHC	GB	Nursing home facility and community characteristics	Imputed using k-NN
Ye et al[43]	USA	COVID-19 risk indices	WHO, CDC, Local DOH, Census Bureau, Google Maps, Reddit	cGAN, LSTM	Disease related data, demographic, mobility and social media data	No
Region differences (clustering)						
Aydin et al[44]	Worldwide	Performances against COVID-19	Self-curated, Kaggle	k-means, hierarchic clustering	GDP, Poverty index, population, stringency index, smoking rate, CVD death rate, diabetes prevalence	Imputed with mean values

(continued)

Table 1. continued

Study	Region	Outcome	Data source	Model	Heterogeneous data[a]	Missing data imputation
Bird et al [45](p19)	Worldwide	COVID-19 risk	Worldometers, CIA, WHO	K% binning discretization, SVM, DT, GB, NB, LDA, QDA	Population, medical doctor density, tobacco use, obesity rate, GDP, land, migration, infant mortality, birth rate, death rate	No
Carrillo-Larco et al[46]	Worldwide	COVID-19 incidence	JHU, GBD, UW, GHO, WHO	k-means	Historical data, diseases, environmental factors, sociodemographics, health system factors	No
Lai et al[47]	USA	COVID-19 incidence	NYT, CDC, Census Bureau, USALEEP,	k-means	population census data, GIS data, business pattern censuses, and other sources	No

[a]Data that are heterogeneous in syntax, schema, and semantics.

[b]Available at https://doi.org/10.7910/DVN/JHFOSE.

ADHB: Human Development Atlas of Brazil; AHRQ: Agency for Healthcare Research and Quality; ANFIS: adaptive neuro fuzzy inference system; ANN: artificial neural network; ARIMA: autoregressive integrated moving average; ARPA: Regional Environmental Protection Agency; BBC: British Broadcasting Corporation; BCB: Central Bank of Brazil; BCG: Bacillus Calmette–Guérin; BGFS-PNN: Broyden-Fletcher-Goldfarb-Shanno Optimized Polynomial Neural Network; CDC: Centers for Disease Control and Prevention; cGAN: conditional generative adversarial net; CIA: Central Intelligence Agency; CIES: Centre International d'Etude du Sport (International Centre for Sports Studies); CMA: China Meteorological Administration; CMS: Centers of Medicare and Medicaid Services; CNIC: Chinese National Influenza Center; CPC-NN: Multivariate clustering based partial curve nearest neighbor; CRC: Coronavirus Resource Center; CSSE: Center for Systems Science and Engineering; CVD: cardiovascular disease; DCP: Department of Civil Protection; DE: differential evolution algorithm; DNN: deep neural network; DOH: Departments of Health; QDA:quadratic discriminant analysis; DT: decision tree; ECDC: European Centre for Disease Prevention and Control; EDEM: Ensemble-based Dynamic Emission Model; EN: Elastic net; ENAC: Ente Nazionale per l'Aviazione Civile (Italian Civil Aviation Authority); EPA: Environmental Protection Agency; ESPN: Entertainment and Sports Programming Network; FAO: Food and Agriculture Organisation of the United Nations; GB: gradient boosting; GBD: global burden of disease; GDP: gross domestic product; GHDx: Global Health Data Exchange; GHO: Global Health Observatory; GHS: Global Health Security; GIS: geographical information systems; GLEAM: global epidemic and mobility model; GSAID: global initiative on sharing all influenza data; IBGE: Brazilian Institute of Geography and Statistics; IDIS: Infectious Disease Information System of China; IEDCR: Institute of Epidemiology, Disease Control and Research; JHU: Johns Hopkins University; KFF: Kaiser Family Foundation; KG: Köppen–Geiger climate classification; k-NN: k-nearest neighbors; LDA: linear discriminant analysis; LR: logistic regression; LSTM: long short-term memory; LTCF: long-term care focus; MLP: multilayer perceptron; MoH: Ministry of Health; MOHME: Ministry of Health and Medical Education; NB: Naïve Bayes; NBS: National Bureau of Statistics of China; NC: nearest centroid; NCAR: National Center for Atmospheric Research; NCEP: National Centers for Environmental Prediction; NHC: National Health Commissions; NICSHC: National Investment Center for Seniors Housing and Care; NLP: natural language processing; NOAA: National Oceanic and Atmospheric Administration; NYT: New York Times; OECD: Organisation for Economic Co-operation and Development; OICA: Organisation Internationale des Constructeurs d'Automobiles (International Organization of Motor Vehicle Manufacturers); PC-NN: partial curve nearest neighbor; PCR: principal components regression; PLSR: partial least squares regression; PSO: particle swarm optimization algorithm; RF: random forest; RT: regression tree; SEIR: susceptible-exposed-infected-recovered model; SGD: stochastic gradient descent; SIP: shelter-in-place; SIR: susceptible-infected-recovered model; SOM: self-organizing maps; SUS: Sistema Único de Saúde (Brazil's publicly funded healthcare system); SVM: support vector machine; UN: United Nations; USALEEP: Small-Area Life Expectancy Estimates Project; USDA ERA: United States Department of Agriculture, Economic Research Service; UW: Washington University; WEO: World Economic Outlook database; WHO: World Health Organization; WHR: World Health Rankings; WPR: world population review.

used simple mean or median imputation, while the third study used the k-nearest neighbor (k-NN) method (Table 1).

All 8 clustering studies used unsupervised machine learning models, with the most popular model being the k-means. These studies aimed to group and compare countries or regions based on COVID-19 incidence, risks, and preparedness or performance. Half of the studies (n = 4, 50.0%) integrated heterogeneous data for modeling (Table 1). One of the 4 studies imputed missing data with mean values (Table 1).

The 4 studies on future demands predicted the need for intensive care unit (ICU) beds or medical consumables (eg, face masks) using data on COVID-19 cases or on consumable sales or production. All 4 studies used ANN (eg, MLP) or RNN (eg, LSTM), with some studies also building machine learning models. None of the studies integrated heterogeneous data for modeling.

Medical imaging-based diagnosis and prognosis

A total of 322 studies described the use of AI for analyzing medical imaging data for COVID-19 diagnosis and prognosis (Supplementary Table S1). All studies analyzed either computed tomography or chest X-ray data, except for 5 studies that analyzed images of lung ultrasound[48–51] or skin lesions.[52] The most common sources of medical images were local hospitals or healthcare systems and image datasets published on public domains, such as GitHub or Kaggle. In these imaging studies, roughly half used the convolutional neural network (CNN)-based models. More than 90% of these studies predicted COVID-19 outcomes using medical imaging data alone. Only 29 out of the 322 studies (9.0%) considered data from heterogenous sources for AI modeling (Table 2). In addition to imaging data, these studies considered influences from demographics (eg, age, sex, etc), clinical characteristics (eg, symptoms, lab results, disease history, etc), and other human factors (eg, exposure history) on COVID-19 outcomes. Five of these studies imputed missing data using simple mean or median imputation (Table 2).

Early detection and prognosis (nonimaging)

A total of 152 studies described the use of AI for COVID-19 early detection (n = 52) and prognosis (n = 100) (Supplementary Table S1). The vast majority of the studies on COVID-19 early detection analyzed COVID-19 positivity (+ vs −, determined by the reverse transcription polymerase chain reaction test) as the study outcome using patient data from hospitals or healthcare systems. A wide range of AI models were used for prediction, although machine learning models (eg, random forest, GBM) were used more often than deep learning models. Furthermore, most studies used a single type of data for COVID-19 detection, such as lab test data (eg, blood cell counts or inflammatory biomarkers) or clinical symptoms. Only 8 out of the 47 studies (17.0%) integrated heterogenous data for modeling (Table 3). In addition to lab and symptom data, these studies considered data on comorbidity, medications, travel/contact history, etc.

The vast majority of the studies on COVID-19 prognosis examined hospitalization, ICU admission, mechanical ventilation requirements, and/or death in COVID-19 patients using data from hospitals or healthcare systems. Traditional machine learning models were preferred over deep learning models, with the most popular model being random forest. Only 21 out of the 92 studies (22.8%) integrated heterogenous data for modeling (Table 3). These heterogenous data included demographics, clinical data (eg, lab, disease and medication history, and symptoms), genetic sequencing data, exposure history, etc.

In the early detection and prognosis studies that integrated heterogenous data (Table 3), 8 studies imputed missing data. Most studies performed simple imputation based on mean, mode, or median values, while 2 studies performed multivariate imputation by chained equations,[100,104] and 1 study imputed missing values using bagging trees.[96]

Drug repurposing and early drug discovery

A total of 53 studies described the use of AI for drug repurposing (36 studies) or early COVID-19 drug discovery (18 studies) (Supplementary Table S1). The majority of the studies focused on screening for candidate drugs in biomolecule or drug databases. Popular data sources included DrugBank (Food and Drug Administration [FDA]-approved and experimental drugs),[110] ChEMBL (bioactivity database for drug discovery),[111] PubChem (substance and compound databases),[112] ZINC (commercially available compounds for virtual screening),[113] BindingDB (experimentally determined protein-ligand binding affinities).[114] Deep learning models (eg, CNN, RNN) were used more often than the machine learning models. Furthermore, 5 out of the 36 drug repurposing studies mined the literature for repurposable drugs.[115–119] All 5 studies used NLP-based methods to mine scientific literature or other relevant data. For example, 1 study examined the description of over 1.2 million bioassays in the ChEMBL database to identify COVID-19-related bioassays.[115]

The 18 studies on early drug discovery mainly focused on screening for potential biomolecules (eg, virtual ligand screening) in ligand or compound databases (eg, ChEMBL, PubChem, ZINC, BindingDB) that could target SARS-CoV-2 functional domains. Similarly, deep learning models were preferred over the machine learning models. None of drug repurposing or early drug discovery studies integrated heterogeneous data for modeling.

Social media data analysis

A total of 44 studies described the use of AI for analyzing social media data (Supplementary Table S1). In these studies, Twitter was the single most popular data source, with 32 studies analyzing tweets from all over the world. The other 12 studies used data from Facebook, Reddit, YouTube, Weibo, etc. Most social media studies adopted a similar analytic approach: NLP methods and tools for text extraction and processing, followed by topic modeling and/or a sentiment analysis. The most common method for topic modeling was the latent Dirichlet allocation, whereas a range of machine learning models were used for sentiment analysis including SVM, Naïve Bayes, k-NN, random forest, etc. None of the social media studies integrated heterogeneous data for modeling.

Genomic, transcriptomic, and proteomic data analysis

A total of 24 studies described the use of AI for analyzing SARS-CoV-2 sequence data (eg, ribonucleic acid [RNA], small interfering RNA [siRNA], or protein sequences) (Supplementary Table S1). One common analysis goal of many of these studies was to determine the unique SARS-CoV-2 RNA or protein features that could potentially be targeted for disease detection and drug or vaccine design. Over half of these studies analyzed the SARS-CoV-2 genome sequences in the National Center for Biotechnology Information GenBank.[120] Other data sources included the Protein Data Bank,[121] National Genomics Data Center of China,[122] or self-generated sequence data. A wide variety of AI models were used in these studies,

Table 2. Studies on medical imaging-based COVID-19 detection or prognosis using heterogeneous data

Study	Region	Outcome	Data source	Model	Heterogeneous data[a]	Missing data imputation
Cai et al[53]	China	RT-PCR negativity	Single hospital	Unspecified DL, LR	CT image data, clinical data	Replaced by median
Cai et al[54]	China	Need and duration of ICU, duration of oxygen inhalation, duration of hospitalization, duration of sputum NAT-positive, clinical prognosis	Single hospital	3DQI platform, U-Net, RF	CT image data, clinical data	No
Chao et al[55]	USA, Iran, Italy	ICU admission	3 hospitals	DNN, RF	CT image data, demographics, vitals, lab data	Imputed by mean values
Chassagnon et al[56]	France	COVID-19 staging and prognosis (mechanical ventilation)	8 hospitals	CNN, DT, Linear SVM, XGBoosting, AdaBoost, Lasso	CT image data, clinical and biological markers	No
Cheng et al [57]	China	Severe vs. nonsevere COVID-19	Single hospital	CNN (uAI Discover-2019nCoV)	CT image data, clinical data	No
D'Ambrosia et al[58]	USA	RT-PCR confirmed SARS-CoV-2 infection	Single hospital	BN, SC, DML, LR	Symptoms, local SARS-CoV-2 prevalence, CXR imaging, molecular diagnostic performance	No
Ebrahimian et al[59]	USA, South Korea	Death vs. recovery, need for mechanical ventilation	Tertiary care hospitals	CNN (U-Net), LR	CXR image data, Demographics, Lab data	No
Fu et al[60]	China	Stable vs progressive COVID-19	Unspecified hospitals	SVM	CT image data, clinical and lab data	No
Grodecki et al[61]	USA, Italy	Clinical deterioration vs death	3 hospitals	CNN (U-Net), LR	CT image data, clinical data	No
Guo et al[62]	China	COVID-19 vs seasonal flu	2 hospitals	RF	CT image data, symptoms, blood tests, RT-PCR results	No
Hahm et al[63]	South Korea	Worsening oxygenation event	Single hospital	DL software (MEDIP)	CT severity score, Demographics, Comorbidity, Lab data	No
Hermans et al[64]	The Netherlands	COVID-19 positivity by RT-PCR	2 hospitals	LR	CT image data, demographics, symptoms, vitals, lab	No
Ho et al[65]	South Korea	Severe vs nonsevere COVID-19	5 hospitals	ANN, CNN, ACNN	CT image data, demographic, clinical, and lab data	No
Jeong et al[66]	South Korea	Severe vs nonsevere COVID-19	Single hospital	AI software (syngo.via Frontier)	CT severity score, demographics, symptoms, comorbidity, lab	No
Kimura-Sandoval et al[67]	Mexico	Need mechanical ventilation, death	Single hospital	AI software (Siemens healthcare)	CT variables, demographics, clinical, lab	No
Lang et al[68]	USA	Acute neuroimaging findings	Single hospital	Unspecified ML, LR	CT severity score, demographics, clinical data	No
Lassau et al[69]	French	Severe vs nonsevere COVID-19	2 hospitals	CNN (EfficientNet-B0, ResNet50, U-Net), LR	CT variables, AI-severity score (5 clinical, biological variables)	Imputed with the average
Li et al[70]	China	Severe vs nonsevere COVID-19	Single hospital	CNN (U-net), RF, GB, XGBoost, LR, SVM	CT outcomes, clinical biochemical indexes	Imputed with mean values

(continued)

Table 2. continued

Study	Region	Outcome	Data source	Model	Heterogeneous data[a]	Missing data imputation
Liu et al[71]	China	COVID-19 vs. non-COVID-19 pneumonia	Single hospital	CT image software (pyradiomics), LR, LASSO	CT outcomes, clinical data	No
Mei et al [72]	USA	COVID-19 positivity by RT-PCR	18 hospitals	CNN, SVM, RF, MLP	CT findings, clinical symptoms, exposure history, Lab	No
Meng et al[73]	China	Death within 14 days	4 hospitals	CNN, LR	CT image features, clinical information	No
Mushtaq et al[74]	Italy	Death, ICU admission	Single hospital	CNN (AI system qXR), Cox PH	CXR severity, demographics, clinical data	No
Ning et al[75]	China	Morbidity, mortality	2 hospitals	CNN, DNN, Ridge LR	CT features, 130 types of clinical features	No
Quiroz et al[76]	China	Severe vs nonsevere COVID-19	2 hospitals	CNN (U-Net), LR, XGBoost	CT features, demographics, clinical data	Imputed with mean values
Salvatore et al[77]	Italy	COVID-19 severity (discharge, hospitalization, ICU, or death)	Single hospital	AI tool (Thoracic VCAR), LR	CT parameters, clinical and lab data	No
Varble et al[78]	China, Japan	Asymptomatic vs pre-symptomatic patients with SARS-CoV-2	2 hospitals	CNN (AH-Net), LASSO LR	CT characteristics, clinical and lab data	No
Xia et al[79]	China	COVID-19 vs. influenza A/B	2 hospitals	DNN	CXR and CT features, 56 clinical features	No
Xu et al[80]	China	Healthy or COVID-19 pneumonia or non-COVID pneumonia	Single hospital	CNN, SVM, KNN, RF	CT features, 23 clinical features, 10 lab testing features	No
Xue et al[51]	China	4-level COVID-19 severity	Multiple hospitals	DSA-MIL, MA-CLR	LUC features, age, medical history, symptoms	No

[a]Data that are heterogeneous in syntax, schema, and semantics.

3DQI: 3D quantitative imaging; ACNN: artificial convolutional neural network; AI: artificial intelligence; ANN: artificial neural networks; BN: Bayesian inference network; CNN: convolutional neural network; DL: deep learning; DML: distance metric-learning; DNN: deep neural network; DSA-MIL: dual-level supervised attention-based multiple; DT: decision tree; GB: gradient boosting; ICU: intensive care unit; LR: logistic regression; LUC: lung ultrasound; MA-CLR: modality alignment contrastive learning of representation instance learning; ML: machine learning; MLP: multilayer perceptron; NAT: nucleic acid testing; RF: random forest; SC: Information-theoretic Set Cover; SVM: support vector machine.

Table 3. Studies on COVID-19 detection or prognosis using heterogeneous data

Study	Region	Outcome	Data source	Model	Heterogeneous data[a]	Missing data imputation
Early detection						
Ahamad et al[81]	China	Confirmed vs. suspected COVID-19 cases	Multiple hospitals	DT, RF, XGBoost, GB, SVM	Structured EHR data (Demographics, symptoms), Structured EHR data (Isolation treatment status, Travel history)	Imputed gender with random values based on male/female ratio; impute age with random values within IQR
Langer et al[82]	Italy	COVID-19 positivity by RT-PCR	Single hospital	ANN	Demographics, Comorbidity, Medications, Signs and Symptoms, Lab, Vitals, CXR	No
Martin et al[83]	Worldwide	COVID-19 positivity	Literature (British Medical Journal)	AI system (Symptoma)	Keywords and symptoms, Age and sex, Symptom occurrence frequency rates, Country-specific disease incidences	No
Obinata et al[84]	Japan	COVID-19 positivity by RT-PCR	2 hospitals	RF	Demographics, Vitals, Lab, Symptoms, Contact history	No
Otoom et al[85]	Worldwide	COVID-19 positivity	CORD-19 repository	SVM, ANN, NB, k-NN, decision table, decision stump, OneR, ZeroR	Symptoms, travel history to suspicious areas, contact history	No
Shimon et al[86]	Israel	COVID-19 positivity	Multiple hospitals	CNN, SVM, RF	Voice samples (acoustic features), self-reported symptoms	No
Wintjens et al[87]	The Netherlands	COVID-19 positivity by RT-PCR	Single hospital	ANN, RF, LR	Breath features (CO, NO2, VOC), clinical and demographic variables	No
Zoabi et al[88]	Israel	COVID-19 positivity by RT-PCR	The Israeli Ministry of Health	GB	Demographics, clinical symptoms, known contact with an infected individual	No
Prognosis						
Al-Najjar et al[89]	South Korea	mortality	KCDC	ANN	Demographics, infection reason and date	No
An et al[90]	South Korea	mortality	KNHIS	LASSO, SVM, RF, k-NN	Sociodemographic and medical information	No
Burian et al[91]	Germany	ICU admission	1 hospital	RF	Demographic, clinical, lab, and imaging data	Imputed with mean or mode
Cheng et al[92]	USA	ICU transfer in 24 hours	1 hospital	RF	Demographics, time-series of the admission–discharge–transfer events, clinical assessments, vital signs, lab and ECG results	Imputed with median value

(continued)

Table 3. continued

Study	Region	Outcome	Data source	Model	Heterogeneous data[a]	Missing data imputation
Das et al[93]	South Korea	mortality	KCDC	LR, SVM, k-NN, RF, GB	Demographic and exposure features	No
Ge et al[94]	China	Ventilator parameters	1 hospital	Unspecified	Demographics, clinical data, Ventilator parameters	No
Haimovich et al[95]	USA	early respiratory decompensation	8 EDs	RF, LASSO, GB, XGBoost	Demographics, medical histories, vitals, outpatient medications, chest radiograph reports, Lab	No
Hu et al[96]	China	mortality	1 hospital	LR, PLS regression, EN, RF, bagged FDA	Demographics, CT features, lab	Imputed using bagging trees
Iwendi et al[97]	Worldwide	Severity, recovery, death	Kaggle (WHO, JHU)	RF	Demographics, symptoms, travel data	No
Josephus et al[98]	Worldwide	mortality	Kaggle (WHO, JHU)	LR	Demographics, symptoms, travel data	Imputed (unspecified)
Li et al[99]	Worldwide	mortality	Github and Wolfram dataset	LR, RF, SVM	Demographics, location, symptoms, travel history, market exposure, chronic disease	No
Liang et al[100]	China	ICU admission, requiring mechanical ventilation, death, etc	Chinese NHC	CPH, ANN	Demographic, clinical, lab, and imaging data	Imputed with multivariate imputation by chained equation
Ma et al[101]	China	mortality	1 hospital	RF, XGboost	Symptoms, comorbidity, demographic, vitals, CT scans results, lab	No
Metsker et al[102]	Russia	mortality	Russian government, Single hospital	ANN	Demographics, comorbidity, lab, treatment, travel history	No
Mountantonakis et al[103]	USA	AF and mortality	13 hospitals	NLP	Demographics, medical history, lab, NLP extracted atrial fibrillation	No
Nakamichi et al[104]	USA	Hospitalization and mortality	Multiple hospitals	AdaBoost, ET, GB, RF	Demographics, comorbidity, SARS-CoV-2 sequence clades	Multiple imputation by chained equations
Neuraz et al[105]	France	in-hospital mortality	39 hospitals	NLP, Cox	Demographics, comorbidity, NLP extracted use of calcium channel blockers	No
Patel et al[106]	USA	Severity	3 hospitals	RF, ANN (MLP), SVM, GB, ET classifier, AdaBoost	Demographics, international travel, contact history, comorbidity, symptoms, blood panel profile	No

(continued)

Table 3. continued

Study	Region	Outcome	Data source	Model	Heterogeneous data[a]	Missing data imputation
Planchuelo-Gómez et al[107]	Spain	headache	1 hospital	GLM, PCA	Intensity and self-reported disability caused by headache, quality and topography of headache, migraine features, COVID-19 symptoms, lab.	No
Schwartz et al[108]	Canada	mortality	iPHIS, CORES, The COD, CCMtool, CCM	NLP, LR	Demographics, comorbidities, symptoms, NLP extracted long-term care home exposure	Imputed by weekly median value
Wu et al[109]	China, Italy, Belgium	ICU admission, death, etc	Multiple hospitals	RF, LR	Demographic, clinical, lab, and imaging data	No

[a]Data that are heterogeneous in syntax, schema, and semantics.

AF: atrial fibrillation; ANN: artificial neural networks; CCM: Public Health Case and Contact Management Solution; CCMtool: Middlesex-London COVID-19 Case and Contact Management tool; CO: carbon monoxide; COD: the Ottawa Public Health COVID-19 Ottawa Database; CORD-19: COVID-19 Open Research Dataset; CORES: Toronto Public Health Coronavirus Rapid Entry System; CPH: Cox proportional hazard; CT: computed tomography; CXR: chest x-ray; DT: decision tree; ECG: electrocardiogram; ED: emergency department; EHR: electronic health record; EN: elastic net; ET: extra trees; FDA: flexible discriminant analysis; GB: gradient boosting; GLM: generalized linear model; ICU: intensive care unit; iPHIS: integrated Public Health Information System; IQR: interquartile range; JHU: John Hopkins University; KCDC: Korea Centers for Disease Control and Prevention; KNHIS: Korean National Health Insurance Service; k-NN: k-nearest neighbors; LR: linear regression; MLP: multilayer perceptron; NB: Naïve Bayes; NHC: National Health Commission; NLP: natural language processing; NO2: nitrogen dioxide; PCA: principal component analysis; PLS: partial least squares; RBF: radial basis function; RF: random forest; SHAP: Shapley additive explanation; SVM: support vector machine; VOC: volatile organic compound; WHO: World Health Organization.

Table 4. Other COVID-19 studies using heterogeneous data

Study	Region	Outcome	Data source	Model	Heterogeneous data[a]	Missing data imputation
Literature mininng						
Reese et al[128]	N/A	Knowledge Graphs for COVID-19 Response	13 knowledge sources	Traditional or graph-based ML	Scientific literature, COVID-19 cases and mortality, Drug, Genome sequence, Diseases, Chemicals	N/A
Surveillance						
Franchini et al[129]	Italy	Individualized COVID-19 risk	Survey, medical records	RF, SVM, GBM	Demographic, Heath status, Other health and social information	No
Miscellaneous topics						
Abdalla et al[130]	USA	Social distancing	NYT, Census Bureau, USDA ERS, CDC, Google Community Mobility Reports	Elastic net	43 socio-demographic variables	No

[a]Data that are heterogeneous in syntax, schema, and semantics.

CDC: Centers for Disease Control and Prevention; GBM: gradient boosting machine; ML: machine learning; NYT: New York Times; RF: random forest; SVM: support vector machine; USDA ERA: US Department of Agriculture Economic Research Service.

including the deep learning models (CNN, RNN) and the traditional machine learning models (k-NN, SVM, random forest, GBM). None of the studies integrated heterogeneous data for modeling.

Other COVID-19 research studies

Survey studies

A total of 14 survey studies used AI models for studying COVID-19-related topics in various populations around world (Supplementary Table S1). The most common study outcomes were self-reported fear, stress, anxiety, and depression related to the pandemic. The majority of the studies used machine learning models, including random forest, XGBoost, SVM, and Naïve Bayes. Two of the studies,[123,124] which were based on the same online survey, collected text data using open-ended questions. These studies performed a sentiment analysis that involved sentiment scores calculation and clustering using the k-mean algorithm. None of the survey studies integrated heterogeneous data for modeling.

Literature mining

A total of 10 studies described the use of AI for mining COVID-19 literature (Supplementary Table S1). Literature mining studies on drug repurposing were summarized in a previous section. These 10 studies focused on summarizing topics and trends in COVID-19 research and identifying future research needs. All but 2 studies mined either PubMed or the COVID-19 Open Research Dataset.[125] Of the other 2 studies, 1 mined *ClinicalTrials.gov* to extract data on COVID-19-related trials,[126] while the other searched the Scopus database for a bibliometric analysis.[127] All of the studies involved NLP methods and tools (eg, word2vec, doc2vec). Some studies performed topic modeling and/or sentiment analysis. The only study that performed heterogeneous data integration was Reese et al (Table 4),[128] in which data from 13 heterogeneous knowledge sources (eg, scientific literature, COVID-19 cases, drug, genome sequences, chemicals, etc) were downloaded, transformed, and integrated to create the KG-COVID-19 knowledge graph.

Surveillance

A total of 6 studies described the use of AI for social distancing or syndromic surveillance (Supplementary Table S1). Three of these studies analyzed data from surveillance cameras for monitoring social distancing using well-known deep learning models for object detection,[131–133] including the single-shot detector, YOLO (you only look once), and/or the regional CNN detector. Two other studies focused on analyzing Bluetooth signal strength data with linear and logistic models for contact tracing[134] or developing NLP and deep learning-based pipeline for sentinel syndromic surveillance of COVID-19 using medical records.[135] The remaining study developed a Telegram Bot that could model individualized COVID-19 risk by integrating heterogenous data, including user responses and health/social data in medical records (Table 4).[129] This lone study involving heterogenous data used machine learning models random forest, SVM, and GBM.

Clinical trials

Two studies described the use of AI models in noninterventional clinical trials on COVID-19 patients (Supplementary Table S1). The 2 trials, namely the READY (NCT04390516) and IDENTIFY (NCT04423991),[136,137] were conducted by the same group of investigators based on the same machine learning algorithm (an XGBoost classifier) designed to predict mechanical ventilation and mortality within 24 hours upon hospital admission using inputs from clinical data. The READY trial evaluated the performance of the algorithm,[136] while the IDENTIFY trial identified a subpopulation of COVID-19 patients who had improved survival from taking hydroxychloroquine.[137] Neither study integrated heterogenous data for modeling.

Miscellaneous topics

A total of 6 studies did not fall under any of the previous research topics (Supplementary Table S1). In the lone study that integrated heterogeneous data for modeling, Abdalla et al integrated 43 sociodemographic variables from multiple sources (eg, Census Bureau, US Department of Agriculture, Centers for Disease Control and Prevention) and built elastic net models to examine how sociodemo-

graphics impacted county-level social distancing (Table 4).[130] Of the remaining studies, 1 used ANN to perform a drive-through mass vaccination simulation,[138] while the other 4 used NLP methods and tools on various research topics, including cross-lingual clinical deidentification in electronic health records (EHRs),[139] dream reports analysis,[140] drug safety analysis by mining the FDA adverse event system,[141] COVID-19 clinical concept (signs and symptoms) identification, and normalization in EHRs.[142]

DISCUSSION

As governments, research communities, and healthcare industries are actively attempting to address the COVID-19 pandemic, we are tasked to identify quick yet reliable solutions for screening, diagnosis, forecasting, surveillance, the development of vaccine or drugs, and so on. On the other hand, with large amounts of COVID-19-related data being collected in novel surveillance systems, AI methods have been widely employed in assisting medical experts and researchers in addressing COVID-19 challenges. In this article, we reviewed 1338 recent studies that applied AI methods or technologies in COVID-19 research. In the 794 studies included in our final qualitative analysis, we identified 7 key areas in which AI was applied. We also found that a wide range of machine learning and deep learning algorithms were used for modeling, although some were used more frequently than others depending on the area of research.

It is not at all surprising that AI methods have been used extensively in many areas of COVID-19 research. AI has been revolutionary for many analytics challenges in medicine and public health. For example, just shy of half of the studies we reviewed were studies of medical imaging analysis for assisting COVID-19 diagnosis. In fact, the use of AI in diagnostic medical imaging has been extensively explored for many diseases, such as cancer,[143] cardiovascular diseases,[144,145] lung diseases,[146] and brain diseases.[147] In these applications, AI has shown impressive sensitivity—similar to or better than expert interpretation—in identifying patterns and abnormalities in medical images that can aid diagnosis. Another major AI application in COVID-19 research is disease forecasting, with one-fifth of the studies we reviewed being in this category. Compared to popular statistical time series models such as the ARIMA, AI models such as the LSTM have been proven to have superior precision and accuracy when predicting time series data,[148] without making explicit assumptions (eg, stationarity) about the data. In several other areas of COVID-19 research, AI methods are the preferred data analysis tools because of their ability to handle large amounts of heterogenous data, including text data such as those in clinical narratives or on social media. For example, in drug discovery and genomic research, AI is ideal for analyzing massive amounts of sequence data (eg, proteomic or genomic data).[149,150]

One limitation of the AI applications included in our scoping review is the lack of integration of data from heterogenous sources for modeling. In the era of precision health, it is critical to examine a comprehensive list of determinants of COVID-19 outcomes, including biological, clinical, social, behavioral, and environmental factors, that exist in various heterogeneous data sources. However, most studies we reviewed used data from a single source to perform the AI-driven tasks. For instance, over 90% of the imaging studies included in this review used data from radiological images only to build AI models for COVID-19 diagnosis. This single-sourced approach ignores other important risk factors such as clinical symptoms, exposure history, lab test results, and so on, leading to algorithms with bias (eg, confounding bias)[151] and suboptimal performance. In fact, many of the medical imaging studies that integrated heterogenous data have shown that data integration led to AI models with better performance compared to models built with imaging data alone.[53–55,62,65,69,76–78] Furthermore, although some data are difficult to get due to privacy issues or simply being unavailable, there are still a range of public data on risk factors that could be easily obtained for modeling. Many studies we reviewed leveraged the "free" data sources, such as the huge amounts of environmental data from the National Oceanic and Atmospheric Administration or the socioeconomic data from the Census Bureau. Overall, integrating heterogenous but relevant data for modeling will help realize the full potential of AI algorithms, and thus improve precision and reduce bias. Our review highlights the need for a multilevel AI framework that supports the analysis of heterogenous data from difference sources.

Our scoping review has several limitations. First, our search strategy is not as comprehensive as that of a systematic review. For example, our keyword list did not include "AI." Articles that used the abbreviation "AI" without mentioning "artificial intelligence" were not included in this review. Although we do not expect a large amount of articles being omitted, we do acknowledge this limitation in keywords. Second, we searched 2 major COVID-19 literature databases rather than the traditional databases used in systematic literature reviews. Relevant articles were often indexed in these 2 COVID-19 databases with a delay of a few days up to months. Third, we did not perform a risk of bias assessment given this is a scoping review.

CONCLUSION

Huge amounts of novel data related to COVID-19 have emerged quickly during the pandemic. As a result, AI methods and technologies have been widely applied in efforts to overcome COVID-19 challenges. In this scoping review (date of literature search: March 9, 2021), we show that a broad range of AI algorithms are used for COVID-19 research, and these algorithms are primarily used in 7 major research areas. We also show that there is a lack of data integration in these AI applications and a need for a multilevel AI framework that supports the analysis of heterogenous data from difference sources.

FUNDING

Drs Guo and Bian were funded in part by the National Institutes of Health (NIH) (Award number: R01 CA246418, R21 CA245858, R21 AG068717, R21 CA253394) and Centers for Disease Control and Prevention (Award number: U18 DP006512).

AUTHOR CONTRIBUTIONS

JB and YG conceived the project. YZ and TL performed the literature search and article screening, with YG being the third reviewer. YZ and TL performed the information extraction and created the initial tables. YG drafted the manuscript. MP, FW, HX, and JB assisted in writing. All authors read and approved the manuscript.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

ACKNOWLEDGMENTS

None.

DATA AVAILABILITY

No new data were generated in support of this research.

CONFLICT OF INTEREST STATEMENT

None declared.

REFERENCES

1. World Health Organization. Coronavirus disease (COVID-19). https://www.who.int/emergencies/diseases/novel-coronavirus-2019Accessed November 12, 2020
2. Centers for Diseas Control and Prevention.1918 Pandemic (H1N1 virus) 2020. https://www.cdc.gov/flu/pandemic-resources/1918-pandemic-h1n1.htmlAccessed November 26, 2020
3. Chen Q, Allot A, Lu Z. Keep up with the latest coronavirus research. *Nature* 2020; 579 (7798): 193.
4. Russell S, Norvig P.. *Artificial Intelligence: A Modern Approach*. 3rd ed. Upper Saddle River: Pearson; 2009.
5. Mitchell TM. *Machine Learning*. 1st ed. New York: McGraw-Hill Education; 1997.
6. Jiang F, Jiang Y, Zhi H, *et al*. Artificial intelligence in healthcare: past, present and future. *Stroke Vasc Neurol* 2017; 2 (4): 230–43.
7. Hamet P, Tremblay J. Artificial intelligence in medicine. *Metabolism* 2017; 69: S36–S40.
8. Benke K, Benke G. Artificial intelligence and Big Data in public health. *Int J Environ Res Public Health* 2018; 15 (12): 2796.
9. Gambhir SS, Ge TJ, Vermesh O, Spitler R. Toward achieving precision health. *Sci Transl Med* 2018; 10 (430): eaao3612. doi:10.1126/scitranslmed.aao3612.
10. Hekler E, Tiro JA, Hunter CM, Nebeker C. Precision health: the role of the social and behavioral sciences in advancing the vision. *Ann Behav Med Publ Med* 2020; 54 (11): 805–26.
11. Bragazzi NL, Dai H, Damiani G, Behzadifar M, Martini M, Wu J. How Big Data and artificial intelligence can help better manage the COVID-19 pandemic. *Int J Environ Res Public Health* 2020; 17 (9): 3176.
12. Lalmuanawma S, Hussain J, Chhakchhuak L. Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: a review. *Chaos Solitons Fractals* 2020; 139: 110059.
13. Chen J, See KC. Artificial intelligence for COVID-19: rapid review. *J Med Internet Res* 2020; 22 (10): e21476.
14. Tayarani-N M-H. Applications of artificial intelligence in battling against Covid-19: a literature review. *Chaos Solitons Fractals* 2020; 142: 110338.doi:10.1016/j.chaos.2020.110338.
15. Albahri OS, Zaidan AA, Albahri AS, *et al*. Systematic review of artificial intelligence techniques in the detection and classification of COVID-19 medical images in terms of evaluation and benchmarking: taxonomy analysis, challenges, future solutions and methodological aspects. *J Infect Public Health* 2020; 13 (10): 1381–96.
16. World Health Organization. Global research on coronavirus disease (COVID-19). https://www.who.int/emergencies/diseases/novel-coronavirus-2019/global-research-on-novel-coronavirus-2019-ncovAccessed September 13, 2020
17. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Ann Intern Med* 2009; 151 (4): 264–9, W64. doi:10.7326/0003-4819-151-4-200908180-00135
18. Lenzerini M. Data integration: a theoretical perspective. In: *Proceedings of the Twenty-First ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*. PODS '02. Association for Computing Machinery; June 3–5, 2002; Madison, WI, USA. doi:10.1145/543613.543644
19. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect Dis* 2020; 20 (5): 533–4.
20. Brooks NA, Puri A, Garg S, *et al*. The association of Coronavirus Disease-19 mortality and prior bacille Calmette-Guerin vaccination: a robust ecological analysis using unsupervised machine learning. *Sci Rep* 2021; 11 (1): 774.
21. Cao Z, Tang F, Chen C, *et al*. Impact of systematic factors on the outbreak outcomes of the novel COVID-19 disease in China: factor analysis study. *J Med Internet Res* 2020; 22 (11): e23853.
22. Cazzolla Gatti R, Velichevskaya A, Tateo A, Amoroso N, Monaco A. Machine learning reveals that prolonged exposure to air pollution is associated with SARS-CoV-2 mortality and infectivity in Italy. *Env Pollut* 2020; 267: 115471. doi:10.1016/j.envpol.2020.115471.
23. Chakraborti S, Maiti A, Pramanik S, *et al*. Evaluating the plausible application of advanced machine learnings in exploring determinant factors of present pandemic: a case for continent specific COVID-19 analysis. *Sci Total Environ* 2020; 765: 142723. doi:10.1016/j.scitotenv.2020.142723.
24. Gujral H, Sinha A. Association between exposure to airborne pollutants and COVID-19 in Los Angeles, United States with ensemble-based dynamic emission model. *Environ Res* 2021; 194: 110704. doi:10.1016/j.envres.2020.110704.
25. Shaffiee Haghshenas S, Pirouz B, Shaffiee Haghshenas S, *et al*. Prioritizing and analyzing the role of climate and urban parameters in the confirmed cases of COVID-19 based on artificial intelligence applications. *Int J Environ Res Public Health* 2020; 17 (10): 3730. doi:10.3390/ijerph17103730.
26. Kasilingam D, Sathiya Prabhakaran SP, Rajendran DK, Rajagopal V, Santhosh Kumar T, Soundararaj A. Exploring the growth of COVID-19 cases using exponential modelling across 42 countries and predicting signs of early containment using machine learning. *Transbound Emerg Dis* 2020; 68 (3): 1001–18. doi:10.1111/tbed.13764.
27. Khan IM, Haque U, Zhang W, *et al*. COVID-19 in China: risk factors and R0 revisited. *Acta Trop* 2021; 213: 105731. doi:10.1016/j.actatropica.2020.105731.
28. Kuo C-P, Fu JS. Evaluating the impact of mobility on COVID-19 pandemic with machine learning hybrid predictions. *Sci Total Environ* 2021; 758: 144151. doi:10.1016/j.scitotenv.2020.144151.
29. Li M, Zhang Z, Cao W, *et al*. Identifying novel factors associated with COVID-19 transmission and fatality using the machine learning approach. *Sci Total Environ* 2020; 764: 142810. doi:10.1016/j.scitotenv.2020.142810.
30. Mollalo A, Rivera KM, Vahedi B. Artificial neural network modeling of novel coronavirus (COVID-19) incidence rates across the continental United States. *Int J Environ Res Public Health* 2020; 17 (12): 4204. doi:10.3390/ijerph17124204.
31. Nikolopoulos K, Punia S, Schafers A, Tsinopoulos C, Vasilakis C. Forecasting and planning during a pandemic: COVID-19 growth rates, supply chain disruptions, and governmental decisions. *Eur J Oper Res* 2021; 290 (1): 99–115. doi:10.1016/j.ejor.2020.08.001
32. Pourghasemi HR, Pouyan S, Farajzadeh Z, *et al*. Assessment of the outbreak risk, mapping and infection behavior of COVID-19: application of the autoregressive integrated-moving average (ARIMA) and polynomial models. *PLoS One* 2020; 15 (7): e0236238.
33. Torrats-Espinosa G. Using machine learning to estimate the effect of racial segregation on COVID-19 mortality in the United States. *Proc Natl Acad Sci USA* 2021; 118 (7): e2015577118. doi:10.1073/pnas.2015577118.
34. Zawbaa H, El-Gendy A, Saeed H, *et al*. A study of the possible factors affecting COVID-19 spread, severity and mortality and the effect of social distancing on these factors: machine learning forecasting model. *Int J Clin Pract* 2021; 75 (6): e14116. doi:10.1111/ijcp.14116.
35. Cobb JS, Seale MA. Examining the effect of social distancing on the compound growth rate of COVID-19 at the county level (United States) using statistical analyses and a random forest machine learning model. *Public Health* 2020; 185: 27–9.
36. Galvan D, Effting L, Cremasco H, Adam Conte-Junior C. Can socioeconomic, health, and safety data explain the spread of COVID-19 outbreak

on Brazilian federative units? *Int J Environ Res Public Health* 2020; 17 (23): 8921. doi:10.3390/ijerph17238921.
37. Hasan KT, Rahman MM, Ahmmed MM, Chowdhury AA, Islam MK. 4P model for dynamic prediction of COVID-19: a statistical and machine learning approach. *Cogn Comput* 2021; 17: 1–14. doi:10.1007/s12559-020-09786-6.
38. Liu F, Wang J, Liu J, *et al.* Predicting and analyzing the COVID-19 epidemic in China: based on SEIRD, LSTM and GWR models. *PLoS One* 2020; 15 (8): e0238280.
39. Mehta M, Julaiti J, Griffin P, Kumara S. Early stage machine learning-based prediction of US county vulnerability to the COVID-19 pandemic: machine learning approach. *JMIR Public Health Surveill* 2020; 6 (3): e19446.
40. Pandit B, Bhattacharjee S, Bhattacharjee B. Association of clade-G SARS-CoV-2 viruses and age with increased mortality rates across 57 countries and India. *Infect Genet Evol* 2021; 90: 104734. doi:10.1016/j.meegid.2021.104734.
41. Roy S, Ghosh P. Factors affecting COVID-19 infected and death rates inform lockdown-related policymaking. *PLoS One* 2020; 15 (10): e0241165.
42. Sun CLF, Zuccarelli E, Zerhouni EGA, *et al.* Predicting coronavirus disease 2019 infection risk and related risk drivers in nursing homes: a machine learning approach. *J Am Med Dir Assoc* 2020; 21 (11): 1533–8.e6. doi:10.1016/j.jamda.2020.08.030.
43. Ye Y, Hou S, Fan Y, *et al.* alpha-Satellite: An AI-driven system and benchmark datasets for dynamic COVID-19 risk assessment in the United States. *IEEE J Biomed Health Info* 2020; 24 (10): 2755–64. doi:10.1109/JBHI.2020.3009314.
44. Aydin N, Yurdakul G. Assessing countries' performances against COVID-19 via WSIDEA and machine learning algorithms. *Appl Soft Comput* 2020; 97: 106792. doi:10.1016/j.asoc.2020.106792.
45. Bird JJ, Barnes CM, Premebida C, Ekart A, Faria DR. Country-level pandemic risk and preparedness classification based on COVID-19 data: a machine learning approach. *PLoS One* 2020; 15 (10): e0241332.
46. Carrillo-Larco RM, Castillo-Cara M. Using country-level variables to classify countries according to the number of confirmed COVID-19 cases: an unsupervised machine learning approach. *Wellcome Open Res* 2020; 5: 56.
47. Lai Y, Charpignon M-L, Ebner DK, Celi LA. Unsupervised learning for county-level typological classification for COVID-19 research. *Intell Based Med* 2020; 1: 100002. doi:10.1016/j.ibmed.2020.100002
48. Arntfield R, VanBerlo B, Alaifan T, *et al.* Development of a convolutional neural network to differentiate among the etiology of similar appearing pathological B lines on lung ultrasound: a deep learning study. *BMJ Open* 2021; 11 (3): e045120.
49. Muhammad G, Shamim Hossain M. COVID-19 and non-COVID-19 classification using multi-layers fusion from lung ultrasound images. *Info Fusion* 2021; 72 :80–8. doi:10.1016/j.inffus.2021.02.013.
50. Roy S, Menapace W, Oei S, *et al.* Deep learning for classification and localization of COVID-19 markers in point-of-care lung ultrasound. *IEEE Trans Med Imaging* 2020; 39 (8): 2676–87.
51. Xue W, Cao C, Liu J, *et al.* Modality alignment contrastive learning for severity assessment of COVID-19 from lung ultrasound and clinical information. *Med Image Anal* 2021; 69: 101975. doi:10.1016/j.media.2021.101975
52. Mathur J, Chouhan V, Pangti R, Kumar S, Gupta S. A convolutional neural network architecture for the recognition of cutaneous manifestations of COVID-19. *Dermatol Ther* 2021; 34 (2): e14902. doi:10.1111/dth.14902
53. Cai Q, Du S-Y, Gao S, *et al.* A model based on CT radiomic features for predicting RT-PCR becoming negative in coronavirus disease 2019 (COVID-19) patients. *BMC Med Imaging* 2020; 20 (1): 118. doi:10.1186/s12880-020-00521-z.
54. Cai W, Liu T, Xue X, *et al.* CT quantification and machine-learning models for assessment of disease severity and prognosis of COVID-19 patients. *Acad Radiol* 2020; 27 (12): 1665–78. doi:10.1016/j.acra.2020.09.004.
55. Chao H, Fang X, Zhang J, *et al.* Integrative analysis for COVID-19 patient outcome prediction. *Med Image Anal* 2021; 67: 101844. doi:10.1016/j.media.2020.101844.
56. Chassagnon G, Vakalopoulou M, Battistella E, *et al.* AI-driven quantification, staging and outcome prediction of COVID-19 pneumonia. *Med Image Anal* 2021; 67: 101860. doi:10.1016/j.media.2020.101860.
57. Cheng Z, Qin L, Cao Q, *et al.* Quantitative computed tomography of the coronavirus disease 2019 (COVID-19) pneumonia. *Radiol Infect Dis* 2020; 7 (2): 55–61. doi:10.1016/j.jrid.2020.04.004.
58. D'Ambrosia C, Christensen H, Aronoff-Spencer E. Computing SARS-CoV-2 infection risk from symptoms, imaging, and test data: diagnostic model development. *J Med Internet Res* 2020; 22 (12): e24478.
59. Ebrahimian S, Homayounieh F, Rockenbach MABC, *et al.* Artificial intelligence matches subjective severity assessment of pneumonia for prediction of patient outcome and need for mechanical ventilation: a cohort study. *Sci Rep* 2021; 11 (1): 858. doi:10.1038/s41598-020-79470-0.
60. Fu L, Li Y, Cheng A, Pang P, Shu Z. A novel machine learning-derived radiomic signature of the whole lung differentiates stable from progressive COVID-19 infection: a retrospective cohort study. *J Thorac Imaging* 2020; 35 (6): 361–8. doi:10.1097/RTI.0000000000000544.
61. Grodecki K, Lin A, Razipour A, *et al.* Epicardial adipose tissue is associated with extent of pneumonia and adverse outcomes in patients with COVID-19. *Metabolism* 2021; 115: 154436.
62. Guo X, Li Y, Li H, *et al.* An improved multivariate model that distinguishes COVID-19 from seasonal flu and other respiratory diseases. *Aging* 2020; 12 (20): 19938–44.
63. Hahm CR, Lee YK, Oh DH, *et al.* Factors associated with worsening oxygenation in patient with nonsevere COVID-19 pneumonia. *Tuberc Respir Seoul* 2021; 84 (2): 115–24. doi:10.4046/trd.2020.0139.
64. Hermans JJR, Groen J, Zwets E, *et al.* Chest CT for triage during COVID-19 on the emergency department: myth or truth? *Emerg Radiol* 2020; 27 (6): 641–51.
65. Ho TT, Park J, Kim T, *et al.* Deep learning models for predicting severe progression in COVID-19-infected Patients. *JMIR Med Info* 2021; 9 (1): e24973. doi:10.2196/24973.
66. Jeong YJ, Nam BD, Yoo JY, *et al.* Prognostic implications of CT feature analysis in patients with COVID-19: a nationwide cohort study. *J Korean Med Sci* 2021; 36 (8): e51. doi:10.3346/jkms.2021.36.e51.
67. Kimura-Sandoval Y, Arevalo-Molina ME, Cristancho-Rojas CN, *et al.* Validation of chest computed tomography artificial intelligence to determine the requirement for mechanical ventilation and risk of mortality in hospitalized coronavirus disease-19 Patients in a tertiary care center in Mexico City. *Rev Invest Clin* 2021; 73 (2): 111–9. doi:10.24875/RIC.20000451.
68. Lang M, Li MD, Jiang KZ, *et al.* Severity of chest imaging is correlated with risk of acute neuroimaging findings among patients with COVID-19. *AJNR Am J Neuroradiol* 2021; 42 (5): 831–7. doi:10.3174/ajnr.A7032.
69. Lassau N, Ammari S, Chouzenoux E, *et al.* Integrating deep learning CT-scan model, biological and clinical variables to predict severity of COVID-19 patients. *Nat Commun* 2021; 12 (1): 634. doi:10.1038/s41467-020-20657-4.
70. Li D, Zhang Q, Tan Y, *et al.* Prediction of COVID-19 severity using chest computed tomography and laboratory measurements: evaluation using a machine learning approach. *JMIR Med Inform* 2020; 8 (11): e21604. doi:10.2196/21604.
71. Liu H, Ren H, Wu Z, *et al.* CT radiomics facilitates more accurate diagnosis of COVID-19 pneumonia: compared with CO-RADS. *J Transl Med* 2021; 19 (1): 29. doi:10.1186/s12967-020-02692-3.
72. Mei X, Lee H-C, Diao K-Y, *et al.* Artificial intelligence-enabled rapid diagnosis of patients with COVID-19. *Nat Med* 2020; 26 (8): 1224–8.
73. Meng L, Dong D, Li L, *et al.* A Deep learning prognosis model help alert for COVID-19 patients at high-risk of death: a multi-center study. *IEEE J Biomed Health Inform* 2020; 24 (12): 3576–84. doi:10.1109/JBHI.2020.3034296.

74. Mushtaq J, Pennella R, Lavalle S, *et al.* Initial chest radiographs and artificial intelligence (AI) predict clinical outcomes in COVID-19 patients: analysis of 697 Italian patients. *Eur Radiol* 2021; 31 (3): 1770–9.
75. Ning W, Lei S, Yang J, *et al.* Open resource of clinical data from patients with pneumonia for the prediction of COVID-19 outcomes via deep learning. *Nat Biomed Eng* 2020; 4 (12): 1197–207.
76. Quiroz JC, Feng Y-Z, Cheng Z-Y, *et al.* Automated severity assessment of COVID-19 based on clinical and imaging data: algorithm development and validation. *JMIR Med Inform* 2021; 9 (2): e24572. doi:10.2196/24572.
77. Salvatore C, Roberta F, Angela de L, *et al.* Clinical and laboratory data, radiological structured report findings and quantitative evaluation of lung involvement on baseline chest CT in COVID-19 patients to predict prognosis. *Radiol Med* 2021; 126 (1): 29–39.
78. Varble N, Blain M, Kassin M, *et al.* CT and clinical assessment in asymptomatic and pre-symptomatic patients with early SARS-CoV-2 in outbreak settings. *Eur Radiol* 2021; 31 (5): 3165–76.
79. Xia Y, Chen W, Ren H, *et al.* A rapid screening classifier for diagnosing COVID-19. *Int J Biol Sci* 2021; 17 (2): 539–48. doi:10.7150/ijbs.53982.
80. Xu M, Ouyang L, Han L, *et al.* Accurately differentiating between patients with COVID-19, patients with other viral infections, and healthy individuals: multimodal late fusion learning approach. *J Med Internet Res* 2021; 23 (1): e25535.
81. Ahamad MM, Aktar S, Rashed-Al-Mahfuz M, *et al.* A machine learning model to identify early stage symptoms of SARS-Cov-2 infected patients. *Expert Syst Appl* 2020; 160: 113661. doi:10.1016/j.eswa.2020.113661.
82. Langer T, Favarato M, Giudici R, *et al.* Development of machine learning models to predict RT-PCR results for severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) in patients with influenza-like symptoms using only basic clinical data. *Scand J Trauma Resusc Emerg Med* 2020; 28 (1): 113. doi:10.1186/s13049-020-00808-8.
83. Martin A, Nateqi J, Gruarin S, *et al.* An artificial intelligence-based first-line defence against COVID-19: digitally screening citizens for risks via a chatbot. *Sci Rep* 2020; 10 (1): 19012. doi:10.1038/s41598-020-75912-x.
84. Obinata H, Yokobori S, Ogawa K, *et al.* Indicators of acute kidney injury as biomarkers to differentiate heatstroke from coronavirus disease 2019: a retrospective multicenter analysis. *J Nippon Med Sch* 2021; 88 (1): 80–6.
85. Otoom M, Otoum N, Alzubaidi MA, Etoom Y, Banihani R. An IoT-based framework for early identification and monitoring of COVID-19 cases. *Biomed Signal Process Control* 2020; 62: 102149. doi:10.1016/j.bspc.2020.102149.
86. Shimon C, Shafat G, Dangoor I, Ben-Shitrit A. Artificial intelligence enabled preliminary diagnosis for COVID-19 from voice cues and questionnaires. *J Acoust Soc Am* 2021; 149 (2): 1120. doi:10.1121/10.0003434.
87. Wintjens AGWE, Hintzen KFH, Engelen SME, *et al.* Applying the electronic nose for pre-operative SARS-CoV-2 screening. *Surg Endosc* 2020; 1–8. doi:10.1007/s00464-020-08169-0.
88. Zoabi Y, Deri-Rozov S, Shomron N. Machine learning-based prediction of COVID-19 diagnosis based on symptoms. *NPJ Digit Med* 2021; 4 (1): 3. doi:10.1038/s41746-020-00372-6.
89. Al-Najjar H, Al-Rousan N. A classifier prediction model to predict the status of Coronavirus COVID-19 patients in South Korea. *Eur Rev Med Pharmacol Sci* 2020; 24 (6): 3400–3. doi:10.26355/eurrev_202003_20709.
90. An C, Lim H, Kim D-W, Chang JH, Choi YJ, Kim SW. Machine learning prediction for mortality of patients diagnosed with COVID-19: a nationwide Korean cohort study. *Sci Rep* 2020; 10 (1): 18716. doi:10.1038/s41598-020-75767-2.
91. Burian E, Jungmann F, Kaissis GA, *et al.* Intensive care risk estimation in COVID-19 pneumonia based on clinical and imaging parameters: experiences from the Munich cohort. *J Clin Med* 2020; 9 (5): 1514. doi:10.3390/jcm9051514.
92. Cheng F-Y, Joshi H, Tandon P, *et al.* Using machine learning to predict ICU transfer in hospitalized COVID-19 patients. *J Clin Med* 2020; 9 (6): 1668. doi:10.3390/jcm9061668.
93. Das AK, Mishra S, Saraswathy Gopalan S. Predicting CoVID-19 community mortality risk using machine learning and development of an online prognostic tool. *PeerJ* 2020; 8: e10083.
94. Ge F, Zhang D, Wu L, Mu H. Predicting psychological state among Chinese undergraduate students in the COVID-19 epidemic: a longitudinal study using a machine learning. *Neuropsychiatr Treat* 2020; 16: 2111–8. doi:10.2147/NDT.S262004.
95. Haimovich AD, Ravindra NG, Stoytchev S, *et al.* Development and validation of the quick COVID-19 severity index: a prognostic tool for early clinical decompensation. *Ann Emerg Med* 2020 76 (4): 442–53. doi:10.1016/j.annemergmed.2020.07.022.
96. Hu C, Liu Z, Jiang Y, *et al.* Early prediction of mortality risk among patients with severe COVID-19, using machine learning. *Int J Epidemiol* 2020; 49 (6): 1918–29. doi:10.1093/ije/dyaa171.
97. Iwendi C, Bashir AK, Peshkar A, *et al.* COVID-19 patient health prediction using boosted random forest algorithm. *Front Public Health* 2020; 8: 357. doi:10.3389/fpubh.2020.00357.
98. Josephus BO, Nawir AH, Wijaya E, Moniaga JV, Ohyver M. Predict mortality in patients infected with COVID-19 virus based on observed characteristics of the patient using logistic regression. *Procedia Comput Sci* 2021; 179: 871–7. doi:10.1016/j.procs.2021.01.076.
99. Li Y, Horowitz MA, Liu J, *et al.* Individual-level fatality prediction of COVID-19 patients using AI methods. *Front Public Health* 2020; 8: 587937. doi:10.3389/fpubh.2020.587937.
100. Liang W, Yao J, Chen A, *et al.* Early triage of critically ill COVID-19 patients using deep learning. *Nat Commun* 2020; 11 (1): 3543. doi:10.1038/s41467-020-17280-8.
101. Ma X, Ng M, Xu S, *et al.* Development and validation of prognosis model of mortality risk in patients with COVID-19. *Epidemiol Infect* 2020; 148: e168. doi:10.1017/S0950268820001727.
102. Metsker O, Kopanitsa G, Yakovlev A, Veronika K, Zvartau N. Survival analysis of COVID-19 patients in Russia using machine learning. *Stud Health Technol Inform* 2020; 273: 223–7. doi:10.3233/SHTI200644.
103. Mountantonakis SE, Saleh M, Fishbein J, *et al.* Atrial fibrillation is an independent predictor for in-hospital mortality in patients admitted with SARS-CoV-2 infection. *Heart Rhythm* 2021; 18 (4): 501–7.
104. Nakamichi K, Shen JZ, Lee CS, *et al.* Hospitalization and mortality associated with SARS-CoV-2 viral clades in COVID-19. *Sci Rep* 2021; 11 (1): 4802. doi:10.1038/s41598-021-82850-9.
105. Neuraz A, Lerner I, Digan W, *et al.*; AP-HP/Universities/INSERM COVID-19 Research Collaboration; AP-HP COVID CDR Initiative. Natural language processing for rapid response to emergent diseases: case study of calcium channel blockers and hypertension in the COVID-19 pandemic. *J Med Internet Res* 2020; 22 (8): e20773.
106. Patel D, Kher V, Desai B, *et al.* Machine learning based predictors for COVID-19 disease severity. *Sci Rep* 2021; 11 (1): 4673. doi:10.1038/s41598-021-83967-7.
107. Planchuelo-Gomez A, Trigo J, de Luis-Garcia R, Guerrero AL, Porta-Etessam J, Garcia-Azorin D. Deep phenotyping of headache in hospitalized COVID-19 patients via principal component analysis. *Front Neurol* 2020; 11: 583870. doi:10.3389/fneur.2020.583870.
108. Schwartz KL, Achonu C, Buchan SA, *et al.* Epidemiology, clinical characteristics, household transmission, and lethality of severe acute respiratory syndrome coronavirus-2 infection among healthcare workers in Ontario, Canada. *PLoS One* 2020; 15 (12): e0244477.
109. Wu G, Zhou S, Wang Y, *et al.* A prediction model of outcome of SARS-CoV-2 pneumonia based on laboratory findings. *Sci Rep* 2020; 10 (1): 14042. doi:10.1038/s41598-020-71114-7
110. Wishart DS, Knox C, Guo AC, *et al.* DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res* 2008; 36 (Database issue): D901–906.
111. Gaulton A, Bellis LJ, Bento AP, *et al.* ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res* 2012; 40 (Database issue): D1100–D1107.
112. Kim S, Thiessen PA, Bolton EE, *et al.* PubChem substance and compound databases. *Nucleic Acids Res* 2016; 44 (D1): D1202–D1213.

113. Irwin JJ, Shoichet BK. ZINC – a free database of commercially available compounds for virtual screening. *J Chem Inf Model* 2005; 45 (1): 177–82.
114. Liu T, Lin Y, Wen X, Jorissen RN, Gilson MK. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res* 2007; 35 (Database issue): D198–201.
115. Duran-Frigola M, Bertoni M, Blanco R, *et al.* Bioactivity profile similarities to expand the repertoire of COVID-19 Drugs. *J Chem Inf Model* 2020; 60 (12): 5730–4. doi:10.1021/acs.jcim.0c00420.
116. Gates LE, Hamed AA. The anatomy of the SARS-CoV-2 biomedical literature: introducing the CovidX network algorithm for drug repurposing recommendation. *J Med Internet Res* 2020; 22 (8): e21169.
117. Khan JY, Khondaker MTI, Hoque IT, *et al.* Toward preparing a knowledge base to explore potential drugs and biomedical entities related to COVID-19: automated computational approach. *JMIR Med Inform* 2020; 8 (11): e21648. doi:10.2196/21648.
118. Zeng X, Song X, Ma T, *et al.* Repurpose open data to discover therapeutics for COVID-19 using deep learning. *J Proteome Res* 2020; 19 (11): 4624–36. doi:10.1021/acs.jproteome.0c00316.
119. Zhang R, Hristovski D, Schutte D, Kastrin A, Fiszman M, Kilicoglu H. Drug repurposing for COVID-19 via knowledge graph completion. *J Biomed Inform* 2021; 115: 103696. doi:10.1016/j.jbi.2021.103696.
120. GenBank Overview. https://www.ncbi.nlm.nih.gov/genbank/Accessed April 6, 2021
121. Berman HM, Westbrook J, Feng Z, *et al.* The protein data bank. *Nucleic Acids Res* 2000; 28 (1): 235–42.
122. CNCB-NGDC Members and Partners. Database Resources of the National Genomics Data Center, China National Center for Bioinformation in 2021. *Nucleic Acids Res* 2021; 49 (D1): D18–D28.
123. Flint SW, Brown A, Tahrani AA, Piotrkowicz A, Joseph A-C. Cross-sectional analysis to explore the awareness, attitudes and actions of UK adults at high risk of severe illness from COVID-19. *BMJ Open* 2020; 10 (12): e045309.
124. Flint SW, Piotrkowicz A, Watts K. Use of artificial intelligence to understand adults' thoughts and behaviours relating to COVID-19. *Perspect Public Health* 2021; 1757913920979332. doi:10.1177/1757913920979332.
125. Lu Wang L, Lo K, Chandrasekhar Y, *et al.* CORD-19: The Covid-19 open research dataset. *ArXiv* 2020; arXiv: 2004.10706v2.
126. Alag S. Analysis of COVID-19 clinical trials: a data-driven, ontology-based, and natural language processing approach. *PLoS One* 2020; 15 (9): e0239694.
127. De Felice F, Polimeni A. Coronavirus disease (COVID-19): a machine learning bibliometric analysis. *In Vivo* 2020; 34 (3 suppl): 1613–7.
128. Reese JT, Unni D, Callahan TJ, *et al.* KG-COVID-19: a framework to produce customized knowledge graphs for COVID-19 response. *Patterns (N Y)* 2021; 2 (1): 100155.
129. Franchini M, Pieroni S, Martini N, *et al.* Shifting the paradigm: The Dress-COV telegram bot as a tool for participatory medicine. *Int J Environ Res Public Health* 2020; 17 (23): 8786. doi:10.3390/ijerph17238786.
130. Abdalla M, Abar A, Beiter ER, Saad M. Asynchrony between individual and government actions accounts for disproportionate impact of COVID-19 on vulnerable communities. *Am J Prev Med* 2020; 60 (3): 318–26. doi:10.1016/j.amepre.2020.10.012.
131. Saponara S, Elhanashi A, Gagliardi A. Implementing a real-time, AI-based, people detection and social distancing measuring system for Covid-19. *J Real Time Image Process* 2021; 1–11. doi:10.1007/s11554-021-01070-6.
132. Shorfuzzaman M, Hossain MS, Alhamid MF. Towards the sustainable development of smart cities through mass video surveillance: a response to the COVID-19 pandemic. *Sustain Cities Soc* 2021; 64: 102582. doi:10.1016/j.scs.2020.102582.
133. Szczepanek R. Analysis of pedestrian activity before and during COVID-19 lockdown, using webcam time-lapse from Cracow and machine learning. *PeerJ* 2020; 8: e10132.
134. Sattler F, Ma J, Wagner P, *et al.* Risk estimation of SARS-CoV-2 transmission from bluetooth low energy measurements. *NPJ Digit Med* 2020; 3: 129. doi:10.1038/s41746-020-00340-0.
135. Wen A, Wang L, He H, *et al.* An aberration detection-based approach for sentinel syndromic surveillance of COVID-19 and other novel influenza-like illnesses. *J Biomed Inform* 2020; 113: 103660. doi:10.1016/j.jbi.2020.103660.
136. Burdick H, Lam C, Mataraso S, *et al.* Prediction of respiratory decompensation in Covid-19 patients using machine learning: The READY trial. *Comput Biol Med* 2020; 124: 103949. doi:10.1016/j.compbiomed.2020.103949.
137. Burdick H, Lam C, Mataraso S, *et al.* Is Machine learning a better way to identify COVID-19 patients who might benefit from hydroxychloroquine treatment?-The IDENTIFY Trial. *J Clin Med* 2020; 9 (12): 3834. doi:10.3390/jcm9123834.
138. Asgary A, Valtchev SZ, Chen M, Najafabadi MM, Wu J. Artificial intelligence model of drive-through vaccination simulation. *Int J Environ Res Public Health* 2020; 18 (1): 268. doi:10.3390/ijerph18010268
139. Catelli R, Gargiulo F, Casola V, De Pietro G, Fujita H, Esposito M. Crosslingual named entity recognition for clinical de-identification applied to a COVID-19 Italian data set. *Appl Soft Comput* 2020; 97: 106779. doi:10.1016/j.asoc.2020.106779.
140. Mota NB, Weissheimer J, Ribeiro M, *et al.* Dreaming during the Covid-19 pandemic: computational assessment of dream reports reveals mental suffering related to fear of contagion. *PLoS One* 2020; 15 (11): e0242903.
141. Shan W, Hong D, Zhu J, Zhao Q. Assessment of the potential adverse events related to ribavirin-interferon combination for novel coronavirus therapy. *Comput Math Methods Med* 2020; 2020: 1391583. doi:10.1155/2020/1391583
142. Wang J, Abu-El-Rub N, Gray J, *et al.* COVID-19 SignSym: a fast adaptation of a general clinical NLP tool to identify and normalize COVID-19 signs and symptoms to OMOP common data model. *J Am Med Inf Assoc* 2021; ocab015. doi:10.1093/jamia/ocab015.
143. Dana J, Agnus V, Ouhmich F, Gallix B. Multimodality imaging and artificial intelligence for tumor characterization: current status and future perspective. *Semin Nucl Med* 2020; 50 (6): 541–8.
144. Slomka PJ, Miller RJ, Isgum I, Dey D. Application and translation of artificial intelligence to cardiovascular imaging in nuclear medicine and noncontrast CT. *Semin Nucl Med* 2020; 50 (4): 357–66.
145. Xu B, Kocyigit D, Grimm R, Griffin BP, Cheng F. Applications of artificial intelligence in multimodality cardiovascular imaging: a state-of-the-art review. *Prog Cardiovasc Dis* 2020; 63 (3): 367–76.
146. Ma J, Song Y, Tian X, Hua Y, Zhang R, Wu J. Survey on deep learning for pulmonary medical imaging. *Front Med* 2020; 14 (4): 450–69.
147. Zhu G, Jiang B, Tong L, Xie Y, Zaharchuk G, Wintermark M. Applications of deep learning to neuro-imaging techniques. *Front Neurol* 2019; 10: 869.
148. Siami-Namini S, Tavakoli N, Namin AS. A Comparison of ARIMA and LSTM in Forecasting Time Series. In: proceedings of the 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA); December 17–20, 2018; Orlando, FL, USA. doi:10.1109/ICMLA.2018.00227
149. Swan AL, Mobasheri A, Allaway D, Liddell S, Bacardit J. Application of machine learning to proteomics data: classification and biomarker identification in postgenomics biology. *OMICS* 2013; 17 (12): 595–610.
150. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. *Nat Rev Genet* 2015; 16 (6): 321–32.
151. Sackett DL. Bias in analytic research. *J Chronic Dis* 1979; 32 (1-2): 51–63.