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 Bian1,2

1Department of Health Outcomes and Biomedical Informatics, College of Medicine, University of Florida, Gainesville, Florida, USA, 2Cancer Informatics Shared Resource, University of Florida Health Cancer Center, Gainesville, Florida, USA, 3Department of Pharmaceutical Outcomes and Policy, College of Pharmacy, University of Florida, Gainesville, Florida, USA, 4Department of Epidemiology, College of Public Health and Health Professions & College of Medicine, University of Florida, Gainesville, Florida, USA, 5Department of Population Health Sciences, Weill Cornell Medicine, New York, New York, USA, and 6School 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 heterogenous 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). 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.3 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.3 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 heterogeneous 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) and the World Health Organization (WHO) COVID-19 database 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.
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, 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 for modeling (Table 1). Three of these studies imputed missing data. Two studies...
| Study | Data source | Model | Heterogeneous data | Missing data imputation |
|-------|-------------|-------|-------------------|------------------------|
| Table 1. Studies on COVID-19 forecasting that integrated heterogeneous data | | | | |
| Brooks et al. | World Bank, Worldometer, Index Mundi, Wikipedia, Our World in Data, JHU, BCG Atlas, WHO, Oxford, GHC Index, China Health & Family Planning Statistical Yearbook, China City Yearbook, China National Network, China National Malaria, China National Health, China National Institute of Environmental Health, China National Statistical Bureau, China National Meteorological Bureau, China National Health, China National Science & Technology Planning, China National Cancer Institute | k-means, linear regression | Socioeconomic, health system readiness, environment, emerging disease response, the pandemic, and disease spread | Imputed with mean values |
| Cao et al. | World Bank, Worldometer, Index Mundi, Wikipedia, Our World in Data, JHU, BCG Atlas, WHO, Oxford, GHC Index, China Health & Family Planning Statistical Yearbook, China City Yearbook, China National Network, China National Malaria, China National Health, China National Science & Technology Planning, China National Cancer Institute | XGBoost | Environmental, health, socioeconomic factors | No |
| Cazzolla-Gatti et al. | Italian Civil Protection, ARPA, I.Stat, EpiCentro, Italian MoH, ENAC, ACI.it | RF | Environmental, health, socioeconomic factors | No |
| Chakraborti et al. | ECDC, World Bank, Google RF, GB Natural (climatic, environmental) and human (socioeconomic, demographic) factors | RF, GB | Infrastructural, environmental, population density, and infection-related factors | No |
| Gujral et al. | JHU, US EPA, EDEM Air pollution, meteorological data, county-level demographics | | | No |
| Haghshenas et al. | Unspecified | ANN (PSO, DE) | Historical data, climate and urban factors | No |
| Kasilingam et al. | WHO, World Bank, Weather Underground, China National Health, China National Science & Technology Planning, China National Cancer Institute, China National Institute of Environmental Health, China National Health, China National Science & Technology Planning, China National Cancer Institute | LR, DT, RF, SVM | Infrastructure, environment, policies, and infection-related factors | No |
| Khan et al. | Chinese NHC, IDIS, NBS, NCEP/NCAR | K-means, SIR | Temperature, population density, and demographic information | No |
| Kuo et al. | NYT, USDA ERA, gridMET, Google, Federal Reserve Bank of Dallas | EN, PCR, LR, k-NN, RF, GB, 2-layer ANN | County-level demographic, environmental and mobility data | Imputed with median values |
| Li et al. | JHU, NOAA, AG system, CIA, Wikipedia, ESPON, CES, Hub, BRC, UN, WHO, WEO, World Bank, WHO, OECD | ANN (ISO, DE) | Natural climate, environmental, demographic factors | No |
| Mollalo et al. | USAFacts (CDC, JHU CSSE), US Census, GHDx | ANN (MLP) | Historical data, socioeconomic factors, disease mortality | No |
| Nikolopoulou et al. | WHO, JHU, Behan University, University of Texas, Major Brown, WHO, World Bank, OECD, Google | RF | Climate, travel restrictions and curfew, population density, disease rates (lung, heart, diabetes), GDP spent on healthcare, air pollution, import, trade, and financial data | No |
| Pourghasemi et al. | Iran | | Climate, sociocultural and demographic factors | No |
| Study               | Region                                | Outcome                                      | Data source                                      | Model                        | Heterogeneous dataa                                      | Missing data imputation |
|--------------------|---------------------------------------|----------------------------------------------|-------------------------------------------------|------------------------------|---------------------------------------------------------|-------------------------|
| Torrats-Espinosa   | USA                                   | COVID-19 incidence and death rate            | Unspecifiedb                                    | Double-Lasso Regression     | County-level demographics, density and potential for public interaction, social capital, risk factors, capacity of the healthcare system, air pollution, employment in essential businesses, and political views | No                      |
| Zawbaa et al       | 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                            | USA                                          | COVID-19 incidence                              | RF                          | SIP orders, county metrics                              | No                      |
| Galvan et al       | Brazil                                | COVID-19 incidence and deaths                | Brazil MoH, IBGE, SUS, BCB, ADHB                | ANN (SOM)                   | Socioeconomic, health, and safety data                  | No                      |
| Hasan et al        | Bangladesh                            | COVID-19 incidence                           | WHO, IEDCR, survey                              | LSTM, ANFIS, ANN (MLP)      | Governing authorities, compliance, probability of infection and test positivity | No                      |
| Liu et al          | 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        | USA                                   | COVID-19 incidence                           | NYT, CDC, GHDx                                 | XGBoost                     | County-level population statistics, county-level disease rate and mortality | No                      |
| Pandit et al       | Worldwide                             | COVID-19 mortality rate                      | WHO, GSAID                                      | LogitBoost, Ada-boostM1     | Age, SARS-CoV-2 clade information                       | No                      |
| Roy et al          | 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          | USA                                   | COVID-19 incidence                           | Local DOH, CMS, LTCF, NICSCH                   | GB                          | Nursing home facility and community characteristics    | Imputed using k-NN     |
| Ye et al           | 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                           | Performs against COVID-19                   | Self-curated, Kaggle                            | k-means, hierarchical clustering | GDP, Poverty index, population, stringency index, smoking rate, CVD death rate, diabetes prevalence | Imputed with mean values |

(continued)
| Study          | Region     | Outcome     | Data source                                | Model          | Heterogeneous dataa | Missing data imputation |
|---------------|------------|-------------|--------------------------------------------|----------------|---------------------|-------------------------|
| Bird et al45  | 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 al46 | Worldwide | COVID-19 incidence | JHU, GBD, UW, GHO, WHO                | k-means        | Historical data, diseases, environmental factors, sociodemographics, health system factors | No                      |
| Lai et al47   | USA        | COVID-19 incidence | NYT, CDC, Census Bureau, USALEEP,         | k-means        | population census data, GIS data, business pattern censuses, and other sources | No                      |

aData that are heterogeneous in syntax, schema, and semantics.

bAvailable at https://doi.org/10.7910/DVN/JHFOSE.
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 ultrasound48–51 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 heterogeneous 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 heterogeneous 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 heterogeneous data for modeling (Table 3). These heterogeneous 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 heterogeneous 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.115–117 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.113

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,
| Study          | Region                | Outcome                                                   | Data source                                                                 | Model                                      | Heterogeneous data | Missing data imputation |
|---------------|-----------------------|-----------------------------------------------------------|----------------------------------------------------------------------------|--------------------------------------------|--------------------|-------------------------|
| Cai et al⁵³   | China                 | RT-PCR negativity                                        | Single hospital                                                            | Unspecified DL, LR                         | CT image data      | Replaced by median      |
| Cai et al⁵⁴   | 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 al⁵⁵  | USA, Iran, Italy      | ICU admission                                             | 3 hospitals                                                                | DNN, RF                                    | CT image data, demographics, vitals, lab data | Imputed by mean values |
| Chassagnon et al⁵⁶ | 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 al⁵⁷ | China                 | Severe vs. nonsevere COVID-19                             | Single hospital                                                            | CNN (uAI Discover-2019nCoV)                | Symptoms, local SARS-CoV-2 prevalence, CXR imaging, molecular diagnostic performance | No                      |
| D’Ambrosia et al⁵⁸ | USA                   | RT-PCR confirmed SARS-CoV-2 infection                    | Single hospital                                                            | BN, SC, DML, LR                            | No                 | No                      |
| Ebrahimian et al⁵⁹ | 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 al⁶⁰    | China                 | Stable vs progressive COVID-19                            | Unspecified hospitals                                                      | SVM                                        | CT image data, clinical and lab data        | No                      |
| Grodecki et al⁶¹ | USA, Italy            | Clinical deterioration vs death                           | 3 hospitals                                                                | CNN (U-Net), LR                            | CT image data, clinical data                | No                      |
| Guo et al⁶²    | China                 | COVID-19 vs seasonal flu                                 | 2 hospitals                                                                | RF                                         | CT image data, symptoms, blood tests, RT-PCR results | No                      |
| Hahm et al⁶³   | South Korea           | Worsening oxygenation event                              | Single hospital                                                            | DL software (MEDIP)                        | CT severity score, Demographics, Comorbidity, Lab data | No                      |
| Hermans et al⁶⁴ | The Netherlands       | COVID-19 positivity by RT-PCR                            | 2 hospitals                                                                | LR                                         | CT image data, demographics, symptoms, vitals, lab | No                      |
| Ho et al⁶⁵    | South Korea           | Severe vs nonsevere COVID-19                             | 5 hospitals                                                                | ANN, CNN, ACNN                             | CT image data, demographic, clinical, lab data | No                      |
| Jeong et al⁶⁶   | 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 al⁶⁷ | Mexico          | Need mechanical ventilation, death                       | Single hospital                                                            | AI software (Siemens healthcare)          | CT variables, demographics, clinical, lab   | No                      |
| Lang et al⁶⁸    | USA                   | Acute neuroimaging findings                              | Single hospital                                                            | Unspecified ML, LR                         | CT severity score, demographics, clinical data | No                      |
| Lassau et al⁶⁹  | 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 al⁷⁰    | 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*                                                                 | Missing data imputation |
|------------------|--------------|-------------------------------------------------------------------------|------------------------------------------|-------------------------------|------------------------------------------------------------------------------------|-------------------------|
| Liu et al⁷¹      | China        | COVID-19 vs. non-COVID-19 pneumonia                                     | Single hospital                          | CT image software (pyradiomics), LR, LASSO | CT outcomes, clinical data                                                         | No                      |
| Mei et al⁷²       | USA          | COVID-19 positivity by RT-PCR                                            | 18 hospitals                             | CNN, SVM, RF, MLP              | CT findings, clinical symptoms, exposure history, Lab                              | No                      |
| Meng et al⁷³      | China        | Death within 14 days                                                    | 4 hospitals                              | CNN, LR                       | CT image features, clinical information                                             | No                      |
| Mushtaq et al⁷⁴  | Italy        | Death, ICU admission                                                     | Single hospital                          | CNN (AI system qXR), Cox PH    | CXR severity, demographics, clinical data                                           | No                      |
| Ning et al⁷⁵      | China        | Morbidity, mortality                                                    | 2 hospitals                              | CNN, DNN, Ridge LR             | CT features, 130 types of clinical features                                        | No                      |
| Quiroz et al⁷⁶    | China        | Severe vs nonsevere COVID-19                                             | 2 hospitals                              | CNN (U-Net), LR, XGBoost       | CT features, demographics, clinical data                                           | Imputed with mean values|
| Salvatore et al⁷⁷ | 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 al⁷⁸    | 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 al⁷⁹       | China        | COVID-19 vs. influenza A/B                                               | 2 hospitals                              | DNN                            | CXR and CT features, 56 clinical features                                          | No                      |
| Xu et al⁸⁰        | 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 al⁸¹       | China        | 4-level COVID-19 severity                                               | Multiple hospitals                       | DSA-MIL, MA-CLR                | LUC features, age, medical history, symptoms                                        | No                      |

*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.
| Study                          | Region          | Outcome                                      | Data source                          | Model                                      | Heterogeneous dataa | Missing data imputation |
|-------------------------------|-----------------|----------------------------------------------|--------------------------------------|--------------------------------------------|---------------------|------------------------|
| **Early detection**          |                 |                                              |                                      |                                            |                     |                        |
| Ahamad et al81                | 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 al82               | Italy           | COVID-19 positivity by RT-PCR                | Single hospital                      | ANN                                        | Demographics, Comorbidity, Medications, Signs and Symptoms, Lab, Vitals, CXR | No                     |
| Martin et al83               | Worldwide       | COVID-19 positivity                          | Literature (British Medical Journal) | AI system (Symptoms)                       | Keywords and symptoms, Age and sex, Symptom occurrence frequency rates, Country-specific disease incidences | No                     |
| Obinata et al84              | Japan           | COVID-19 positivity by RT-PCR                | 2 hospitals                          | RF                                         | Demographics, Vitals, Lab, Symptoms, Contact history, Symptoms, travel history to suspicious areas, contact history | No                     |
| Otoom et al85                | Worldwide       | COVID-19 positivity                          | CORD-19 repository                   | SVM, ANN, NB, k-NN, decision table, decision stump, OneR, ZeroR | No                  |                        |
| Shimon et al86               | Israel          | COVID-19 positivity                          | Multiple hospitals                  | CNN, SVM, RF                               | Voice samples (acoustic features), self-reported symptoms | No                     |
| Wintjens et al87             | 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 al88                | 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 al89            | South Korea     | mortality                                    | KCDC                                 | ANN                                        | Demographics, infection reason and date | No                     |
| An et al90                   | South Korea     | mortality                                    | KNHIS                                 | LASSO, SVM, RF, k-NN                       | Socioeconomic and medical information | No                     |
| Burian et al91               | Germany         | ICU admission                                 | 1 hospital                           | RF                                         | Demographics, clinical, lab, and imaging data | Imputed with mean or mode |
| Cheng et al92                | 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 |
| Study          | Region       | Outcome                          | Data source          | Model                      | Heterogeneous data*                          | Missing data imputation |
|---------------|--------------|----------------------------------|----------------------|----------------------------|----------------------------------------------|-------------------------|
| Das et al93   | South Korea  | mortality                        | KCDC                 | LR, SVM, k-NN, RF, GB       | Demographic and exposure features            | No                      |
| Ge et al94    | China        | Ventilator parameters            | 1 hospital           | Unspecified                | Demographics, clinical data, Ventilator parameters | No                      |
| Haimovich et al95 | USA         | early respiratory decompensation | 8 EDs                | RF, LASSO, GB, XGBoost     | Demographics, medical histories, vitals, outpatient medications, chest radiograph reports, Lab | No                      |
| Hu et al96    | China        | mortality                        | 1 hospital           | LR, PLS regression, EN, RF, bagged FDA | Demographics, CT features, lab | Imputed using bagging trees |
| Iwendi et al97 | Worldwide    | Severity, recovery, death        | Kaggle (WHO, JHU)    | RF                         | Demographics, symptoms, travel data          | No                      |
| Josephus et al98 | Worldwide    | mortality                        | Kaggle (WHO, JHU)    | LR                         | Demographics, symptoms, travel data          | Imputed (unspecified)   |
| Li et al99    | Worldwide    | mortality                        | Github and Wolfram dataset | LR, RF, SVM | Demographics, location, symptoms, travel history, market exposure, chronic disease | No                      |
| Liang et al100 | 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 al101   | China        | mortality                        | 1 hospital           | RF, XGBoost                | Symptoms, comorbidity, demographic, vitals, CT scans results, lab | No                      |
| Metsker et al102 | Russia      | mortality                        | Russian government, Single hospital | ANN | Demographics, comorbidity, lab, treatment, travel history | No                      |
| Mountantonakis et al103 | USA    | AF and mortality                 | 13 hospitals         | NLP                        | Demographics, medical history, lab, NLP extracted atrial fibrillation | No                      |
| Nakamichi et al104 | USA    | Hospitalization and mortality    | Multiple hospitals   | AdaBoost, ET, GB, RF       | Demographics, comorbidity, SARS-CoV-2 sequence clades | Multiple imputation by chained equations |
| Neuraz et al105 | France      | in-hospital mortality            | 39 hospitals         | NLP, Cox                   | Demographics, comorbidity, NL.P extracted use of calcium channel blockers | No                      |
| Patel et al106 | USA          | Severity                        | 3 hospitals          | RF, ANN (MLP), SVM, GB, ET classifier, AdaBoost | Demographics, international travel, contact history, comorbidity, symptoms, blood panel profile | No                      |
| Study                  | Region                | Outcome                  | Data source                          | Model          | Heterogeneous data | Missing data imputation |
|------------------------|-----------------------|--------------------------|--------------------------------------|----------------|--------------------|------------------------|
| Planchuelo-Gómez et al | Spain                 | headache                 | 1 hospital                           | GLM, PCA       |                    | No                     |
|                        |                       |                          |                                      |                | Intensity and self-reported disability caused by headache, quality and topography of headache, migraine features, COVID-19 symptoms, lab. |                        |
| Schwartz et al         | Canada                | mortality                | iPHIS, CORES, The COD, CCMtool, CCM  | NLP, LR        |                    | Imputed by weekly median value |
| Wu et al               | China, Italy, Belgium | ICU admission, death, etc | Multiple hospitals                   | RF, LR         |                    | No                     |

aData 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; 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.
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 Naive Bayes. Two of the studies,\textsuperscript{123,124} which were based on the same online survey, collected data from 13 knowledge sources for modeling. None of the 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.\textsuperscript{125} Of the other 2 studies, 1 mined ClinicalTrials.gov to extract data on COVID-19-related trials,\textsuperscript{126} while the other searched the Scopus database for a bibliometric analysis.\textsuperscript{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),\textsuperscript{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,\textsuperscript{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\textsuperscript{134} or developing NLP and deep learning-based pipeline for sentinel syndromic surveillance of COVID-19 using medical records.\textsuperscript{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).\textsuperscript{129} This lone study involving heterogeneous 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),\textsuperscript{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,\textsuperscript{136} while the IDENTIFY trial identified a subpopulation of COVID-19 patients who had improved survival from taking hydroxychloroquine.\textsuperscript{137} Neither study integrated heterogeneous 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 socio-demographic 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 sociodemographic influences on COVID-19 patient outcomes were modulated by regional features, such as measures of social cohesion and income inequality.
the use of AI in diagnostic medical imaging has been extensively ex-
Medical imaging analysis for assisting COVID-19 diagnosis. In fact,
were used more frequently than others depending on the area of re-
We also found that a wide range of machine learning and
Qualitative analysis, we identified 7 key areas in which AI was ap-
gies in COVID-19 research. In the 794 studies included in our final
related data being collected in novel surveillance systems, AI meth-
and so on. On the other hand, with large amounts of COVID-19-
are actively attempting to address the COVID-19 pandemic, we are
As governments, research communities, and healthcare industries

DISCUSSION
As governments, research communities, and healthcare industries
actively attempting to address the COVID-19 pandemic, we are
tasked to identify quick yet reliable solutions for screening, diagno-
sis, 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 meth-
ods 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 technolo-
gies in COVID-19 research. In the 794 studies included in our final
qualitative analysis, we identified 7 key areas in which AI was ap-
plied. 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 re-
search.

It is not at all surprising that AI methods have been used exten-
sively in many areas of COVID-19 research. AI has been revolution-
ary 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 ex-
plored for many diseases, such as cancer,143 cardiovascular dis-
ases,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 abnor-
malities 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 ex-
plicit 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 het-
erogenous data, including text data such as those in clinical nar-
atives or on social media. For example, in drug discovery and
genomic research, AI is ideal for analyzing massive amounts of se-
quence data (eg, proteomic or genomic data).149,150

One limitation of the AI applications included in our scoping re-
view is the lack of integration of data from heterogeneous sources for
modeling. In the era of precision health, it is critical to examine a
comprehensive list of determinants of COVID-19 outcomes, includ-
ing biological, clinical, social, behavioral, and environmental fac-
tors, 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 ap-
proach ignores other important risk factors such as clinical symp-
toms, 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.53–55,62,64,69,76–78 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 envi-
ronmental data from the National Oceanic and Atmospheric Ad-
ministration 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 mul-
tilevel 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 technolo-
gies 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 inte-
gration in these AI applications and a need for a multilevel AI frame-
work 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 per-
formed 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 Infor-
matics Association online.
74. Mushiaq J, Pennella R, Lavalie 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. *J Med Imag* 2021; 9 (2): e24572. doi:10.2196/24572.

77. Salvatore C, Roberta F, Angela d 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/jibs.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. 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/s10349-020-00808-8.

82. Martin A, Natao J, Gruarin S, et al. 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.

83. Obinata H, Yokobori S, Shomron N. Machine learning-based prediction of COVID-19 outcomes 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.

84. 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/s10349-020-00808-8.

85. 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/s10349-020-00808-8.

86. 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/jibs.53982.

87. 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.

88. 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. *J Med Imag* 2021; 9 (2): e24572. doi:10.2196/24572.
