Revealing transparency gaps in publicly available COVID-19 datasets used for medical artificial intelligence development—a systematic review

Joseph E Alderman, Maria Charalambides*, Gagandeep Sachdeva, Elinor Laws, Joanne Palmer, Elsa Lee, Vaishnavi Menon, Qasim Malik, Sonam Vadera, Melanie Calvert, Marzyeh Ghassemi, Melissa D McCradden, Johan Ordish, Bilal Mateen, Charlotte Summers, Jacqui Gath, Rubeta N Matin, Alastair K Denniston, Xiaoxuan Liu*

During the COVID-19 pandemic, artificial intelligence (AI) models were created to address health-care resource constraints. Previous research shows that health-care datasets often have limitations, leading to biased AI technologies. This systematic review assessed datasets used for AI development during the pandemic, identifying several deficiencies. Datasets were identified by screening articles from MEDLINE and using Google Dataset Search. 192 datasets were analysed for metadata completeness, composition, data accessibility, and ethical considerations. Findings revealed substantial gaps: only 48% of datasets documented individuals' country of origin, 43% reported age, and under 25% included sex, gender, race, or ethnicity. Information on data labelling, ethical review, or consent was frequently missing. Many datasets reused data with inadequate traceability. Notably, historical paediatric chest x-rays appeared in some datasets without acknowledgment. These deficiencies highlight the need for better data quality and transparent documentation to lessen the risk that biased AI models are developed in future health emergencies.

Lancet Digit Health 2024; 6: e827–47

*These authors contributed equally

Queen Elizabeth Hospital, University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK (J E Alderman MB ChB, E Laws MB BCh, J Palmer PhD, V Menon MB ChB, S Vadera MB BS, Prof A K Denniston PhD, X Liu PhD)**; Institute of Inflammation and Ageing** (J E Alderman, E Laws, J Palmer, J Ordish MA, Prof A K Denniston, X Liu)**, NIHR Birmingham Biomedical Research Centre** (J E Alderman, E Laws, J Palmer, Prof M Calvert PhD, Prof A K Denniston, X Liu)**, Birmingham Health Partners Centre for Regulatory Science and Innovation** (Prof M Calvert)**, and NIHR Applied Research Collaboration (ARC) West Midlands** (Prof M Calvert)**, University of Birmingham, Birmingham, UK; University Hospital Southampton NHS Foundation Trust, Southampton, UK** (M Charalambides MB ChB)**; The Royal Wolverhampton NHS Trust, Wolverhampton, UK** (G Sachdeva MB ChB)**; Guy's, King's, & St Thomas' School of Medical Education** (E Lee MSc) **and AI Centre for Value Based Healthcare** (Q Malik MB ChB)**, King's College London, London, UK; Birmingham Women's and Children's NHS Foundation Trust, Birmingham, UK** (Q Malik, Prof A K Denniston)**; University Hospitals of Leicester NHS Trust, Leicester, UK** (S Vadera)**; Centre for Patient Reported Outcomes Research, Institute of Applied Health Research, College of Medical and Dental Sciences, University of Birmingham, Birmingham, UK** (Prof M Calvert)**; NIHR Blood and Transplant Research Unit (BTRU) in Precision Transplant**

Introduction

The second half of 2019 saw the emergence of a novel, highly infectious pneumonia of unknown cause in Wuhan, China.[1] In the subsequent months, the causative pathogen would be uncovered as SARS-CoV-2,[2] the disease would be named COVID-19, and the consequences of a rapidly spreading pandemic would drastically affect lives and livelihoods.[3] Many countries' health-care systems promptly pivoted to address exponentially rising numbers of patients needing treatment.[4–6]

As a result of COVID-19, by November, 2023, over 770 million people had been infected worldwide, and nearly 7 million had died.[7] The world's scientific community responded rapidly to the emerging crisis, with the first full genomes of SARS-CoV-2 released in early January, 2020, identification of dexamethasone as a highly effective treatment by June, 2020, and the first vaccines administered in December, 2020.[8–11] Scientific organisations and governments recognised early in the pandemic that delivering the necessary scientific progress at speed and scale would require information sharing, issuing guidance and legal derogations to reduce restrictions on the flow of data.[12,13] The body of scientific research related to COVID-19 grew substantially in a short period, and as of Nov 7, 2023, more than 2·2 million scientific articles had been published in this area.[14]

The growing interest in medical artificial intelligence (AI) before the pandemic prompted many to investigate whether AI could be used to ease the burden of overloaded health-care systems. Early research used clinical data sourced directly from hospitals to train image classifiers using convolutional neural networks and similar technologies with the intention of diagnosing COVID-19 using thoracic CT image data.[15] Other developers and groups contributed to online challenges hosted by Kaggle and other similar platforms to develop algorithms targeting a plethora of diagnostic and prognostic tasks, in some cases competing for prize money.[16] Despite this initial enthusiasm and the development of several AI health technologies for COVID-19, few health-care providers deployed these products. Moreover, three systematic reviews of predictive models and AI medical devices for COVID-19 have highlighted substantial shortcomings in the methodology underpinning these technologies and the evidence supporting their use.[17–19]

AI health technologies are trained and evaluated using health datasets, and accordingly any biases or limitations present in these datasets can affect the performance of the resultant technology.[20] The consequences vary depending on the clinical context and the nature of the AI health technology and datasets used. Of concern, there is increasing evidence that biases in the underlying data can be replicated or extended by AI health technologies and cause harm to patients.[20–22] A previous systematic review focusing on nine of the most commonly used COVID-19 datasets found a high risk of bias being transmitted to downstream AI models.[23] Systematic reviews of health-care datasets of ophthalmic and skin cancer images have highlighted that demographic attributes of the individuals in datasets are infrequently reported and that most data derives from patients residing in a small number of countries.[24,25] The result is underrepresentation that follows two key patterns: exclusion, when individuals or groups are not present in the dataset; and misrepresentation, when the data available for individuals or groups are inaccurate, incomplete, or limited in other ways. Data biases are generated when underrepresentation and other limitations in data affect some

and Cellular Therapeutics, Birmingham, UK (Prof M Calvert); Department of Electrical Engineering and Computer Science (M Ghassemi PhD) and Institute for Medical Engineering & Science (M Ghassemi), Massachusetts Institute of Technology, Cambridge, MA, USA; Department of Bioethics, The Hospital for Sick Children, Toronto, ON, Canada (M D McCradden PhD); Genetics & Genome Biology, SickKids Research Institute, Toronto, ON, Canada (M D McCradden); Hughes Hall (J Ordish) and Victor Phillip Dahdaleh Heart & Lung Research Institute (Prof C Sumers PhD), University of Cambridge, Cambridge, UK; Roche Diagnostics, Rotkreuz, Switzerland (J Ordish); Institute of Health Informatics, University College London, London, UK (B Mateen MBBS); PATH, Seattle, WA, USA (B Mateen); Wellcome Trust, London, UK (B Mateen); Independent Cancer Patients Voice, London, UK (J Gath); Oxford University Hospitals NHS Foundation Trust, Oxford, UK (R N Matin PhD); NIHR Biomedical Research Centre, Moorfields Eye Hospital and University College London, London, UK (Prof A K Denniston)

Correspondence to:
Dr Xiaoxuan Liu, Queen Elizabeth Hospital, University Hospitals Birmingham NHS Foundation Trust, Birmingham B15 2TH, UK
x.liu.8@bham.ac.uk

For more on the **STANDING TOGETHER programme** see https://www.datadiversity.org

See **Online** for appendix 1

For more on the **STANDING Together recommendations** see https://www.datadiversity.org/recommendations

groups over others. Data-driven technologies developed with datasets showing these biases are less likely to serve the needs of those who are inadequately represented, which is referred to as health data poverty.[26] The need for datasets to adequately represent the populations for whom AI health technologies are developed is highlighted as a guiding principle in the Good Machine Learning Principles coauthored by the Food and Drug Administration (USA), Medicines and Healthcare products Regulatory Agency (UK), and Health Canada.[27]

The negative consequences of COVID-19 have disproportionately affected minoritised individuals, in particular those grouped by their ethnicity or race, socioeconomic status, sex, gender, or age (among others).[28–32] These trends are likely influenced by differences in patterns of comorbidities (including obesity and cardiovascular disease), and by biases inherent to the way health care is delivered (including the recognition that pulse oximeter devices underestimate dangerous hypoxaemia in people with darker skin tones).[33] It is therefore critical that datasets used for AI development are diverse and inclusive, and in particular that they are representative of at-risk groups. This systematic review aimed to characterise the composition and reporting of publicly available datasets related to COVID-19 that were used to develop AI models across the duration of the pandemic. Specifically, the reporting of demographic attributes was assessed to characterise who was represented in these datasets and how, and which groups were underrepresented. This Review forms part of the STANDING Together programme—an international multistakeholder collaboration to produce recommendations encouraging datasets and AI health technologies that are inclusive and equitable.[34]

Methods

This systematic review was conducted in adherence with the PRISMA guidelines (appendix 1, pp 12–16).[35] The review did not evaluate direct health-related outcomes and therefore did not meet the criteria for registration of the protocol with PROSPERO.[36] Ethical approval was not required for this systematic review.

Search strategy and selection criteria

Searches were conducted using MEDLINE and Google Dataset Search (Alphabet, Mountain View, CA, USA). The initial MEDLINE search was conducted using the Ovid interface for articles published from database inception until Oct 15, 2021; this search was then repeated on Feb 27, 2024, to include all additional articles published until the end of the global health emergency (declared by WHO as May 5, 2023). MeSH terms and keywords used included: "dataset" OR "database" OR "collection" OR "repository" OR "artificial intelligence" OR "neural networks" OR "machine learning", AND "COVID-19" OR "coronavirus", AND "diagnosis" OR "risk" OR "severity" OR "prediction" OR "outcome" OR "imaging" OR "image" OR "score" OR "scoring". The full search strategy was based on that used in previous dataset reviews[24,25] and is available in appendix 1 (pp 2–3). There was no restriction of articles based on investigated population characteristics, geographical origin, or patient population, but only articles in English were included for screening.

Selection process and data extraction

The primary purpose of this review was to identify relevant datasets rather than research articles; accordingly, there was an additional dataset screening stage following the article full-text screening. Two reviewers (MCh, GS, QM, or VM) independently screened titles and abstracts using Rayyan (Cambridge, MA, USA).[37] The inclusion criteria for articles was that they detailed COVID-19 datasets within the article (including review articles), or the article described the use of datasets to train or test machine-learning algorithms for COVID-19. Articles exploring only population health, epidemiology, or COVID-19 forecasting were excluded. A third reviewer (XL or JEA) was consulted to arbitrate any conflicts between the two other reviewers and held the casting vote. A single reviewer (MCh, GS, XL, ELe, QM, VM, ELa, or JEA) reviewed each full-text article against the eligibility criteria. The reviewer simultaneously identified any datasets referred to in the article and collated these into an online spreadsheet with details of the source article. Duplicated datasets were aggregated into a single record and the number of citing articles was recorded. During data extraction, a URL linking to each dataset or a link to the citing article was recorded. The full text of all but two articles identified via searches were accessible to the reviewers for screening.

An additional search for datasets was conducted by GS using Google Dataset Search on May 12, 2022, to identify additional datasets that might have been missed via literature searching. Structured search queries cannot be used with this platform, and pilot searches with narrow search terms found that relevant datasets were missed, and so a broad search term of "COVID 19 dataset" was used. All search results were screened independently by two reviewers (GS and JEA).

Two independent reviewers extracted data from any descriptive documentation or dataset metadata available in the same repository, or otherwise immediately linked to the dataset. Data were extracted using a standardised data collection form by one of ELa, XL, GS, MCh, QM, or VM, and was then checked by JEA. This form was created based on the STANDING Together recommendations with Google Forms (Alphabet, Mountain View, CA, USA; appendix 1, pp 17–59). These data included: details of the dataset's identity; the composition of the data; any hosting; any restrictions applied to data access (eg, whether an application process was in place, or whether payment was required); demographic attributes of included individuals based on the groupings set out in

	Date dataset first published	Open or managed access	Hosting platform	What is this dataset composed of?	If medical images are included in the dataset and are labelled, does the dataset specify how these labels were assigned?	How was COVID-19 diagnosis confirmed?	How many images or individuals are represented in this dataset?	Did the dataset receive ethical approval or waiver by an ethics committee, IRB, or similar?	Did patients give their consent for data to be included in this dataset?
3DLSC-COVID	Jan 30, 2021	Open access	Other	CT chest images	Reported by radiology expert	..	..	..	..
A blood atlas of COVID-19 defines hallmarks of disease severity and specificity	Aug 17, 2021	Open access	Zenodo	Genomics data; other laboratory research data	The dataset does not contain images	PCR test	116 individuals	Yes	Yes
A COVID multiclass dataset of CT scans	Unknown	Open access	Kaggle	CT chest images	Images are unlabelled	PCR test	4173 images from 210 individuals	Yes	..
A diagnostic host response biosignature for COVID-19 from RNA profiling of nasal swabs and blood	Dec 15, 2020	Open access	Other	Other laboratory research data (transcriptomics)	The dataset does not contain images	PCR test	333 individuals	Yes	No
A large dataset of real patients CT scans for COVID-19 identification	June 19, 2020	Open access	Other	CT chest images	..	PCR test	4173 CT images from 210 individuals	Yes	No
A single-cell atlas of the peripheral immune response to severe COVID-19	May 18, 2020	Open access	Other	Other laboratory research data	The dataset does not contain images	PCR test	13 samples from 12 individuals	Yes	Yes
Actualmed COVID-19 CHEST X-RAY DATASET INITIATIVE	May 5, 2020	Open access	Github	Chest x-ray images	Images are unlabelled	..	239 rows in metadata file, but unclear if duplicates	..	..
AIforCOVID imaging archive	Dec 1, 2021	Managed access	Other	Chest x-ray images	Images are unlabelled	PCR test	820 images, unknown number of individuals	Yes	..
Amsterdam Medical Data Science—Dutch Datawarehouse COVID-19	March 1, 2020	Managed access	Other	EHR or medical records data	The dataset does not contain images	PCR test; reporting and data system score with clinical suspicion	..	..	..
Augmented COVID-19 X-Ray Images Dataset	March 25, 2020	Open access	Mendeley data	Chest x-ray images	Images are unlabelled	..	1826 images, unknown number of individuals	..	..
Automatic Detection Of COVID-19 From Ultrasound Data	April 25, 2020	Open access	Github	Ultrasound thorax images	Reported by health-care professional, but speciality or experience not stated	Usually PCR, but in some cases not stated as multiple data sources combined	..	Yes	Yes
Balanced Augmented COVID CXR Dataset	Unknown	Open access	Kaggle	Chest x-ray images	..	..	..	..	..
BIMCV-COVID19+	Unknown	Managed access	Other	Chest x-ray images; CT chest images	Automatically labelled via AI model or similar	PCR, IgG, or IgM	..	Yes	No
Bravi et al, 2020	June 24, 2020	Open access	Other	No images	The dataset does not contain images	..	1603 individuals	..	..
BrixIA COVID-19 project	Jul 1, 2021	Managed access	Other	Chest x-ray images	Reported by radiology expert	Same as covid-chestxray-dataset	..	Yes	..
Chest Imaging	March 28, 2020	Open access	Social network posts	Chest x-ray images	Images are unlabelled	..	50 individuals	..	..
Chest Imaging Appearance of COVID-19 Infection	Feb 13, 2020	Open access	Other	Chest x-ray images; CT chest images; EHR or medical records data	Reported by health-care professional, but speciality or experience not stated	..	3 individuals	..	..

(Table continues on next page)

	Date dataset first published	Open or managed access	Hosting platform	What is this dataset composed of?	If medical images are included in the dataset and are labelled, does the dataset specify how these labels were assigned?	How was COVID-19 diagnosis confirmed?	How many images or individuals are represented in this dataset?	Did the dataset receive ethical approval or waiver by an ethics committee, IRB, or similar?	Did patients give their consent for data to be included in this dataset?
(Continued from previous page)									
Chest x-ray (COVID-19 & Pneumonia)	Unknown	Open access	Kaggle	Chest x-ray images	Images are unlabelled	Chest x-ray images	6432 images, unknown number of individuals	..	..
Chest x-ray images with three classes: COVID-19, Normal, and Pneumonia	June 11, 2020	Open access	Mendeley data	Chest x-ray images	..	..	603 images, unknown number of individuals	..	..
Chest xray for COVID-19 detection	Unknown	Open access	Kaggle	Chest x-ray images	..	..	371 images, unknown number of individuals	..	..
Chest Xray Images PNEUMONIA and Covid-19	Unknown	Open access	Kaggle	Chest x-ray images; CT chest images	..	..	6118 individuals	..	..
Chest: COVID-19 index	Unknown	Open access	Other	CT chest images	Reported by health-care professional, but speciality or experience not stated	..	..	..	..
Clinical and immunological benefits of convalescent plasma therapy in severe COVID-19: insights from a single center open label randomised control trial	Feb 17, 2022	Open access	Other	Proteomics data	The dataset does not contain images	..	..	..	..
CO-IRv2	June 21, 2021	Open access	Figshare	CT chest images	..	..	..	..	..
Combined COVID-19 Dataset	May 13, 2020	Open access	Mendeley data	Chest x-ray images; CT chest images	..	..	6130 images, unknown number of individuals	..	..
Computer-Aided diagnostic for classifying Chest X-Ray Images using Deep Ensemble Learning	June 13, 2022	Open access	Zenodo	Chest x-ray images	..	..	..	..	..
CORONACASES.org	Unknown	Open access	Other	CT chest images	..	..	10 images, unknown number of individuals	..	..
CoronaHack-Chest X-Ray-Dataset	Unknown	Open access	Kaggle	Chest x-ray images	..	..	..	..	..
Coronavirus (COVID-19) CC-19 dataset	Jul 10, 2020	Open access	Github	CT chest images	..	..	89 individuals	..	..
CoroNet	Unknown	Open access	Other	Chest x-ray images	..	..	..	..	..
Coswara	Oct 12, 2022	Open access	Zenodo	Respiratory sounds	The dataset does not contain images	PCR; antigen test	2746 individuals	Yes	Yes
COUGHVID	Sept 24, 2020	Open access	Zenodo	Cough sounds	The dataset does not contain images	..	..	..	Yes
Cov-Caldas	Dec 5, 2022	Open access	Figshare	Chest x-ray images	Reported by radiology expert	PCR test	7303 images from 2821 individuals	Yes	Yes
COVID 19 CT Scan Dataset	Unknown	Open access	Kaggle	CT chest images	..	..	..	..	..
COVID Dataset (China Consortium of Chest CT Image Investigation)	April 24, 2020	Open access	Other	CT chest images	Images are unlabelled	PCR test	6752 images from 4154 individuals	Yes	Yes
COVID laboratory values	Jan 25, 2021	Open access	Zenodo	Chest x-ray images; EHR or medical records data	..	PCR test	488 individuals	Yes	No
covid_19_2020	Unknown	Open access	Kaggle	Chest x-ray images	..	Chest x-ray images	16 147 images	..	..

(Table continues on next page)

	Date dataset first published	Open or managed access	Hosting platform	What is this dataset composed of?	If medical images are included in the dataset and are labelled, does the dataset specify how these labels were assigned?	How was COVID-19 diagnosis confirmed?	How many images or individuals are represented in this dataset?	Did the dataset receive ethical approval or waiver by an ethics committee, IRB, or similar?	Did patients give their consent for data to be included in this dataset?
(Continued from previous page)									
Covid_MyDataset	June 25, 2020	Open access	Other	Chest x-ray images	..	..	..	..	..
COVID-19 (Automated Detection of Covid-19 Cases Using Deep Neural Networks with X-Ray Images)	April 28, 2020	Open access	Github	Chest x-ray images	..	..	1000 images	..	..
COVID-19 & Normal Posteroanterior (PA) X-Rays	Unknown	Open access	Kaggle	Chest x-ray images	..	..	280 images, unknown number of individuals	..	..
COVID-19 and common pneumonia chest CT dataset	June 24, 2020	Open access	Mendeley data	CT chest images	..	PCR test	828 images from 618 individuals	..	..
COVID-19 Case Surveillance Public Use Data With Geography	Feb 3, 2021	Open access	Other	EHR or medical records data	The dataset does not contain images	Laboratory confirmed	105 million cases, could be duplicates if people were reinfected	Yes	No
COVID-19 Cell Atlas	Unknown	Open access	Other	Other laboratory research data	The dataset does not contain images	..	..	..	..
COVID-19 Chest Ct Image Augmentation GAN Dataset	Oct 26, 2020	Open access	Kaggle	CT chest images	..	..	..	..	..
COVID-19 Chest X Rays	Unknown	Open access	Kaggle	Chest x-ray images	..	..	148 images, unknown number of individuals	..	..
COVID-19 Chest X-Ray Image Repository	May 20, 2021	Open access	Figshare	Chest x-ray images	Images are unlabelled	PCR test	..	..	..
COVID-19 Chest Xray	April 28, 2020	Open access	Kaggle	Chest x-ray images	..	..	205 individuals	..	..
COVID-19 CT Lung And Infection Segmentation Dataset	April 20, 2020	Open access	Zenodo	CT chest images	Reported by radiology expert	PCR test	20 images	..	..
COVID-19 CT scans	Unknown	Open access	Kaggle	CT chest images	Reported by radiology expert	..	20 images	..	..
COVID-19 CT segmentation dataset	April 2, 2020	Open access	Other	CT chest images	Reported by radiology expert	..	100 images from 43 individuals	..	..
COVID-19 DATASET 3 CLASSES	March 18, 2020	Open access	Other	Chest x-ray images	..	..	..	..	..
COVID-19 Detection X-Ray Dataset	Unknown	Open access	Kaggle	Chest x-ray images	..	..	5863 images, unknown number of individuals	..	..
COVID-19 diagnostic	Unknown	Open access	Kaggle	EHR or medical records data	The dataset does not contain images	..	5644 individuals	..	..
COVID-19 image data collection	June 10, 2020	Open access	Github	Chest x-ray images; CT chest images	..	..	216 individuals with CT images, numbers of chest x-ray images not reported	..	..
COVID-19 Image Dataset	Unknown	Open access	Kaggle	Chest x-ray images	..	..	..	..	..
COVID-19 IMAGING DATABASE	March 25, 2020	Open access	Other	Chest x-ray images; CT chest images	Reported by health-care professional, but speciality or experience not stated	PCR test	59 individuals	..	..
COVID-19 immune features revealed by a large-scale single-cell transcriptome atlas	Aug 22, 2021	Managed access	Other	Other laboratory research data	The dataset does not contain images	..	Not specified	..	..

(Table continues on next page)

	Date dataset first published	Open or managed access	Hosting platform	What is this dataset composed of?	If medical images are included in the dataset and are labelled, does the dataset specify how these labels were assigned?	How was COVID-19 diagnosis confirmed?	How many images or individuals are represented in this dataset?	Did the dataset receive ethical approval or waiver by an ethics committee, IRB, or similar?	Did patients give their consent for data to be included in this dataset?
(Continued from previous page)									
COVID-19 LUNG CT LESION SEGMENTATION CHALLENGE—2020	Nov 2, 2020	Open access	Other	CT chest images	Reported by radiology expert	PCR test	295 individuals	..	..
COVID-19 Lung CT Scans	Unknown	Open access	Kaggle	CT chest images	..	..	8439 images	Yes	..
COVID-19 Numeric Dataset	Jul 28, 2021	Open access	Kaggle	EHR or medical records data	The dataset does not contain images	..	319 rows (not stated if each row relates to a unique individual)	..	..
COVID-19 Patients Lungs X Ray Images 10000	Unknown	Open access	Kaggle	Chest x-ray images	..	..	98 images	..	..
COVID-19 pneumonia new dataset (available on request)	April 1, 2020	Open access	Github	Chest x-ray images	..	..	..	..	..
COVID-19 POSTEROANTERIOR CHEST X-RAY FUSED (CPCXR) DATASET	Oct 20, 2020	Open access	Github	Chest x-ray images	Reported by radiology expert	..	1214 images	No	No
COVID-19 Radiography Database	July 20, 2020	Open access	Kaggle	Chest x-ray images, lung mask images, and associated metadata	..	..	32138 images	..	..
COVID-19 Radiography Dataset	Unknown	Open access	Kaggle	Chest x-ray images	Images are unlabelled	..	21165 images	..	..
COVID-19 Symptom Tracker Dataset (CVST)	April 3, 2020	Managed access	Other	Symptoms submitted by the public via an app	The dataset does not contain images	..	..	No	Yes
COVID-19 tissue atlases reveal SARS-COV-2 pathology and cellular targets	April 9, 2021	Open access	Other	Other laboratory research data (RNA transcriptomics)	The dataset does not contain images	PCR test	16 individuals	..	..
COVID-19 Wearables Data (2020)	Unknown	Open access	Other	Wearable device data	The dataset does not contain images	..	..	..	..
COVID-19 X rays	Jan 1, 2020	Open access	Kaggle	Chest x-ray images; CT chest images	..	..	95 images	..	..
COVID-19 X-ray Images5	Unknown	Open access	Kaggle	Chest x-ray images	..	..	1443 images	..	..
COVID-19 xray dataset	Jul 29, 2020	Open access	Other	Chest x-ray images	Reported by radiology expert	..	6500 images	..	..
COVID-19 Xray Dataset (Train & Test Sets)	Unknown	Open access	Kaggle	Chest x-ray images	Images are unlabelled	..	40 images, unknown number of individuals	..	..
Covid-19 Xray images using Cnn	Unknown	Open access	Kaggle	Chest x-ray images	..	..	284 images, unknown number of individuals	..	..
COVID-19_dataset	Unknown	Open access	Kaggle	Chest x-ray images	..	Chest x-ray images	50 images, unknown number of individuals	..	..
COVID-19_Radiography_Database.zip	Unknown	Open access	Other	Chest x-ray images	..	..	..	..	..
COVID-19-AR	Jul 3, 2020	Open access	Other	Chest x-ray images; CT chest images	..	PCR test	105 individuals	..	..
COVID-19-CT-CXR	June 11, 2020	Open access	Github	Chest x-ray images; CT chest images	..	..	263 images, unknown number of individuals	..	..
COVID-19-NY-SBU	Nov 8, 2021	Open access	Other	Chest x-ray images; CT chest images; brain MRI	..	PCR test; clinical features	1384 individuals	..	..

(Table continues on next page)

	Date dataset first published	Open or managed access	Hosting platform	What is this dataset composed of?	If medical images are included in the dataset and are labelled, does the dataset specify how these labels were assigned?	How was COVID-19 diagnosis confirmed?	How many images or individuals are represented in this dataset?	Did the dataset receive ethical approval or waiver by an ethics committee, IRB, or similar?	Did patients give their consent for data to be included in this dataset?
(Continued from previous page)									
Covid-19-Patient-Health-Analytics	Unknown	Open access	Github	EHR or medical records data	The dataset does not contain images	..	1085 individuals	..	..
COVID-19&Normal&Pneumonia_CT_Images	Unknown	Open access	Kaggle	CT chest images	..	..	..	..	..
covid-chestxray-dataset	March 28, 2020	Open access	Github	Chest x-ray images	..	..	481 (from metadata file)	Yes	Yes
COVID-Classifier	May 18, 2020	Open access	Github	Chest x-ray images	..	..	420 images, unknown number of individuals	..	..
COVID-CS dataset	Feb 18, 2021	Open access	Github	Features extracted from CT images	..	..	750 individuals	..	..
COVID-CT	June 17, 2020	Open access	Github	CT chest images	Reported by radiology expert	..	349 images from 216 individuals	..	..
COVID-CT-MD	April 26, 2021	Open access	Figshare	CT chest images	Reported by radiology expert	PCR test	305 individuals	Yes	Yes
COVID-CT-Rate	Jan 14, 2022	Open access	Figshare	CT chest images	Reported by radiology expert	..	433 images from 82 individuals	Yes	..
COVID-CTset	Jul 23, 2020	Open access	Github	CT chest images	Images are unlabelled	..	63 849 images from 377 individuals	..	..
COVID-CXNet	Jul 30, 2020	Open access	Github	Chest x-ray images	..	Chest x-ray images of patients with (mostly) PCR-positive COVID-19 are collected from different publicly available sources, such as SIRM	900 images, unknown number of individuals	..	..
COVID-CXR-AI	Sept 10, 2021	Open access	Github	Chest x-ray images	..	..	..	..	..
Covid-GAN and Covid-Net mini Chest X-ray	Unknown	Open access	Kaggle	Chest x-ray images	GAN generated synthetic images	GAN generated synthetic images	1000 GAN generated images	..	..
COVID-LDCT.zip	Unknown	Managed access	Other	CT chest images	Reported by radiology expert	PCR test for 36% of cases; 64% of cases were obtained by consensus between three experienced radiologists with 95·6% agreement	260 individuals	..	..
COVID-Net	Dec 3, 2020	Open access	Kaggle	Chest x-ray images; CT chest images; Ultrasound thorax images	..	..	~23 000 individuals, plus 29 000 ultrasound images	..	..
COVID-QU-Ex Dataset	Unknown	Open access	Kaggle	Chest x-ray images	..	..	..	..	..
COVID-XRay-5K DATASET	April 24, 2020	Open access	Github	Chest x-ray images	Reported by radiology expert	Chest x-ray images	5000 images, unknown number of individuals	..	..
COVID-19 image repository	May 25, 2020	Open access	Other	Chest x-ray images; EHR or medical records data	..	..	243 individuals	..	..

(Table continues on next page)

	Date dataset first published	Open or managed access	Hosting platform	What is this dataset composed of?	If medical images are included in the dataset and are labelled, does the dataset specify how these labels were assigned?	How was COVID-19 diagnosis confirmed?	How many images or individuals are represented in this dataset?	Did the dataset receive ethical approval or waiver by an ethics committee, IRB, or similar?	Did patients give their consent for data to be included in this dataset?
(Continued from previous page)									
COVID19 unseen dataset	Unknown	Open access	Other	Chest x-ray images; CT chest images	..	PCR test	51 individuals	..	..
COVID19_Pneumonia_Normal_Chest_Xray_PA_Dataset	Unknown	Open access	Kaggle	Chest x-ray images	..	..	6939 images, unknown number of individuals	..	..
COVID19_xp	Oct 16, 2020	Open access	Github	Chest x-ray images	..	Same as covid-chestxray-dataset	1248 images, unknown number of individuals	No	No
COVID19-CT-Dataset	Feb 19, 2021	Open access	Other	CT chest images	..	..	1013 individuals	..	..
COVID19-image-classification	Jan 1, 2020	Open access	Github	Chest x-ray images	..	..	..	..	..
Covid19-Pneumonia-Normal Chest X-Ray Images	June 14, 2022	Open access	Mendeley data	Chest x-ray images	..	..	..	..	..
covid19-transcriptomics-pathogenesis-diagnostics-results	Nov 13, 2020	Open access	Github	Genomics data	The dataset does not contain images	PCR test	238 samples (not stated if each is for an individual patient)	Yes	No
COVID19-vs-Normal Dataset	Nov 21, 2020	Open access	Github	Chest x-ray images	..	..	800 images	..	..
COVID19-xray	Unknown	Open access	Kaggle	Chest x-ray images	Images are unlabelled	Chest x-ray images	1161 images, unknown number of individuals	..	..
COVID19ML	Sept 14, 2020	Open access	Github	EHR or medical records data	The dataset does not contain images	PCR test	1485 individuals	..	..
COVIDGR datasets	Sept 22, 2020	Open access	Github	Chest x-ray images	Images are unlabelled	PCR test	852 images, unknown number of individuals	..	..
CovidStudy	Unknown	Open access	Github	EHR or medical records data	The dataset does not contain images	PCR test	895 records (not stated if each is for an individual patient)	Yes	Yes
COVIDx CT Dataset	Dec 3, 2020	Open access	Github	CT chest images	Images are unlabelled	Two sub-datasets exist: COVIDx CT-A (PCR or radiologist confirmed) and COVIDx CT-B (unverified, assumed COVID diagnoses)	4501 individuals	..	..
COVIDx CXR-4	Unknown	Open access	Kaggle	Chest x-ray images	..	..	84 818 images from 45 342 individuals	..	..
COVIDx Dataset	March 18, 2020	Open access	Github	Chest x-ray images; CT chest images	..	..	~95 240 images	..	..
COVIDx-US	March 17, 2021	Open access	Github	Ultrasound thorax images; ultrasound thorax videos	..	..	..	..	..
CovidXrayNet	April 15, 2021	Open access	Github	Chest x-ray images	..	..	15 348 images, unknown number of individuals	..	..
CRI dataset 10,000 images	Jul 21, 2021	Managed access	Other	Chest x-ray images	..	Same as covid-chestxray-dataset	10 046 images	No	..

(Table continues on next page)

	Date dataset first published	Open or managed access	Hosting platform	What is this dataset composed of?	If medical images are included in the dataset and are labelled, does the dataset specify how these labels were assigned?	How was COVID-19 diagnosis confirmed?	How many images or individuals are represented in this dataset?	Did the dataset receive ethical approval or waiver by an ethics committee, IRB, or similar?	Did patients give their consent for data to be included in this dataset?
(Continued from previous page)									
CT Images in COVID-19	Aug 31, 2020	Open access	Other	CT chest images	Reported by health-care professional, but speciality or experience not stated	PCR test	661 individuals	..	..
CT Scans for COVID-19 Classification	April 7, 2020	Open access	Kaggle	CT chest images	Automatically labelled via AI model or similar	PCR test	19 685 images	Yes	..
CT-Angel	Nov 5, 2020	Open access	Github	CT chest images	Reported by radiology expert	CT chest images	46 096 images from 106 individuals	..	..
CT-COV19	Unknown	Open access	Github	CT chest images	..	..	>1000 individuals	..	..
Curated Chest X-Ray Image Dataset for COVID-19	Unknown	Open access	Kaggle	Chest x-ray images	..	..	9208 images, unknown number of individuals	..	..
Curated Dataset for COVID-19 Posterior-Anterior Chest Radiography Images (X-Rays)	Dec 6, 2022	Open access	Mendeley data	Chest x-ray images	..	..	..	..	..
Deep learning enables accurate diagnosis of novel coronavirus (COVID-19) with CT images	Unknown	Open access	Github	CT chest images	..	PCR test	274 individuals	Yes	No
Detecting-COVID-19-in-X-ray-images	March 20, 2020	Open access	Github	Chest x-ray images	..	..	..	..	..
Detection of COVID-19 Infection from Routine Blood Exams with Machine Learning: a Feasibility Study	April 25, 2020	Open access	Zenodo	EHR or medical records data; other laboratory research data	The dataset does not contain images	PCR test	280 individuals	Yes	Yes
Development, evaluation, and validation of machine learning models for COVID-19 detection based on routine blood tests	Oct 12, 2020	Open access	Zenodo	EHR or medical records data	The dataset does not contain images	PCR; lateral flow	1736 individuals	Yes	Yes
Diagnosis of COVID-19 and its clinical spectrum	Unknown	Open access	Kaggle	EHR or medical records data	The dataset does not contain images	PCR test	5644 individuals	..	..
DLAI3 Hackathon Phase3 COVID-19 CXR Challenge	Jul 13, 2020	Open access	Kaggle	Chest x-ray images	..	..	..	..	..
ECG Images dataset of Cardiac and COVID-19 Patients	Nov 12, 2020	Open access	Mendeley data	Electrocardiogram	The dataset does not contain images	..	..	..	..
Epidemiological data from the COVID-19 outbreak, real-time case information	March 13, 2020	Open access	Figshare	Epidemiological data specifically relating to COVID-19 infection	The dataset does not contain images	Not applicable	..	..	..
Epidemiology of COVID-19 inpatients	Unknown	Open access	Github	EHR or medical records data; other laboratory research data	The dataset does not contain images	PCR test	5320 individuals	Yes	..
Eurorad	Unknown	Open access	Other	Chest x-ray images; CT chest images; EHR or medical records data	Dataset is a platform for sharing clinical cases, and each image is associated with a vignette submitted by authors	Various methods	..	..	..
Extensive and Local Phase Enhanced COVID-19 X-Ray	Unknown	Open access	Kaggle	Chest x-ray images	..	..	4038 images from 2006 individuals	..	..

(Table continues on next page)

	Date dataset first published	Open or managed access	Hosting platform	What is this dataset composed of?	If medical images are included in the dataset and are labelled, does the dataset specify how these labels were assigned?	How was COVID-19 diagnosis confirmed?	How many images or individuals are represented in this dataset?	Did the dataset receive ethical approval or waiver by an ethics committee, IRB, or similar?	Did patients give their consent for data to be included in this dataset?
(Continued from previous page)									
Extensive COVID-19 X-Ray and CT Chest Images Dataset	June 5, 2020	Open access	Mendeley data	Chest x-ray images; CT chest images	Images are unlabelled	..	9544 chest x-rays and 8055 CTs, unknown number of individuals	..	..
Figure 1 COVID-19 Chest X-ray Dataset Initiative	May 9, 2020	Open access	Github	Chest x-ray images	..	..	55 images from 48 individuals	..	..
GeneSignaturesCOVID19	Unknown	Open access	Github	Genomics data	The dataset does not contain images	..	..	..	..
Genome-wide DNA methylation analysis of COVID-19 severity	May 31, 2021	Open access	Other	Genomics data	The dataset does not contain images	PCR test	128 individuals	Yes	Yes
Host methylation predicts SARS-CoV-2 infection and clinical outcome	Sept 20, 2021	Open access	Other	Other laboratory research data	The dataset does not contain images	PCR test	525 samples (not stated if each sample is from an individual patient)	Yes	Yes
Host transcriptomic profiling of COVID-19 patients with mild, moderate, and severe clinical outcomes	Dec 19, 2020	Open access	Other	Other laboratory research data	The dataset does not contain images	PCR test	50 individuals	Yes	No
HUST-19	Aug 18, 2020	Open access	Other	CT chest images	Reported by health-care professional, but speciality or experience not stated	CT chest images	104 individuals	..	..
Images of COVID-19 positive and negative pneumonia patients	May 20, 2020	Open access	Mendeley data	CT chest images	Reported by radiology expert	CT chest images	8367 rows (assume CT slices), unknown number of individuals	..	..
Impact of lung segmentation on the diagnosis and explanation of COVID-19 in chest X-ray images	Unknown	Open access	Github	Chest x-ray images	..	..	..	No	No
In vivo antiviral host response to SARS-CoV-2 by viral load, sex, and age (four datasets in one)	Sept 8, 2020	Open access	Other	Genomics data	The dataset does not contain images	PCR test	493 samples from 430 individuals	Yes	No
Inf-Net (Automatic COVID-19 Lung Infection Segmentation from CT Images)	May 22, 2020	Open access	Github	CT chest images	Reported by health-care professional, but speciality or experience not stated	..	1700 images	..	..
LA-DNN for COVID-19 diagnosis	May 7, 2020	Open access	Github	CT chest images	Reported by health-care professional, but speciality or experience not stated	..	746 images	..	..
Large COVID-19 CT scan slice dataset	Unknown	Open access	Kaggle	CT chest images	Reported by radiology expert	..	17 104 images from 1130 individuals	..	..
Large-scale single-cell analysis reveals critical immune characteristics of COVID-19 patients	Unknown	Open access	Other	Other laboratory research data; transcriptomics	The dataset does not contain images	..	..	..	..
Lipták et al, 2022	March 21, 2022	Open access	Other	EHR or medical records data	The dataset does not contain images	PCR test	680 individuals	Yes	Yes
MIDRC-RICORD-1A	Dec 18, 2020	Open access	Other	CT chest images	Reported by radiology expert	PCR test	110 individuals	..	..

(Table continues on next page)

	Date dataset first published	Open or managed access	Hosting platform	What is this dataset composed of?	If medical images are included in the dataset and are labelled, does the dataset specify how these labels were assigned?	How was COVID-19 diagnosis confirmed?	How many images or individuals are represented in this dataset?	Did the dataset receive ethical approval or waiver by an ethics committee, IRB, or similar?	Did patients give their consent for data to be included in this dataset?
(Continued from previous page)									
MIDRC-RICORD-1B	Feb 5, 2021	Open access	Other	CT chest images	Reported by radiology expert	PCR test	120 images from 117 individuals	..	..
MIDRC-RICORD-1C	Jan 15, 2021	Open access	Other	Chest x-ray images	Reported by radiology expert	PCR test	998 images from 361 individuals	..	..
mmc2.xlsx	Jul 11, 2022	Open access	Other	EHR or medical records data; other laboratory research data	The dataset does not contain images	PCR test	455 individuals	Yes	Yes
Mosmed COVID-19 CT Scans	Unknown	Open access	Kaggle	CT chest images	Images are unlabelled	CT chest images	1000 images from 1000 individuals	..	..
MosMedData: Chest CT Scans with COVID-19 Related Findings	April 25, 2020	Open access	Other	CT chest images	..	CT chest images	1110 CT studies from 1110 individuals	No	No
Multi-omic profiling reveals widespread dysregulation of innate immunity and hematopoiesis in COVID-19	May 8, 2021	Open access	Other	Genomics data; proteomics data; other laboratory research data; transcriptomics	The dataset does not contain images	PCR test	70 samples from 14 individuals	Yes	Yes
No dataset name (COVID-19 training and test dataset [appendix 1 and 2])	Dec 23, 2020	Open access	Other	EHR or medical records data	The dataset does not contain images	..	Training (375 individuals); testing (106 individuals)	Yes	No
Novel COVID-19 Chestxray Repository	Unknown	Open access	Kaggle	Chest x-ray images	..	..	3975 images, unknown number of individuals	..	..
Olink Proteomics Dataset	May 18, 2021	Managed access	Other	Proteomics data	The dataset does not contain images	..	384 individuals	..	..
Open COVID-19 Data Working Group	Unknown	Open access	Github	EHR or medical records data	The dataset does not contain images	PCR test	..	..	..
Per-COVID-19: A Benchmark Dataset for COVID-19 Percentage Estimation from CT-Scans	Aug 13, 2021	Open access	Github	CT chest images	..	..	3986 images, unknown number of individuals	No	No
Pneumonia (virus) vs COVID-19	Unknown	Open access	Kaggle	Chest x-ray images	Images are unlabelled	..	1563 images	..	..
Pre_Surv_COVID_19	April 25, 2020	Open access	Github	EHR or medical records data; other laboratory research data; LDH; lymphocyte; CRP	The dataset does not contain images	..	485 individuals	..	..
PROCOVID dataset	Aug 1, 2022	Open access	Zenodo	EHR or medical records data; other laboratory research data	The dataset does not contain images	..	..	..	..
Proteomic and Metabolomic Characterization of COVID-19 Patient Sera	May 27, 2020	Open access	Other	Genomics data; proteomics data	The dataset does not contain images	PCR test	99 individuals	Yes	No
Proteomics and Metabolomics for COVID-19 patient sera	May 27, 2020	Open access	Other	Genomics data; proteomics data	The dataset does not contain images	PCR test	99 individuals	Yes	No
public_dataset-clustered-ct-scans-for-lobes-ensemble	Unknown	Managed access	Other	CT chest images	..	..	..	..	..

(Table continues on next page)

	Date dataset first published	Open or managed access	Hosting platform	What is this dataset composed of?	If medical images are included in the dataset and are labelled, does the dataset specify how these labels were assigned?	How was COVID-19 diagnosis confirmed?	How many images or individuals are represented in this dataset?	Did the dataset receive ethical approval or waiver by an ethics committee, IRB, or similar?	Did patients give their consent for data to be included in this dataset?
(Continued from previous page)									
QaTa-COV19 Dataset	Jan 1, 2021	Open access	Kaggle	Chest x-ray images	..	..	..	..	..
Radiologists' Annotations on COVID-19+ X-rays	Oct 6, 2020	Open access	Other	Annotations of radiology images (images located in separate place on the internet; BIMCV)	Reported by radiology expert	..	..	..	..
Radiopaedia*	Unknown	Open access	Other	Chest x-ray images; CT chest images	Reported by health-care professional, but speciality or experience not stated	Documented on a case-by-case basis	..	No	..
Radiopaedia COVID-19	Unknown	Open access	Other	Chest x-ray images	..	CT chest images	17 images	..	..
Radiopeadia_Covid_Data	Unknown	Open access	Github	Written case report data from radiopedia website	The dataset does not contain images	..	..	..	..
RAIG COVID19 compiled dataset	March 3, 2021	Open access	Figshare	Chest x-ray images; CT chest images	..	..	3900 images	..	..
RYDLS-20	Unknown	Open access	Other	Chest x-ray images	..	..	..	..	..
SARS-COV-2 Ct-Scan Dataset	May 14, 2020	Open access	Kaggle	CT chest images	..	..	120 images, unknown number of individuals	Yes	No
SARS-CoV-2-RBV1.sav	Jan 18, 2023	Open access	Mendeley data	EHR or medical records data; other laboratory research data	The dataset does not contain images	PCR test	5296 individuals	Yes	No
SARS-CoV-2-RBV2.sav	Jan 19, 2023	Open access	Mendeley data	EHR or medical records data; other laboratory research data	The dataset does not contain images	PCR test	3899 individuals	Yes	No
SARS-CoV-2-RBV3.sav	Jan 20, 2023	Open access	Mendeley data	EHR or medical records data; other laboratory research data	The dataset does not contain images	PCR test	2597 individuals	Yes	No
SDY1708	March 1, 2020	Open access	Other	Genomics data; proteomics data; other laboratory research data (transcriptomics)	The dataset does not contain images	PCR test	64 individuals	..	..
SegmentedLungCXRs	April 22, 2020	Open access	Github	Chest x-ray images	..	..	723 images	Yes	..
SIIM-FISABIO-RSNA COVID-19 Detection	Unknown	Open access	Kaggle	Chest x-ray images	Reported by radiology expert	RSNA and BIMCV used PCR test	10 178 images	No	..
Single cell transcriptomics of PBMCs during severe COVID-19	Jan 27, 2022	Open access	Other	Other laboratory research data (transcriptomics)	The dataset does not contain images	..	30 samples from 18 individuals	Yes	Yes
SIRM COVID-19 Database	March 3, 2020	Open access	Other	Chest x-ray images; CT chest images	Reported by radiology expert	No standardised method exists for whole dataset. Some cases are confirmed by nasal swabs, others by PCR	115 images	..	..
Spectrum of Viral Load and Host Response Seen in Autopsies of SARS-CoV-2 Infected Lungs	May 11, 2020	Open access	Other	Other laboratory research data	The dataset does not contain images	PCR test	24 individuals	Yes	No

(Table continues on next page)

	Date dataset first published	Open or managed access	Hosting platform	What is this dataset composed of?	If medical images are included in the dataset and are labelled, does the dataset specify how these labels were assigned?	How was COVID-19 diagnosis confirmed?	How many images or individuals are represented in this dataset?	Did the dataset receive ethical approval or waiver by an ethics committee, IRB, or similar?	Did patients give their consent for data to be included in this dataset?
(Continued from previous page)									
SPGC-COVID	Sept 18, 2021	Open access	Figshare	CT chest images	Reported by radiology expert	CT chest images	130 individuals	..	..
The interferon landscape along the respiratory tract impacts the severity of COVID-19	Aug 25, 2021	Open access	Other	Other laboratory research data (transcriptomics)	The dataset does not contain images	..	..	..	..
Thread of COVID-19 CXR (all SARS-CoV-2 PCR+)	April 28, 2020	Open access	Social network posts	Chest x-ray images	..	..	50 individuals	..	..
TSG-ICU	Unknown	Open access	Github	EHR or medical records data	The dataset does not contain images	..	..	Yes	No
Type I interferon pathways in SARS-CoV-2-infected individuals from Pakistan	Dec 1, 2021	Open access	Other	Other laboratory research data (transcriptomics)	The dataset does not contain images	PCR test	..	Yes	Yes
UESTC-COVID-19 Dataset	Dec 31, 2020	Managed access	Other	CT chest images	Automatically labelled via AI model or similar	..	120 individuals	..	..
Ultrasound in COVID-19	Unknown	Open access	Other	Ultrasound thorax images	Reported by health-care professional, but speciality or experience not stated	..	..	..	..
underdiagnosis_of_covid_19_cases_in_brazil	Unknown	Open access	Github	Unable to determine from dataset	The dataset does not contain images	..	..	..	..
VAERS	NA (dataset released annually)	Open access	Other	Vaccine adverse event data	The dataset does not contain images	..	..	..	..
Welltory COVID-19 and Wearables Open Data Research	Unknown	Open access	Github	EHR or medical records data; heart rate variability measurements from wearable devices	The dataset does not contain images	..	185 individuals	Yes	Yes
X-ray images three levels	June 9, 2021	Open access	Figshare	Chest x-ray images	Images are unlabelled	Chest x-ray images	5935 images	..	..
Yan et al, 2020	May 14, 2020	Open access	Other	EHR or medical records data	The dataset does not contain images	..	485 individuals	Yes	..

References to the included datasets can be found in appendix 2. Names of datasets are exactly as given on the hosting platform (including case, spelling, and punctuation), or are defined by the title or the author name and year of the study in which the dataset is reported. BIMCV=El Banco digital de Imagen Médica de la Comunidad Valenciana. CRP=C-reactive protein. EHR=electronic health record. GAN=generative adversarial network. IRB=institutional review board. LDH=lactate dehydrogenase. NA=not applicable. SIRM=Società Italiana di Radiologia Medica e Interventistica. *This is a non-exhaustive list as not every case on Radiopaedia is inspected.

***Table*: Details of the 192 COVID-19 datasets analysed for this systematic review, listed in alphabetical order by name**

See Online for appendix 2

the Good Machine Learning Principles and the UK Equality Act 2010;[27,38] clinical data (including how any ground truths were established); the formats and types of data included; details of the dataset's creators; and ethical considerations. If some form fields could not be completed because data were not available, these fields were left blank during data capture and analysis. Any unclear data were summarised in a comment box within the data extraction form; all of these were reviewed by JEA and are provided verbatim with other data extracted from studies in the table and appendix 2.

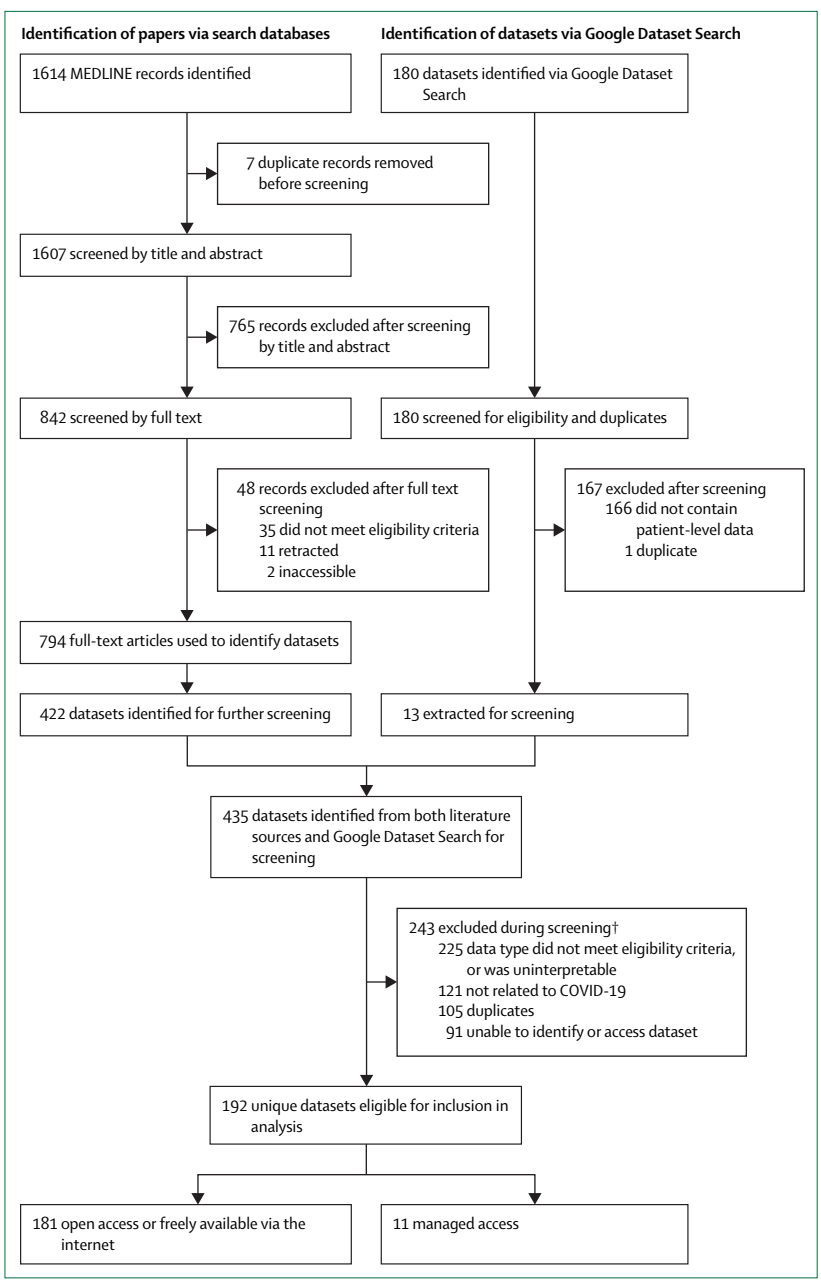

Figure 1: **PRISMA diagram showing identification of datasets from search databases and Google Dataset Search**

Articles which detailed COVID-19 datasets or described the use of datasets to train or test machine-learning algorithms for COVID-19 were included. Articles exploring only population health, epidemiology, or COVID-19 forecasting were excluded. *Multiple papers cited the same datasets, and some papers cited multiple datasets. †Some datasets had more than one reason for exclusion.

Datasets were excluded if: they were inaccessible despite attempts made by the reviewers to locate the dataset, including emailing the first and senior authors of the article describing the dataset; data contained within the dataset were uninterpretable despite translation using Google Translate; they were found to be entirely composed of data from another dataset (in which case the records were merged); the data did not to relate to COVID-19; or the dataset contained only the following data types: continuous waveform data, compilations of published papers or other written information only, population surveillance data not relating to individual participants, data relating only to non-humans, or basic science laboratory data relating to cultured cell lines rather than pathology samples. We did not consider the risk of bias in screened or included articles themselves; instead, this review focused on bias and limitations at the dataset level. Risk of bias was assessed via analysis of datasets' documentation and completeness of metadata reporting rather than using specific toolkits. Data synthesis was not relevant for this systematic review, although simple summary statistics (frequencies and percentages) are reported. Figures were created using Google Drawings and Google Sheets (Alphabet, Mountain View, CA, USA), and R version 4.1.1 for Mac and the packages tidyverse, magrittr, and readxl.

Patient and public involvement

Two patient and public representatives assisted with the design of this project, including developing the initial application for funding. One of these members remained involved throughout the project as a co-investigator. A patient and public involvement and engagement (PPIE) group was established as part of the overarching STANDING Together programme in April, 2022, and this group began regular quarterly meetings from June, 2022 onwards. The researcher team presented project updates at these meetings, and patient and public involvement and engagement group members were asked to contribute to decision making about the conduct of the review (for instance, which characteristics to include in data extraction). Patient and public involvement and engagement group members also contributed to the dissemination plan for this project and for STANDING Together more broadly.

Results

Datasets identified from the literature search

The initial MEDLINE search identified 962 initial records, and a further 652 after the search was updated. After screening by title and abstract, 765 records were excluded as they did not meet inclusion criteria. Following this, 422 datasets were identified from 794 unique full-text literature sources. A targeted search

for datasets via Google Dataset Search identified 180 datasets, with 13 meeting inclusion criteria. A total of 435 datasets were screened; 243 were duplicates, inaccessible, or did not meet eligibility criteria and were excluded, leaving 192 datasets for final review (figure 1). A summary of the included datasets is provided in the table, with additional information for each provided in appendix 2.

Documentation accompanying datasets

Most datasets (179 [93%] of 192) provided some form of documentation describing the dataset (figure 2). Of these 179 datasets, 120 (67%) were documented in a single resource (eg, a readme file, or a scientific article), and 59 (33%) had dataset documentation that spanned multiple resources. The most common format for dataset documentation was a readme file or similar electronic information (157 [88%]). 69 (39%) datasets were accompanied by a scientific article (published paper or preprint), but in many cases the article described a trained model rather than the dataset itself, and in some cases the dataset had no link back to the published article. No datasets were accompanied by a datasheet, healthsheet, or related transparency artifact.[39,40] In many cases datasets' documentation was not a single document, and instead relevant information was split across multiple sources.

Composition of datasets

Reporting of the geographical location that included data was obtained from was variable. Of 192 datasets, 92 (48%) provided specific information on the geographical origin of their data, with the remaining 100 (52%) providing no information, or insufficient information to determine precisely where data originated (figure 2). Of the 92 datasets providing precise location information, 14 (15%) contained data from multiple countries, and 78 (85%) from a single country. In total, 72 countries were represented in the datasets analysed, with China the country most represented (figure 3). Overall, six countries accounted for over 50% of the available datasets (China, USA, Italy, Spain, Brazil, and Iran; appendix 1 pp 4–6).

No information was provided about the specific location of the dataset curation team for 84 (44%) datasets, and of the 108 datasets that did provide this information, 94 (87%) were created by teams in a single country, and 14 (13%) by teams spanning multiple countries (figure 2).

The reporting of demographic attributes of individuals represented within datasets varied widely (figure 2). Most of the 192 datasets did not report any attributes, with only 82 (43%) datasets reporting age, 41 (21%) reporting sex, 28 (15%) reporting gender, and under 10% reporting race (16 [8%]) and ethnicity (14 [7%]). We do not provide summary statistics for these attributes because the high proportion of missing data would mean such values are likely to be misleading.

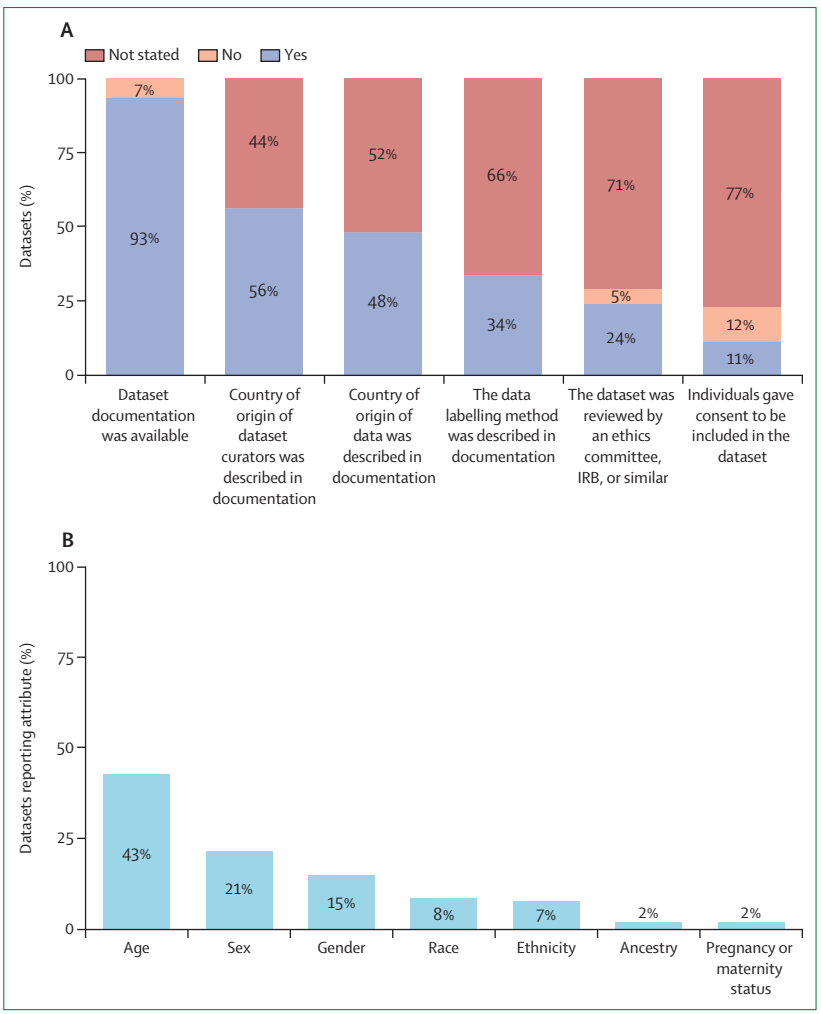

Figure 2: **Summary of the documentation and composition of COVID-19 datasets**
(A) Although most of the 192 datasets analysed had some form of documentation, only around half included information about the geographical location of dataset curation teams or the origin of the data. Most datasets did not describe how data were labelled, whether the dataset had been ethically reviewed, or whether individuals whose data were included had given consent for this. (B) Of the 192 datasets analysed, 82 reported the age of individuals included (43%), 41 reported sex (21%), and 28 reported gender (15%). Only one dataset reported both sex and gender, and 23 datasets (12%) reported information that could have pertained to either sex or gender but did not specify which was reported. Under 10% of datasets reported race or ethnicity, and there was substantial heterogeneity in the groupings used, meaning that comparison between datasets was not possible. No datasets reported data relating to sexual orientation, disability, marriage or civil partnership, or religion or beliefs. IRB=institutional review board.

75 (39%) of the 192 included datasets presented data that were sourced from other previously published datasets, in some cases having been processed (eg, segmentation of a previously uploaded CT imaging dataset). Documentation for 23 (12%) of the 192 datasets was insufficient to assess whether the data were novel or not. In many cases, precise details of the provenance and origin of the data were incompletely reported, or there were inconsistencies between the information listed in different places relating to the same dataset.

At least nine datasets reproduced images from the Chest X-ray Images (Pneumonia) dataset hosted on

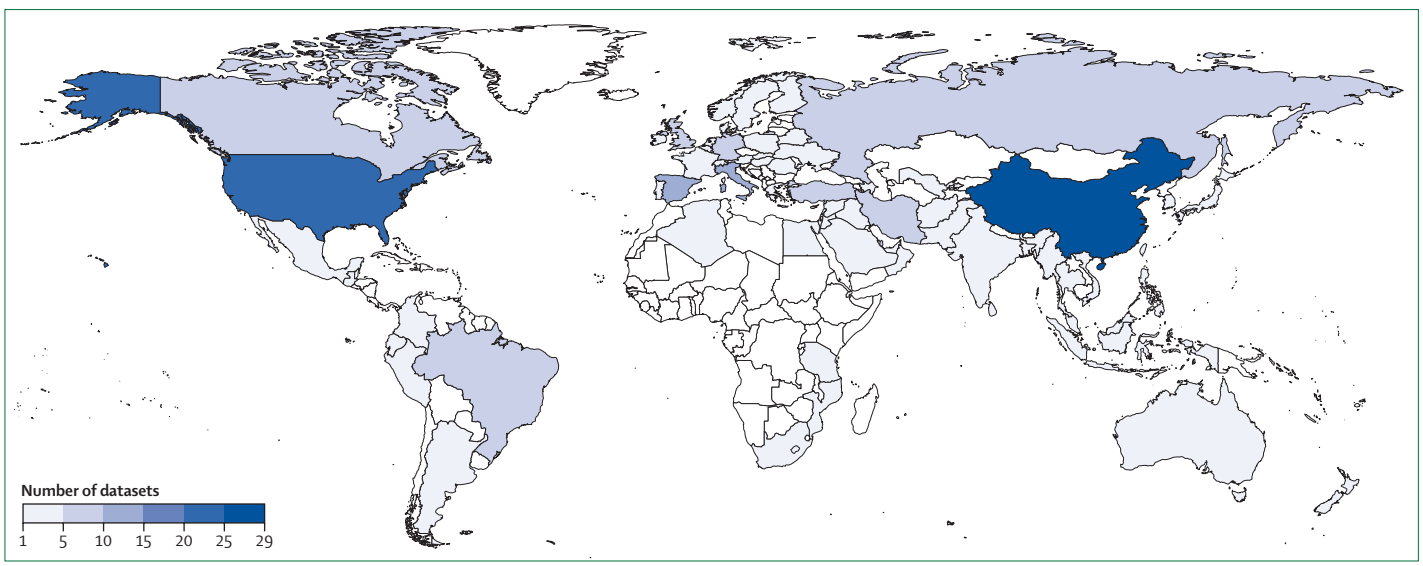

Figure 3: **World map showing the numbers of datasets identified by country**
Datasets reporting data from multiple countries were counted once for each country listed. China was the most represented country (29 datasets), followed by the USA (25 datasets), and Italy and Spain (12 datasets each). In total, individuals from 72 countries were represented in the datasets. A full breakdown of the number of datasets by country is provided in appendix 1 (pp 4–6).

Kaggle,[41] which was created before the onset of the COVID-19 pandemic (figure 4). The chest x-ray images in this dataset originated from a paper published in 2018 by Kermany and colleagues,[42] and are entirely composed of paediatric chest x-ray images. None of the nine datasets that we identified as reproducing these images included this information, despite the documentation for this dataset stating that images, "were selected from retrospective cohorts of pediatric patients of one to five years old from Guangzhou Women and Children's Medical Center, Guangzhou". Many of these nine datasets were themselves widely reused in other databases: for example, the COVID-19 Radiography Database, one of the databases that cited the Chest X-Ray Images (Pneumonia) dataset, was itself cited by 74 of the 794 full-text articles we reviewed.[43] A summary of deficiencies in the documentation of these nine datasets is provided in appendix 1 (pp 7–11). One dataset reported a date of publication, seven acknowledged that data had been reused, and six provided a link or citation to the dataset's source. There were several examples of data duplication, where datasets containing the same data were combined, risking instance amplification and train-test contamination in downstream aggregated datasets. Five datasets also reported sampling from the covid-chestxray-dataset,[44] which was itself widely reused, increasing the likelihood of unrecognised data duplication.

Data access and hosting information

Most datasets were freely available via the internet (181 [94%] of 192), in some cases with specific open-access licenses. The most used platforms to host these datasets were Kaggle (46 [25%] of 181) and Github (45 [25%] of 181).

11 (6%) datasets managed access to their data, requiring data access requests or other forms of approval. In one case access to the data would only be permissible after submitting a proposal for a governance review process, signing a data-sharing agreement, and by using a dedicated trusted research environment, the cost of which needed to be paid by the data user. For this review, we did not need to access the data itself.

Nature of data within datasets

Most datasets (134 [70%] of 192) were composed entirely or in part of thoracic imaging data. Of these, 88 (66%) included chest x-ray images, 61 (46%) CT images of the thorax, and four (3%) thoracic ultrasound images, with some datasets including images from multiple imaging modalities. Of the 134 chest imaging datasets, 116 (87%) were labelled, but of these 77 (66%) gave no information about how labelling was performed (figure 2). Labelling was performed by radiologists in 23 (17%) of the 134 thoracic imaging datasets (17·1%), by unspecified health-care professionals in 12 (9%) datasets, and via automated approaches (including AI models) in three datasets (2%). Other non-imaging data types included in datasets included electronic health record data, electrocardiogram data, genomics, proteomics, transcriptomics, and other basic science data.

Consent and ethical considerations

55 (29%) of 192 datasets included information on ethical approval or waiver provided by an ethics committee, institutional review board, or similar, with nine

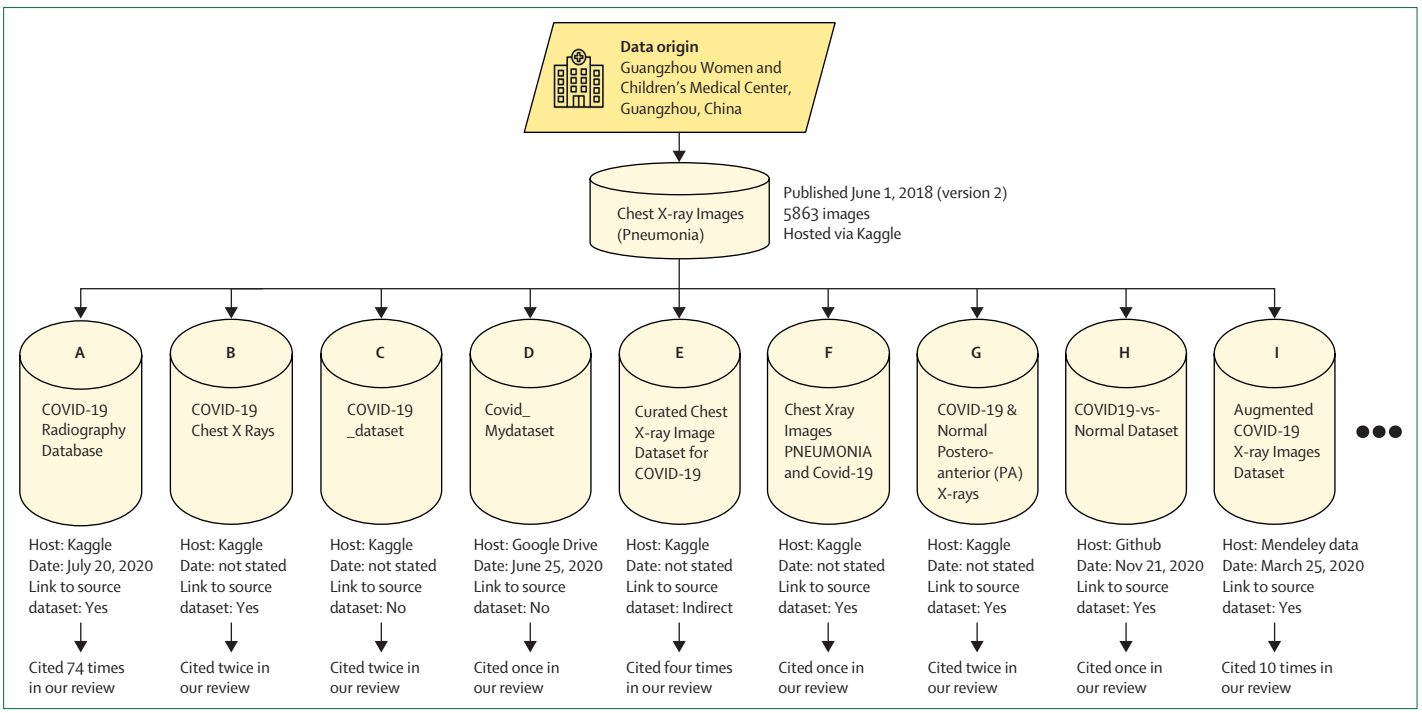

Figure 4: **Re-use of the Chest X-Ray Images (Pneumonia) dataset and its derivatives**
Nine datasets (represented from A–I) were identified as containing chest x-ray images sourced from the Chest X-ray Images (Pneumonia) dataset hosted on Kaggle.[41] None of these datasets reproduced information relating to the origin of the chest x-ray data, specifically reporting that the images were obtained before the COVID-19 pandemic, and that they were all paediatric images, despite this information being provided in the documentation for the Chest X-ray Images (Pneumonia) dataset. Six datasets (A,B,F,G,H, and I) provide a direct link back to the Chest X-ray Images (Pneumonia) dataset in their documentation, and two (C and D) provide no information about the source data. Dataset E is duplicated, with copies both on Kaggle and Mendeley Data. Metadata is provided on the Mendeley data page that are missing from the Kaggle page. Aside from dataset H, all other datasets mixed images from multiple sources, so it is not possible to determine how many of the 5863 images were reproduced in each of the downstream datasets (appendix 1 pp 7–11). Overall, these nine datasets were cited 97 times by included articles in our review.

(5%) stating that no ethical approval or waiver had been granted. The remaining 137 (71%) of 192 provided no information about ethical review. 21 (11%) datasets reported that people whose data were included had given consent for data sharing, 23 (12%) reported that no consent had been given, and the remaining 148 (77%) did not include any information regarding consent (figure 2).

Discussion

This systematic review aimed to characterise the composition and reporting of publicly available datasets related to COVID-19 that were used to develop AI models. Specifically, the reporting of demographic attributes was assessed to characterise who was represented in these datasets and how, and which groups were under-represented. We found that for most datasets the documentation was inadequate to answer these questions because the demographic attributes of individuals included were not reported. We additionally highlight several themes observed frequently amongst the datasets reviewed.

Data remixing

Data reuse was common, with a substantial proportion of datasets either entirely or partly composed of data from other datasets—a trend termed remixing.[23] When data were reproduced, metadata were sometimes removed, meaning that individual data instances were not traceable back to their origin. Accordingly, the combination of multiple datasets into a remix dataset, each of which might also have originally been a remix, could amplify particular data instances. A related risk is that the same instances can appear in seemingly independent datasets used for training and testing a model. Together, these could result in overfitting that goes undetected, impairing real-world performance and affecting safety of regulated medical devices that are developed using these data. Notably, independence of training and test datasets is one of the guiding principles underpinning good machine learning practice for medical device development.[27]

The most reproduced dataset was the Chest X-ray Images (Pneumonia) dataset,[41] which was created from a study published by Kermany and colleagues,[42] and is entirely composed of paediatric images. None of the nine datasets that reproduced these images included this information, and in some cases combined normal paediatric X-rays with abnormal adult X-rays to create a new remix dataset. Paediatric and adult chest x-ray images are fundamentally different, and models trained on such a dataset might learn features distinguishing adult from

paediatric X-rays, rather than identifying lung injury caused by COVID-19—resulting in underperformance if they were used in practice. The unexpected presence of paediatric images would be immediately obvious to most health-care professionals, highlighting the need for diverse teams, including domain experts, in the creation of datasets.

Although intuitively it might seem efficient to remix datasets with the intention of compensating for biases or boosting representation of rare classes, doing so risks any biases compounding and multiplying rather than cancelling out.[45] In addition to these technical risks, dataset remixing could carry legal risk. For some datasets a specific licence was specified stating the conditions under which data were made available (eg, those specified by Creative Commons), whereas for other datasets no such licence was provided. We identified remix datasets that were made available with an open-access licence that was different to the licence of the sources they had been created from. In many jurisdictions datasets can be protected by copyright and related law; such protections mean that in the absence of a formal licence, data reuse could be contested by the owner (usually the original dataset creator). When existing discrete datasets are insufficient for a particular purpose it will often be optimal to collect new data; however, doing so comes with substantial resource requirements and might not be feasible or possible. Dataset remixing might be appropriate in some cases if conducted transparently with clear documentation of the means of selection, sampling, processing, and aggregating the data sources used. Preservation of any dataset documentation, use of file types with embedded metadata, and adherence to FAIR (findable, accessible, interoperable, and reusable)[46] principles might also reduce the risk inherent to remixing when its use cannot be avoided.

Under-representation and other data limitations

AI models created using datasets that under-represent particular groups are less likely to perform effectively for these groups; if datasets do not report these attributes this under-representation will be invisible, masking any resulting model under-performance.[20] Although most datasets were accompanied by some form of documentation, limitations and biases were rarely acknowledged. Demographic attributes of individuals (including age, sex, gender, race, and ethnicity) were inconsistently reported, and often missing entirely. Less than 25% of datasets reported sex or gender, and several of these conflated these related but separate social constructs. Categories used for race and ethnicity vary between countries, and therefore also varied between datasets, rendering comparisons in reporting of these attributes between datasets impossible—however, very few datasets reported race or ethnicity at all. The health consequences of COVID-19 disproportionately affected minoritised racial and ethnic groups and those experiencing greatest socioeconomic deprivation; had AI health technologies been deployed at scale any underperformance would likely have disproportionately affected these same groups, compounding the health inequity they already experience. The findings in this Review are consistent with similar reviews relating to other health conditions, highlighting that deficiencies in reporting of demographic attributes are not unique to COVID-19.[24,25]

Infectious diseases such as COVID-19 have inherently unstable outbreak patterns, and countries approve and deploy interventions at different speeds. As such, metadata relating to the geographical and temporal origin of a dataset is crucial to contextualising and accounting for data and distribution shifts. This review found that over half of datasets did not report which country data were collected in, and many did not report their date of publication (including every dataset hosted on Kaggle, which reports only how old a dataset is, rather than when it was published. In some cases it was possible to ascertain a precise publication date from other sources of information, for example a linked scientific paper). Some datasets did provide this information, either for the whole dataset or for each individual instance within it (eg, giving the timestamp and location where a particular instance was recorded). When information was provided about a dataset's publication date, the date format was often not specified—differing formats are used internationally. Data and distribution shifts are likely to be particularly pronounced for remix datasets, in which data aggregated from multiple sources and epochs might have fundamentally different contexts, particularly if data from before the pandemic were included.

Data labelling can be an important source of bias, but most labelled datasets did not report how labels were acquired. When information was provided there was substantial heterogeneity in practice, and in many cases, very large numbers of images were labelled by a small team of labellers. Minimal information was provided about their expertise or whether interobserver or intraobserver agreement had been assessed. Also, over half of the datasets did not describe how COVID-19 was tested for—the ground truth against which models' predictions are evaluated. Many laboratories will have used RT-PCR, as this is established as the definitive diagnostic test for respiratory viral pathogens. However, specific laboratory protocols and reagents will have varied, and in some cases, positive diagnosis might have been made using self-administered antigen tests or using clinical features alone. Different methods of diagnosing COVID-19 will have different error rates, and so the reliability of this ground-truth label might vary across studies.

Data access

Of the 192 datasets included in this review, 181 were freely available via the internet and hosted across a range

of platforms including Kaggle, Github, Mendeley Data, Figshare, Zenodo, and others. Some articles used data from continually updated public platforms, but did not specify the search string used, or the time and date stamp for when data were extracted. Some datasets were hosted via file-sharing websites (including Dropbox and Google Drive), and in many cases, the URL to the dataset provided in a scientific article was no longer accessible. Many platforms used to share datasets were not specifically designed for health-care purposes, which could contribute to the gaps in medical metadata reporting we show in this Review. In addition, many platforms do not track changes made to dataset documentation—in some cases available information about a dataset had been updated or altered between this Review being submitted and the final version being published. When datasets were inaccessible, the first and senior authors of the citing paper were emailed asking for assistance—in all but one of these scenarios, we received no response to this email or to a repeat email sent 1 month later. We also found that numerous URLs to Kaggle datasets were broken. This issue was readily solved by manually adding "/datasets/" to the URL, but is still an example of the potential fragility associated with online open-access dataset hosting platforms.

Some of the freely available online datasets had very similar names and many had no permanent identifier (eg, a digital object identifier), making it difficult to distinguish one dataset from another. We found instances in which the same datasets were fully or partly replicated by the same curation teams across multiple repositories, often with similar or identical names. In many cases the primary intent of those sharing datasets was to enable reproduction of their published research rather than to promote data reuse for other purposes—poor reproducibility has been widely recognised as a problem in the machine learning literature.[47] Nonetheless, our review found that these datasets were often reused by other researchers. We found multiple examples of scientific articles and datasets that were related, but where one failed to adequately refer to the other, meaning that those using a dataset might not know that crucial context is provided in an unlinked accompanying paper. A key message is that regardless of the original intent for data being shared, it has the potential to be used for multiple different purposes. Accordingly, including comprehensive documentation is always needed wherever a dataset is published.

Ethical considerations

Many datasets included a suggested citation format in their accompanying documentation; however, a minority requested that data users cite one or more papers that were unrelated to the dataset, with the resultant effect being that the citation count for these unrelated papers could be inflated. Most datasets reviewed did not refer to ethical review, data sharing agreements, consent from those whose data were included, or adherence to data protection laws. Although international collaboration between high-incoe countries and those with fewer resources is well intentioned, data sharing across borders requires safeguards to prevent data colonialism—the exploitation of marginalised and disempowered individuals and communities whose data could be acquired and used without their knowledge or explicit consent.[48,49]

Limitations of this study, and future directions

Although this review of COVID-19 datasets used a robust systematic search strategy combining resources from both MEDLINE and Google Dataset Search, the list of datasets we reviewed is likely not a complete list of all those available. The reasons for why this list might not be complete include that MEDLINE does not provide total coverage of biomedical literature, because we limited our search to articles in the English language, and because datasets that were not cited in scientific articles would not have been identified by our search. However, the intention of this Review was to highlight key limitations in the composition and reporting of datasets used in real-world research, rather than to give a precise analysis of the numbers involved. All search strategies reflect a trade-off between minimising the risk of missing relevant articles versus increasing the number of articles requiring screening. Our search strategy (including the choice of databases and search terms used) was designed to achieve a pragmatic balance. The narrative nature of this Review means that the effect of a small number of missed articles on our results is likely to be small. The data presented in this Review paper reflect the information available in datasets' documentation, which in some cases might not be an accurate or contemporaneous summary of the dataset itself. The lessons we highlight have been shown across datasets related to several other health conditions and are consistent with another review of COVID-19 datasets.[23] Accordingly, we advocate that there are lessons to be learned from the data presented in this Review, notwithstanding its methodological limitations. Qualitative or mixed methods studies could enable greater understanding of the factors that contribute to the dataset limitations we have described herein.

Despite the problems with datasets that we have discussed, we found many examples of good practice, including comprehensive and transparent documentation practices, but in many cases these datasets were not the most frequently cited. We believe that transparent communication of dataset composition, including biases and limitations, can mean AI health technologies are developed with the most appropriate (rather than the most popular) datasets. In times of crisis (such as in the pandemic, when usual safeguards and protections for patients might be less easy to enforce), it is particularly important that all stakeholders involved in dataset

creation and use, act in accordance with best practices.[50] The findings from this Review have contributed to the development of the STANDING Together recommendations for the documentation and use of health-care datasets. We advocate that both dataset creators and data users contribute to comprehensive dataset characterisation. Accompanying a health-care dataset with a Healthsheet artifact[40] and adhering to the STANDING Together recommendations[51] will provide data users assurance that a dataset is fit for their purpose, and enable the development of AI health technologies that are safe and effective for everyone in society.

Contributors

XL and AKD led the initial idea development and acquired funding for this work in collaboration with MCa, RNM, MDM, and MG. The protocol was authored by XL, MCh, GS, JEA, JP, and AKD. Searches and screening of articles and datasets were conducted by GS, MCh, JEA, XL, ELa, SV, VM, QM, and ELe. Data extraction was completed by GS, MCh, JEA, XL, ELa, SV, VM, QM, and ELe. Data analysis was conducted by all authors, and led by JEA, MCh, and XL. The initial manuscript was drafted by JEA, MCh, GS, ELa, JP, AKD, and XL. The submitted version of the manuscript was edited by all authors. JG is a public contributor to this project as well as being co-investigator. JEA and XL are responsible for the content as guarantors. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Declaration of interests

JEA is a named researcher on grants from the Medical Research Council (MRC) and the Engineering & Physical Sciences Research Council with payments made to his institution for the delivery of other research projects; and is the co-organiser of Alan Turing Institute Clinical AI interest group (unpaid). ELe has received grants from the NHS AI Lab. MCa has received grants from the National Institute for Health and Care Research (NIHR), Health Data Research UK (HDR-UK), Innovate UK, Macmillan Cancer Support, GlaxoSmithKline, UCB Pharma, Research England as part of United Kingdom Research and Innovation (UKRI), European Commission, European Federation of Pharmaceutical Industries and Associations, Brain Tumour Charity, Gilead, Janssen, UKRI, and Merck for the delivery of other research projects; has received payment for delivering lectures from the University of Maastricht; has received a speaker fee from Cochrane Portugal; payments for reviewing from the South-Eastern Norway Regional Health Authority and Singapore National Medical Research Council; consulting fees from Aparito, CIS Oncology, Takeda, Merck, Daiichi Sankyo, Glaukos, GlaxoSmithKline, Patient-Centered Outcomes Research Institute, Genentech, Vertex, ICON, Halfloop, and Pfizer; and is associated with Proteus Consortium. MG has received grants from the Canadian Institute for Advanced Research, Helmsley Trust, Wellcome Trust, Moore Foundation, Volkswagen Foundation, I-Clinic, IBM-AI, Janssen research and development, Takeda, Quanta Computing, and Microsoft Research; and has acted as an advisor for the Symposium on Artificial Intelligence for Learning Health Systems and Conference on Health, Inference, and Learning. MDM has received grants from the Canadian Institutes of Health Research, AMS Healthcare, and the SickKids Foundation. JO works or has previously worked with AdvaMed AI Framework Group member (unpaid), AdvaMed Software Working Group member (unpaid), MedTech Europe AI Working Group member (unpaid), and MedTech Europe Digital Health Committee member (unpaid). CS has received grants from NIHR, UKRI, MRC, and NIHR Cambridge Biomedical Research Centre. RNM has received grants from MRC, British Heart Foundation, United States Agency for International Development, and HDR-UK. AKD has received grants from MRC and NIHR. XL has received grants from NIHR, Wellcome Trust, Research England, Moorfields Eye Hospital Charity; has received consulting fees from Hardian Health; has received payment or honoraria from the University of Turku; and is employed by Apple as a health scientist. All other authors declare no competing interests.

Acknowledgments

Detailed information relating to each included dataset is provided in appendix 2. The study protocol is available on request by emailing the corresponding author of this article. This Review is part of the STANDING Together initiative, funded by the NHS AI Lab and The Health Foundation, and supported by the NIHR (AI_HI200014). Grant funding for the current project was awarded by the NHS AI Lab and The Health Foundation to University Hospitals Birmingham NHS Foundation Trust and partnered academic institutions (applicants: XL, AKD, MCa, RNM, MDM, MG). The funders had no role in the study design or in its delivery. All researchers were independent from the funders. We wish to acknowledge the contributions of the STANDING Together working group, and the patient and public involvement and engagement group: Aaishah Aslam, Adewale O Adebajo, Alan Karthikesalingam, Cassandra H Leung, Darren Treanor, Elizabeth Sapey, Francis McKay, Heather Cole-Lewis, Jude Beng, Maxine Mackintosh, Negar Rostamzadeh, Neil Sebire, Russell Pearson, Samina Begum, and Stephen Pfohl. Some members of the working group and patient and public involvement and engagement group preferred to be acknowledged anonymously.

Editorial note: The Lancet Group takes a neutral position with respect to territorial claims in published maps.

References

1 Wang C, Horby PW, Hayden FG, Gao GF. A novel coronavirus outbreak of global health concern. *Lancet* 2020; **395:** 470–73.
2 Zhu N, Zhang D, Wang W, et al. A novel coronavirus from patients with pneumonia in China, 2019. *N Engl J Med* 2020; **382:** 727–33.
3 Wellcome. From equality to global poverty: the COVID-19 effects on societies and economies. June 29, 2021. https://wellcome.org/news/equality-global-poverty-how-covid-19-affecting-societies-and-economies (accessed Dec 21, 2023).
4 Miller IF, Becker AD, Grenfell BT, Metcalf CJE. Disease and healthcare burden of COVID-19 in the United States. *Nat Med* 2020; **26:** 1212–17.
5 Salyer SJ, Maeda J, Sembuche S, et al. The first and second waves of the COVID-19 pandemic in Africa: a cross-sectional study. *Lancet* 2021; **397:** 1265–75.
6 Sun J, He W-T, Wang L, et al. COVID-19: epidemiology, evolution, and cross-disciplinary perspectives. *Trends Mol Med* 2020; **26:** 483–95.
7 WHO. WHO coronavirus (COVID-19) dashboard. https://covid19.who.int/ (accessed Nov 7, 2023).
8 Enserink M. Dispute simmers over who first shared SARS-CoV-2's genome. *Science* 2023; **380:** 16–17.
9 RECOVERY Trial chief investigators. Low-cost dexamethasone reduces death by up to one third in hospitalised patients with severe respiratory complications of COVID-19. June 16, 2020. https://www.recoverytrial.net/news/low-cost-dexamethasone-reduces-death-by-up-to-one-third-in-hospitalised-patients-with-severe-respiratory-complications-of-covid-19 (accessed Dec 21, 2023).
10 Horby P, Lim WS, Emberson JR, et al. Dexamethasone in hospitalized patients with COVID-19. *N Engl J Med* 2021; **384:** 693–704.
11 Watson OJ, Barnsley G, Toor J, Hogan AB, Winskill P, Ghani AC. Global impact of the first year of COVID-19 vaccination: a mathematical modelling study. *Lancet Infect Dis* 2022; **22:** 1293–302.
12 Wellcome. Sharing research data and findings relevant to the novel coronavirus (COVID-19) outbreak. Jan 31, 2020. https://wellcome.org/press-release/sharing-research-data-and-findings-relevant-novel-coronavirus-ncov-outbreak (accessed Dec 21, 2023).
13 Secretary of State for Health and Social Care. COVID-19—notice under regulation 3(4) of the Health Service Control of Patient Information Regulations 2002. March 20, 2020. https://www.england.nhs.uk/wp-content/uploads/2022/07/COPI-notice-to-nhs-england-improvement-covid-19.pdf (accessed Dec 21, 2023).
14 Dimensions. COVID-19 report: publications, clinical trials, funding. https://reports.dimensions.ai/covid-19/ (accessed Nov 7, 2023).
15 Mei X, Lee H-C, Diao K-Y, et al. Artificial intelligence-enabled rapid diagnosis of patients with COVID-19. *Nat Med* 2020; **26:** 1224–28.

16 Kaggle, Society for Imaging Informatics in Medicine. SIIM-FISABIO-RSNA COVID-19 Detection. https://www.kaggle.com/competitions/siim-covid19-detection (accessed Nov 7, 2023).

17 Wynants L, Van Calster B, Collins GS, et al. Prediction models for diagnosis and prognosis of COVID-19: systematic review and critical appraisal. *BMJ* 2020; **369:** m1328.

18 Mann S, Berdahl CT, Baker L, Girosi F. Artificial intelligence applications used in the clinical response to COVID-19: a scoping review. *PLoS Digit Health* 2022; **1:** e0000132.

19 Carobene A, Milella F, Famiglini L, Cabitza F. How is test laboratory data used and characterised by machine learning models? A systematic review of diagnostic and prognostic models developed for COVID-19 patients using only laboratory data. *Clin Chem Lab Med* 2022; **60:** 1887–901.

20 Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nat Med* 2021; **27:** 2176–82.

21 Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 2019; **366:** 447–53.

22 Cao J, Zhang X, Shahinian V, et al. Generalizability of an acute kidney injury prediction model across health systems. *Nat Mach Intell* 2022; **4:** 1121–29.

23 Garcia Santa Cruz B, Bossa MN, Sölter J, Husch AD. Public COVID-19 x-ray datasets and their impact on model bias—a systematic review of a significant problem. *Med Image Anal* 2021; **74:** 102225.

24 Khan SM, Liu X, Nath S, et al. A global review of publicly available datasets for ophthalmological imaging: barriers to access, usability, and generalisability. *Lancet Digit Health* 2021; **3:** e51–66.

25 Wen D, Khan SM, Ji Xu A, et al. Characteristics of publicly available skin cancer image datasets: a systematic review. *Lancet Digit Health* 2022; **4:** e64–74.

26 Ibrahim H, Liu X, Zariffa N, Morris AD, Denniston AK. Health data poverty: an assailable barrier to equitable digital health care. *Lancet Digit Health* 2021; **3:** e260–65.

27 US Food and Drug Administration, Health Canada, Medicines and Healthcare products Regulatory Agency. Good machine learning practice for medical device development: guiding principles. October, 2021. https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles (accessed Nov 29, 2023).

28 Platt L, Warwick R. Are some ethnic groups more vulnerable to COVID-19 than others? May 1, 2020. https://ifs.org.uk/publications/are-some-ethnic-groups-more-vulnerable-covid-19-others (accessed Nov 7, 2023).

29 Magesh S, John D, Li WT, et al. Disparities in COVID-19 outcomes by race, ethnicity, and socioeconomic status: a systematic-review and meta-analysis. *JAMA Netw Open* 2021; **4:** e2134147.

30 Roth GA, Emmons-Bell S, Alger HM, et al. Trends in patient characteristics and COVID-19 in-hospital mortality in the United States during the COVID-19 pandemic. *JAMA Netw Open* 2021; **4:** e218828.

31 Flor LS, Friedman J, Spencer CN, et al. Quantifying the effects of the COVID-19 pandemic on gender equality on health, social, and economic indicators: a comprehensive review of data from March, 2020, to September, 2021. *Lancet* 2022; **399:** 2381–97.

32 Patel JA, Nielsen FBH, Badiani AA, et al. Poverty, inequality and COVID-19: the forgotten vulnerable. *Public Health* 2020; **183:** 110–11.

33 Sjoding MW, Dickson RP, Iwashyna TJ, Gay SE, Valley TS. Racial bias in pulse oximetry measurement. *N Engl J Med* 2020; **383:** 2477–78.

34 Ganapathi S, Palmer J, Alderman JE, et al. Tackling bias in AI health datasets through the STANDING Together initiative. *Nat Med* 2022; **28:** 2232–33.

35 Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 2021; **372:** n71.

36 Booth A, Clarke M, Dooley G, et al. The nuts and bolts of PROSPERO: an international prospective register of systematic reviews. *Syst Rev* 2012; **1:** 2.

37 Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan—a web and mobile app for systematic reviews. *Syst Rev* 2016; **5:** 210.

38 Legislation.gov.uk. Equality Act 2010. https://www.legislation.gov.uk/ukpga/2010/15/contents (accessed Dec 21, 2023).

39 Gebru T, Morgenstern J, Vecchione B, et al. Datasheets for datasets. *Commun ACM* 2021; **64:** 86–92.

40 Rostamzadeh N, Mincu D, Roy S, et al. Healthsheet: development of a transparency artifact for health datasets. In: Proceedings of the 2022 ACM conference on fairness, accountability, and transparency. New York, NY: Association for Computing Machinery, 2022: 1943–61.

41 Mooney P. Chest x-ray images (pneumonia). 2018. https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia (accessed Dec 21, 2023).

42 Kermany DS, Goldbaum M, Cai W, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* 2018; **172:** 1122–31.

43 Rahman T, Chowdhury M, Khandakar A. COVID-19 radiography database. https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database (accessed March 29, 2024).

44 Cohen JP. Covid-chestxray-dataset. https://github.com/ieee8023/covid-chestxray-dataset (accessed Dec 21, 2023).

45 Huang JY. Representativeness is not representative: addressing major inferential threats in the UK biobank and other big data repositories. *Epidemiology* 2021; **32:** 189–93.

46 Wilkinson MD, Dumontier M, Aalbersberg IJJ, et al. The FAIR guiding principles for scientific data management and stewardship. *Sci Data* 2016; **3:** 160018.

47 McDermott MBA, Wang S, Marinsek N, Ranganath R, Foschini L, Ghassemi M. Reproducibility in machine learning for health research: still a ways to go. *Sci Transl Med* 2021; **13:** eabb1655.

48 Carroll SR, Garba I, Figueroa-Rodríguez OL, et al. The care principles for Indigenous data governance. *Data Sci J* 2020; **19:** 43.

49 The First Nations Information Governance Centre. The First Nations principles of OCAP. https://fnigc.ca/ocap-training/ (accessed Nov 29, 2023).

50 London AJ, Kimmelman J. Against pandemic research exceptionalism. *Science* 2020; **368:** 476–77.

51 The STANDING Together collaboration. Recommendations for diversity, inclusivity, and generalisability in artificial intelligence health technologies and health datasets. Oct 30, 2023. https://zenodo.org/records/10048356 (accessed Feb 6, 2024).