



OPEN

DATA DESCRIPTOR

# Chest imaging representing a COVID-19 positive rural U.S. population

Shivang Desai<sup>1,2</sup>, Ahmad Baghal<sup>3</sup>, Thidathip Wongsurawat<sup>3</sup>, Piroon Jenjaroenpun<sup>3</sup>, Thomas Powell<sup>3</sup>, Shaymaa Al-Shukri<sup>3</sup>, Kim Gates<sup>3</sup>, Phillip Farmer<sup>3</sup>, Michael Rutherford<sup>3</sup>, Geri Blake<sup>3</sup>, Tracy Nolan<sup>3</sup>, Kevin Sexton<sup>3,4</sup>, William Bennett<sup>3</sup>, Kirk Smith<sup>3</sup>, Shorabuddin Syed<sup>3</sup> & Fred Prior<sup>1,3</sup>✉

As the COVID-19 pandemic unfolds, radiology imaging is playing an increasingly vital role in determining therapeutic options, patient management, and research directions. Publicly available data are essential to drive new research into disease etiology, early detection, and response to therapy. In response to the COVID-19 crisis, the National Cancer Institute (NCI) has extended the Cancer Imaging Archive (TCIA) to include COVID-19 related images. Rural populations are one population at risk for underrepresentation in such public repositories. We have published in TCIA a collection of radiographic and CT imaging studies for patients who tested positive for COVID-19 in the state of Arkansas. A set of clinical data describes each patient including demographics, comorbidities, selected lab data and key radiology findings. These data are cross-linked to SARS-COV-2 cDNA sequence data extracted from clinical isolates from the same population, uploaded to the GenBank repository. We believe this collection will help to address population imbalance in COVID-19 data by providing samples from this normally underrepresented population.

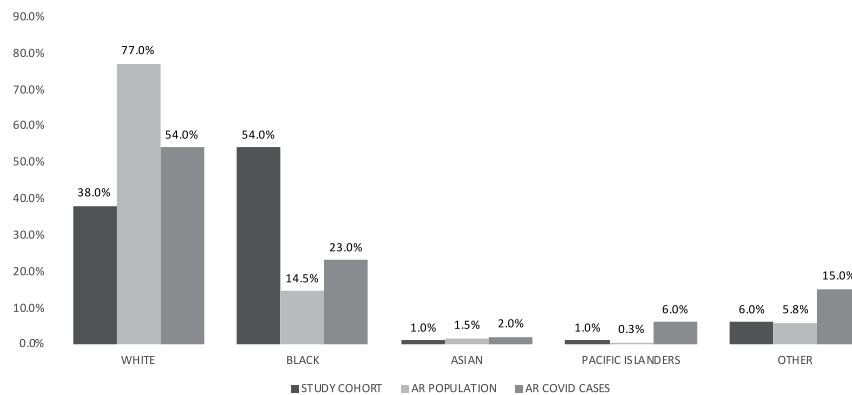
## Background & Summary

Rural Americans are at greater health risk and are more susceptible to five leading causes of death than urban Americans<sup>1</sup>. Bolin *et al.*<sup>2</sup> summarize findings of the Rural Healthy People 2020 survey and identify an array of health issues that increase the health risks faced by the rural population of the United States. Significant among these are higher rates of hypertension, obesity, heart disease, stroke, and substance abuse, coupled with limited access to quality healthcare and public health infrastructure. Arkansas (AR) is a rural state, with 42% of Arkansans living in rural counties, (<https://worldpopulationreview.com/states/arkansas-population/>) compared to 15% in the country as a whole. Arkansas' health problems mirror those of the rest of the rural U.S.<sup>3</sup>

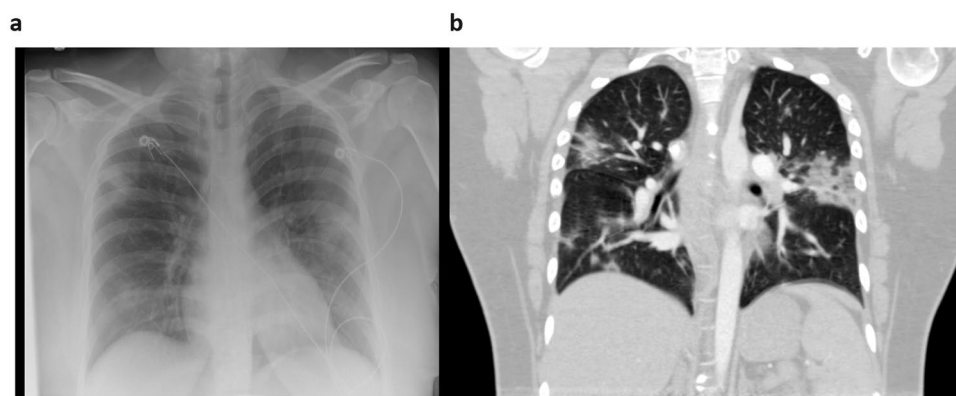
According to recent American Census Survey (ACS) data, the composition of Arkansas' population is white 77.0%; black, 14.5%; Asian, 1.5%; pacific islanders, 0.3%; and other 5.8% (other race, 2.6%; two or more races, 2.5%; native American, 0.7%), (Fig. 1). The study cohort reported here includes 105 COVID-19 patients seen at the University of Arkansas for Medical Sciences (UAMS) with the following demographics characteristics: 50% (n = 53) female and 50% (n = 52) male; age range 19 to 91 years; and white 38% (n = 40), black 54% (n = 57), Asian 1% (n = 1), pacific islanders 1% (n = 1), other 6% (n = 6), (Fig. 1). The study population is representative of neighboring, rural counties in the proximity of UAMS. Figure 1 also illustrates the demographics of the COVID-19 infected population in Arkansas as of July 20, 2020 based on statistics reported daily by the Arkansas Department of Health on their web site <https://experience.arcgis.com/experience/c2ef4a4fcbef2e48a21e4f9e9>.

Radiologic imaging has proven to be an essential tool for screening, diagnosis, and management of patients with COVID-19 infection<sup>4,5</sup>. In March 2020 the American College of Radiology published guidelines that recommended the use of chest radiography as the "first-line test to diagnose COVID-19" with CT being used "sparingly and reserved for hospitalized, symptomatic patients"<sup>6</sup>. Largely because of these guidelines, the bulk of imaging

<sup>1</sup>Department of Radiology, University of Arkansas for Medical Sciences, Little Rock, Arkansas, USA. <sup>2</sup>Department of Radiology, University of Texas, Southwestern Medical Center, Dallas, Texas, USA. <sup>3</sup>Department of Biomedical Informatics, University of Arkansas for Medical Sciences, Little Rock, Arkansas, USA. <sup>4</sup>Department of Surgery, University of Arkansas for Medical Sciences, Little Rock, Arkansas, USA. ✉e-mail: [FWPrior@uams.edu](mailto:FWPrior@uams.edu)



**Fig. 1** Study cohort population distribution. Racial characteristics of study cohort in comparison to Arkansas total population and the state's Covid-19 cases.



**Fig. 2** Chest Radiograph (a) and Computed Tomography (CT) image in the sagittal plane (b) of the same patient (Covid-19-Ar-16434363) taken on the same day. Patchy airspace opacities and ground-glass changes are seen in both lungs.

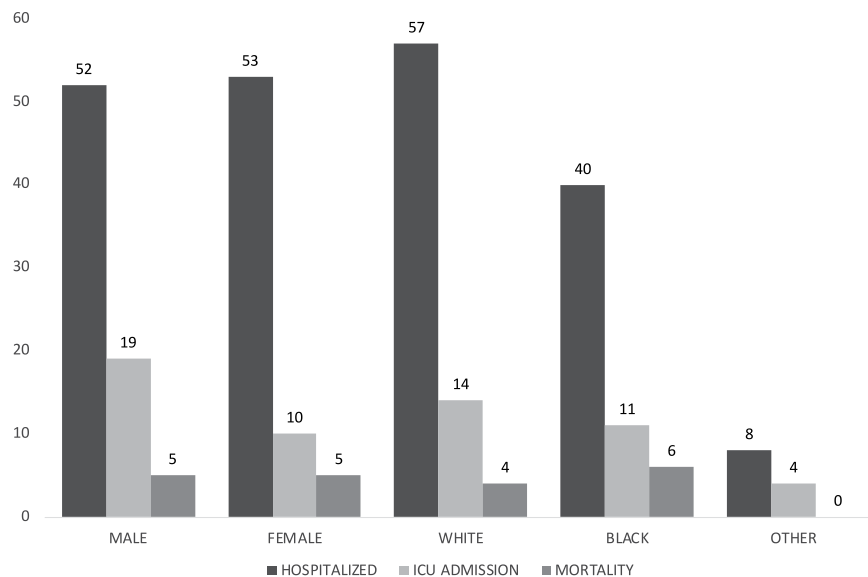
studies performed thus far in the US are chest radiographs. The image collection presented here is consistent with these guidelines.

The ability to sequence the whole genome of the SARS-CoV-2 virus is essential for tracking the evolution of the virus and the resulting disease. GenBank<sup>7</sup> contains 22,542 SRA runs and 11,142 Nucleotide records (as of July 20, 2020) and is growing daily. We have contributed to this repository sequences of the virus strains infecting our local population.

## Methods

**Patient selection.** Images and clinical correlate data were obtained with approval of the UAMS Institutional Review Board (IRB number 228350) which includes open access publication of this data in de-identified form. All data were collected from hospitalized patients with a positive COVID-19 lab test (PCR) verified diagnosis who had imaging studies within eight days prior to diagnosis and at least one imaging study post diagnosis. Emphasis was placed on patients with multiple post diagnosis imaging studies. All imaging studies of the head were excluded due to issues of potential patient identification from volumetric CT reconstructions<sup>8,9</sup>.

**Imaging.** Imaging techniques were standard of care, primarily portable (97%), digital chest radiographs collected either by computed radiography, CR, (19 patients, 26 image series) or direct digital capture, DX, (100 patients, 236 series). CT (computed tomography) studies were performed on 21% (23) of patients depending on the severity of symptoms and clinical status (resulting in 199 CT image series). Radiographs were performed using a single view AP (anteroposterior) technique or PA (posteroanterior) and lateral views depending on patient mobility and intubation status. CT was performed on multi-slice, spiral scanners. Contiguous 1 mm axial images were obtained through the chest (Online-only Table 1 provides acquisition parameters for each of the 199 CT series). Axial (3 mm spatial resolution) and 2 mm sagittal and coronal reconstructions were performed and subsequently reviewed on a dedicated Picture Archive and Communication System (PACS) workstation. A total of 256 chest imaging studies were identified, including 233 radiographs and 23 CT studies performed on a total of 105 patients. Figure 2 is an illustrative example of a chest radiograph and CT for the same patient taken on the same day. Final radiology reports for these examinations signed by board-certified radiologists were analyzed for key imaging findings.



**Fig. 3** Disease Progression (by Sex, Race). Breakdown of the rates of hospitalization, ICU admission, and mortality by sex and race.

**Summary of key imaging findings.** The most frequent pattern of imaging findings is ‘Organizing pneumonia’ which is essentially a pattern of lung changes as a response to inflammation. The reported key imaging findings on radiographs and CT include confluent and patchy multifocal airspace opacification, either ground-glass opacities or consolidation. These airspace opacities were predominantly bilateral, multifocal, and peripheral. Although more commonly, these changes were diffuse and bilateral, lower lobe predominance was seen in cases that were more focal. A total of 59/232 (25%) radiographs and 8/23 (35%) CT were negative for airspace opacification. In our patient population, when present, ground glass changes were more common than consolidation, likely suggesting an earlier presentation in the disease course. Pleural effusions were rare and, when present, were only trace to mild. Other atypical findings like mediastinal lymphadenopathy, cavitation, and pneumothorax were not identified in our patient population. One of the 23 CT studies showed bilateral segmental pulmonary emboli.

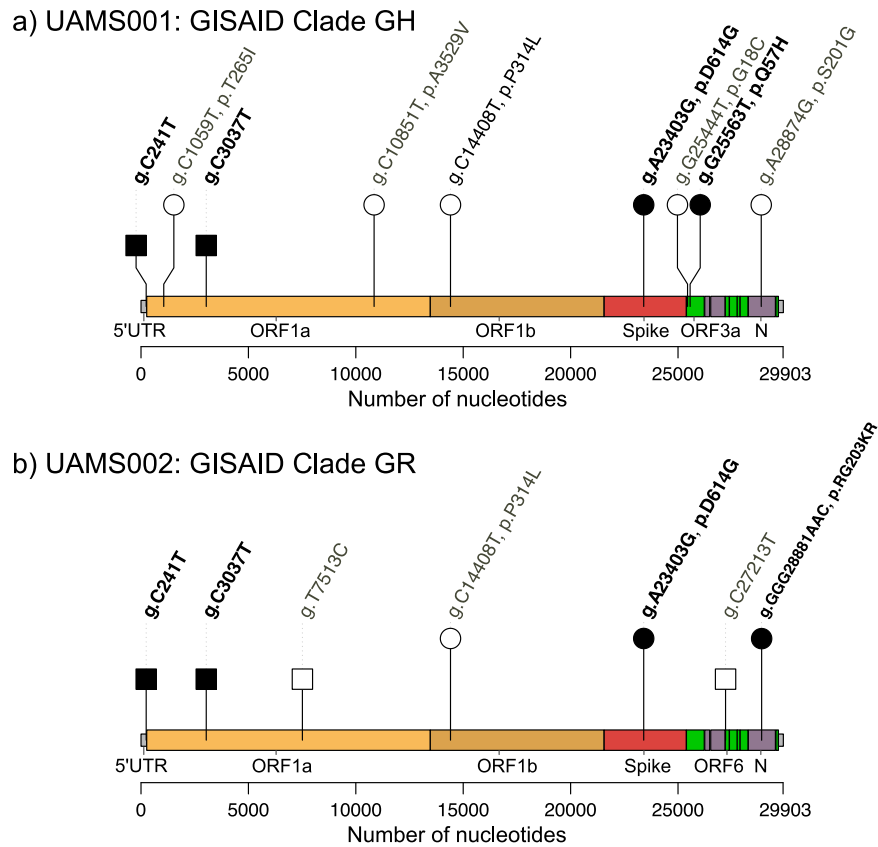
The majority of ICU patients (28/29) showed radiographic AP changes, with 59% showing bilateral diffuse changes with left lower lobe involvement. One Radiographic study for one ICU patient had no significant findings. Seventy percent (7/10) of ICU mortality occurred in patients showing left lower lobe opacities. In contrast, one of these patients had no significant findings on CR or DX radiographic studies, 1 had general atelectasis, and 1 had diffuse disease without left lower lobe opacity.

**Clinical correlates.** The average age in the cohort was 54.3 years old (range 19–91) with an even sex distribution (52 Male, 53 Female). The worldwide incidence is reported to be close to 1:1, with a 50% increase in hospitalizations, ICU stays, and mortality among males<sup>10</sup>. The racial characteristics of the cohort are presented in Fig. 1 in comparison to the total population of Arkansas and the current characteristics of the state-wide infected population. The average BMI in the cohort is 33.1 (18.7–64.9), well within obese range (30.0 or higher). Key Comorbidities include burns (2%), malnutrition (3%), pregnancy (4%), chronic kidney disease (11%), diabetes (21%), organ transplant (3%), and cancer (24%).

The overall ICU admission rate was 28% (29/105). The Average age among those admitted to the ICU was 58 (range 25–89), slightly higher than the average for the cohort as a whole. The racial breakdown of ICU admissions included 28% of the white patients, 25% of the black, 50% of other, and 100% of Pacific islanders. The ICU population was 66% male and 33% female and included 1 pregnant patient. Forty three percent of patients admitted to the ICU had BMI greater than 30. Of the black patients, 21% had diabetes and 29% chronic renal disease, while among white patients the highest percentage comorbidities were cancer (10%) and diabetes (10%). The ICU mortality rate was 34.4% (10/29) which is 1.5 times the national average of 23.6%<sup>11</sup>. The overall mortality rate was 10% (10/105) and all 10 patients in the mortality group were first admitted to ICU. Arkansan males were 1.95 times more likely to go to ICU (19/52 vs. 10/53). Our data shows an almost even overall mortality rate among males (5/53) and females (5/52). The data also suggest that once in the ICU, female mortality occurs at a rate 1.9 times that of males (5/19 vs. 5/10). Figure 3 summarizes rates of hospitalization, ICU admission and mortality stratified by sex and race.

**COVID-19 testing samples.** A set of residual, de-identified nasopharyngeal samples testing positive for SARS-CoV-2 by quantitative reverse-transcriptase PCR, was obtained from the clinical molecular diagnostics lab at the University of Arkansas for Medical Sciences (UAMS). All samples were obtained with approval of the UAMS Institutional Review Board (IRB number 260840). This approval includes the right to publish viral sequences with all references to the specific human participants who provided the samples being removed.

**SARS-CoV-2 genome sequencing.** RNA extracted from nasopharyngeal samples was reverse transcribed, as described in protocol published by ARTIC Network (<https://doi.org/10.17504/protocols.io.bdp715rn>).



**Fig. 4** The illustration of SARS-CoV-2 genome organization and a plot of mutations. Two isolates of UAMS SARS-CoV-2 genome, UAMS001 (a) and UAMS002 (b), are grouped into clade G (D614G mutation) but different sub-group (GH and GR, respectively). Genome size of both isolates is 29,903-nucleotides. The leader sequence is represented by a small grey square at the 5' terminus of the genome. Open reading frames (ORFs) 1a (yellow) and 1b (brown) encode the nonstructural proteins. Spike (red), envelope, membrane, and nucleocapsid are shown in purple and the accessory proteins in green. Vertical lines represent the presence of a mutation. Above each line, the variants are annotated as the nucleic acid change (such as g.C241T) and amino-acid change (such as p.D614G) at that specific site. The non-synonymous mutations are presented by circles and the synonymous mutations squares. Black circles and black squares are the locations used for clade identification characterized by Global Initiative on Sharing All Influenza Data (GISAID).

SARS-CoV-2 specific primer set version 3 (218 primers) was kindly provided by Dr. Joshua Quick, University of Birmingham. The PCR amplification and extension condition of SARS-CoV-2 genome was performed using the COVID-19 ARTIC v3 Illumina library construction and sequencing protocol V.4 (<https://doi.org/10.17504/protocols.io.bgxjxkn>). The library was prepared and loaded on a R9.4.1/FLO-MIN106 flow cell, and sequenced with the MinION Mk1B (Oxford Nanopore Technology, ONT). Base-calling of the resulting FAST5 files was performed using Guppy (version 3.4.5)<sup>12</sup>, and the RAMPART software (v1.0.5) from the ARTIC Network<sup>13</sup> was used to monitor sequencing in real-time. The ARTIC Network bioinformatics protocol was used for all genome assembly and variant calling steps<sup>14</sup>. Figure 4 illustrates the resultant SARS-CoV-2 genome organization and identified mutations. Two isolates of UAMS SARS-CoV-2 genome were identified and grouped into a clade<sup>15</sup>. Black circles and black squares in Fig. 4 identify the locations used for clade identification characterized by Global Initiative on Sharing All Influenza Data (GISAID)<sup>15</sup>. TrackViewer software was used for visualization<sup>16</sup>.

### Data Records

**COVID-19-AR image collection.** Image data were extracted from the clinical PACS (Sectra AB, Linköping, Sweden) at the University of Arkansas for Medical Sciences using Digital Imaging and Communications in Medicine (DICOM) query/retrieve software. All image data were de-identified and stored in DICOM standard format<sup>17</sup> on TCIA as collection COVID-19-AR<sup>18</sup>.

**Clinical data structure.** All clinical data were obtained from the Arkansas Clinical Data Repository (AR-CDR)<sup>19</sup>. The AR-CDR is a research data warehouse that provides a single and secure source of data for use in clinical and translational research; housing data for this project that are extracted from the EPIC (Epic Systems Inc, Verona, WI) electronic health record (EHR) system and legacy systems.

Data are provided on TCIA<sup>18</sup> in a spreadsheet format (Microsoft excel) with one tab for patient data and a second for imaging study related data. Patient demographics, outcomes, co-morbidities and other clinical information are provided. Online-only Table 2 defines categorical variables (column headings) that are included.

**Genomic data structure.** Viral genomes are stored in the GenBank FASTA format<sup>20</sup> which stores both a nucleotide sequence and its corresponding quality scores. All required metadata, including collection dates, location, nucleotide sequence, amino acid sequences, and gene annotations, as described in the GenBank data model<sup>21</sup> were provided for each sequence.

Data are available from the NCBI Sequence Read Archive<sup>22</sup> and via direct links to each SARS-CoV-2 viral genome<sup>23,24</sup>. An NCBI BioProject was created (<https://identifiers.org/bioproject:PRJNA643530>) and also provides links to the viral genomes. Links to these data are also provided on TCIA<sup>18</sup>.

## Technical Validation

**Imaging and clinical data.** All image data were processed using standard TCIA curation workflows. TCIA uses a standards-based approach for de-identification of images stored in the Digital Imaging and Communications in Medicine (DICOM) format. DICOM Standard PS3.15 2016a - Security and System Management Profiles<sup>25</sup> defines how to correctly de-identify DICOM objects. The image data described in this data descriptor were de-identified using the “Basic Application Confidentiality Profile,” amended by inclusion of profile options: Clean Pixel Data Option (removal of information burned into the pixel data), Clean Descriptors Option (removal of PHI from data elements of type text or string), Retain Longitudinal With Modified Dates Option (all dates are shifted in a consistent manner by a random number so that relative temporal information is retained but absolute dates are masked), Retain Patient Characteristics Option (descriptive patient information is retained), Retain Device Identity Option (acquisition system information of potential scientific value is retained), and Retain Safe Private Option (retain scientifically valuable information stored in vendor private data elements) as is standard TCIA practice. Additional details on the de-identification process including the process for modifying dates by shifting using a random number are provided on the TCIA web site ([wiki.cancerimagingarchive.net/display/Public/Submission+and+De-identification+Overview](http://wiki.cancerimagingarchive.net/display/Public/Submission+and+De-identification+Overview)).

The TCIA curation team verifies completeness of the received collection, full removal of all PHI, proper labeling of all information to facilitate retrieval, and proper linkages among components of the collection<sup>26–28</sup>. TCIA curation procedures assure image quality and data integrity. TCIA uses the Posda open-source framework to implement its curation process<sup>29,30</sup>. Posda-based TCIA curation workflows ensure the scientific utility of data and eliminate protected health information. After Posda processing is complete, a data curator visually inspects every DICOM image.

Patient demographics, outcomes, and other clinical information were also validated during the TCIA curation process. Prior to this review, two of the authors (AB, FP) manually reviewed all clinical data for consistency, accuracy, and completeness, while SD and FP reviewed all image data and radiology reports. Images were reviewed on the clinical PACS and using TCIA’s radiology image viewer<sup>31</sup>.

**SARS-CoV-2 genome sequencing.** To validate the ARTIC protocol for both experimental and computational methods, we included the Washington strain SARS-CoV-2 genomics RNA (2019-nCoV/USA-WA1/2020, MN985325.1) acquired from the American Type Culture Collection (ATCC) as the positive control and nuclease-free water as the negative control. The positive control and negative control were sequenced simultaneously with the clinical sequences and analyzed using the ARTIC protocol to obtain SARS-CoV-2 genomes. The SARS-CoV-2 genome sequence from the positive control showed no differences relative to the original genome sequence, MN985325.1. For the negative control, no viral sequence was obtained.

Received: 23 July 2020; Accepted: 27 October 2020;

Published online: 24 November 2020

## References

- Garcia, M. C. *et al.* Reducing potentially excess deaths from the five leading causes of death in the rural United States. *MMWR Surveill. Summ.* **66**, 1 (2017).
- Bolin, J. N. *et al.* Rural healthy people 2020: new decade, same challenges. *J. Rural Health.* **31**, 326–333, <https://doi.org/10.1111/jrh.12116> (2015).
- Smith, N. Arkansas’s Big Health Problems and How We Plan to Solve Them, *Arkansas Department of Health* <https://www.healthy.arkansas.gov/images/uploads/pdf/ARHealthReportHealthProblems.pdf> (2013).
- Li, Y. & Xia, L. Coronavirus disease 2019 (COVID-19): role of chest CT in diagnosis and management. *Am J Roentgenol.* **214**, 1280–1286 (2020).
- Bai, H. X. *et al.* Performance of radiologists in differentiating COVID-19 from viral pneumonia on chest CT. *Radiology.* **296**, E46–E54, <https://doi.org/10.1148/radiol.2020200823> (2020).
- American College of Radiology. ACR Recommendations for the use of Chest Radiography and Computed Tomography (CT) for Suspected COVID-19 Infection <https://www.acr.org/Advocacy-and-Economics/ACR-Position-Statements/Recommendations-for-Chest-Radiography-and-CT-for-Suspected-COVID19-Infection> (2020).
- Sayers, E. W. *et al.* GenBank. *Nucleic Acids Res.* **48**, D84–D86, <https://doi.org/10.1093/nar/gkz956> (2019).
- Prior, F. W. *et al.* Facial recognition from volume-rendered magnetic resonance imaging data. *IEEE T. Inf. Technol. B.* **13**, 5–9 (2008).
- Schwarz, C. G. *et al.* Identification of anonymous MRI research participants with face-recognition software. *N. Engl. J. Med.* **381**, 1684–1686 (2019).
- Gebhard, C., Regitz-Zagrosek, V., Neuhauser, H. K., Morgan, R. & Klein, S. L. Impact of sex and gender on COVID-19 outcomes in Europe. *Biol. Sex Differ.* **11**, 1–13 (2020).
- Quah, P., Li, A. & Phua, J. Mortality rates of patients with COVID-19 in the intensive care unit: a systematic review of the emerging literature. *Crit. Care.* **24**, 285, <https://doi.org/10.1186/s13054-020-03006-1> (2020).
- Wick, R. R., Judd, L. M. & Holt, K. E. Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biol.* **20**, 129 (2019).
- Artic Network. Source code for: Read assignment, mapping and phylogenetic analysis in real time (RAMPART). *GitHub* <https://github.com/artic-network/rampart> (2020).
- Loman, N., Rowe, W. & Rambaut, A. nCoV-2019 novel coronavirus bioinformatics protocol. *Artic Network* <https://artic.network/ncov-2019/ncov2019-bioinformatics-sop.html> (2020).
- Mercatelli, D. & Giorgi, F. M. Geographic and Genomic Distribution of SARS-CoV-2 Mutations. *Front. Microbiol.* **11**, <https://doi.org/10.3389/fmicb.2020.01800> (2020).

16. Ou, J. & Zhu, L. J. trackViewer: a Bioconductor package for interactive and integrative visualization of multi-omics data. *Nat. Methods*. **16**, 453–454 (2019).
17. DICOM. In *NEMA PS3 / ISO 12052 Digital Imaging and Communications in Medicine (DICOM) Standard*, (NEMA, Rosslyn, VA, 2020).
18. Desai, S. *et al.* Data from Chest Imaging with Clinical and Genomic Correlates Representing a Rural COVID-19 Positive Population. *The Cancer Imaging Archive* <https://doi.org/10.7937/tcia.2020.py71-5978> (2020).
19. Baghal, A., Zozus, M., Baghal, A., Al-Shukri, S. & Prior, F. Factors Associated with Increased Adoption of a Research Data Warehouse, in *Improving Usability, Safety and Patient Outcomes with Health Information Technology* Vol. 257 (eds Lau, F. *et al.*) 31–35 (IOS Press BV, 2019).
20. Mills, L. *Common File Formats*. *Curr. Protoc. Bioinform.* **45**, A.1B.1–A.1B.18, <https://doi.org/10.1002/0471250953.bia01bs45> (2003).
21. Ostell, J. M., Wheelan, S. J. & Kans, J. A. In *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins* (ed. Baxeavanis, A. & Ouellette, B.) Ch. 2 (Wiley, 2001).
22. *NCBI Sequence Read Archive* <https://identifiers.org/ncbi/insdc.sra:SRP272745> (2020).
23. Jenjaroenpun, P., Ussery, D., Wongsurawat, T. Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/USA/AR-UAMS001/2020, complete genome. *GenBank* <https://identifiers.org/ncbi/insdc:MT766907.1> (2020).
24. Jenjaroenpun, P., Ussery, D., Wongsurawat, T. Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/USA/AR-UAMS002/2020, complete genome. *GenBank* <https://identifiers.org/ncbi/insdc:MT766908.1> (2020).
25. DICOM. In *PS3.15 2016a - Security and System Management Profiles* (NEMA, Rosslyn, VA, 2016).
26. Clark, K. *et al.* The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *J. Digit. Imaging*. **26**, 1045–1057 (2013).
27. Moore, S. M. *et al.* De-identification of Medical Images with Retention of Scientific Research Value. *RadioGraphics*. **35**, 727–735, <https://doi.org/10.1148/rg.2015140244> (2015).
28. Prior, F. *et al.* The public cancer radiology imaging collections of The Cancer Imaging Archive. *Sci. Data*. **4**, 170124 (2017).
29. Bennett, W., Smith, K., Jarosz, Q., Nolan, T. & Bosch, W. Reengineering workflow for curation of DICOM datasets. *J. Digit. Imaging*. **31**, 783–791 (2018).
30. Grossberg, A. J. *et al.* Imaging and clinical data archive for head and neck squamous cell carcinoma patients treated with radiotherapy. *Sci. Data*. **5**, 180173 (2018).
31. Ziegler, E. *et al.* Open Health Imaging Foundation Viewer: An Extensible Open-Source Framework for Building Web-Based Imaging Applications to Support. *Cancer Research. JCO Clin. Cancer Info*. **4**, 336–345, <https://doi.org/10.1200/cci.19.00131> (2020).

## Acknowledgements

This project has been funded in whole or in part with federal funds from the National Center for Advancing Translational Sciences UL1 TR003107 and the National Cancer Institute, Contract No. 75N91019D00024, Subcontract 20X023F. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government. The authors wish to acknowledge and thank Dr. Sean Adams and Dr. Kikumi Ono-Moore of the Arkansas Children's Nutrition Center for graciously sharing their BLS2 laboratory, Dr. Joshua Kennedy of the Department of Allergy and Immunology of UAMS for providing residual patient samples and Dr. Joshua Quick, University of Birmingham for providing a SARS-CoV-2 specific primer set.

## Author contributions

Desai worked with Prior to analyze radiologist findings and contributed to the manuscript. Baghal directed the collection of clinical data and along with Powell and Sexton analyzed this data and contributed to the definition of the set of clinical parameters to be published. Wongsurawat and Jenjaroenpun sequenced the viral genomes, performed validation and contributed to the manuscript. Syed, Al-Shukri and Gates collected the clinical data while Farmer, Smith and Rutherford collected the imaging data. Blake, Nolan and Bennett performed data curation and validation. Prior and Smith conceived of the idea and Prior was the lead contributor to the manuscript. All authors reviewed and contributed to the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Correspondence** and requests for materials should be addressed to F.P.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

The Creative Commons Public Domain Dedication waiver <http://creativecommons.org/publicdomain/zero/1.0/> applies to the metadata files associated with this article.

© The Author(s) 2020