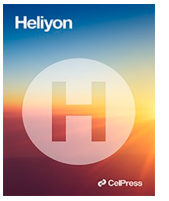
Heliyon 9 (2023) e17934

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Review article

Machine learning with multimodal data for COVID-19

Weijie Chen [a,b,*], Rui C. Sá [a,c], Yuntong Bai [a,b], Sandy Napel [a,d], Olivier Gevaert [a,e], Diane S. Lauderdale [a,f], Maryellen L. Giger [a,g]

[a] Medical Imaging and Data Resource Center (MIDRC), USA
[b] Division of Imaging, Diagnostics, and Software Reliability, Office of Science and Engineering Laboratories, Center for Devices and Radiological Health, USA
[c] Department of Medicine, University of California, San Diego, USA
[d] Department of Radiology, Stanford University, USA
[e] Department of Medicine and Department of Biomedical Data Science, Stanford University, USA
[f] Department of Public Health Sciences, University of Chicago, USA
[g] Department of Radiology, University of Chicago, USA

ARTICLE INFO

Keywords:
COVID-19
Multimodal data
Machine learning

ABSTRACT

In response to the unprecedented global healthcare crisis of the COVID-19 pandemic, the scientific community has joined forces to tackle the challenges and prepare for future pandemics. Multiple modalities of data have been investigated to understand the nature of COVID-19. In this paper, MIDRC investigators present an overview of the state-of-the-art development of multimodal machine learning for COVID-19 and model assessment considerations for future studies. We begin with a discussion of the lessons learned from radiogenomic studies for cancer diagnosis. We then summarize the multi-modality COVID-19 data investigated in the literature including symptoms and other clinical data, laboratory tests, imaging, pathology, physiology, and other omics data. Publicly available multimodal COVID-19 data provided by MIDRC and other sources are summarized. After an overview of machine learning developments using multimodal data for COVID-19, we present our perspectives on the future development of multimodal machine learning models for COVID-19.

1. Introduction

The coronavirus disease 2019 (COVID-19) pandemic caused by Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) has claimed over 6 million lives with over 630 million confirmed cases around the world to date [1]. Besides the primary target organ, i.e., the respiratory system, the virus has exhibited deleterious impacts on other organs, including brain, heart, kidney, liver, and the endocrine system [2]. Moreover, while the majority of COVID-19 patients recover within weeks, some patients suffer from post-acute COVID-19 syndrome with long-term complications [3]. In response to this unprecedented global healthcare crisis, the scientific community has joined forces to tackle the challenges and prepare for future pandemics. The purpose of this paper is to present our perspective on the use of multimodal data and machine learning technologies for assessing infection and progression of COVID-19.

Multi-modality data are under investigation for a variety of COVID-19 clinical tasks, e.g., detection/diagnosis, prognostication, severity characterization, and prediction/monitoring of treatment response. The gold standard for detection of the SARS-CoV-2 virus is

* Corresponding author. Medical Imaging and Data Resource Center (MIDRC), USA.
E-mail address: weijie.chen@fda.hhs.gov (W. Chen).

https://doi.org/10.1016/j.heliyon.2023.e17934
Received 17 May 2023; Accepted 3 July 2023
Available online 5 July 2023

the Reverse Transcription Polymerase Chain Reaction (RT-PCR) test. Multiple imaging modalities, such as chest x-ray radiography (CXR), computed tomography (CT), magnetic resonance imaging (MRI), and ultrasound (US), are being used or investigated for diagnosis (especially when RT-PCR is not available), for characterization of disease severity, and for monitoring treatment response [4,5]. Plasma multi-omic profiles and clinical data are also being investigated for the characterization of COVID-19 severity [6]. Additionally, measurements of gas exchange, hemodynamics, and lung mechanics, as well as chest CT scans, are enabling the building of computational models on ventilation-perfusion inequality, and thus shedding light on the pathophysiology of the disease [7].

The availability of multimodal data could enable the comprehensive understanding of disease thereby potentially informing better patient care than might a single modality, which is especially beneficial when investigating a new disease such as COVID-19. Clinically, critical decisions are often made during multidisciplinary consultation meetings (e.g., tumor boards) with experts from multiple specialties, as interpretation of multimodal data may be beyond any single physician's capability. Machine learning (ML) holds great promise in the analysis and integration of large amounts of complex multimodal data for specified clinical tasks, as has been demonstrated in cancer diagnosis/prognosis such as radiogenomics [8,9], fusion of pathology and genomics data [10,11], among many others. However, despite the large number of publications on using ML techniques for COVID-19, a recent review [12] of prediction models using multimodal data for COVID-19 concluded that "almost all published prediction models are poorly reported and at high risk of bias." Another review [13] of ML methods using CXR and CT concluded that "none of the models identified are of potential clinical use due to methodological flaws and/or underlying biases." As more multimodal data become available and more ML models are developed for COVID-19, we believe that an overview and critical appraisal of the state-of-the-art techniques are crucial to inform better development strategies and assessment of study designs. We emphasize that our focus is on multimodal data and ML models that integrate different types of data. Reviews of general ML development in COVID-19 can be found in the literature [12–14]; however, these reviews either focus on imaging based studies [13,14] or include models based on different types of data but not integration of multimodal data [12].

In this paper, we begin with a discussion of lessons learned from radiogenomic studies for cancer diagnosis. This actively-researched field, although on a different clinical task, investigates techniques and assessment methods to study the correlation and integration of multimodal data, which can shed light on a new disease such as COVID-19. We then summarize the types of data and ML techniques relevant to COVID-19. Finally, we present our perspectives on future development of ML techniques using multimodal data for COVID-19.

1.1. Lessons learned from radiogenomics studies for cancer diagnosis

Radiogenomic studies in cancer have generally made use of multimodal data to better understand and potentially improve cancer diagnosis, prognosis, and treatment. Historically, radiogenomic studies have combined transcriptomic data from tissue or blood samples with features extracted from images (a) to explore relationships between the underlying cancer biology and the disease presentation on medical images, and (b) to combine image and molecular features for better diagnosis and treatment planning. While this approach has not yet been leveraged for COVID-19, we describe it below as it has many parallels that may be explored in the future.

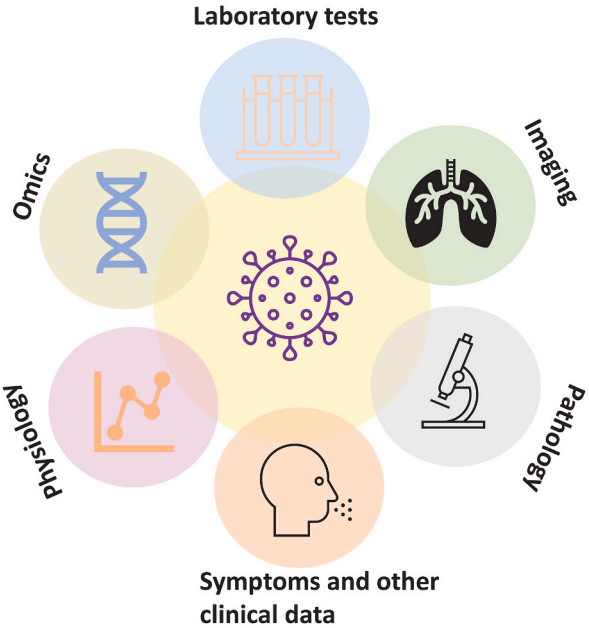


Fig. 1. Multimodal data for COVID-19.

Early studies of radiogenomics in cancer [8,15] focused on hepatocellular carcinoma [16–18], lung cancer [19–23], glioma [24, 25], breast cancer [26–28] and head and neck squamous carcinoma [22,29]. Early work in lung cancer radiogenomics showed that lung cancer radiographic images reflect important molecular properties of lung cancer patients including EGFR mutation status [30–32] and relevant gene expression pathways [33]. For example, an initial radiogenomic map of lung cancer patients showed that 56 gene expression metagenes can be predicted from CT image features with mean accuracy of 72% [21]. Similarly, radiogenomics analysis of head and neck squamous carcinoma showed that radiomic patterns reflect mutation status, DNA methylation patterns, histopathological diagnosis, and clinical outcome [29,34].

Breast cancer researchers collaborated in the MRI mapping of tumors to various clinical, molecular, and genomics markers of prognosis and risk of recurrence, including gene expression profiles [35–39]. The MRI mapping had been conducted on de-identified datasets of invasive breast carcinomas on which radiomic algorithms yielded computer-extracted quantitative lesion features from dynamic contrast-enhanced breast MRIs from The Cancer Imaging Archive (TCIA) [35] and clinical and genomic data from The Cancer Genome Atlas (TCGA) [36]. The investigators showed some specific imaging-genomic associations, such as a positive correlation between the transcriptional activities of genetic pathways and a blurred tumor margin indicating potentially tumor invasion into surrounding tissue. From association studies, they showed that MicroRNA (miRNA) expressions were associated with breast cancer tumor size and enhancement texture; indicating that miRNAs might mediate tumor growth and tumor heterogeneity of angiogenesis [35].

While an abundance of literature in radiogenomics has shown promising results, it should be noted that many of the studies involve small numbers of subjects and lack external validation, thereby limiting the generalizability of the ML technologies to patient populations at large and warranting further research [8,40,41].

1.2. Types of data for COVID-19

Multiple types of data have been investigated and used for developing ML algorithms for COVID-19 diagnosis/prognosis, including symptoms and other clinical data, laboratory tests, imaging, omics data, and pathology/physiology (Fig. 1).

Typical COVID-19 symptoms as listed by the US Center of Disease Control (CDC) [42] include fever, cough, shortness of breath, fatigue, muscle or body aches, headache, new loss of taste or smell, sore throat, congestion or runny nose, nausea or vomiting, and diarrhea. While the symptoms may range from mild to severe, they are usually recorded as binary features (e.g., presence or absence) in ML models. Besides patient symptoms, other important clinical data include patient demographics (age/gender/race), comorbidities, admission to ICU, use of assisted ventilation, and final clinical outcomes such as discharge (recover) or death, which can be used as severity and prognosis endpoints in ML models. For Clinical Laboratory Tests (CLTs), RT-PCR is currently the gold standard for detection of COVID-19 and is often used as the clinical endpoint for ML training and prediction. Besides RT-PCR, CLTs usually include complete blood count, a comprehensive metabolic panel and a coagulation panel, and additional tests such testing for inflammatory markers such as ESR, *C*-reactive protein, D-dimer etc [43].

The role of imaging has evolved during the pandemic. Chest radiography and CT imaging were initially suggested in some regions as an alternative and possibly superior testing method especially when RT-PCR was not widely available. Then, imaging applications evolved into different roles, such as for characterization of disease severity, examining COVID-19 manifestations in other organs, and investigation of the long-term sequelae of COVID-19 [5]. Moreover, lung ultrasound (LUS) was more widely employed by emergency and intensive care physicians in the UK during the COVID-19 pandemic, despite lacking patient outcome data associated with LUS findings [44]. Clinical research approaches using pulmonary magnetic resonance imaging (MRI) have been reported to be concordant with chest CT in manifestation of typical features of COVID-19 pneumonia and thus have been suggested as potential alternatives for patients who should avoid exposure to ionizing radiation [4]. However, MRI applications remain limited in clinical practice. Multimodal brain imaging studies have shown evidence for brain-related abnormalities in COVID-19; e.g., a recent study with longitudinal MRI scans showed significant longitudinal effects including reductions in gray matter thickness and global brain size in COVID-19 positive patients [45].

Besides sequencing of the virus itself, identification of variants, and their evolution and differential risk, genomic analysis of blood draws is the most common strategy to generate molecular data for COVID-19 patients [6]. Both the plasma and Peripheral Blood Mononuclear Cells (PBMCs) harbor important signals about health and disease. The plasma can be analyzed for cell free DNA & cell free RNA using sequencing, and proteins can be analyzed using proteomic and metabolomic technologies. PBMCs include lymphocyte, monocyte, and dendritic cells that can be analyzed with single cell multi-omics technologies. More specifically, by using innovative sequencing techniques the whole transcriptome, surface protein levels and TCR sequences can be simultaneously analyzed from the same cell. A recent study performed a comprehensive analysis of both the plasma and PBCMs using these omics technologies in 139 COVID-19 patients representing the spectrum of infection severities, and showed that a major immunological shift is present between mild and moderate disease [6]. While a full multi-omics analysis of both plasma and PBMCs might be prohibitive, prior studies can inform new specific omics analyses that can be customized to the clinical question to be answered of COVID-19 patients.

The physiology and pathophysiology of COVID-19 is still poorly understood in its entirety. The SARS-CoV-2 virus binds to the Angiotensin Converting Enzyme 2 (ACE2) receptors present in the lung [46] (also in heart, blood vessels, brain, intestine, Kidney, testis [46]) and results in generalized vasoconstriction in the lung, impacting gas exchange [47]. ACE2 binding also impacts diuretic and anti-inflammatory pathways, and may contribute to a cascade of vascular and vascular endothelial effects, increased cytokine production, increasing vascular responsiveness to inflammatory cytokines across multiple organs and systems [46]. Consistently, the relevant pathophysiological data for COVID-19 spans multiple organs and multiple spatial and temporal scales and spans both the immediate post-infection period and the Post-Acute Sequelae of Covid-19 (PASC). Pathophysiological mechanistic understanding of

PASC is even more incipient. PASC affects multiple organs and systems, with risk that seems independent of acute disease severity [48] resulting in lingering inflammation, increased cardiovascular risk, increased risk of clot formation, as well as changes in brain structure [45], erectile dysfunction [49], kidney damage [50], and long-term lung damage [51].

The combination of multimodal data, comprising clinical data, data from well controlled physiological experiments, data from electronic health records (EHR), pathology, real-world data (wearable sensors) with imaging and machine learning may prove essential to the understanding of the underlying physiology spanning multiple scales and organ systems, as well as explain the observed heterogeneity, and contribute to identifying therapeutic targets and interventions.

In Table 1, we list several publicly available datasets that contain both medical imaging and other types of data for COVID-19 studies. The number of patients in these datasets ranges from 100 to 3000. The most commonly used clinical imaging modalities in publicly available multimodal datasets for COVID-19 are chest CT and X-ray imaging. Non-imaging data in these datasets are mainly clinical data including patient demographics, symptoms, treatment(s) received, and clinical laboratory tests. Other types of data, while relevant to COVID-19, have rarely been found in public multimodal datasets. It is also worth noting that many of these multimodal datasets are not fully paired, i.e., missing data are common. For example, in the "integrative CT images and Clinical Features for COVID-19 (iCTCF)" dataset [52], out of the 1521 subjects, 1342 subjects had both CT and clinical data. In the Stony Brook University (SBU) dataset, out of 1384 subjects, for imaging data, only 458 subjects had CT scans, and 1365 subjects had X-ray images; for clinical measurements, 271 subjects had no records for any significant comorbidities or symptoms, and for different CLTs, the number of missing values ranges from 208 to 1170.

1.3. Overview of statistical and machine learning techniques

This section provides an overview of machine learning techniques utilizing multimodal data for COVID-19, which broadly include statistical analysis using semantic features, hand-crafted (human-engineered) features, and deep learning models. Table 2 summarizes the technological characteristics of these techniques with references that are overviewed in this section.

Generally, machine learning using medical images relies on the ability to ascribe features to images. There are two basic categories of features: (1) semantic features refer to words or phrases, which may have numerical codes associated with them, that describe images or regions of interest (ROIs) within them, and (2) computational features are numerical variables that can be computed directly from the images. Semantic features can be ascribed to images by experts or assigned based on machine learning using images and curated associated features for training. Hand-crafted computational features can be defined in advance as mathematical combinations of image pixel locations and intensity values that relate to visual characteristics, e.g., object shape, margin sharpness, gray or color value distributions, and texture, or they can be discovered by machine learning algorithms. Finally, it is often desirable to limit observation to certain parts of the image (e.g., a tumor in cancer studies, the lungs in a chest CT), giving rise to the need for image segmentation. While this has shown to be extremely important in cancer applications, regions of the lungs impacted by COVID may be hard for humans and computers to segment. End-to-end deep learning models learn features and identify regions automatically in a seamless pipeline and thereby do not rely on segmentation and human-engineered features. However, this flexibility is associated with the need for large amount of training data and poor interpretability.

Several studies [53–55,62] have focused on analyzing the statistical relationship between CT characteristics, clinical measurements, and laboratory findings in COVID-19 patients. In all of these, CT image features were extracted by multiple radiologists

Table 1
A list of Public COVID-19 datasets containing multimodal data (imaging and clinical data).

Dataset	Hosts	Sample size (#subjects)	Phenotype	Imaging modality	Non-imaging data
BIMCV COVID-19+[52]	https://bimcv.cipf.es/bimcv-projects/bimcv-covid19/	1131	COVID-19	CT, CR, DX	demographic data, radiological reports, CLTs
iCTCF [51]	https://ngdc.cncb.ac.cn/ictcf/;	1521	COVID-19 and non-COVID-19	CT	demographic data, medical history, CLTs
V2-COVID19-NII [53]	https://data.uni-hannover.de/dataset/cov-19-img	243	COVID-19 and non-COVID-19	CR, DX	demographic data, treatment, CLTs
Available in MIDRC (https://www.midrc.org; https://data.midrc.org)(~105,000 medical imaging studies, >44,000 patients, COVID-19+/−, CR, DX, CT and limited clinical data)					
PETAL RED CORAL[54,*]	https://data.midrc.org/explorer	1480	COVID-19+	CR, DX, CT	Demographic data, medical history, clinical lab tests
N3C-MIDRC**	https://ncats.nih.gov/n3c	3000	COVID-19	CR, DX	Electronic Health records
SBU dataset [54]	TCIA: https://doi.org/10.7937/TCIA.BBAG-2923 MIDRC: https://data.midrc.org/explorer	1384	COVID-19	CT, CR, DX, MRI, PET	demographic data, medical history, symptoms, CLTs
COVID-19-AR [55]	TCIA: https://doi.org/10.7937/tcia.2020.py71-5978 MIDRC: https://data.midrc.org/explorer	105	COVID-19	CT, CR, DX	demographic data, medical history, treatment

*Interoperability of two platforms: MIDRC for medical images, and BioData Catalyst for Electronic Health Record (EHR) data; ** Interoperability of two platforms: MIDRC for medical images, and N3C for EHR data.

Table 2
Overview of statistical and machine learning techniques for COVID-19 using multimodal data.

Type of data	Technological characteristics	References
CT images, clinical measurements, laboratory tests	Traditional statistical methods, correlation analysis	Xiong et al. [56], Qin et al. [57], Sun et al. [58], Dane et al. [59]
Physiology measurements and physical activities	Random forest model for early detection of COVID-19 infection	Mason et al. [60]
demographics, medications, laboratory tests, CPT and ICD codes	Fusion machine learning model to predict severity of COVID-19	Tariq et al. [61]
CT images and clinical features	Integrative deep learning model for prediction of COVID-19 morbidity and mortality	Ning et al. [51]

following clinical guidelines. Such manually-extracted CT features may include the following: lesion location (e.g., peripheral, central, both central and peripheral); lesion attenuation (e.g., ground glass opacities (GGOs), consolidations, and mixed GGOs); lesion pattern (e.g., patchy, oval); distribution of affected lobes; pleural effusion, air bronchogram, interstitial thickening or reticulation, tree-in-bud signs, etc. Statistical analyses have been applied to analyze the relationship between CT image features, clinical characteristics and COVID-19. Statistical analyses have, for example, included, chi-square or Fisher exact test for categorical features, Mann-Whitney *U* test or Student t-test for quantitative features, linear regression [53] and logistic regression [62], and Spearman or Pearson correlation analysis [53,54]. These analyses may shed light on the design of more complicated ML models for the integration of multimodal data or help with interpretation of ML models.

Mason et al. [56] developed random forest models for early detection of COVID-19 infection using multimodal data from a wearable device including photoplethysmography-based measurement of physiology parameters and accelerometry-based measurement of physical activities of patients. This unique effort yielded promising results with cross-validation on a limited dataset, thereby warranting further development in diverse populations. Tariq et al. [57] compared early, middle, and late fusion ML models using multi-modal data including demographics, medications, laboratory tests, CPT and ICD codes to predict the severity of COVID-19 at the time of testing and a COVID-19 positive patient's need for hospitalization. While promising results were reported, the authors admit that the study has limited data from a highly integrated academic healthcare system putting the generalizability of the models into question.

While most machine learning models use either Clinical Features (CFs) or imaging features as predictors, Ning et al. [52] developed an integrative CT images and CFs for COVID-19 (iCTCF) model that integrates the CFs with chest CT images for prediction of COVID-19 morbidity and mortality outcomes using a deep learning based algorithm trained with data from 1170 patients. In their model, a 13-layer Convolutional Neural Network (CNN) model was first used to classify individual CT slices into three types (non-informative, positive, negative) and the positive images were input to the second 13-layer CNN to predict patient outcomes. In the meantime, the CFs were combined by a 7-layer Deep Neural Network (DNN). Finally, the predictions using CT images and CFs were integrated using the penalized logistic regression model to output final predictions on morbidity or mortality outcomes of patients.

1.4. Perspective on future studies

We believe that development of machine learning models using multimodal data has great potential to help understand COVID-19 more comprehensively, provide clinical tools in the detection, diagnosis, and triage of patients as well as detection and treatment of long COVID complications. To inform future development of multimodal ML models for COVID-19, we present our perspective on several important aspects: data curation, model strategy, clinical relevance, and assessment.

Curation of high-quality and diverse multimodal datasets plays a vital role. It is critically important that a dataset for ML

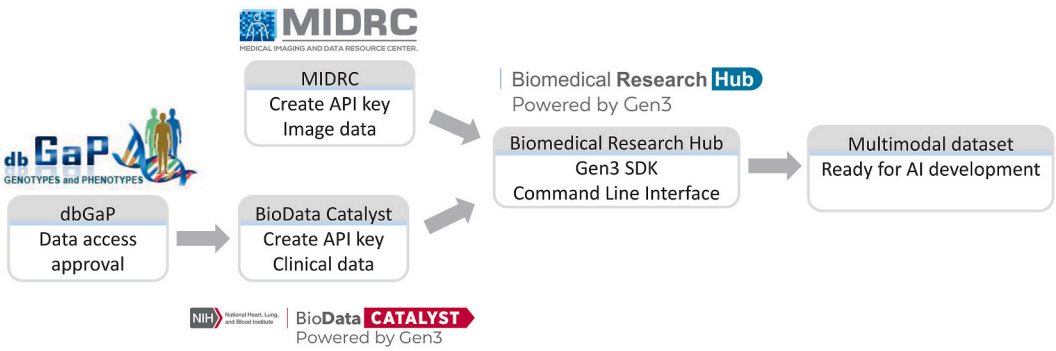


Fig. 2. Curation of the PETAL multimodal dataset involving two sources. The clinical data is hosted by Biodata Catalyst and access requires approval through dbGaP. The imaging data is hosted by MIDRC. Both data servers are powered by Gen3 and matching of the data from the two sources is achieved by a command line interface using the Gen3 Software Development Kit (SDK) on the Biomedical Research Hub, which requires Application Programming Interface (API) keys created on MIDRC and BioData catalyst.

development is diverse enough to represent the intended patient population and minimize bias. It is also crucial that a dataset is large enough to train a complex model such as deep neural networks to avoid overfitting. The curation of multimodal dataset is generally more challenging than a single-modality dataset due to its scarcity and the possible need for effort to match data from multiple sources. As an example, Fig. 2 shows the workflow of the curation of a multimodal dataset of US patients hospitalized with COVID-19 originally collected by the Prevention and Early Treatment of Acute Lung Injury (PETAL) Network [58]. The clinical data of 1480 patients hosted by BioData Catalyst (BDCat) [59] include patient demographics, medical history, symptoms, and laboratory results. While the dataset is publicly available, the access to these data needs an approval through dbGaP (https://www.ncbi.nlm.nih.gov/gap/). A subset of these patients have imaging data that are hosted by MIDRC including chest radiographs and CTs. Both data servers are powered by Gen3 (https://gen3.org/), which provides a command line interface using the Gen3 SDK to query/retrieve data. One can match the data from MIDRC and BDCat by using the API keys created from the two sources and a crosswalk service provided by the Biomedical Research Hub (BRH) (https://brh.data-commons.org/login).

Interoperability is also available between MIDRC and the National COVID Cohort Collaborative (N3C, https://ncats.nih.gov/n3c). MIDRC currently hosts imaging data from 3000 patients with the corresponding EHR data available in N3C. Medical imaging and EHR data can be associated using Privacy Preserving Record Linkage tokens through a linkage honest broker, and a crosswalk service provided by the Biomedical Research Hub (BRH). Access to the EHR data requires access to N3C, as set out in https://ncats.nih.gov/n3c/about/applying-for-access.

The availability of public datasets from multiple sources is beneficial to help improve data diversity. On the other hand, merging datasets from multiple sources may be challenging due to variations and biases in patient populations, data collection and image acquisition protocols. Great attention must be paid to the image quality differences across different sources that may lead to shortcut learning [60,61], i.e., models learning institution-specific markers or annotations instead of pathology in the image. Data harmonization techniques may help mitigate this issue. Representativeness of data is key, as this is a potentially significant source of bias. For COVID-19, different tasks will require different target population distributions; typically based on 'reference' distributions such as data from the US Census, COVID-19 diagnosis, hospitalization, and mortality present significant differences across age and self-reported racial groups, e.g., which need to be accounted for when creating training, tuning and test cohorts. Cross-repository cohort building, availability of desired data types and specific data may further limit sample size, or result in significant missing data. Indeed, missing data are common and oftentimes only a subset of patients in a multimodal dataset have data from all modalities. Moreover, even when multimodal data are available, they may be acquired at different times; such temporal misalignment must be considered for a disease such as COVID-19 that is evolving quickly. Multimodal machine learning model development must take these into consideration.

Multimodal machine learning has many design strategies to address medical problems, particularly for COVID-19. *Representation* refers to techniques for representing and summarizing multimodal data in a way that exploits the complementarity and redundancy of multiple modalities. *Correlation* studies investigate the quantitative correlation and relations between different modalities for a certain clinical task, which provides basis for translation or replacement of one modality to another (e.g., use of non-invasive testing to replace invasive ones). *Fusion* models aim to join information from multiple modalities thereby taking advantage of the complementarity to improve accuracy. Huang et al. [63] provide a systematic review and implementation guidelines on the fusion of medical imaging and EHR data using deep learning, in which, like earlier reviews in more general AI applications [64], pros and cons of different fusion strategies are discussed including early fusion (concatenating multimodal features), late fusion (integration decisions of models from individual models), and hybrid fusion models (allowing joint learning and optimizing features).

In addition to data driven associations, the ability to use established biology and physiology knowledge to create quantitative imaging or clinical features may contribute to improve model accuracy. Physiology driven mechanistic features and models can enrich phenomenological data-driven models through cohort normalization, missing data imputation, etc. [65]. Probably the most noteworthy example is the use of digital twins [66,67], where computational physiological models create constrained, plausible, individualized artificial data instances of data, including human physiology. Synthetic digital twins, the generation of which involves fusion of multimodal data, may be used to enrich datasets, probe for potential bias, or assess sensitivity to parameter changes in a multimodal machine learning model.

For a multimodal machine learning model to be clinically useful, it is crucial to target the right clinical question. In relation to the current stage of COVID-19, some of the most pressing questions are linked to PASC (long-COVID), especially who is likely to develop PASC; what mechanistic pathways may be common to multiple PASC presentations; and who is likely to respond to specific therapeutical interventions, e.g. from depression [68]. Considering the complexity of PASC and the large ongoing efforts to collect and curate imaging, omics, clinical and EHR data in long-COVID (the NIH Recover Initiative, MIDRC-N3C interoperability, for example), PASC may prove to be an important ground for the multimodal, multi-data type, machine learning approaches that are the focus of this manuscript.

Assessment methods play a critical role in the development of multimodal ML models to avoid pitfalls and enhance quality in the design, analysis, and reporting of ML studies. Statistical methods are essential for study design to appropriately collect and/or partition data for training, tuning, internal and external validations to avoid bias, for analysis of results to yield generalizable results with assessment of uncertainty. Tools, such as the PROBAST tool [69], are available to facilitate the development of ML models with minimized bias and better applicability. Many machine-learning-good-practice checklists/guidelines are also generally useful tools such as the checklist for artificial intelligence in medical imaging (CLAIM) [70] and the guidelines for publication of AI in medical physics [71].

2. Conclusion

Machine learning technologies using multimodal data have great potential to help better understand COVID-19 and provide clinical tools in the diagnosis, prognosis, and triage of patient as well as prediction and treatment of long COVID complications. Experience gained and lessons learned from radiogenomics and general multimodal machine learning technologies have and will continue to facilitate the development for COVID-19. Conversely, the experience gained through the development of multimodal COVID-19 machine learning will contribute to other efforts, such as cancer diagnosis and precision medicine [72]. For example, the MIDRC infrastructure and data curation techniques for data sharing, the interoperability of MIDRC with other data repositories as shown in this paper, and MIDRC developed tools such as the bias identification and mitigation tool [73] can be applied to ML models in other applications as well. We recognize that multimodal machine learning, whether for general medical applications or for COVID-19 specifically, is still in its infancy. A major bottleneck is the scarcity of high-quality multimodal data. Our MIDRC projects have put significant efforts into the curation of multimodal data. It can be expected that the availability of multimodal data and the development of ML models will greatly advance our understanding of COVID-19 and potentially lead to clinically impactful tools.

Author contribution statement

All authors listed have significantly contributed to the development and the writing of this article.

Data availability statement

This is a review paper and datasets under review were public and their sources were included in the paper.

Additional information

Supplementary content related to this article has been published online at [URL].

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Research reported is part of MIDRC and was made possible by the National Institute of Biomedical Imaging and Bioengineering (NIBIB) of the National Institutes of Health under contracts 75N92020C00008 and 75N92020C00021. This work is supported by NIH through the Data and Technology Advancement (DATA) National Service Scholar program.

References

[1] World Health Organization, Weekly Epidemiological Update on COVID-19 - 30 November 2022, 2022.
[2] R. Rana, A. Tripathi, N. Kumar, N.K. Ganguly, A comprehensive overview on COVID-19: future perspectives, Front. Cell. Infect. Microbiol. 11 (2021).
[3] A. Nalbandian, et al., Post-acute COVID-19 syndrome, Nat. Med. 27 (2021) 601–615.
[4] B.K.K. Fields, N.L. Demirjian, H. Dadgar, A. Gholamrezanezhad, Imaging of COVID-19: CT, MRI, and PET, Semin. Nucl. Med. 51 (2021) 312–320.
[5] J.P. Kanne, et al., COVID-19 imaging: what we know now and what remains unknown, Radiology 299 (2021) E262–E279.
[6] Y. Su, et al., Multi-omics resolves a sharp disease-state shift between mild and moderate COVID-19, Cell 183 (2020) 1479–1495.e1420.
[7] M. Busana, et al., The impact of ventilation–perfusion inequality in COVID-19: a computational model, J. Appl. Physiol. 130 (2021) 865–876.
[8] S. Napel, M. Giger, Special section guest editorial:radiomics and imaging genomics: quantitative imaging for precision medicine, J. Med. Imaging 2 (2015), 041001.
[9] E. Trivizakis, et al., Artificial intelligence radiogenomics for advancing precision and effectiveness in oncologic care, Int. J. Oncol. 57 (2020) 43–53.
[10] R.J. Chen, et al., Pan-cancer integrative histology-genomic analysis via multimodal deep learning, Cancer Cell 40 (2022) 865–878.e866.
[11] A. Cheerla, O. Gevaert, Deep learning with multimodal representation for pancancer prognosis prediction, Bioinformatics 35 (2019) i446–i454.
[12] L. Wynants, et al., Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal, BMJ 369 (2020) m1328.
[13] M. Roberts, et al., Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans, Nat. Mach. Intell. 3 (2021) 199–217.
[14] T.E. Komolafe, et al., Diagnostic test accuracy of deep learning detection of COVID-19: a systematic review and meta-analysis, Acad. Radiol. 28 (2021) 1507–1523.
[15] R. Colen, et al., NCI workshop report: clinical and computational requirements for correlating imaging phenotypes with genomics signatures, Transl Oncol 7 (2014) 556–569.
[16] M.D. Kuo, J. Gollub, C.B. Sirlin, C. Ooi, X. Chen, Radiogenomic analysis to identify imaging phenotypes associated with drug response gene expression programs in hepatocellular carcinoma, J. Vasc. Intervent. Radiol. 18 (2007) 821–831.
[17] M. Renzulli, et al., Can current preoperative imaging Be used to detect microvascular invasion of hepatocellular carcinoma? Radiology 279 (2016) 432–442.
[18] A. Sagir Kahraman, Radiomics in hepatocellular carcinoma, J. Gastrointest. Cancer 51 (2020) 1165–1168.
[19] A.K. Das, M.H. Bell, C.S. Nirodi, M.D. Story, J.D. Minna, Radiogenomics predicting tumor responses to radiotherapy in lung cancer, Semin. Radiat. Oncol. 20 (2010) 149–155.
[20] V.S. Nair, et al., Prognostic PET 18F-FDG uptake imaging features are associated with major oncogenomic alterations in patients with resected non-small cell lung cancer, Cancer Res. 72 (2012) 3725–3734.

[21] O. Gevaert, et al., Non-small cell lung cancer: identifying prognostic imaging biomarkers by leveraging public gene expression microarray data–methods and preliminary results, Radiology 264 (2012) 387–396.
[22] H.J. Aerts, et al., Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach, Nat. Commun. 5 (2014) 4006.
[23] S. Napel, W. Mu, B.V. Jardim-Perassi, H. Aerts, R.J. Gillies, Quantitative imaging of cancer in the postgenomic era: radio(geno)mics, deep learning, and habitats, Cancer 124 (2018) 4633–4649.
[24] P.O. Zinn, et al., Radiogenomic mapping of edema/cellular invasion MRI-phenotypes in glioblastoma multiforme, PLoS One 6 (2011), e25451.
[25] O. Gevaert, et al., Glioblastoma multiforme: exploratory radiogenomic analysis by using quantitative image features, Radiology 273 (2014) 168–174.
[26] S. Yamamoto, D.D. Maki, R.L. Korn, M.D. Kuo, Radiogenomic analysis of breast cancer using MRI: a preliminary study to define the landscape, AJR Am. J. Roentgenol. 199 (2012) 654–663.
[27] S. Yamamoto, et al., Breast cancer: radiogenomic biomarker reveals associations among dynamic contrast-enhanced MR imaging, long noncoding RNA, and metastasis, Radiology 275 (2015) 384–392.
[28] L.J. Grimm, Breast MRI radiogenomics: current status and research implications, J. Magn. Reson. Imag. 43 (2016) 1269–1278.
[29] C. Huang, et al., Development and validation of radiomic signatures of head and neck squamous cell carcinoma molecular features and subtypes, EBioMedicine 45 (2019) 70–80.
[30] S. Wang, et al., Predicting EGFR mutation status in lung adenocarcinoma on computed tomography image using deep learning, Eur. Respir. J. 53 (2019).
[31] R. Minamimoto, et al., Prediction of EGFR and KRAS mutation in non-small cell lung cancer using quantitative (18)F FDG-PET/CT metrics, Oncotarget 8 (2017) 52792–52801.
[32] O. Gevaert, et al., Predictive radiogenomics modeling of EGFR mutation status in lung cancer, Sci. Rep. 7 (2017), 41674.
[33] M. Zhou, et al., Non-small cell lung cancer radiogenomics map identifies relationships between molecular and imaging phenotypes with prognostic implications, Radiology 286 (2018) 307–315.
[34] P. Mukherjee, et al., CT-Based radiomic signatures for predicting histopathologic features in head and neck squamous cell carcinoma, Radiol Imaging Cancer 2 (2020), e190039.
[35] Y. Zhu, et al., Deciphering genomic underpinnings of quantitative MRI-based radiomic phenotypes of invasive breast carcinoma, Sci. Rep. 5 (2015), 17787.
[36] W. Guo, et al., Prediction of clinical phenotypes in invasive breast carcinomas from the integration of radiomics and genomics data, J. Med. Imaging 2 (2015), 041007.
[37] H. Li, et al., Quantitative MRI radiomics in the prediction of molecular classifications of breast cancer subtypes in the TCGA/TCIA data set, npj Breast Cancer 2 (2016), 16012.
[38] H. Li, et al., MR imaging radiomics signatures for predicting the risk of breast cancer recurrence as given by research versions of MammaPrint, oncotype DX, and PAM50 gene assays, Radiology 281 (2016) 382–391.
[39] E.S. Burnside, et al., Using computer-extracted image phenotypes from tumors on breast magnetic resonance imaging to predict breast cancer pathologic stage, Cancer 122 (2016) 748–757.
[40] H.X. Bai, et al., Imaging genomics in cancer research: limitations and promises, Br. J. Radiol. 89 (2016), 20151030.
[41] K. Pinker, et al., Background, current role, and potential applications of radiogenomics, J. Magn. Reson. Imag. 47 (2018) 604–620.
[42] Centers for Disease Control and Prevention of USA (CDC). Symptoms of COVID-19. https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html, 2021.
[43] M. Cascella, A. Aleem, et al., Features, Evaluation, and Treatment of Coronavirus (COVID-19), 2021. https://www.ncbi.nlm.nih.gov/books/NBK554776/.
[44] K. Jackson, R. Butler, A. Aujayeb, Lung ultrasound in the COVID-19 pandemic, Postgrad. Med. 97 (2021) 34.
[45] G. Douaud, et al., SARS-CoV-2 is associated with changes in brain structure in UK Biobank, Nature (2022).
[46] A.R. Bourgonje, et al., Angiotensin-converting enzyme 2 (ACE2), SARS-CoV-2 and the pathophysiology of coronavirus disease 2019 (COVID-19), J. Pathol. 251 (2020) 228–248.
[47] H. Karmouty-Quintana, et al., Emerging mechanisms of pulmonary vasoconstriction in SARS-CoV-2-induced acute respiratory distress syndrome (ARDS) and potential therapeutic targets, Int. J. Mol. Sci. 21 (2020) 8081.
[48] A. Subramanian, et al., Symptoms and risk factors for long COVID in non-hospitalized adults, Nat. Med. 28 (2022) 1706–1714.
[49] M. Kaynar, A.L.Q. Gomes, I. Sokolakis, M. Gül, Tip of the iceberg: erectile dysfunction and COVID-19, Int. J. Impot. Res. 34 (2022) 152–157.
[50] S. Yende, C.R. Parikh, Long COVID and kidney disease, Nat. Rev. Nephrol. 17 (2021) 792–793.
[51] J.T. Grist, et al., Lung abnormalities detected with hyperpolarized 129Xe MRI in patients with long COVID, Radiology 305 (2022) 709–717.
[52] W. Ning, et al., Open resource of clinical data from patients with pneumonia for the prediction of COVID-19 outcomes via deep learning, Nature biomedical engineering 4 (2020) 1197–1207.
[53] Y. Xiong, et al., Clinical and high-resolution CT features of the COVID-19 infection: comparison of the initial and follow-up changes, Invest. Radiol. (2020).
[54] D. Sun, et al., CT quantitative analysis and its relationship with clinical features for assessing the severity of patients with COVID-19, Korean J. Radiol. 21 (2020) 859.
[55] B. Dane, G. Brusca-Augello, D. Kim, D.S. Katz, Unexpected findings of coronavirus disease (COVID-19) at the lung bases on abdominopelvic CT, Am. J. Roentgenol. 215 (2020) 603–606.
[56] A.E. Mason, et al., Detection of COVID-19 using multimodal data from a wearable device: results from the first TemPredict Study, Sci. Rep. 12 (2022) 3463.
[57] A. Tariq, et al., Patient-specific COVID-19 resource utilization prediction using fusion AI model, npj Digital Medicine 4 (2021) 94.
[58] I.D. Peltan, et al., Characteristics and outcomes of US patients hospitalized with COVID-19, Am. J. Crit. Care 31 (2022) 146–157.
[59] Lung National Heart, Blood Institute, National Institutes of Health, U.S, Department of health and human services. The NHLBI BioData catalyst, in: (National Heart, Lung, and Blood Institute, National Institutes of Health, U.S. Department of Health and Human Services, 2020.
[60] J.D. López-Cabrera, R. Orozco-Morales, J.A. Portal-Díaz, O. Lovelle-Enríquez, M. Pérez-Díaz, Current limitations to identify covid-19 using artificial intelligence with chest x-ray imaging (part ii). The shortcut learning problem, Health Technol. (2021) 1–15.
[61] A.J. DeGrave, J.D. Janizek, S.-I. Lee, AI for radiographic COVID-19 detection selects shortcuts over signal, Nat. Mach. Intell. 3 (2021) 610–619.
[62] L. Qin, et al., A predictive model and scoring system combining clinical and CT characteristics for the diagnosis of COVID-19, Eur. Radiol. 30 (2020) 6797–6807.
[63] S.C. Huang, A. Pareek, S. Seyyedi, I. Banerjee, M.P. Lungren, Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines, NPJ Digit Med 3 (2020) 136.
[64] T.a.A. Baltrusaitis, Chaitanya, Louis-Philippe Morency, Multimodal machine learning: a survey and taxonomy, IEEE Trans. Pattern Anal. Mach. Intell. 41 (2019) 423–443.
[65] E. Jones, et al., Phenotyping heart failure using model-based analysis and physiology-informed machine learning, J. Physiol. 599 (2021) 4991–5013.
[66] G. Coorey, et al., The health digital twin to tackle cardiovascular disease—a review of an emerging interdisciplinary field, npj Digital Medicine 5 (2022) 126.
[67] E.A. Stahlberg, et al., Exploring approaches for predictive cancer patient digital twins: opportunities for collaboration and innovation, Front Digit Health 4 (2022), 1007784.
[68] T. Oakley, J. Coskuner, A. Cadwallader, M. Ravan, G. Hasey, EEG Biomarkers to Predict Response to Sertraline and Placebo Treatment in Major Depressive Disorder, IEEE Trans Biomed Eng, 2022.
[69] R.F. Wolff, et al., PROBAST: a tool to assess the risk of bias and applicability of prediction model studies, Ann. Intern. Med. 170 (2019) 51–58.
[70] J. Mongan, L. Moy, Charles E. Kahn, J. Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers, Radiology: Artif. Intell. 2 (2020), e200029.

[71] I. El Naqa, et al., AI in medical physics: guidelines for publication, Med. Phys. 48 (2021) 4711–4714.
[72] K.M. Boehm, P. Khosravi, R. Vanguri, J. Gao, S.P. Shah, Harnessing multimodal data integration to advance precision oncology, Nat. Rev. Cancer 22 (2022) 114–126.
[73] D. Karen, et al., Toward fairness in artificial intelligence for medical image analysis: identification and mitigation of potential biases in the roadmap from data collection to model deployment, J. Med. Imag. 10 (2023), 061104.