The Utility of the Virtual Imaging Trials Methodology for Objective Characterization of AI Systems and Training Data

Fakrul Islam Tushar[1,2], Lavsen Dahal[1,2], Saman Sotoudeh-Paima[1,2], Ehsan Abadi[1,2], William P. Segars[1,2], Joseph Y. Lo[1,2], Ehsan Samei[1,2]

[1]Center for Virtual Imaging Trials, Carl E. Ravin Advanced Imaging Laboratories, Dept. of Radiology, Duke University School of Medicine, Durham, NC 27705, USA

[2]Department of Electrical & Computer Engineering, Pratt School of Engineering, Duke University, Durham, NC 27705, USA

Corresponding author: Fakrul Islam Tushar (e-mail: tushar.ece@duke.edu).

ABSTRACT

Purpose: The credibility of Artificial Intelligence (AI) models for medical imaging continues to be a challenge, affected by the diversity of models, the data used to train the models, and applicability of their combination to produce reproducible results for new data.

Approach: In this work we aimed to explore if the emerging Virtual Imaging Trials (VIT) methodologies can provide an objective resource to approach this challenge. The study was conducted for the case example of COVID-19 diagnosis using clinical and virtual computed tomography (CT) and chest radiography (CXR) processed with convolutional neural networks. Multiple AI models were developed and tested using 3D ResNet-like and 2D EfficientNetv2 architectures across diverse datasets.

Results: The performance differences were evaluated in terms of the area under the curve (AUC) and the DeLong method for AUC confidence intervals. The models trained on the most diverse datasets showed the highest external testing performance, with AUC values ranging from 0.73-0.76 for CT and 0.70-0.73 for CXR. Internal testing yielded higher AUC values (0.77 -0.85 for CT and 0.77-1.0 for CXR), highlighting a substantial drop in performance during external validation, which underscores the importance of diverse and comprehensive training and testing data. Most notably, VIT approach provided objective assessment of the utility of diverse models and datasets while further providing insight into the influence of dataset characteristics, patient factors, and imaging physics on AI efficacy.

Conclusions: The VIT approach can be used to enhance model transparency and reliability, offering nuanced insights into the factors driving AI performance and bridging the gap between experimental and clinical settings.

KEYWORDS

Virtual Imaging trials, In Silico, Computed tomography, Chest Radiography, Machine Learning, Deep Learning, AI, Image Processing, Generalizability, Explainability, Pneumonia, COVID-19

I. INTRODUCTION

Radiology artificial intelligence (AI) models often struggle to generalize, resulting in limited clinical applicability [1]. This is primarily due to the existential limits on the diversity of data upon which a model can be trained. Obviously, models trained on small datasets, such as from a single center, do not represent the general population. But even large, multi-center datasets can still be plagued by inconsistency across scanner vendors, acquisition protocols, reconstruction algorithms, pre/post-processing, or patient inclusion criteria. Consequently, AI developers continuously strive for massive amounts of data, hoping that the large magnitude can overcome the generalizability problem.

Failure of medical imaging AI models to generalize is a pervasive problem. The crisis of reproducibility was starkly evident during the COVID-19 pandemic when chest radiography (CXR) and computed tomography (CT) were initially employed for detecting and managing lung infections [2, 3]. In the rush to develop AI aides for radiologists, however, many studies reported unrealistic, near-perfect performances that dropped almost to chance upon external testing [4-10]. The availability of numerous large public datasets of medical images, including those from the Medical Imaging and Data Resource Center (MIDRC), has led to a plethora of AI models for the diagnosis of COVID-19. Yet a review of 62 studies asserted that none of these models were fit for clinical use due to methodological flaws and underlying biases [19]. While the focus on COVID-19 is waning and imaging is no longer used for the primary diagnosis of the disease, with some exception [11], the rare combination of so much data accompanied by widespread problems in reproducibility offers our field a rare opportunity to understand how best we can appropriate AI methods, for both clinical practice as well as future health crises.

A promising solution to this ongoing reproducibility crisis in AI lies in the use of the Virtual Imaging Trial (VIT) approach. Simulating the three key components of an imaging trial, patients, scanners, and readers [27], VITs offer control over input variables to generate virtual images representing a diverse range of patient characteristics and imaging techniques. Diseases can be simulated with pixel-level ground truth in terms of their location, size, and characteristic features. Simulations can also encompass different scanner technologies or acquisition protocols. Given the precise controls used to generate this data, VITs can elucidate which factors drive model performance. This approach facilitates not only truly independent external validation but also rigorous and unbiased testing across diverse scenarios. The VIT approach has been applied to a wide range of diseases and modalities, including lesion detection in mammography and breast tomosynthesis [28-30], nodule detection, and COPD quantification in chest CT. Previous work involved VITs for validating deep-learning models for COVID-19 detection using clinical and virtual datasets [33, 34]. Arun et al. [35] highlighted the limitations of Grad-CAM due to repeatability and reproducibility issues.

This study aimed to ascertain how the VIT methodology can add objectivity, explainability, and overall generalizability to the AI process. The study was done in the context of COVID-19 diagnosis, given its associated diversity of models and data. Building upon our prior validated, open-source deep-learning models for case-level COVID-19 detection with CT and CXR images [33, 34], the study modeled virtual patients replicating a diverse range of anatomies and manifestations of COVID-19 pneumonia [31, 32]. The virtual patients were imaged using simulated scanners, replicating the physical and technical characteristics of actual medical imaging devices. The AI model “readers,” representing radiologists, read the images, allowing us to evaluate the diagnostic performance of AI models under consistent conditions. Comparing with results from multiple clinical datasets and multiple AI models, we aimed to unpack the interplay of dataset-model matching and mismatching on the results, to evaluate the influence of patient- and physics-based factors on the generalizability of the results, and to assess the utility of VIT as an independent validation to provide a controlled environment for evaluating AI models.

II. METHODS

Institutional Review Board approval was obtained for this exempt study. The study, detailed below, deployed anonymized clinical image datasets as well as simulated virtual data. The clinical datasets were multiple [12, 14-16, 18, 23, 36-39], varying in size, diversity, demographics, and class definitions. The virtual data were from a population of 4D-XCAT models with varying COVID-19 size and distribution, imaged using virtual CT and CXR scanners (DukeSim, CVIT, Duke University) [31].

Multiple convolutional neural network (CNN) models (detailed below) with residual connections were developed to process CT and CXR images efficiently. These lightweight CNNs, designed to reduce computational complexity while maintaining high accuracy, were used to classify cases as positive or negative for COVID-19. The CNN models were trained using single and various combinations of clinical datasets. In parallel experiments, CT or CXR clinical data were analyzed for internal and external performance shift.

The virtual data were reserved as a separate external validation. By varying the virtual imaging trial parameters, they were also used to assess how performance may be affected by factors pertaining to the patients (i.e., infection size) or imaging physics (i.e., effective dose and modalities). An illustration of the overall workflow of the analysis is presented in Figure 1.

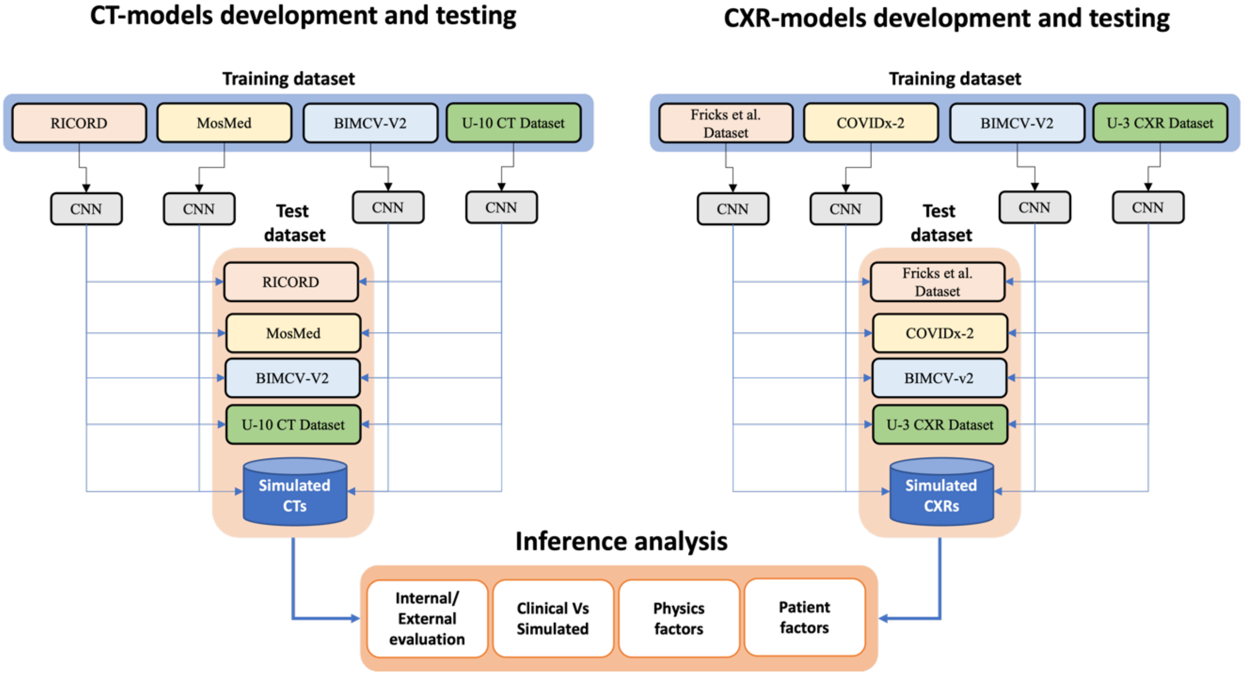


FIGURE 1. Study design overview. 12,844 CT scans and 25,219 CXR images for COVID-19 diagnosis were drawn from 13 clinical datasets comprising single or multiple centers (Supplement Fig. 1-2). Multiple deep-learning-based models were developed using these clinical datasets. All models underwent internal testing (held-out from the same training dataset) and external testing (all other datasets). Further external testing was performed using virtually simulated CT and CXR images to analyze effect of patient and imaging physics factors.

II.A. Clinical Dataset

Define abbreviations and The clinical CT data included a total of 12,844 volumes of 7,452 patients from 10 datasets: RICORD [18], MosMed [15] BIMCV-COVID-19 +/- (BIMCV-V2) [14], COVID-CT-MD [12], CT Images in COVID-19 [13], PleThora [39], COVID19-CT-dataset [36], Stony Brook University COVID-19 Positive Cases (COVID-19-NY-SBU) [16], A Large-Scale CT and PET/CT Dataset for Lung Cancer Diagnosis (Lung-PET-CT-Dx) [37], and Lung Image Database Consortium / Image Database Resource Initiative (LIDC-IDRI) [40]. These datasets had different prevalences of COVID-19 positive and negative images (Figure 2a) and demographics. Summary statistics of the CT datasets are detailed in Table 1.

Furthermore, all ten clinical CT datasets above were combined to create the U-10 CT dataset, which provides a more diverse dataset for factors such as patient population and demographics, disease appearances, CT systems, and imaging protocols. Figure 4 shows the inclusion and exclusion criteria followed in the curation of the clinical CT data.

CXR analysis included 25,219 clinical CXR images collected from three datasets: Fricks et al. [23], BIMCV [14], and COVIDx-CXR-2 [38]. These datasets also had different prevalences of COVID-19 positive and negative images (Figure 2b) and demographics. All three clinical CXR datasets were also combined to form the U-3 CXR dataset, with one important caveat. In one of the datasets, COVIDx-CXR-2, positive images were from different sources, but the negative class was much larger and mainly from one source, namely the RSNA Pneumonia Detection Challenge [41] (Figure 2b). To ensure a balanced training and validation process for the unified U-3 dataset, the negative cases were randomly subsampled to achieve an equal distribution between the two classes, which also provided a more balanced contribution from this particular dataset. Summary statistics of the CXR datasets are detailed in Table 2.

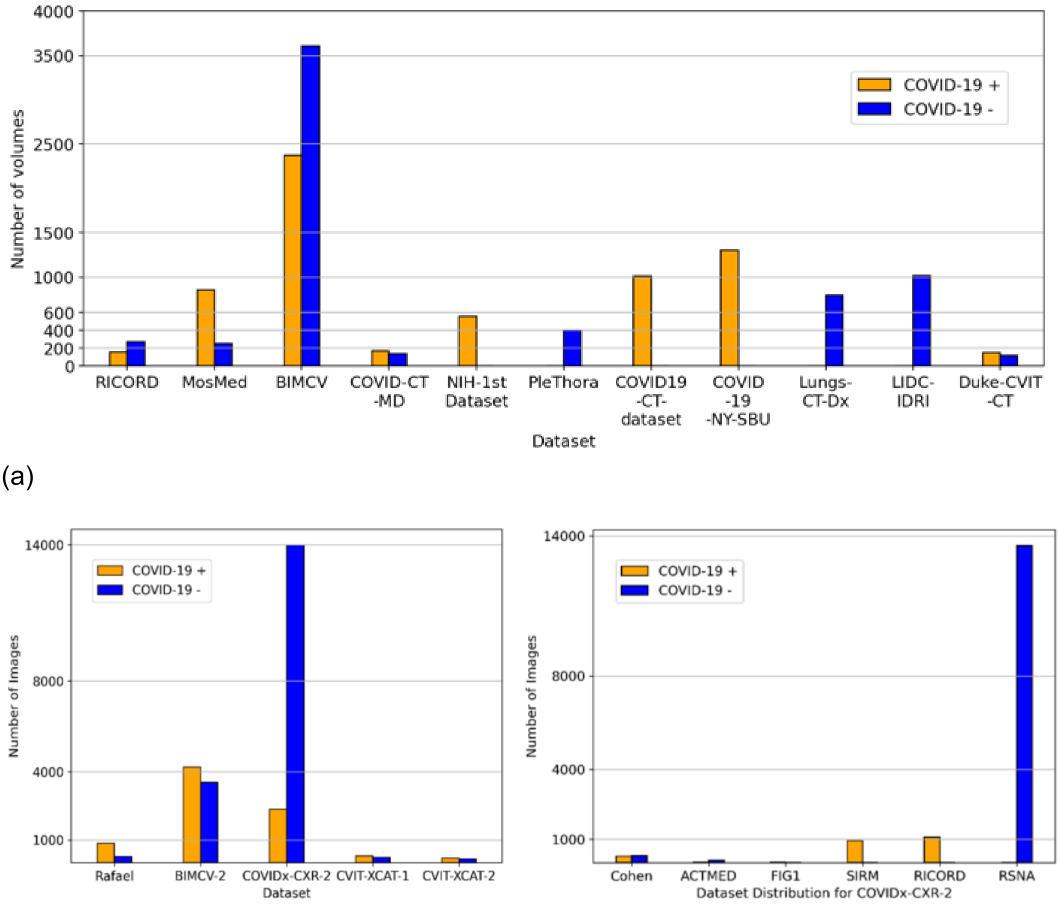
TABLE 1. Clinical CT patient datasets utilized in model development and testing. The combination of all ten constitutes the U-10 CT dataset. Demographic values are reported as the percentage of patient sex and mean of patient age.

No	Dataset	Source	Demographics	Category	Train*	Validation*	Test*
1.	RICORD [18] (1b,1b)	Turkey, USA, Canada, Brazil	44% women	COVID+	66 (90)	22 (32)	22 (33)
			Age 54 ±17	COVID-	70 (72)	23 (23)	24 (25)
				Total	**136 (162)**	**45 (55)**	**46 (58)**
2.	MosMed [15]	Russia	56% women Age 47	COVID+	512 (512)	170 (170)	174(174)
				COVID-	152 (152)	50 (50)	52 (52)
				Total	**664 (664)**	**220 (220)**	**226 (226)**
3.	BIMCV-V2 [14]	Spain	42% women.	COVID+	455 (1421)	152(484)	152(470)
			Age 64 ±16	COVID-	728 (2077)	239(706)	268(823)
				Total	**1183 (3498)**	**391 (1190)**	**420 (129)**
4.	COVID-CT-MD [12]	Iran	40% women.	COVID+	101(101)	33 (33)	35 (35)
			Age 51 ±16	COVID-	81 (81)	27 (27)	28 (28)
				Total	**182 (182)**	**60 (60)**	**63 (63)**
5.	An et al. [13]	Multi-center	N/A	**COVID+**	379 (391)	126 (129)	127 (130)
6.	PleThora[39]	USA	31% women. Age 68 ± 10	**COVID-**	241 (241)	80 (80)	81 (81)
7.	COVID19-CT [36]	Iran	39.1% women Age: 47 ± 16	**COVID+**	604 (604)	201 (201)	202 (202)
8.	COVID-19-NY-SBU[16]	USA	43% women. (Age: ranges between 18-90 years)	**COVID+**	251 (739)	84 (278)	84 (282)
9.	Lungs-CT-Dx [37]	China	46% women, Age 61 ± 10	**COVID-**	207 (479)	69 (154)	70 (164)
10.	LIDC-IDRI [40]	USA	N/A	**COVID-**	606 (611)	202 (204)	202 (203)
	Total / U-10 CT				**4453 (7571)**	**1478 (2571)**	**1521 (2702)**

Note-* Number of patients (number of scans), COVID+= COVID-19 positive, COVID-= COVID-19 negative, COVID-19-NY-SBU = Stony Brook University COVID-19 Positive Cases, Lungs-CT-Dx= A Large-Scale CT and PET/CT Dataset for Lung Cancer Diagnosis.

TABLE 2. Clinical CXR Patient Cohorts utilized in model development and testing. Demographic values are reported as the percentage of patient sex and mean of patient age.

No	Dataset	Source	Demographics	Category	Train	Validation	Test
1.	Fricks et al.[23]	Iran, Italy, USA	N/A	COVID+	544	136	171
				COVID-	174	44	55
				Total	**718**	**180**	**226**
2.	BIMCV-V2 [14]	Spain	46% Women Age 63 ± 17	COVID+	2694	674	843
				COVID-	2265	566	708
				Total	**4959**	**1240**	**1551**
3.	COVIDx-CXR-2 [38]	Multi-center	N/A	COVID+	1727	431	200
				COVID-	11034	2759	200
				Total	**12761**	**3190**	**400**
	Total		N/A		**18438**	**4610**	**2177**
4	**U-3 CXR dataset**		N/A	COVID+	4965	1241	1214
				COVID-	4965	1241	963
				Total	**9930**	**2482**	**2177**



(a)

(b)

FIGURE 2. Histograms showing distribution of COVID-19 positive (+) and negative (-) cases among different datasets (clinical and simulated) (a) CTs and (b) CXRs. In the latter, COVID-CXR-2 is further decomposed into its subsets. Log scale is used to show the large variation in numbers of exams. Note that the prevalence varies greatly, and some datasets contain only one class.

II.B. Virtual Dataset

The XCAT computational phantoms used in this study were based on the method described in detail by Abadi et al. [31]. An overview of the method is illustrated in Figure 3. Creating computational phantoms for COVID-19 is a process that unfolds in four distinct stages: constructing the body framework, detailing the morphological characteristics of lung abnormalities, replicating the texture and composition of affected lung tissues, and performing simulated scans to generate the virtual images.

Body Framework Construction: The process began with the development of the normal anatomy using the 4D extended cardiac-torso (XCAT) model from Duke University [42, 43]. The XCAT model provided a comprehensive foundation with detailed anatomy, dynamic organ motions, and textured tissues, built from real patient data spanning a range of patient characteristics such as sex, body size, and lung volume. Fifty separate phantoms were used for this study.

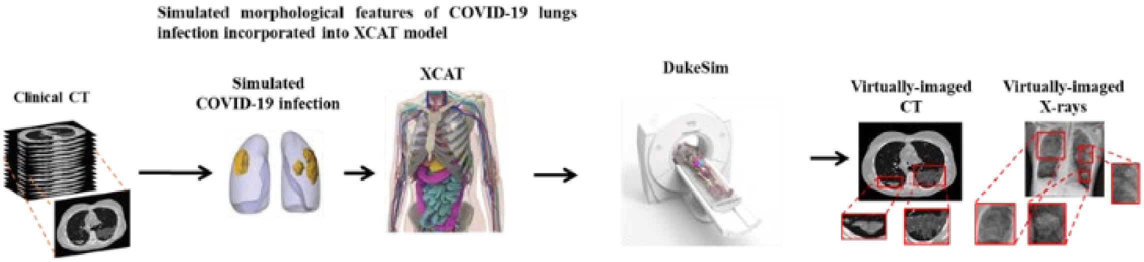

Simulated morphological features of COVID-19 lungs infection incorporated into XCAT model
Clinical CT
Simulated COVID-19 infection
XCAT
DukeSim
Virtually-imaged CT
Virtually-imaged X-rays

FIGURE 3. An overview COVID-19 computation phantoms development and simulated CT and CXR images.

Detailing Lung Abnormalities: The second stage involved the meticulous detailing of lung abnormalities typical of COVID-19, such as ground-glass opacities (GGO) and consolidations. This was achieved by examining CT scans from clinically confirmed COVID-19 patients (N=20), where the abnormalities were manually segmented and modeled in a series of surfaces mimicking the morphology [31]. These modeled features were then integrated into the XCAT phantoms, ensuring a match in body dimensions, sex, and age, to represent the disease's manifestations within the computational models accurately.

Replicating Lung Tissue Composition: The last phase involved fine-tuning the phantom's lung textures and materials to mirror the properties of the lung tissues affected by COVID-19 within the phantoms. This involved adjusting the lung parenchyma's texture in the computational model to reflect the changes observed in actual CT images, such as the addition of fluids in the case of GGO or the uniform texture seen in consolidations. These adjustments ensured that the simulated lung tissues closely mimicked the radiological features of COVID-19, allowing for realistic simulation outcomes.

Simulating the Imaging Process: Virtual CT and CXR datasets were generated by scanning the virtual patient models with or without the disease using an X-ray image acquisition simulator (DukeSim, CVIT, Duke University) [31, 32]. DukeSim is designed to replicate the physical processes in x-ray imaging with CT and CXR, including modeling of x-ray tube spectra, scanner geometry, and detector configuration. DukeSim combines ray tracing for rapid image generation and Monte Carlo techniques for accurate modeling of scatter, attenuation, and detector noise. The virtual framework allowed the scanning of the same virtual patient with both modalities without other confounding factors. Virtual scans were repeated at different effective doses (0.01, 0.1, 0.3, 1.6, 5.6, and 11.2 mSv). The dose settings were selected to represent a wide range of clinical applicability, as well as a direct comparison of CT and CXR images at the same hypothetical dose and motion state. For the CXR acquisitions, two commercial post-processing algorithms (denoted as algorithms A and B to maintain confidentiality) were applied to examine the effects of vendor heterogeneity. Table 3 shows the characteristics of the generated CT and CXR images.

TABLE 3. Virtual (CVIT-COVID) dataset attributes, including imaging protocols and disease distributions.

	Number of virtual exams	
Effective dose (mSv)	**COVID-19**	**Negative**
	CVIT-COVID-CT	
0.3	50	40
1.6	50	40
5.6	50	40
11.2	50	40
Total (CT)	**200**	**160**
	CVIT-COVID-CXR	
0.01	50	40
0.10	50	40
0.3	50	40
Total (CXR)	**150**	**120**

II.C. Pre-Processing

Standard preprocessing was performed on both CT and CXR images. Each CT volume was resampled to voxel dimensions of 2 mm × 2 mm × 5 mm (w, h, d). Intensities were clipped between -1000 to 500 HU, then standardized to a mean of 0 and standard deviation of 1. To reduce computational cost and the influence of background organs, three-dimensional (3D) patches of size 160×160×96 (w, h, d) were centered about the lungs. The patch size was based on average lung size plus a margin to allow for patient variability. CXR images were resized and randomly cropped to a size of 300x384 pixels, then standardized to 0.5 mean and 0.5 standard deviation to maintain consistency with the pre-trained dataset.

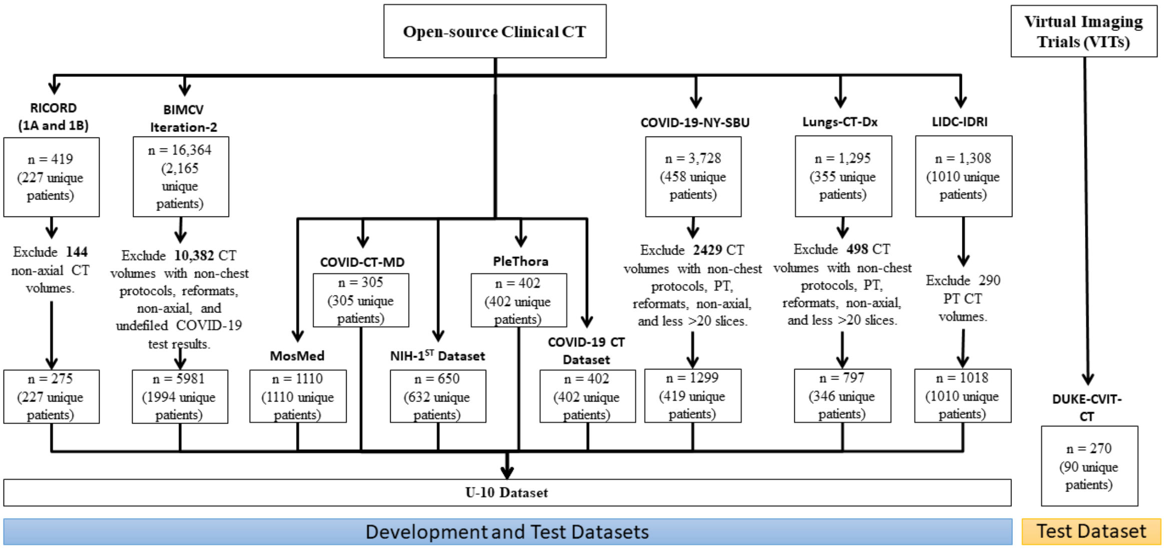

Open-source Clinical CT
Virtual Imaging Trials (VITs)
RICORD (1A and 1B)
n = 419 (227 unique patients)
Exclude 144 non-axial CT volumes.
n = 275 (227 unique patients)
BIMCV Iteration-2
n = 16,364 (2,165 unique patients)
Exclude 10,382 CT volumes with non-chest protocols, reformats, non-axial, and undefiled COVID-19 test results.
n = 5981 (1994 unique patients)
MosMed
n = 1110 (1110 unique patients)
COVID-CT-MD
n = 305 (305 unique patients)
NIH-1ST Dataset
n = 650 (632 unique patients)
PleThora
n = 402 (402 unique patients)
COVID-19 CT Dataset
n = 402 (402 unique patients)
COVID-19-NY-SBU
n = 3,728 (458 unique patients)
Exclude 2429 CT volumes with non-chest protocols, PT, reformats, non-axial, and less >20 slices.
n = 1299 (419 unique patients)
Lungs-CT-Dx
n = 1,295 (355 unique patients)
Exclude 498 CT volumes with non-chest protocols, PT, reformats, non-axial, and less >20 slices.
n = 797 (346 unique patients)
LIDC-IDRI
n = 1,308 (1010 unique patients)
Exclude 290 PT CT volumes.
n = 1018 (1010 unique patients)
DUKE-CVIT-CT
n = 270 (90 unique patients)
U-10 Dataset
Development and Test Datasets
Test Dataset

Figure 4. Flowchart of inclusion and exclusion criteria for the chest CT scans. n= number of CT volumes. A total of 16,949 CT scans of 11,166 patients were used for model development and testing. There were ten clinical datasets: RICORD [18], MosMed [15], BIMCV-COVID-19 +/- (BIMCV-V2),[14] COVID-CT-MD [12], CT Images in COVID-19 [13], PleThora [39], COVID19-CT-dataset,[36] Stony Brook University COVID-19 Positive Cases (COVID-19-NY-SBU) [16], A Large-Scale CT and PET/CT Dataset for Lung Cancer Diagnosis (Lungs-CT-Dx) [37], and Lung Image Database Consortium / Image Database Resource Initiative (LIDC-IDRI) [40], These ten clinical datasets were united into the U-10 CT Dataset. Additionally, simulated data were from the Center for Virtual Imaging Trials CT Dataset, Duke-CVIT-CT [31].

II.D. Model Development and Training

As noted above, previous studies have shown that complex deep learning models can reproduce non-generalizable near-perfect performance due to fundamental overtraining [33, 34]. To minimize this effect, we intentionally selected lightweight ResNet-like models [44, 45] and trained four separate CT-based models using the RICORD, MosMed, BIMCV, and U-10 CT datasets. The ResNet architecture has also shown consistent performance across various medical imaging tasks [33, 44, 45]. Similarly, for CXR, we trained four different EfficientNetv2 [46] models using the data from Fricks et al., BIMCV, COVIDx-CXR-2, and U-3 CXR datasets, respectively. Each dataset was randomly divided by the patient into subsets of training (60%), validation (20%), and testing (20%). No cross-validation was performed; instead, we utilized a train-validation-test split. As we aimed to assess the utility of virtual data for clinically trained algorithms, no training was applied to the virtual data. Instead, the model trained on clinical data was applied to the entire dataset for testing.

CT models used a simple 3D CNN inspired by ResNet [47], the architecture is shown in Figure 5. After initial convolution, features were learned across two resolution scales, then halved by max-pooling (pooling size 2×2×2) while doubling the number of filters. The last R-block features underwent batch normalization, rectified linear unit (ReLu), global max-pooling, dropout (dropout rate 0.5), and finally, a dense classification layer with sigmoid activation for binary case-level COVID-19 detection. Additionally, we applied L2 regularization with a coefficient of 0.001 to prevent overfitting. The stochastic gradient descent (SGD) optimizer was used to optimize the weights with decay learning rate, and weighted binary cross-entropy was used as the loss function. Weights were initialized to a uniform distribution. To retain the natural prevalence, no class balancing was performed during training.

The hyperparameters for the CT models were set as follows: initial learning rate of 1e-6, maximum learning rate of 1e-4, learning rate decay of 1e-2, batch size of 24, and 300 training epochs. CXR models were based on Efficientnetv2 with the original architecture [46], SGD was selected as the optimizer with the learning rate scheduler, [48] initial learning rate of 0.01, and cross-entropy loss. All models were developed using Python TensorFlow v2.6 and PyTorch deep learning frameworks.

Using a parallel computing cluster with eight 48 GB GPUs, we achieved an average of 36 virtual CT scans per hour and generated each CXR scan in under one minute. Model training utilized lightweight architectures, requiring ~8–16 GB VRAM to balance efficiency and performance.

All model weights, initial hyperparameters, and code are made publicly accessible [49].

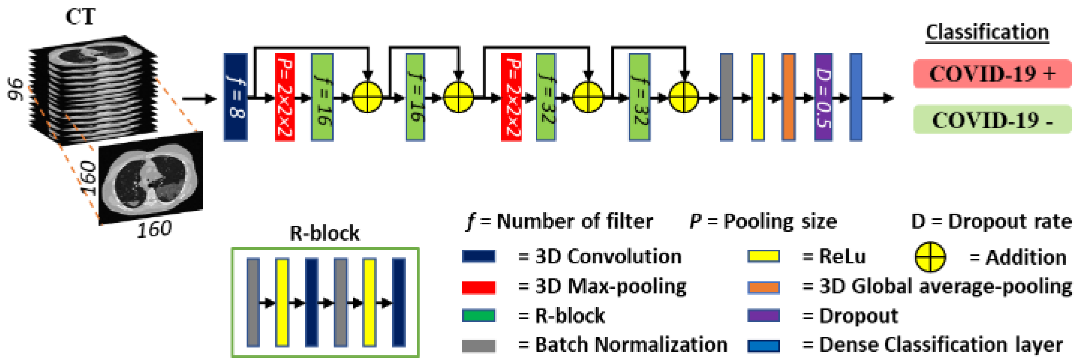


Figure 5. 3D CNN architecture for CT classification of COVID-19. The classification module is a 3D Resnet-like model with 2 R-Blocks in each resolution. The number of filters is denoted as f. The final output is a tensor of the probability of being COVID-19 positive or negative.

II.E. Evaluation and Statistical Analysis

We conducted a series of evaluations to assess the model performance on clinical and virtual datasets. For each clinical dataset, we followed the procedure of the prior studies of performing binary classification of the presence or absence of COVID-19 for the patient as a whole. For the virtual dataset, the controlled simulation process allowed us to evaluate further the influence of input variables related to the patient as well as the image acquisition.

We first evaluated the effect of the patient-related factor of infection size to understand the impact of infection severity on model performance. The virtual COVID-19 pneumonia cases were divided into two groups: "higher" infection (above the median value of 2.6% of total lung volume) and "lower" infection (below this median value). This approach helps in assessing how well the AI models perform across a spectrum of disease severity and identifying any performance biases or limitations. Additionally, we conducted evaluations based on the physics of image acquisition, specifically the imaging modality of CXR vs. CT, as well as a wide range of effective radiation doses.

To support our findings and assess the significance of the results, all performances were evaluated using the receiver operating characteristic area under the curve (AUC) with 95% confidence interval (CI) calculated by the DeLong method as implemented by pROC 1.16.2 in R 3.6.1 with 2000 bootstrapping samples [50].

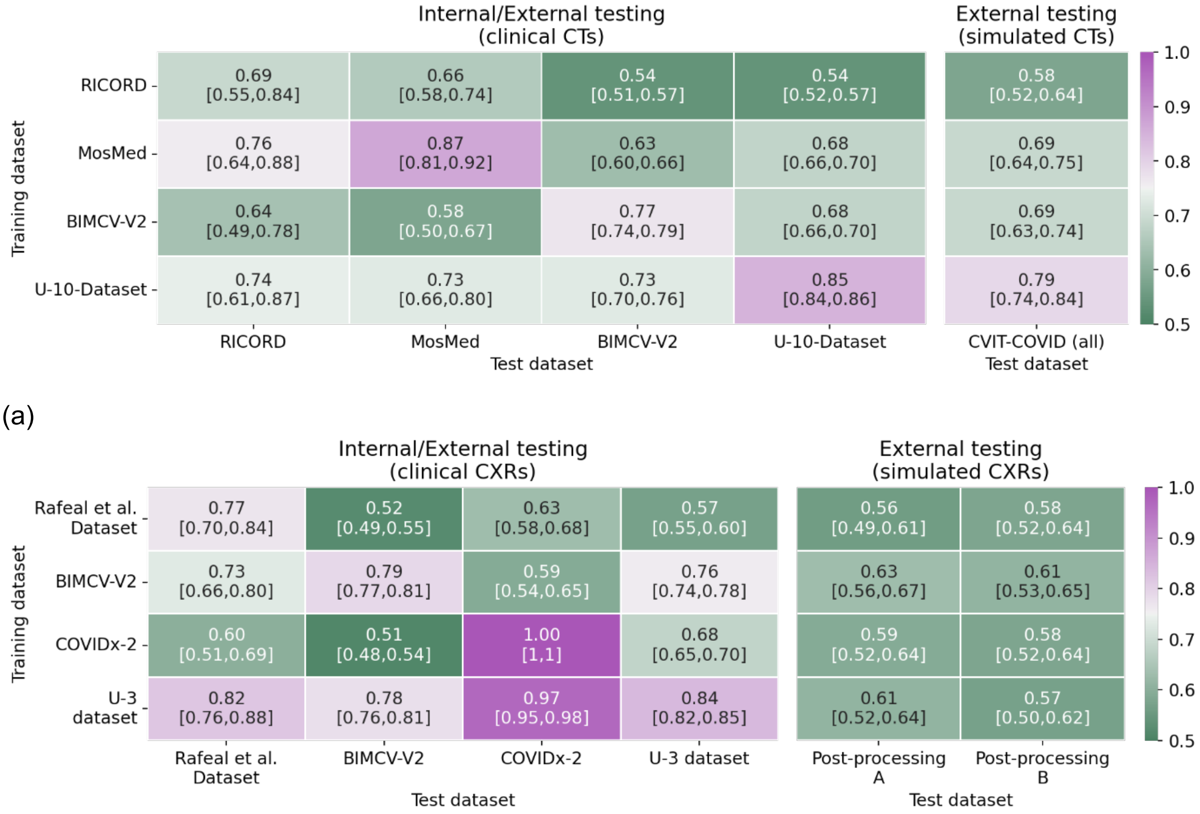
III. RESULTS

III.A. Evaluation of Models' Performance on Clinical Data

As depicted in Figure 6, clinical CT and CXR models exhibited a consistent drop in performance from internal to external testing, and those differences often exceeded the confidence intervals. While some loss of performance is expected in external testing, these remarkably consistent differences indicate systemic differences across these datasets. The CT models showed an internal validation AUC range of 0.69 to 0.85, whereas external testing consistently dropped to between 0.54 and 0.76. Similarly, for CXR models, internal performance ranged from an AUC of 0.77 to 1, while external testing AUC again dropped to a range of 0.51 to 0.73. Models trained on the most diverse datasets (U-3 CXR and U-10 CT) consistently yielded a testing performance that was the highest or second highest. Notably, despite its size, the COVIDx-CXR-2 dataset for CXR was very biased, resulting in perfect internal validation and near-perfect external testing even for the U-3 model that was trained on all three datasets.

III.B. Evaluation of Models' Performance on Virtual Data

As shown in Figure 6, all CT models achieved intermediate AUC values on the virtual data, consistent with their performance on the combined clinical training data. In other words, virtual data outperformed some of the actual clinical data, suggesting that virtual data are adequately realistic and often less biased. Among the CT models, training with the most diverse U-10 CT dataset yielded the highest testing performance on the virtual CT images, outperforming all three of the clinical datasets. This is remarkable since those three clinical datasets contributed to the U-10 CT training dataset, whereas the virtual data were completely independent. A similar pattern was observed with the CXR models, further supporting the robustness of the virtual datasets.



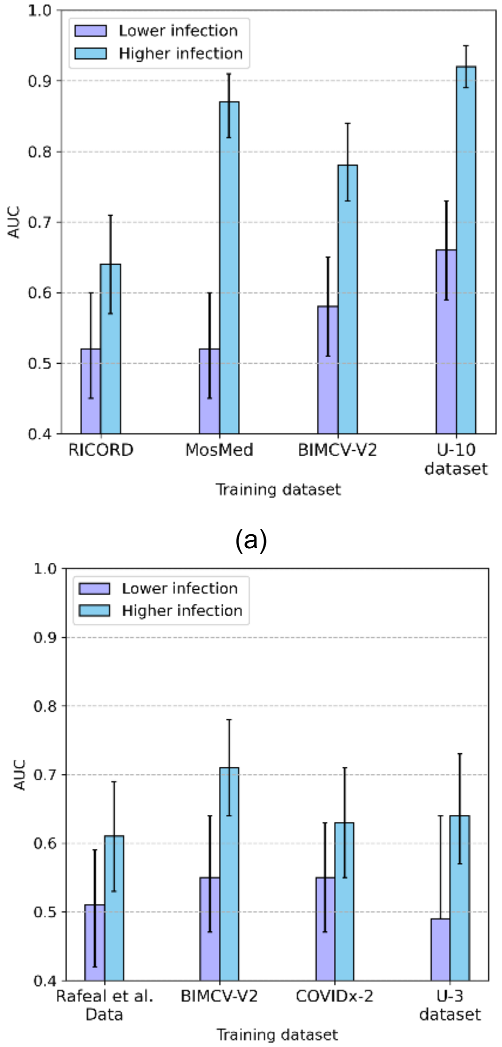
(a)

(b)

Figure 6. Confusion matrix of case-level COVID-19 detection performance of (a) CT and (b) CXR models. Training dataset is shown in rows and testing dataset in columns; diagonal represents internal validation, while off-diagonal entries are external testing. Additional external testing on simulated images is shown on the right. Performance is reported as receiver operating characteristic area under the curve with 95% confidence interval. All models generally performed worse on external testing with both clinical and simulated data. However, models trained with the union datasets (U-10 CT and U-3 CXR) consistently yielded the highest external testing performance. Internal validation AUC values ranged from 0.69 to 0.85 for CT models and 0.77 to 1.0 for CXR models, with external testing dropping to 0.54–0.76 and 0.51–0.73, respectively. Furthermore, simulation testing consistently provided intermediate results that may be more indicative of true performance.

III.C. Evaluation per Patient and Disease

Assessing the effect of infection size on the performance of models, Figure 7 shows all models performed better on both CT and CXR images with higher infection compared to images with lower infection. These results demonstrate the utility of VIT towards explainability.



(a)

(b)

Figure 7. Both (a) CT and (b) CXR models each trained on four datasets (represented on the x-axis), consistently demonstrated superior performance in "higher infection" cases, where the pneumonia volume exceeded the median, compared to "lower infection" cases that fell below the median. For CXR, results were almost identical for the two post-processing algorithms, so only algorithm A is shown. Error bars represent the 95% confidence interval.

III.D. Evaluations per Image Acquisition

For the same virtual patients, we assessed the performance of models over a wide, overlapping range of effective doses for the virtual CT and CXR acquisitions. As shown in Figure 8, the 3D CT models consistently outperformed the 2D CXR models, but the confidence intervals for the AUCs overlapped. Within each modality, although the effective dose (mSv) varied by 30-fold to represent the widest possible range of clinical use, there was no statistically significant change in performance [43, 51].

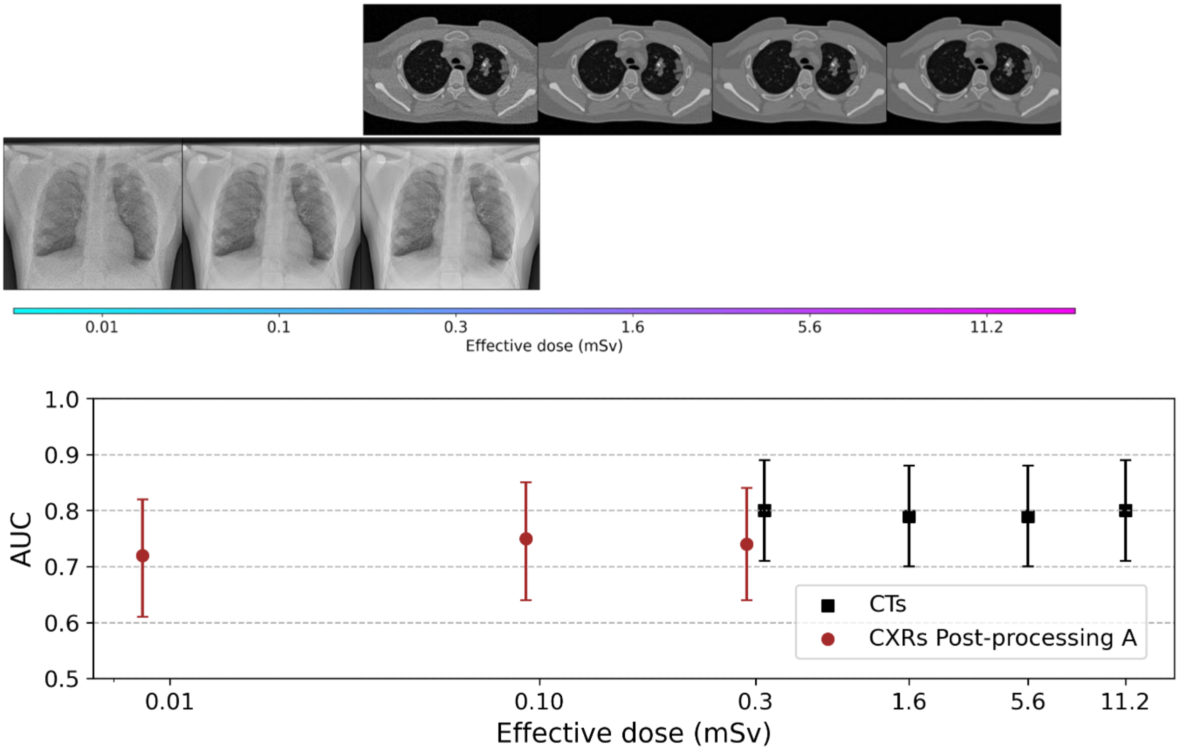


Figure 8. Simulated images were used to evaluate physics-based factors. Although models consistently performed better on CT over CXR, the differences were not significant at the shared dose of 0.3 mSv. Within each modality, performances were also not significantly different across a wide range of effective dose. Error bars correspond to 95% confidence interval.

IV. DISCUSSION

There has been considerable research to develop AI models to improve radiology diagnosis. However, the practical application of these models in clinical practice has been hindered by two related challenges. First, models often underperform when applied to a new dataset with different attributes. Those attributes include, but are not limited to, patient demographics, disease characteristics, scanner makes and models, and acquisition protocols. Second, most models function as “black boxes” that lack interpretability, making it difficult to determine which factors may contribute to an outcome or poor performance. These issues became particularly evident during the urgent scientific response to COVID-19, when many early studies reported high performances that did not generalize [21, 24, 26, 33, 34, 52]. Although biases in AI models for healthcare may be unavoidable, a comprehensive understanding of such factors, supported by effective external testing, can raise confidence that such models are trustworthy [19, 21, 53]. This study addresses the problem of biases in medical imaging AI models by leveraging clinical and virtual data for independent testing, thus enabling the evaluation of both generalizability and interpretability. While the study used COVID-19 diagnosis as its target task, given its rich resources of models and data, the lessons learned can apply to other AI-targeted tasks. By predicting the reliability of AI models in diverse clinical settings, this research facilitates the effective integration of AI into clinical practice.

We compiled a large cohort of clinical CT and CXR images from dataset resources representing over 22,000 patients. Despite the large amount of training data, however, model performance was still impaired due to class imbalance and confounding issues such as radiographic markers, incorrect image orientation, and collimator edges [21, 26, 54]. Proper data curation is time-consuming and requires domain expertise in medical imaging, rendering this process prohibitively costly.[55] Therefore, external validation of AI models is essential to rule out biases [19, 21]. Towards that goal, this study explored the use of virtual data [27, 31], which proved to offer two crucial advantages. First, the virtual image data provided external validation that is not only truly independent but also controlled. Second, the VIT framework allowed the evaluation of the models under different patient- and physics-based factors. This provided an opportunity for interpretability with respect to clinical or technical dependencies. Our study demonstrates the utility of VIT simulations to conduct AI imaging studies in a trustworthy, reproducible, and practicable manner.

One of our primary objectives was to analyze the impact of dataset variability on model generalization. To minimize overfitting, we intentionally used lightweight networks [44, 45]. Even so, all models still dropped in performance substantially from internal to external testing, consistent with other studies [21, 26, 33, 34]. This generalizability gap likely reflects inherent biases in the existing datasets with regard to institutional bias, patient demographics, disease appearances, and image quality [21, 24, 54]. To address such bias, we trained models on the combined U-10 CT and U-3 CXR datasets, which incorporated greater diversity. These models demonstrated improved external testing performance compared to the single-dataset models. The model trained on the diverse U-10 CT dataset demonstrated a very consistent AUC of approximately 0.73 across all three clinical datasets, suggesting that combining diverse data yields more credible and representative performance for this challenging clinical task. These general trends were also observed for the CXR datasets, but with considerable residual bias due to the disproportionate influence of the COVIDx-CXR-2 dataset, which is much larger than other datasets and leads to confounding bias as its positive and negative cases come from different institutions. This quandary shows that despite rigorous training and external testing, AI models can still be affected by fundamental data biases.

The VIT process proved to deliver a more realistic portrayal of true clinical performance. When many models were tested on virtual images, their performance fell consistently within the middle of the range of external testing on clinical datasets, suggesting that the simulations presented data with an appearance that was realistic and relevant. This is highly encouraging considering the models were applied to the virtual data without even being trained on them, highlighting the potential generalizability of virtual datasets to evaluate AI-based diagnosis algorithms. Unlike clinical datasets, the virtual images are further free of institutional bias or other confounding factors, because the VIT framework offers precisely reproducible controls in terms of patient sampling as well as physical image formation. This enabled us to compare identical virtual patients with and without the disease and also to conduct virtual imaging of each patient using both CT and CXR. The degree of experimental control provided by VITs is not physically possible in real clinical trials.

By integrating virtual datasets with clinical datasets, we aimed to enhance the generalizability and reliability of AI systems in medical imaging, ensuring their applicability in diverse clinical scenarios. The concept of a virtual dataset is integral to our study, providing a robust alternative to conventional datasets. These datasets offer precise control over imaging parameters, including patient anatomy, disease characteristics, and imaging conditions, ensuring consistency and reproducibility. Unlike conventional datasets, which often suffer from variability in patient demographics and imaging protocols, virtual datasets enable a controlled and repeatable generation of imaging data. As shown in Table 4, virtual datasets possess advanced features such as comprehensive patient-level, slice-level, and pixel-level annotations, and the ability to image the same virtual patient with both CT and CXR at multiple doses. These features facilitate rigorous evaluation and validation of AI models, allowing for systematic studies of the effects of various factors on model performance.

TABLE 4. Attributes of CT and CXR datasets. Note that virtual data are the only ones that contain all attributes, including the advanced features where the same virtual patient can be imaged with both CT and CXR at multiple doses, with multiple CXR post-processing. X= available.

	Class Type		Label Level			Advanced features
Datasets	COVID-19 positive	COVID-19 negative	Patient-level	Slice-level	Pixel-level	
CT datasets						
RICORD[18]	X	X	X			
MosMed[15]	X	X	X			
BIMCV-Iteration 2[14]	X	X	X			
COVID-CT-MD[12]	X	X	X	X		
An et al. dataset[13]	X		X			
COVID19-CT-dataset[36]	X		X			
COVID-19-NY-SBU[16]	X		X			
Lungs-CT-Dx[37]		X	X			
LIDC-IDRI[40]		X	X			
Duke-CVIT-CT	X	X	X	X	X	X
CXR datasets						
Fricks *et al*. dataset[23]	X	X	X	N/A		
BIMCV-2[14]	X	X	X	N/A		
COVIDx-CXR-2[38]	X	X	X	N/A		
Duke-CVIT-CXR	X	X	X	N/A	X	X

Our VIT analysis further provided intriguing insights into the effects of patient- and physics-based factors driving AI performance. Regardless of the training datasets for both the CT and CXR models, there was a noticeable increase in performance when the COVID-19 infection size was larger than the median value. For both imaging modalities, performances stayed consistent even across a 30-fold range in effective dose (which well exceeds the range in clinical practice), suggesting that dose may not be as relevant for the AI detection of diffuse diseases such as pneumonia. In stark contrast to the model evaluation on clinical data, our analysis confirmed that CT outperformed CXR, which was consistent with expectations since 3D CT scans provide superior spatial information over 2D CXR images.

This study had several limitations. Although the virtual CT and CXR images realistically reproduced both anatomical and physical processes, they were generated from a pool of fifty virtual patients with variable anatomy and severity of the disease. Virtual datasets offer control and reproducibility but must be complemented with real-world validation to ensure ethical and clinically applicable AI models. Consequently, simulation testing showed consistent trends but with large confidence intervals. The minimal impact of imaging dose observed in our study might be influenced by down-sampling during the preprocessing. Additionally, the study did not account for potential variability in scanner-specific imaging characteristics, which could affect model performance in real-world settings. Future work will increase the number of computational phantoms to represent even larger and more diverse patient populations and explore the inclusion of additional imaging parameters to improve realism. In terms of the network architectures, each modality was analyzed using a single lightweight design, foregoing extension experiments with other networks. Expanding the model evaluation to include more complex architectures could provide insights into generalizability across different network types. Models were developed only to conduct case-level detection, which is the only annotation available in almost all datasets. Furthermore, the label of COVID-19 as negative or positive was defined independently per each dataset, and those standards varied widely, including radiologist assessment or different diagnostic tests [2]. Some datasets included both COVID-19 pneumonia and other types of pneumonia, which may not be readily differentiated by imaging alone. Finally, future work should aim to address these limitations by incorporating more detailed multi-class annotations and evaluating model performance under different disease classification scenarios.

V. CONCLUSIONS

AI-based diagnosis models hold the potential to revolutionize healthcare. However, factors contributing to model bias remain underexplored, especially in the medical imaging domain. An essential prerequisite to clinical deployment is a robust external evaluation. The VIT framework plays a crucial role in addressing the ongoing reproducibility crisis in AI models by providing the necessary image data that is objective and controlled. By enabling consistent evaluation across diverse scenarios, VIT not only helps to identify bias but also facilitates improvements in model robustness and generalizability. As emerging AI techniques continue to evolve [56-58], the need for rigorous evaluation frameworks like VIT becomes even more critical to ensure their reliability and clinical relevance. By studying patient- or physics-based factors influencing model performance, the VIT methodology offers potential for interpretability and opportunities for model refinement. Through these contributions, virtual imaging trials can enhance clinical trials, making them faster, more rigorous, and more reproducible.

DISCLOSURES

The authors have no conflicts of interest to declare that are relevant to the content of this article.

APPENDIX A – DATA AVAILABILITY

The clinical data utilized in this study are open-source and can be referenced via the citation in Table 1 and Table 2. The authors are committed to promoting transparency and open science. Reasonable requests for access to an anonymized version of the private datasets (Duke-CVIT-CT and Duke-CVIT-CXR) can be made by contacting the corresponding author. Upon publication, all model weights, initial hyperparameters, and code will be publicly accessible at https://gitlab.oit.duke.edu/cvit-public/cvit_revicovid19

ACKNOWLEDGMENT

This work was funded in part by the Center for Virtual Imaging Trials, NIH/NIBIB P41EB028744 and NIH/NCI R01 CA261457.

REFERENCES

[1] F. Nensa, D. Pinto dos Santos, and M. Dietzel, "Beyond accuracy: Reproducibility must lead AI advances in radiology," *European Journal of Radiology,* vol. 180, 2024, doi: 10.1016/j.ejrad.2024.111703.

[2] G. D. Rubin *et al.*, "The Role of Chest Imaging in Patient Management during the COVID-19 Pandemic: A Multinational Consensus Statement from the Fleischner Society," *Radiology,* vol. 296, no. 1, pp. 172-180, Jul 2020, doi: /10.1148/radiol.2020201365.

[3] J. P. Kanne *et al.*, "COVID-19 Imaging: What We Know Now and What Remains Unknown," *Radiology,* vol. 299, no. 3, pp. E262-E279, Jun 2021, doi: /10.1148/radiol.2021204522.

[4] H. Gunraj, L. Wang, and A. Wong, "COVIDNet-CT: A Tailored Deep Convolutional Neural Network Design for Detection of COVID-19 Cases From Chest CT Images," *Front Med (Lausanne),* vol. 7, p. 608525, 2020, doi: /10.3389/fmed.2020.608525.

[5] S. A. Harmon *et al.*, "Artificial intelligence for the detection of COVID-19 pneumonia on chest CT using multinational datasets," *Nat Commun,* vol. 11, no. 1, p. 4080, Aug 14 2020, doi: /10.1038/s41467-020-17971-2.

[6] T. Javaheri *et al.*, "CovidCTNet: an open-source deep learning approach to diagnose covid-19 using small cohort of CT images," *NPJ Digit Med,* vol. 4, no. 1, p. 29, Feb 18 2021, doi: /10.1038/s41746-021-00399-3.

[7] C. Jin *et al.*, "Development and evaluation of an artificial intelligence system for COVID-19 diagnosis," *Nat Commun,* vol. 11, no. 1, p. 5088, Oct 9 2020, doi: /10.1038/s41467-020-18685-1.

[8] H. X. Bai *et al.*, "Artificial Intelligence Augmentation of Radiologist Performance in Distinguishing COVID-19 from Pneumonia of Other Origin at Chest CT," *Radiology,* vol. 296, no. 3, pp. E156-E165, Sep 2020, doi: /10.1148/radiol.2020201491.

[9] P. Afshar, S. Heidarian, F. Naderkhani, A. Oikonomou, K. N. Plataniotis, and A. Mohammadi, "COVID-CAPS: A capsule network-based framework for identification of COVID-19 cases from X-ray images," *Pattern Recognition Letters,* vol. 138, pp. 638-643, 2020, doi: /10.1016/j.patrec.2020.09.010.

[10] H. Panwar, P. K. Gupta, M. K. Siddiqui, R. Morales-Menendez, and V. Singh, "Application of deep learning for fast detection of COVID-19 in X-Rays using nCOVnet," *Chaos, Solitons & Fractals,* vol. 138, p. 109944, 2020, doi: /10.1016/j.chaos.2020.109944.

[11] D. Kollias, A. Arsenos, and S. Kollias, "Domain adaptation, explainability & fairness in ai for medical image analysis: Diagnosis of covid-19 based on 3-d chest ct-scans," *arXiv preprint arXiv:2403.02192,* 2024.

[12] P. Afshar *et al.*, "COVID-CT-MD, COVID-19 computed tomography scan dataset applicable in machine learning and deep learning," *Sci Data,* vol. 8, no. 1, p. 121, Apr 29 2021, doi: /10.1038/s41597-021-00900-3.

[13] P. An *et al.*, "CT Images in COVID-19 [Data set]," *The Cancer Imaging Archive,* 2020, doi: /10.7937/TCIA.2020.GQRY-NC81.

[14] M. d. l. I. Vayá *et al.*, "Bimcv covid-19+: a large annotated dataset of rx and ct images from covid-19 patients," *arXiv preprint arXiv:2006.01174,* 2020.

[15] S. P. Morozov *et al.*, "MosMedData: data set of 1110 chest CT scans performed during the COVID-19 epidemic," *Digital Diagnostics,* vol. 1, no. 1, pp. 49-59, 2020-05-13 2020.

[16] J. Saltz *et al.*, "Stony Brook University COVID-19 Positive Cases [Data set]," *The Cancer Imaging Archive.,* 2021, doi: /10.7937/TCIA.BBAG-2923.

[17] S. Shakouri *et al.*, "COVID19-CT-dataset: an open-access chest CT image repository of 1000+ patients with confirmed COVID-19 diagnosis," *BMC Research Notes,* vol. 14, no. 1, 2021, doi: /10.1186/s13104-021-05592-x.

[18] E. B. Tsai *et al.*, "The RSNA International COVID-19 Open Radiology Database (RICORD)," *Radiology,* vol. 299, no. 1, pp. E204-E213, Apr 2021, doi: /10.1148/radiol.2021203957.

[19] 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," *Nature Machine Intelligence,* vol. 3, no. 3, pp. 199-217, 2021, doi: /10.1038/s42256-021-00307-0.

[20] "Medical Imaging and Data Resource Center (MIDRC). https://www.midrc.org/," no. 6 July, 2023, 2023. [Online]. Available: https://www.midrc.org/.

[21] A. J. DeGrave, J. D. Janizek, and S.-I. Lee, "AI for radiographic COVID-19 detection selects shortcuts over signal," *Nature Machine Intelligence,* vol. 3, no. 7, pp. 610-619, 2021, doi: 10.1038/s42256-021-00338-7.

[22] D. Driggs *et al.*, "Machine Learning for COVID-19 Diagnosis and Prognostication: Lessons for Amplifying the Signal While Reducing the Noise," *Radiol Artif Intell,* vol. 3, no. 4, p. e210011, Jul 2021, doi: 10.1148/ryai.2021210011.

[23] R. B. Fricks *et al.*, "Deep learning classification of COVID-19 in chest radiographs: performance and influence of supplemental training," *J Med Imaging (Bellingham),* vol. 8, no. 6, p. 064501, Nov 2021, doi: 10.1117/1.JMI.8.6.064501.

[24] J. Sun *et al.*, "Performance of a Chest Radiograph AI Diagnostic Tool for COVID-19: A Prospective Observational Study," *Radiol Artif Intell,* vol. 4, no. 4, p. e210217, Jul 2022, doi: /10.1148/ryai.210217.

[25] D. Khemasuwan and H. G. Colt, "Applications and challenges of AI-based algorithms in the COVID-19 pandemic," *BMJ Innovations,* vol. 7, no. 2, pp. 387-398, 2021, doi: /10.1136/bmjinnov-2020-000648.

[26] D. Nguyen *et al.*, "Deep Learning-Based COVID-19 Pneumonia Classification Using Chest CT Images: Model Generalizability," (in English), *Front Artif Intell,* Original Research vol. 4, no. 87, p. 694875, 2021-June-29 2021, doi: /10.3389/frai.2021.694875.

[27] E. Abadi *et al.*, "Virtual clinical trials in medical imaging: a review," *Journal of Medical Imaging,* vol. 7, no. 4, p. 042805, 2020. [Online]. Available: https://doi.org/10.1117/1.JMI.7.4.042805.

[28] A. Badano *et al.*, "Evaluation of Digital Breast Tomosynthesis as Replacement of Full-Field Digital Mammography Using an In Silico Imaging Trial," *JAMA Network Open,* vol. 1, no. 7, p. e185474, 2018, doi: 10.1001/jamanetworkopen.2018.5474.

[29] F. I. Tushar *et al.*, "Virtual NLST: towards replicating national lung screening trial," in *Medical Imaging 2024: Physics of Medical Imaging*, 2024, vol. 12925: SPIE, pp. 442-447.

[30] S. Sotoudeh-Paima *et al.*, "Development and application of a virtual imaging trial framework for longitudinal quantification of emphysema in CT," in *Proceedings of SPIE--the International Society for Optical Engineering*, 2024, vol. 12925, p. 129251H.

[31] E. Abadi, W. Paul Segars, H. Chalian, and E. Samei, "Virtual Imaging Trials for Coronavirus Disease (COVID-19)," *AJR Am J Roentgenol,* vol. 216, no. 2, pp. 362-368, Feb 2021, doi: /10.2214/AJR.20.23429.

[32] E. Abadi, B. Harrawood, S. Sharma, A. Kapadia, W. P. Segars, and E. Samei, "DukeSim: A Realistic, Rapid, and Scanner-Specific Simulation Framework in Computed Tomography," *IEEE Transactions on Medical Imaging,* vol. 38, no. 6, pp. 1457-1465, 2019, doi: /10.1109/tmi.2018.2886530.

[33] F. I. Tushar *et al.*, *Virtual vs. reality: external validation of COVID-19 classifiers using XCAT phantoms for chest computed tomography* (SPIE Medical Imaging). SPIE, 2022.

[34] L. Dahal *et al.*, *Virtual versus reality: external validation of COVID-19 classifiers using XCAT phantoms for chest radiography* (SPIE Medical Imaging). SPIE, 2022.

[35] N. Arun *et al.*, "Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging," *Radiology: Artificial Intelligence,* vol. 3, no. 6, p. e200267, 2021.

[36] S. Shakouri *et al.*, "COVID19-CT-dataset: an open-access chest CT image repository of 1000+ patients with confirmed COVID-19 diagnosis," *BMC Res Notes,* vol. 14, no. 1, p. 178, May 12 2021, doi: /10.1186/s13104-021-05592-x.

[37] P. Li, S. Wang, T. Li, J. Lu, Y. HuangFu, and D. Wang, "A Large-Scale CT and PET/CT Dataset for Lung Cancer Diagnosis," 2020, doi: /10.7937/TCIA.2020.NNC2-0461.

[38] L. Wang, Z. Q. Lin, and A. Wong, "COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images," *Sci Rep,* vol. 10, no. 1, p. 19549, Nov 11 2020, doi: /10.1038/s41598-020-76550-z.

[39] K. J. Kiser *et al.*, "Data from the Thoracic Volume and Pleural Effusion Segmentations in Diseased Lungs for Benchmarking Chest CT Processing Pipelines PleThora) [Data set]," *The Cancer Imaging Archive,* 2020, doi: /10.7937/tcia.2020.6c7y-gq39.

[40] S. G. Armato *et al.*, "The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A completed reference database of lung nodules on CT scans," *Medical Physics,* vol. 38, no. 2, pp. 915-931, 2011 2011, doi: 10.1118/1.3528204.

[41] "Radiological Society of North America. RSNA pneumonia detection challenge. https://www.kaggle.com/c/rsna-pneumonia-detection-challenge/data," 2019. [Online]. Available: https://www.kaggle.com/c/rsna-pneumonia-detection-challenge/data.

[42] W. P. Segars *et al.*, "Population of anatomically variable 4D XCAT adult phantoms for imaging research and optimization," (in eng), *Med Phys,* vol. 40, no. 4, p. 043701, Apr 2013, doi: 10.1118/1.4794178.

[43] Y. Zhang, X. Li, W. P. Segars, and E. Samei, "Comparison of patient specific dose metrics between chest radiography, tomosynthesis, and CT for adult patients of wide ranging body habitus," *Medical Physics,* vol. 41, no. 2, p. 023901, 2014.

[44] F. I. Tushar *et al.*, "Classification of Multiple Diseases on Body CT Scans Using Weakly Supervised Deep Learning," *Radiology: Artificial Intelligence,* vol. 4, no. 1, p. e210026, 2022, doi: /10.1148/ryai.210026.

[45] F. I. Tushar, V. D'Anniballe, G. Rubin, E. Samei, and J. Lo, *Co-occurring diseases heavily influence the performance of weakly supervised learning models for classification of chest CT* (SPIE Medical Imaging). SPIE, 2022.

[46] M. Tan and Q. Le, "Efficientnetv2: Smaller models and faster training," in *International Conference on Machine Learning*, 2021: PMLR, pp. 10096-10106.

[47] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.

[48] T. He, Z. Zhang, H. Zhang, Z. Zhang, J. Xie, and M. Li, "Bag of tricks for image classification with convolutional neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 558-567.

[49] F. I. Tushar *et al.* "CVIT COVID-19 AI Model Code." Duke University. https://gitlab.oit.duke.edu/cvit-public/cvit_revicovid19 (accessed.

[50] X. Robin *et al.*, "pROC: an open-source package for R and S+ to analyze and compare ROC curves," *BMC Bioinformatics,* vol. 12, no. 1, p. 77, 2011/03/17 2011, doi: /10.1186/1471-2105-12-77.

[51] M. Fujita *et al.*, "Lung cancer screening with ultra-low dose CT using full iterative reconstruction," *Japanese Journal of Radiology,* vol. 35, no. 4, pp. 179-189, 2017/04/01 2017, doi: /10.1007/s11604-017-0618-y.

[52] J. Dhont, C. Wolfs, and F. Verhaegen, "Automatic coronavirus disease 2019 diagnosis based on chest radiography and deep learning - Success story or dataset bias?," *Med Phys,* vol. 49, no. 2, pp. 978-987, Feb 2022, doi: /10.1002/mp.15419.

[53] L. Seyyed-Kalantari, G. Liu, M. McDermott, I. Y. Chen, and M. Ghassemi, "CheXclusion: Fairness gaps in deep chest X-ray classifiers," in *BIOCOMPUTING 2021: Proceedings of the Pacific Symposium*, 2020: World Scientific, pp. 232-243.

[54] L. Seyyed-Kalantari, H. Zhang, M. B. A. McDermott, I. Y. Chen, and M. Ghassemi, "Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations," *Nat Med,* vol. 27, no. 12, pp. 2176-2182, Dec 2021, doi: /10.1038/s41591-021-01595-0.

[55] M. J. Willemink *et al.*, "Preparing Medical Imaging Data for Machine Learning," *Radiology,* vol. 295, no. 1, pp. 4-15, Apr 2020, doi: /10.1148/radiol.2020192224.

[56] P. Guo *et al.*, "Maisi: Medical ai for synthetic imaging," *arXiv preprint arXiv:2409.11169,* 2024.

[57] Y. He *et al.*, "VISTA3D: Versatile Imaging SegmenTation and Annotation model for 3D Computed Tomography," *arXiv preprint arXiv:2406.05285,* 2024.

[58] N. C. Codella *et al.*, "Medimageinsight: An open-source embedding model for general domain medical imaging," *arXiv preprint arXiv:2410.06542,* 2024.