Medical Image Analysis 103 (2025) 103576

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

# Virtual lung screening trial (VLST): An *in silico* study inspired by the national lung screening trial for lung cancer detection

Fakrul Islam Tushar [a,b,*], Liesbeth Vancoillie [a], Cindy McCabe [a,c], Amareswararao Kavuri [a], Lavsen Dahal [a,b], Brian Harrawood [a], Milo Fryling [a], Mojtaba Zarei [a,b], Saman Sotoudeh-Paima [a,b], Fong Chi Ho [a,b], Dhrubajyoti Ghosh [d], Michael R. Harowicz [a], Tina D. Tailor [a], Sheng Luo [d], W. Paul Segars [a,b,c], Ehsan Abadi [a,b,c], Kyle J. Lafata [a,b,c,e], Joseph Y. Lo [a,b,c,1], Ehsan Samei [a,b,c,1]

[a] *Center for Virtual Imaging Trials, Carl E. Ravin Advanced Imaging Laboratories, Dept. of Radiology, Duke University School of Medicine, USA*
[b] *Dept. of Electrical & Computer Engineering, Pratt School of Engineering, Duke University, USA*
[c] *Medical Physics Graduate Program, Duke University, USA*
[d] *Dept. of Biostatistics and Bioinformatics, Duke University School of Medicine, USA*
[e] *Dept. of Radiation Oncology, Duke University School of Medicine, USA*

ARTICLE INFO

*Keywords:*
Virtual lung screening trial (VLST)
National lung screening trials (NLST)
Computer-aided diagnosis
Computed tomography (CT)
Chest radiography (CXR)
Human digital twins
Neural networks

ABSTRACT

Clinical imaging trials play a crucial role in advancing medical innovation but are often costly, inefficient, and ethically constrained. Virtual Imaging Trials (VITs) present a solution by simulating clinical trial components in a controlled, risk-free environment. The Virtual Lung Screening Trial (VLST), an *in silico* study inspired by the National Lung Screening Trial (NLST), illustrates the potential of VITs to expedite clinical trials, minimize risks to participants, and promote optimal use of imaging technologies in healthcare. This study aimed to show that a virtual imaging trial platform could investigate some key elements of a major clinical trial, specifically the NLST, which compared Computed tomography (CT) and chest radiography (CXR) for lung cancer screening. With simulated cancerous lung nodules, a virtual patient cohort of 294 subjects was created using XCAT human models. Each virtual patient underwent both CT and CXR imaging, with deep learning models, the AI CT-Reader and AI CXR-Reader, acting as virtual readers to perform recall patients with suspicion of lung cancer. The primary outcome was the difference in diagnostic performance between CT and CXR, measured by the Area Under the Curve (AUC). The AI CT-Reader showed superior diagnostic accuracy, achieving an AUC of 0.92 (95 % CI: 0.90–0.95) compared to the AI CXR-Reader's AUC of 0.72 (95 % CI: 0.67–0.77). Furthermore, at the same 94 % CT sensitivity reported by the NLST, the VLST specificity of 73 % was similar to the NLST specificity of 73.4 %. This CT performance highlights the potential of VITs to replicate certain aspects of clinical trials effectively, paving the way toward a safe and efficient method for advancing imaging-based diagnostics.

## 1. Introduction

Lung cancer ranks as the leading cause of cancer-related deaths, accounting for approximately 1.8 million fatalities in 2020 (Ferlay et al., 2021; Sung et al., 2021). Projections from the American Cancer Society indicate that an estimated 238 K individuals in the United States are anticipated to be diagnosed with lung cancer in 2023 (Siegel et al., 2023). In the realm of timely detection and diagnosis of lung cancer, imaging modalities like chest radiography (CXR) and Computed Tomography (CT) scans play a crucial role for not only the diagnosis of lung cancer but also those of a wide range of abnormalities (Koning et al., 2011).

Associating the early-state cancer detection to the larger likelihood of cure, multiple lung cancer screening trials worldwide have contributed valuable insights into the efficacy of lung cancer screening, such as National Lung Screening Trial (NLST) (Koning et al., 2011), Nederlands-Leuvens Longkanker Screenings Onderzoek (NELSON) (De Koning et al., 2020), Multicentric Italian Lung Detection (MILD)

* Corresponding author.
*E-mail address:* fakrulislam.tushar@duke.edu (F.I. Tushar).
[1] These authors contributed equally as co-senior authors.

https://doi.org/10.1016/j.media.2025.103576
Received 5 November 2024; Received in revised form 8 March 2025; Accepted 28 March 2025
Available online 5 April 2025
1361-8415/© 2025 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

(Pastorino et al., 2019), British Thoracic Society Lung Cancer Screening Group (BTSLSG) (Field et al., 2016) and International Early Lung Cancer Action Program (IE-LCAP) trial (Welch et al., 2007).

In spite of their benefits, clinical trials are inefficient and costly. Over its duration of 8 years, for example, NLST enrolled over 50,000 subjects at a cost of $256 million (Koning et al., 2011; National Cancer I 2024). At the conclusion of such trials, the diagnosis or treatment being evaluated may already be obsolete. For medical imaging, that risk is particularly high because technology evolves so rapidly. Additionally, clinical trials may involve putting patients at risk, including that of radiation exposure or overdiagnosis leading to unnecessary procedures. Therefore, it is imperative to develop alternative trial procedures that can address these gaps to ensure financial feasibility, timely change of clinical practice, and patient safety (Heleno et al., 2013).

Virtual Imaging Trials (VITs) leverage advances in computational techniques to simulate the workflow of clinical imaging trials, i.e., patients undergoing scans that are interpreted by readers. Compared to clinical trials, VITs may serve as an alternative that is faster, safer, and more cost-effective. In their comprehensive review paper, Abadi et al. delve extensively into VITs using numerous imaging modalities to study myriad applications, such as for coronavirus disease (COVID-19), emphysema, and organ dosimetry (Abadi et al., 2020). There are still very few VIT studies that emulate an end-to-end trial, however, due to the immense complexity of simulating all trial aspects. The most mature VIT studies to date have focused on breast cancer screening, such as the trailblazing study by FDA scientists to emulate a comparative clinical trial of mammography versus breast tomosynthesis (Badano et al., 2018). To advance VIT applications, simulation software modules need to be integrated to facilitate research that is both expedient and rigorous, and the trials need to expand to other clinical settings with high clinical

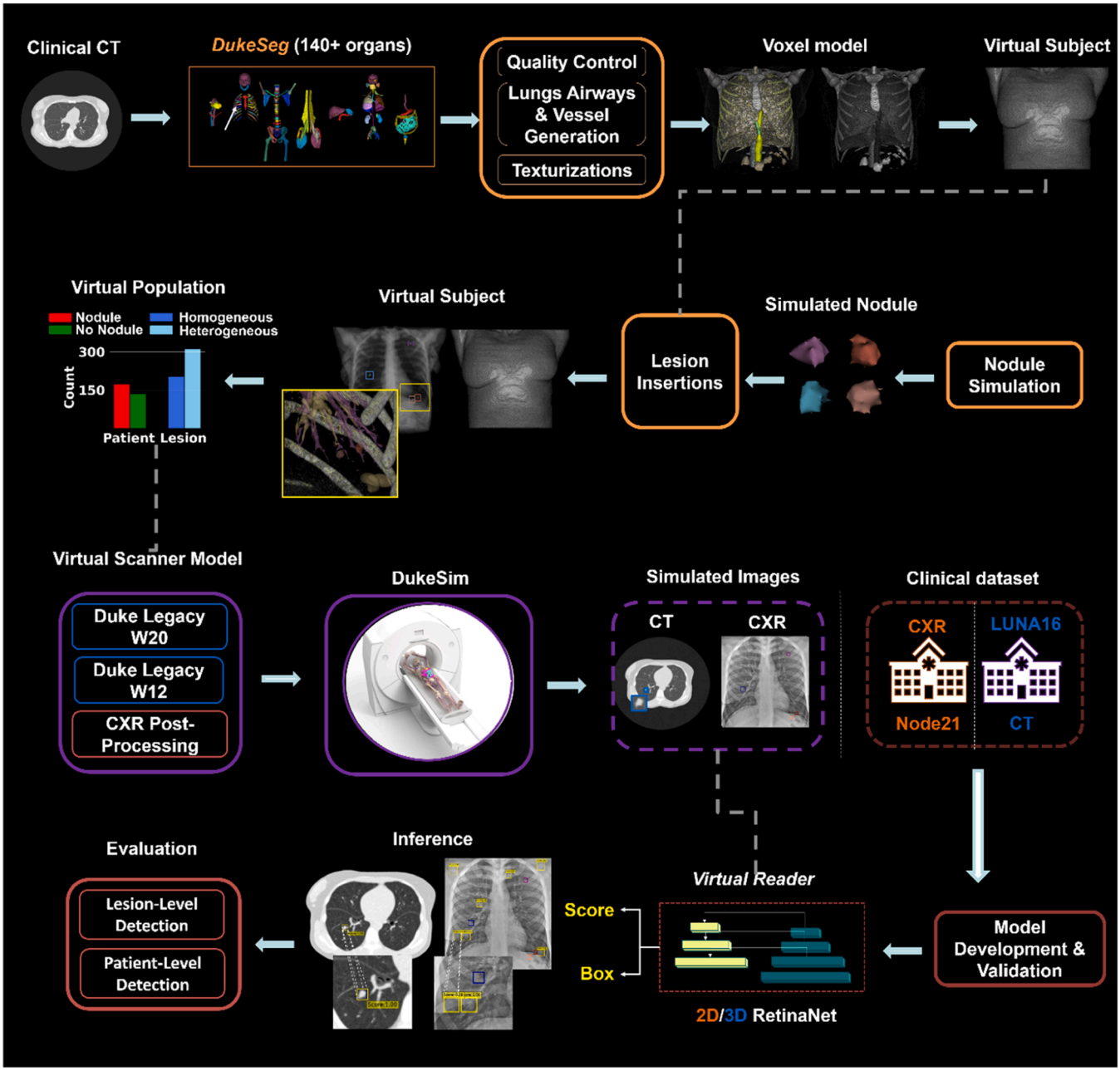



**Fig. 1.** Flowchart summarizing the comprehensive workflow of the Virtual Lung Screening Trial (VLST), from creation of virtual patient models using clinical CT data, through lesion simulation and insertion, to virtual imaging using DukeSim, and concluding with evaluation by a virtual reader employing 2D/3D RetinaNet for both lesion-level and patient-level detection evaluation.

impact.

To address these gaps, this Virtual Lung Screening Trial (VLST) study aimed to take the first steps toward an "*in silico*" emulation to compare efficacy of CXR and CT in lung cancer screening. The NLST was extremely important in medical imaging as it changed clinical practice and led to the adoption of current lung cancer screening programs. Inspired by that trial, our VLST study presents three primary advances. First, a virtual patient cohort was designed to encompass the many anatomical structures of the chest and multiple types of lung nodules. Second, this study implemented virtual scanners for CXR and CT, modalities that have attracted considerable recent attention in machine learning research due to their high clinical volume and impact (Hamamci et al., 2024; Chambon et al., 2024). Finally, image interpretation was performed by virtual readers implemented using reproducible deep learning models. By integrating these software modules, this study presented the first end-to-end virtual imaging trial in the new domain of chest imaging, designed to evaluate an ***in silico*** platform for a task with high clinical relevance.

## 2. Methods

### 2.1. Study Design

Fig. 1 shows an overview of the steps performed to accomplish the VLST consisted of simulating a virtual population, modeling virtual scanners, developing virtual readers, and finally analyzing the detection and diagnosis performance. In the section below, we will briefly explain each of these steps.

### 2.2. Virtual Population

In any clinical trial, the selection of the targeted population and sample groups is of utmost importance, as it directly influences the validity of trial results and outcomes. The creation of a virtual human subject involved two key steps: constructing the normal anatomy followed by insertion of the specific abnormalities. For the VLST, the subjects employed were computational human models generated from full body CT scans from a single health system (Duke University Health System) which encompasses multiple hospitals. The construction of these human models involved a series of steps: beginning with the segmentation of specific organs, ensuring the quality control of these segmentations, followed by the creation of airways and vessels, and culminating in the voxelization process. An outline of this procedure is shown in Fig. 1, with further information available in our previous publication (Lavsen et al., 2023). In this research, we employed both pre-existing and newly developed extended cardiac-torso (XCAT) models representing both sexes with varying age, BMI, and race (Lavsen et al., 2023; Segars et al., 2013; Abadi et al., 2025). Due to computational constraints to generate the complex 3D dataset, this initial study was limited to only 294 virtual subjects. Due to the small sample size, which is discussed in the Limitations section of the discussion, future work will leverage enhanced resources to expand the cohort. Demographically, the mean age was 59 years, with a distribution between 145 men, 116 women, 33 unknown from older phantom participants.

Simulated nodules were generated in two steps. First, a single-density lesion was formed, adhering to the desired morphology. Subsequently, a convolution process created a multi-density structure using a previously published approach (Sauer et al., 2023; McCabe et al., 2024). Several instances of these simulated nodules, featuring diverse characteristics such as size and morphology, are depicted in **Fig. A.1 and Fig. A.2**. Consistent with methods used in previous studies, simulated lesions were designed to represent cancerous nodules (Sauer et al., 2023; McCabe et al., 2024). These simulated nodules were then randomly inserted into the lungs of 174 patients, guided by indications of where nodules have been observed in prior clinical trial studies (Koning et al., 2011). A total of 512 solid lesions were created including 202 (39.4 %) with homogeneous texture and 310 (60.5 %) heterogeneous texture. The remaining 120 patients were designed without any nodules. The ratio of individuals with and without nodules reflects the combination of pre-existing and newly created virtual subjects while ensuring a greatly enriched prevalence of >50 % cancer to facilitate analysis of reader performance. Specifics regarding population demographics, along with inclusion and exclusion criteria, are clarified in Table 1 and Fig. 2.

### 2.3. Virtual Scanners and imaging protocols

The cohort of computational human models with and without nodules was virtually imaged using a validated imaging simulation tool DukeSim (Jadick et al., 2021; Abadi et al., 2019; Abadi et al., Feb 2021). DukeSim generates projection images from voxelized computational models using ray tracing (for primary signal) and Monte Carlo simulation (for scatter signal and radiation dose). Both modalities were designed to align with the acquisition technologies used during the NLST era, ensuring consistency in approach. To create CXR images, raw projection images were post-processed to replicate the contrast, noise, and resolution of the NLST-era images. For CT, projections were reconstructed using the vendor-neutral reconstruction tool MCR Toolkit (Clark & Badea, 2023) with weighted filter back projection configured to mimic the physical and geometrical characteristics of two generic CT scanners, named Duke Legacy W12 and Duke Legacy W20, which are representative of NLST CT systems (Cagnon et al., 2006). The geometry and acquisition configurations of the scanners are listed in **Appendix Table B.1**. Samples of human model and simulated images with lesions are presented in Fig. 3.

### 2.4. Virtual Reader

The virtual reader component of our study deployed deep learning models to emulate the image interpretation by radiologists. Two RetinaNet models were developed: a 2D model for CXR images and a 3D model for CT volumes (Lin et al., 2017). This readily available machine learning architecture was trained with publicly accessible clinical datasets (LUNA16 (Setio et al., 2017), NODE21 (Sogancioglu et al.,

**Table 1**
Cohort Characteristics of the VLST Virtual Population of With Cases Corresponding to with and without simulated nodule.

| Characteristics | All | With Lung Nodule ($n$ = 174) | Without Lung Nodule ($n$ = 87) |
|---|---|---|---|
| **Age (year)** | | | |
| Mean ± Std | 59 ± 14 | 59 ± 15 | 60 ± 13 |
| **Sex** | | | |
| Male | 145 (55.56 %) | 95 (54.60 %) | 50 (57.47 %) |
| Female | 116 (44.44 %) | 79 (45.40 %) | 37 (42.22 %) |
| **Weight (kg)** | | | |
| Mean ± Std | 78± 20 | 77 ± 20 | 80 ± 20 |
| **BMI** | | | |
| Mean ± Std (min-max) | 27± 6 (13–50) | 26± 6 (13–50) | 28 ± 5 (18–41) |
| **Race** | | | |
| White | 196 (75.10 %) | 128 (73.56 %) | 68 (78.16 %) |
| Black or African American | 55 (21.07 %) | 36 (20.69 %) | 19 (21.84 %) |
| Other/Unknown | 10 (3.83 %) | 10 (5.75 %) | |
| **Ethnic** | | | |
| Not Hispanic or Latino | 257 (98.47 %) | 171 (98.28 %) | 86 (98.85 %) |
| Hispanic/ Unknown | 4 (1.53 %) | 3 (1.73 %) | 1 (1.15 %) |

*Note: Characteristics for the existing 33 without lung nodule human models are currently unavailable.

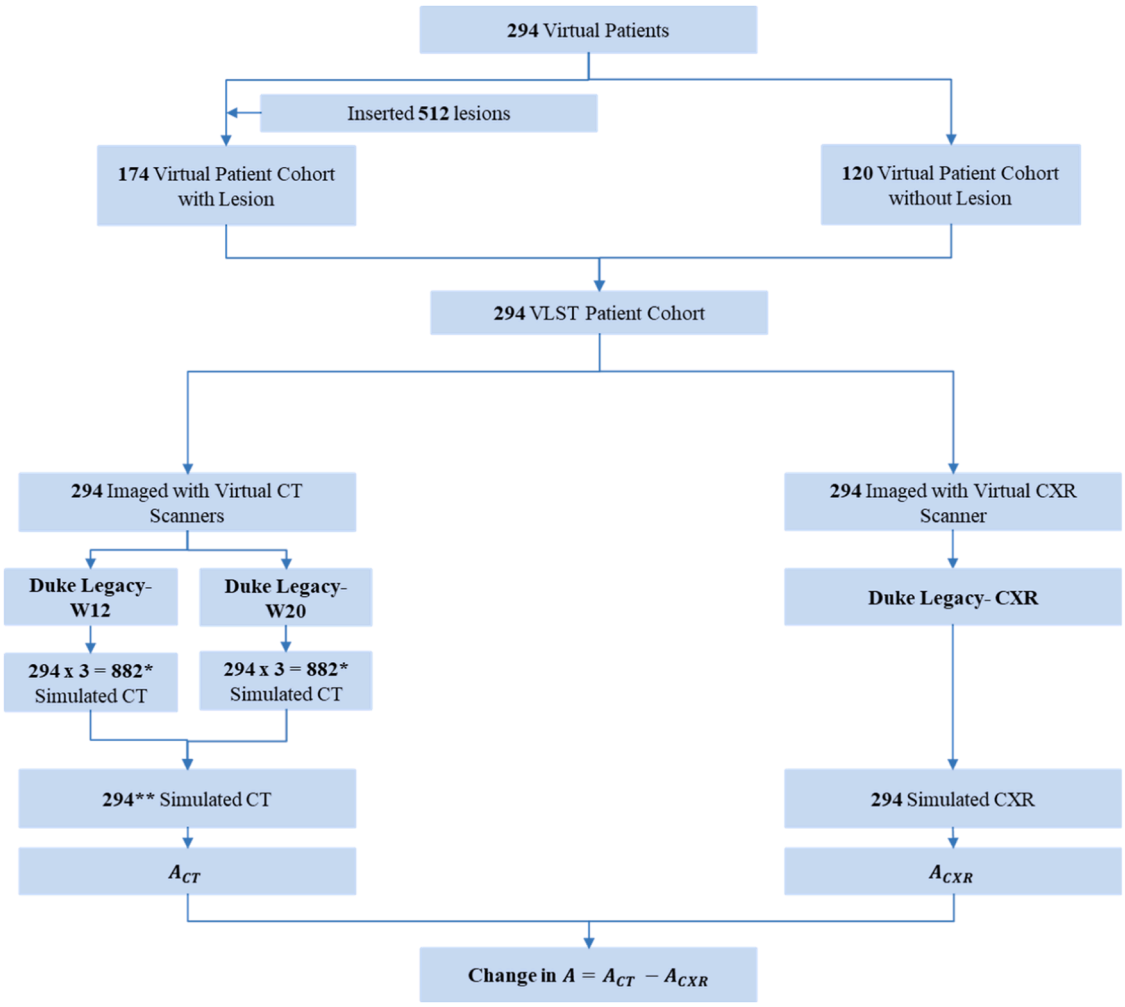
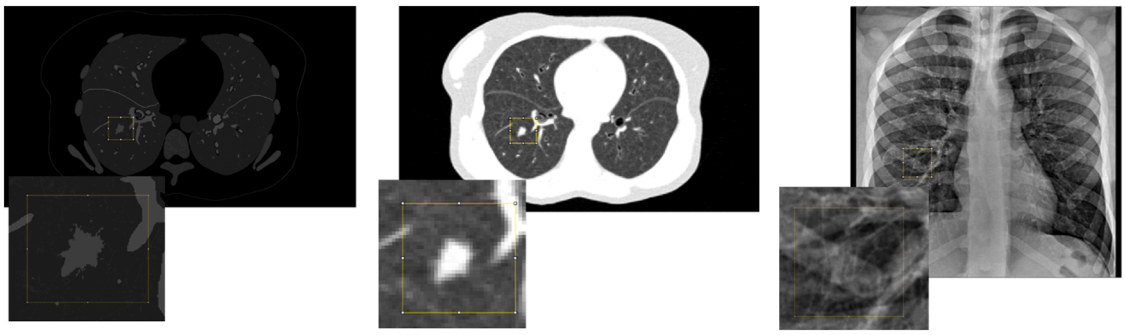



**Fig. 2.** Virtual Patient's inclusion and exclusion criteria and progress through the study. Within the study, 294 virtual patients were assessed, with 174 of them having a total of 512 lesions, varying from homogeneous to heterogeneous in nature. All 294 virtual patients underwent both virtual CT and CXR scans. For the CT cohort, *294 virtual images were processed through two distinct scanners, the Duke Legacy-12 and the Duke Legacy-20, with each scanner producing three unique imaging configurations per patient ($294 \times 3 = 882$). **From these six configurations, one CT image per patient, total 294 was randomly selected for evaluation. As for the CXR cohort, all 294 virtual patients were successfully imaged using the Duke Legacy-CXR scanner. A indicate the area under the receiver operating characteristic curve.



(a) Human Model (b) Simulated CT (c) Simulated CXR

**Fig. 3.** Example of human model and simulated images from the Virtual Lungs Screening Trial. Selected slice of (A) computation human model with a homogenous lesion (B) simulated CT scan image from Duke Legacy W20 scanner C) simulated CXR image using legacy post-processing.

2021)) to ensure ease of replication and minimize virtual reader variability across a variety of diagnostic scenarios for both modalities. The architecture, training methodologies, and patch extraction techniques remained constant across both models, with only the training datasets and data dimensions varying. This approach is akin to having the same radiologist interpret different modalities with consistent training.

The workflow started with image data augmentation and utilized RetinaNet's feature pyramid network to extract multi-scale features (Lin et al., 2017). Anchor boxes generated were matched with reference standard annotations provided with each public dataset, refined through regression, and classified to detect actionable nodules (Zhang et al., 2020). Then for the VLST study, the simulated CXR and CT image data were used for independent, external validation. For each virtual patient, each model detected the candidate locations and corresponding probabilities of being actionable nodules. Additionally, for each patient without nodules, one random location within the lungs was also analyzed by the model to represent negative backgrounds.

Detailed information regarding the development, validation, and clinical datasets used for the virtual reader models is documented in **Appendix C**.

### 2.5. Trial End point

Since each patient exam may receive multiple detections, the lesion with the highest probability value among all lesions in the patient was chosen to represent the aggregated, overall diagnosis for each patient. This patient-level performance replicated the decision-making process of radiologists who search for initial lesion candidates, then use the most suspicious as the index lesion to inform follow-up recommendations. The VLST's primary endpoint was the difference between CT and CXR for the recall for lung cancers at the initial baseline screening exam. Additionally, subgroup analyses were conducted to evaluate the performance difference for the two distinct lesion texture types. In contrast, clinical screening trials like NLST had a much more complex design involving three annual screening rounds and the final endpoint was based on mortality outcomes years later. Reproducing those complexities is beyond current VIT approaches, so the VLST endpoint was focused only on the reader performance for the cancer recall task, reflecting an earlier stage of clinical decision-making.

### 2.6. Statistical Analysis

Performance of VLST models was assessed using the Receiver Operating Characteristic (ROC) Area Under the Curve (AUC) at the patient level, with subgroup analyses for lesion types. The 95 % confidence intervals (CIs) were calculated using the DeLong method with 2000 bootstrapping samples (Boyd et al., 2013).

## 3. Results

In this VLST, virtual readers assessed paired exams from CT and CXR simulated from a cohort of 294 virtual human subjects. This cohort included 174 individuals with lung nodules and 120 without. The mean size (long axis) of these lesions was 10.09 mm, with a standard deviation of 5.09 mm (Fig. A.2). The smallest lesion measured was 4 mm, while the largest lesion was 34 mm. The lesion sizes were distributed such that 25 % of lesions were 6 mm or smaller (first quartile), the median size was 9 mm, and 75 % of the lesions (third quartile) were 12 mm or larger. This measurement method follows the approach used in the NLST, though current practice typically averages the two axial dimensions. Validation performances of the virtual reader models for CT and CXR on public datasets are included in **Appendix C**.

At the patient level, CT outperformed CXR with AUC of 0.92 (95 % CI: 0.89–0.95) compared to 0.72 (95 % CI: 0.67–0.77). Stratifying the results by lesion type revealed significant ($p < 0.001$) performance differences, but CT consistently outperformed CXR in each category. For

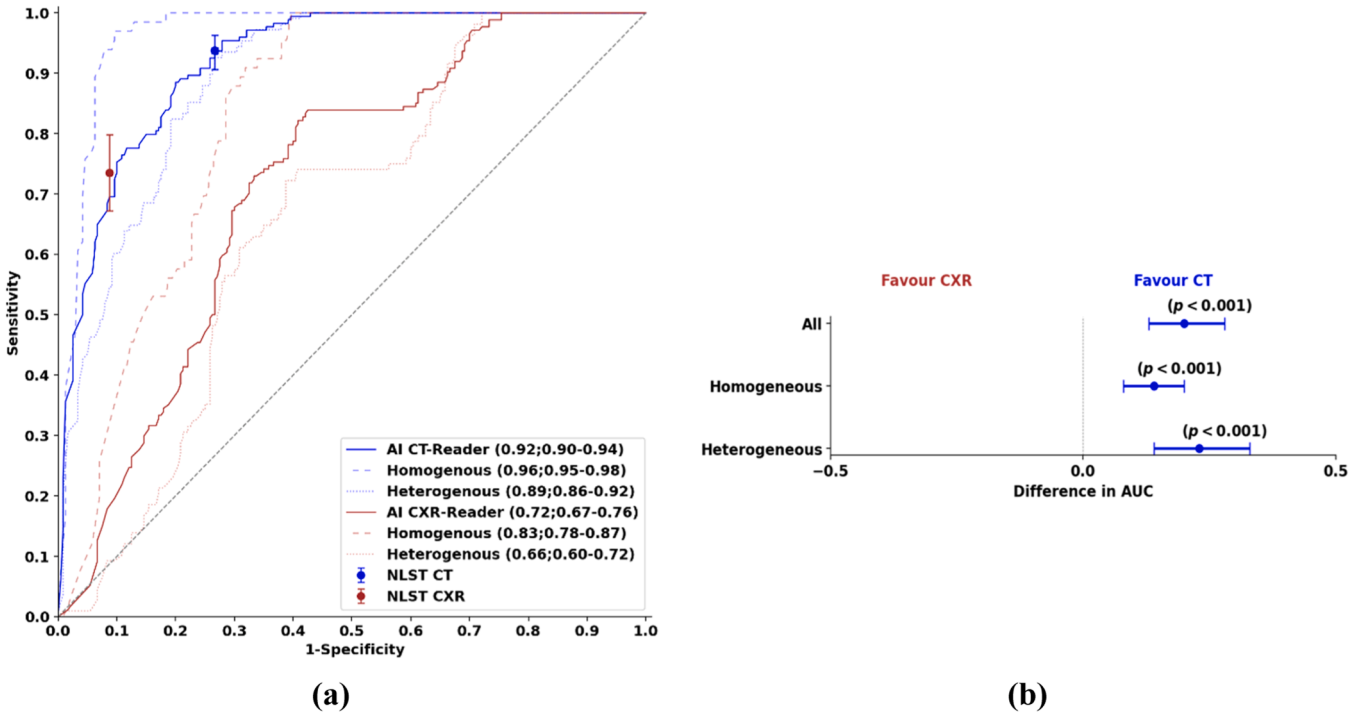



**Fig. 4.** Performance of the virtual readers predicting lung cancer across patient and lesion types. **(a)** Receiver operating characteristic (ROC) curves comparing the performance of the AI CT-Reader and AI CXR-Reader for patient-level predictions, homogeneous lesions, and heterogeneous lesions. Blue and brown lines represent AI CT-Reader and AI CXR-Reader performance, respectively. Homogeneous lesions are represented by a solid line, and heterogeneous lesions by a dashed line. The diagonal dashed line represents the ROC curve for a random classifier with an AUROC of 0.50. NLST CT and NLST CXR results are marked with blue and brown dot (error bar represents 95 % CI of sensitivity and specificity), respectively. **(b)** Forest plot showing the difference in AUROC between AI CT-Reader and AI CXR-Reader for patient-level predictions. The blue and brown markers represent the AUROC differences for CT and CXR, with error bars indicating the 95 % confidence intervals. The p-value for statistical significance is shown ($p < 0.001$).
AI = artificial intelligence. ROC = receiver operating characteristic. AUROC = area under the receiver operating characteristic curve. NLST = National Lung Screening Trial.

the homogeneous lesion type, the AUC for CT was 0.96 (95 % CI: 0.94–0.98), substantially higher than for CXR at 0.83 (95 % CI: 0.78–0.87) (Fig. 4). In contrast, for heterogeneous lesions, both modalities showed lower AUCs: CT 0.89 (95 % CI: 0.86–0.92) versus CXR 0.66 (95 % CI: 0.60–0.72) (Fig. 4). Both NLST and VLST findings reinforce the trend of CT outperforming CXR in lung cancer detection.

## 4. Discussion

The main purpose of this study was to establish a VIT platform capable of emulating certain aspects of major clinical trial in the modalities of CXR and CT imaging. To that end, this Virtual Lung Screening Trial for lung nodule detection represents a preliminary effort to model three fundamental elements of the real-world lung screening trial study: patients, scanners, and readers. By simulating a diverse virtual patient population, utilizing validated radiologic simulators for imaging, and employing machine learning algorithms as standardized virtual readers, our approach provides robust comparison using each virtual subject as their own control, with completely reproducible experimental methods in the end-to-end trial process. Reflecting real-world radiological assessments, CT outperformed CXR in nodule detection.

The composition of the trial population stands as a pivotal element in the execution and success of any clinical study. The VICTRE trial, one of the first published virtual trials in breast imaging, represents a significant milestone within the realm of virtual imaging trials, exhibiting results that are promising when juxtaposed with those derived from human trials (Badano et al., 2018). In comparison, our investigation demonstrated the trial effect using a smaller, strategically designed virtual cohort, in contrast to the larger cohorts used in existing clinical lung screening studies (Koning et al., 2011; De Koning et al., 2020; Field et al., 2016). Despite this, our study extended beyond the scope of prior VIT research, which typically focused on simulating specific pathologies or singular organs., The construction of our trial population posed challenges to replicate the complex thoracic anatomy across a spectrum of age, sex, and race/ethnicity to reflect the diverse population encountered in actual clinical trials (Abadi et al., 2020; Badano et al., 2018; Koning et al., 2011). Furthermore, we calibrated our virtual models to account for a range of lesion sizes and types. This diversity and clinical realism not only enhanced the relevance of our findings but also facilitated future research in complex bodily system simulations.

An advantage of DukeSim (virtual scanner) over physical scanners is its ability to simulate a wide range of imaging conditions and anatomical variations without the limitations of radiation exposure, patient variability, or scanner availability. This allows for more controlled experiments, rapid prototyping of imaging techniques, and optimization of screening protocols, all in a cost-effective and risk-free virtual environment, which is not feasible with physical scanners. While this study sought to mirror NLST-era scanners, other VITs may evaluate the impact of newer-generation technologies on performance. Using virtual patients also allowed scanning the same patient with both modalities, thus using each patient as their own control to provide greater power, which would not be possible in the real world due to the increase in radiation risk.

The virtual readers were designed using established deep-learning libraries (Cardoso et al., 2022) and were trained on publicly available clinical datasets (Setio et al., 2017; Sogancioglu et al., 2021) to ensure that the development process remained reproducible and broadly applicable. Unlike traditional VITs, which typically employ mathematical observer models that evaluate specific imaging regions for the presence or absence of signal (Badano et al., 2018; Samei & Krupinski, 2018), our virtual readers incorporated a search process (Lin et al., 2017). This process mirrored the diagnostic approach of a radiologist to first detect and then characterize lesions (Lin et al., 2017). In previous VIT studies for COVID-19 detection, we showed that deep-learning models can be susceptible to training and testing biases, which were more pronounced for CXR systems with their great diversity of image appearances. To minimize overtraining bias, we have deliberately avoided complex model architectures and elaborate training methodologies, thereby aligning the virtual readers' functionality as closely as possible with that of human readers across different imaging modalities (Dahal et al., 2022; Tushar et al., 2022; Tushar et al., 2023). Reported results demonstrated that CT outperformed CXR in nodule detection, which was consistent with the performance seen in clinical lung cancer screening trials. Our simulated homogenous lesions yielded higher detection performance in both modalities compared to the heterogeneous, demonstrating the effect on trial performance due to lesion texture.

## 5. Limitations

This study has several limitations. It was very time consuming to generate the complex 3D data, as a result the number of patients was much smaller compared to traditional clinical trials. This study serves as a preliminary demonstration of in silico trial capabilities for lung cancer screening, focusing on patient-level detection of cancer, similar to the recall task of clinical trial radiologists. Despite the small sample, our study design allowed efficient comparisons by assessing detection performance using a heavily enriched sample (>50 % cancer prevalence compared to 4 % for NLST) and paired imaging of each patient with both modalities. Future work will expand the virtual population and add lesion types such as semi-solid and ground glass. In terms of the virtual reader, only a single deep learning model was developed for each modality, limiting the assessment of inter-observer variability; this can be addressed by incorporating multiple models with varying architectures and training levels. Finally, variations in CXR performance were observed in VLST, potentially due to virtual scanner artifacts or virtual reader performance, highlighting areas for future improvement. For CT, the results demonstrated strong alignment with expected sensitivity, specificity, and AUC, reinforcing its effectiveness over CXR for cancer recall. These findings underscore the potential of VITs to capture key aspects of real-world clinical scenarios. While this study lacks several features of traditional multi-year screening trials, such as long-term endpoints like mortality or confirmed cancer diagnosis, it represents an early and crucial step toward charting a roadmap for comprehensive in-silico screening trials (Lago et al., 2023; Ria et al., 2024). Acknowledging these limitations, future research will focus on enhancing VITs to better replicate clinical trial scenarios, incorporating outcomes such as progression or survival while improving data efficiency.

## 6. Conclusion

In conclusion, this study presented one of the first end-to-end VITs in the domain of chest imaging, which involve the use of CXR and CT modalities. These modalities are the subject of considerable research interest due to their high clinical volume and impact. Together, the complexity of human models, versatility of scanners, and robustness of readers contribute to the advancement of virtual trials for complex bodily systems and imaging challenges. The transformative potential of virtual imaging trials in advancing evidence-based medicine offers an efficient and ethically conscious approach to medical research and development.

## CRediT authorship contribution statement

**Fakrul Islam Tushar:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Liesbeth Vancoillie:** Writing – review & editing, Supervision. **Cindy McCabe:** Data curation. **Amareswararao Kavuri:** Software, Data curation. **Lavsen Dahal:** Writing – review & editing, Data curation. **Brian Harrawood:** Writing – review & editing, Software. **Milo Fryling:** Software. **Mojtaba Zarei:** Data curation. **Saman Sotoudeh-Paima:** Writing –

review & editing. **Fong Chi Ho:** Data curation. **Dhrubajyoti Ghosh:** Formal analysis. **Michael R. Harowicz:** Writing – review & editing, Methodology, Investigation. **Tina D. Tailor:** Writing – review & editing, Methodology. **Sheng Luo:** Formal analysis. **W. Paul Segars:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Investigation. **Ehsan Abadi:** Writing – review & editing, Supervision, Methodology, Investigation. **Kyle J. Lafata:** Supervision, Investigation. **Joseph Y. Lo:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Investigation, Formal analysis. **Ehsan Samei:** Writing – review & editing, Supervision, Project administration.

## Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

FAKRUL ISLAM TUSHAR reports financial support was provided by National Institutes of Health. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

This work was funded in part by the National Institutes of Health (P41-EB028744, R01EB001838, R01HL155293).

## Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.media.2025.103576.

## Data availability

Data will be made available on request.

## References

Abadi, E., Harrawood, B., Sharma, S., Kapadia, A., Segars, W.P., Samei, E., 2019. DukeSim: a realistic, rapid, and scanner-specific simulation framework in computed tomography. IEEe Trans. Med. ImAging 38 (6), 1457–1465. https://doi.org/10.1109/tmi.2018.2886530.

Abadi, E., Segars, W., Tsui, B.M., et al., 2020. Virtual clinical trials in medical imaging: a review. J. Med. Imag 7 (4), 042805.

Abadi, Ehsan, W. Paul Segars, Nicholas Felice, Saman Sotoudeh-Paima, Eric A. Hoffman, Xiao Wang, Wei Wang et al. "AAPM Truth-based CT (TrueCT) reconstruction grand challenge." *Medical Physics* (2025).

Abadi, E., Paul Segars, W., Chalian, H., Samei, E., Feb 2021. Virtual imaging trials for coronavirus disease (COVID-19). AJR Am. J. Roentgenol. 216 (2), 362–368. https://doi.org/10.2214/AJR.20.23429.

Badano, A., Graff, C.G., Badal, A., et al., 2018. Evaluation of digital breast tomosynthesis as replacement of full-field digital mammography using an In silico imaging trial. JAMa Netw. Open. 1 (7), e185474. https://doi.org/10.1001/jamanetworkopen.2018.5474.

Boyd, K., Eng, K.H., Page, C.D., 2013. Area Under the Precision-Recall curve: Point Estimates and Confidence Intervals. Springer, pp. 451–466.

Cagnon, C.H., Cody, D.D., McNitt-Gray, M.F., Seibert, J.A., Judy, P.F., Aberle, D.R., 2006. Description and Implementation of a Quality Control Program in an Imaging-Based Clinical Trial. Elsevier.

Cardoso M.J., Li W., Brown R., et al. Monai: an open-source framework for deep learning in healthcare. *arXiv preprint arXiv:221102701*. 2022.

Chambon P., Delbrouck J.-.B., Sounack T., et al. CheXpert Plus: hundreds of thousands of aligned radiology texts, images and patients. *arXiv preprint arXiv:240519538*. 2024.

Clark, D.P., Badea, C.T., 2023. MCR toolkit: a GPU-based toolkit for multi-channel reconstruction of preclinical and clinical x-ray CT data. Med. Phys.

Dahal, L., Tushar, F.I., Abadi, E., et al., 2022. Virtual vs. Reality: external validation of COVID-19 classifiers using XCAT phantoms for chest radiography. presented at:. In: Medical Imaging 2022. Physics of Medical Imaging.

De Koning, H.J., Van Der Aalst, C.M., De Jong, P.A., et al., 2020. Reduced lung-cancer mortality with volume CT screening in a randomized trial. England J. Med 382 (6), 503–513. https://doi.org/10.1056/nejmoa1911793.

Ferlay, J., Colombet, M., Soerjomataram, I., et al., 2021. Cancer statistics for the year 2020: an overview. Int. J. Cancer 149 (4), 778–789. https://doi.org/10.1002/ijc.33588.

Field, J.K., Duffy, S.W., Baldwin, D.R., et al., 2016. The UK Lung Cancer Screening Trial: a pilot randomised controlled trial of low-dose computed tomography screening for the early detection of lung cancer. Health Technol. Assess. (Rockv) 20 (40), 1–146. https://doi.org/10.3310/hta20400.

Hamamci I.E., Er S., Almas F., et al. A foundation model utilizing chest CT volumes and radiology reports for supervised-level zero-shot detection of abnormalities. *arXiv preprint arXiv:240317834*. 2024.

Heleno, B., Thomsen, M.F., Rodrigues, D.S., Jørgensen, K.J., Brodersen, J., 2013. Quantification of harms in cancer screening trials: literature review. BMJ 347, f5334. https://doi.org/10.1136/bmj.f5334.

Jadick, G., Abadi, E., Harrawood, B., Sharma, S., Segars, W.P., Samei, E., 2021. A scanner-specific framework for simulating CT images with tube current modulation. Phys. Med. Biol. 66 (18), 185010.

Koning, H.J.de, 2011. Reduced lung-cancer mortality with low-dose computed tomographic screening. England J. Med 365 (5), 395–409. https://doi.org/10.1056/nejmoa1102873.

Lago, M., Sengupta, A., Badano, A., 2023. Design of an in Silico Imaging Trial With Growing Breast Cancer lesions: Comparison Between DM and DBT Detectability, 12463. SPIE. SPIE Medical Imaging.

Lavsen D., Yuqi W., Fakrul Islam T., et al. Automatic quality control in computed tomography volumes segmentation using a small set of XCAT as reference images. 2023:1246342.

Lin T.-.Y., Goyal P., Girshick R., He K., Dollár P. Focal loss for dense object detection. 2017:2980–2988.

McCabe, C., Solomon, J., Segars, W.P., Abadi, E., Samei, E., 2024. Synthesizing Heterogeneous Lung Lesions For Virtual Imaging Trials, 12925. SPIE. SPIE Medical Imaging.

National Cancer I. National lung screening Trial (NLST). 2024/09/03 2024.

Pastorino, U., Sverzellati, N., Sestini, S., et al., 2019. Ten-year results of the Multicentric Italian Lung Detection trial demonstrate the safety and efficacy of biennial lung cancer screening. Eur. J. Cancer 118, 142–148. https://doi.org/10.1016/j.ejca.2019.06.009.

Ria, F., Zhang, A.R., Lerebours, R., et al., 2024. Optimization of abdominal CT based on a model of total risk minimization by putting radiation risk in perspective with imaging benefit. Commun. Med. (Lond) 4 (1), 272. https://doi.org/10.1038/s43856-024-00674-w, 2024/12/19.

Samei, E., Krupinski, E.A., 2018. The Handbook of Medical Image Perception and Techniques. Cambridge University Press.

Sauer, T.J., Bejan, A., Segars, P., Samei, E., 2023. Development and CT image-domain validation of a computational lung lesion model for use in virtual imaging trials. Med. Phys. 50 (7), 4366–4378. https://doi.org/10.1002/mp.16222.

Segars, W.P., Bond, J., Frush, J., et al., 2013. Population of anatomically variable 4D XCAT adult phantoms for imaging research and optimization. Med. Phys. 40 (4), 043701. https://doi.org/10.1118/1.4794178.

Setio, A.A.A., Traverso, A., De Bel, T., et al., 2017. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge. Med. Image Anal. 42, 1–13. https://doi.org/10.1016/j.media.2017.06.015.

Siegel, R.L., Miller, K.D., Wagle, N.S., Jemal, A., 2023. Cancer statistics, 2023. CA Cancer J. Clin. 73 (1), 17–48.

Sogancioglu, E., Murphy, K., Bv, Ginneken, Scholten, E., Schalekamp, S., Hendrix, W., 2021. Data from: Node21. Grand Challenge Competition.

Sung, H., Ferlay, J., Siegel, R.L., et al., 2021. Global Cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J. Clin. 71 (3), 209–249. https://doi.org/10.3322/caac.21660.

Tushar, F.I., Abadi, E., Sotoudeh-Paima, S., et al., 2022. *Virtual* vs. *reality: External Validation of COVID-19 Classifiers Using XCAT Phantoms For Chest Computed Tomography*, 12033. SPIE. SPIE Medical Imaging.

Tushar F.I., Dahal L., Sotoudeh-Paima S., et al. Data diversity and virtual imaging in AI-based diagnosis: a case study based on COVID-19. *arXiv preprint arXiv:230809730*. 2023.

Welch, H.G., Woloshin, S., Schwartz, L.M., et al., 2007. Overstating the evidence for lung cancer screening. Arch. Intern. Med. 167 (21), 2289. https://doi.org/10.1001/archinte.167.21.2289.

Zhang S., Chi C., Yao Y., Lei Z., Li S.Z. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. 2020:9759–9768.