

Deep virtual CT workflow for evaluating AI in low-dose CT lung cancer screening: a 2D demonstration

Rongping Zeng^a, Andreu Badal^a, Ahad Ollah Ezzati^b, Xun Jia^b, Arjun Krishna^c, Klaus Mueller^c, Kyle J Myers^d, Chuang Niu^e, Md Selim^a, Ge Wang^e, and Wenjun Xia^e

^aFood and Drug Administration, Silver Spring, MD

^bDepartment of Radiation Oncology and Molecular Radiation Sciences, John Hopkins University, Baltimore, MD

^cComputer Science Department, Stony Brook University, Stony Brook, NY

^dPuente Solutions LLC, Phoenix, AZ

^eDepartment of Biomedical Engineering, Rensselaer Polytechnic Institute, Troy, NY

ABSTRACT

Over the past decade, most CT systems in the United States have incorporated specialized low-dose CT (LDCT) acquisition protocols for lung cancer screening (LCS). These systems were initially equipped with traditional filtered backprojection (FBP) and statistical/model-based iterative reconstruction (IR) methods. Currently, deep learning (DL)-based reconstruction and denoising (DLR/D) methods are widely available for LDCT, with the potential to generate visually appealing images at reduced dose levels. Yet, the issues of generalizability and instability are recognized, and new evaluation methods need to be developed to ensure that DLR/D images preserve diagnostic quality for various patient characteristics and pathological conditions in lung cancer screening exams. These assessments are likely to require substantial testing data, which may not be possible with patient scans. In this work, we study the feasibility of using a deep learning-based virtual CT workflow to evaluate DLR/D methods in LDCT for LCS. To achieve this goal, our deep virtual CT workflow is designed with the four major components: AI-generated patients, a realistic simulated CT sinogram, deep reconstruction, and performance evaluation in terms of both traditional image quality metrics and AI-based lung nodule detection accuracy. Preliminary results are presented to demonstrate the potential of such a deep workflow for evaluating DLR/D methods in LDCT for LCS.

Keywords: Low-dose computed tomography, lung cancer screening, deep learning, image quality assessment, diagnostic performance, virtual workflow

1. DESCRIPTION OF PURPOSE

The National Lung Screening Trial (NLST) was a randomized multi-center study comparing low-dose CT (LDCT) with chest radiography (CXR) in the screening of older current and former heavy smokers for early detection of lung cancer. The trial indicated a significant (20%) reduction in lung cancer mortality in high-risk smokers who were screened annually (three times) with LDCT versus CXR.¹ Additionally, European studies, including the NELSON and UKLS trials, also showed a similar efficiency of LDCT for LCS.^{2,3} With the demonstrated benefit of LDCT vs CXR, CT vendors integrated LDCT-LCS protocols for their CT systems. These protocols were defined to yield equivalent or better image quality compared to those used in the NLST. Image quality evaluation was based on the standard CT bench testing performance metrics, including CT number accuracy, uniformity, noise, modulation transfer function (MTF), slice thickness, contrast-to-noise ratio (CNR), and noise power spectrum (NPS). Now with the take-up of deep learning (DL)-based reconstruction and denoising (DLR/D) methods, those traditional bench performance metrics may not be sufficient to establish the equivalence of DLR/D methods with the traditional reconstruction options. It is important to evaluate the performance of DLR/D methods for LDCT on tasks relevant to lung cancer screening, such as lung nodule detection and classification. In addition, deep

Send correspondence to Xun Jia, Klaus Mueller, Ge Wang, Rongping Zeng.

E-mails: xunjia@jhu.edu, mueller@cs.stonybrook.edu, wangg6@rpi.edu, rongping.zeng@fda.hhs.gov

learning algorithms may have issues in stability or generalizability, especially when processing data acquired at a very low-dose setting. A large dataset is likely to be required to fully evaluate these AI-specific aspects.

The purpose of this study is to build a virtual CT workflow and evaluate the performance of DLR/D methods for LDCT-LCS. Our virtual workflow incorporates the following components developed by multiple institutes: AI-generated patients, a GPU-based Monte Carlo (MC) CT scanner simulator, deep reconstruction and denoising, and performance assessment in a lung nodule detection task. We present preliminary results to demonstrate the feasibility of such a virtual workflow for evaluating DLR/D methods in the context of LDCT-LCS.

2. METHODS

2.1 AI-generated lung patients

We employed a Denoising Diffusion Probabilistic Model⁴ (DDPM) to train an unconditional large-scale generative model on a large lung-CT dataset. We then performed a conditional generation by sampling via the trained DDPM⁵ and expanded it to encompass multiple reference images simultaneously for guidance, allowing the dynamic application of annotations and conditions in the image generation process. We used our multi-conditioned sampling approach for producing lung patient anatomical images with specific pathologies, such as lung nodules at designated locations.

Our DDPM was trained on a large dataset of 5k lung CT scans, with images extracted at full HU width of 2000. The DDPM iteratively transformed an isotropic Gaussian distribution into a lung CT image distribution. The Markov Chain model learned the reverse of the forward diffusion process, $q(x_t|x_{t-1}) := N(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I)$ with x_t as the latents with added noise and β_t as a fixed variance schedule. The reverse diffusion process $p_\theta(x_{t-1}|x_t) = N(x_{t-1}; \mu_\theta(x_t, t), \sigma_t^2 I)$ was learned via a neural network.

Choi et al. proposed a method to direct the sampling process towards a subset of image distributions surrounding a reference image y , provided that a similarity between the reference image y and the generated image x_0 at a lower-resolution level could be established. Choi et al.'s method consistently enhanced the latent variable x_t within steps $[T, a)$ by incorporating the corresponding low-pass filtered content of the reference image at each sampling step t , denoted by y_t ($y_t = \sqrt{\alpha_t}y + \sqrt{1-\alpha_t}\epsilon$). This ensured that x_t resembled y_t in the low-frequency space. Specifically, the conditional sampling scheme can be formulated as follows:

$$x'_{t-1} = x_{t-1} + \phi_N(y_{t-1}) - \phi_N(x_{t-1}) \quad (1)$$

where $\phi_N(\dots)$ is a low-pass linear filter with N as the downsampling factor.

For a set of M conditional images $y_1, y_2, \dots, y_M$, the downsampling factor of a linear filter ϕ and the number of steps to implement the conditional sampling scheme can be manipulated to fine-tune the DDPM sampling process. This allows for a trained DDPM to generate valid images that share low-resolution features (or similarity) with each of the conditional images. By introducing sets of integers $n_1, n_2, \dots, n_M$ and $a_1, a_2, \dots, a_M$, where n_m denotes the downsampling factor for a linear filter corresponding to each conditional image during the DDPM sampling steps $[T, a_m > 1)$, Eq. (1) can be modified as follows:

$$x'_t = x_t + \sum_{m=1}^M \begin{cases} (\phi_{n_m}(y_{m_t}) - \phi_{n_m}(x_t)), & \text{if } t \geq a_m \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

The integers n_m, a_m for a conditional image depend on the purpose of the conditional image in the generation of images. Our sampling strategy is compiled in Algorithm 1. Specifically for this study, we used two conditional images, one generative lung anatomy map and one lung CT slice with pathology.

Algorithm 1: The DDPM sampling procedure with M conditional images for generating a lung patient image with T time-steps ($T = 4000$ in this study)

```

1 Input: Conditional / guidance images  $y_1, \dots, y_M$ 
2 Output: Generated image  $x$ 
3 Linear filter:  $\phi_{n_1}, \dots, \phi_{n_M}$ 
4 Time-steps (T, a):  $a_1, \dots, a_M$ 
5  $x_T \sim N(0, I)$ 
6 for  $t = T$  to 1 do
7    $z \sim N(0, I)$ 
8   if  $t = 1$  then
9      $z = 0$ 
10   $x_{t-1} = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{1-\alpha_t}{\sqrt{1-\alpha_t}}\epsilon_\theta(x_t, t)) + \sigma_t z$ 
11   $X = 0$ 
12  for  $s = 1$  to  $M$  do
13     $y_{s_{t-1}} \sim q(y_{s_{t-1}}|y_s)$ 
14    if  $t \geq a_s$  then
15       $X = X + \phi_{n_s}(y_{s_{t-1}}) - \phi_{n_s}(x_{t-1})$ 
16     $x_{t-1} \leftarrow x_{t-1} + X$ 
17 return  $x_0$ 

```

2.2 CT system simulation

A 512×512 2D lung patient image generated in the first step was treated as a volume of one slice and assigned a voxel size of $0.9 \times 0.9 \times 6$ mm³. The HU value of each pixel was converted to tissue density and material composition by using the Schneider et al.⁶ correlations. We performed Monte Carlo simulation to model the CT imaging process using a GPU-based MC package gMCDRR,^{7,8} which transports photons based on physics principles and cross sections, and employs the Woodcock algorithm to reduce the simulation time. The code was modified to simulate a ring detector with 16 slices as a real CT scanner. A 120 kVp Tungsten x-ray tube with a focal spot size of 0.4 mm was assumed. The source-to-object distance was 50 cm. The source to detector distance was 95 cm. The simulated x-ray beam size was 540 mm in the fan direction and 5.3 mm in the z-direction at the isocenter to cover all 16 slices. The source emission probability distribution was modified to reflect the effect of a bow-tie filter. Nine hundred 2D projections equally spaced in a full rotation range were computed. The detector pixel pitch was 0.625 mm and the number of pixels in each slice was 1,642. The detector was assumed to be energy integrating (EID), thus once the photons reach the detector, their energy was recorded. After that, Gaussian electronic noise with 350000 eV standard deviation (SD) and zero mean was added. The SD equals to five x-ray photons with 70 keV energy (mean energy of 120 kVp x-rays) that was estimated by averaging the values reported by Fu et al.⁹ and Xu et al.¹⁰ Finally, the fan-beam sinogram of a single-slice volume was extracted from the central row of each 2D projection view over all 900 angles.

In the current simulation, the total number of photons was 5.8×10^{11} to produce approximately 1.5 mSv effective dose as the typical dose used in LDCT LCS.¹¹ This number was increased to 2.3×10^{12} for simulating normal-dose CT (NDCT) with approximately 6 mSv effective dose. The simulation times for LDCT and NDCT cases on a RTX A4000 Nvidia GPU card were approximately 3 and 12 hours, respectively.

2.3 Deep learning-based image reconstruction and denoising

The sinogram simulated in the last step was first reconstructed by the traditional FBP method. The FBP image was then denoised by a pre-trained DDPM CT image denoising model.

Unlike the DDPM described in Section 2.1, this model uses the low-dose counterpart as a conditional input. The reverse process is defined by the distribution: $p_\theta(x_{t-1}|x_t) = N(x_{t-1}; \mu_\theta(x_t, y, t), \sigma_t^2 I)$, where y represents the low-dose image, x is the corresponding clean image to be estimated, μ_θ denotes the mean predicted by the

network, and σ is the standard deviation. The corresponding loss function for training the model is formulated as follows:

$$\mathcal{L} = \mathbb{E}_{x,y} \mathbb{E}_{\epsilon,t} \|\epsilon - \epsilon_{\theta}(\sqrt{\alpha_t}x_0 + \sqrt{1 - \alpha_t}\epsilon, y, t)\|_2^2 \quad (3)$$

where ϵ_{θ} denotes the trained model and $\epsilon \sim N(0, I)$ is the added noise. This model was pre-trained on the Mayo Clinic's public dataset,¹² which includes 2,378 paired LDCT and NDCT 3mm-thickness images from 10 patients. The dose level of the LDCT scan was 25% of the NDCT dose level, matching the LDCT to NDCT dose ratio used in our simulations. We configured the model with a total of 1,000 steps. Training was conducted on a server using four Nvidia Tesla V100 GPUs, involving 500,000 parametric updates. The entire training process took approximately 46 hours. After the training was completed, each slice can be processed within 90 seconds.

2.4 Performance evaluation

A successful implementation of the above three steps provides NDCT, LDCT and DDPM denoised LDCT images of the same lung patient case. We first compared the image quality of the three types of images using traditional metrics including peak-to-signal noise ratio (PSNR), structure similarity index (SSIM), pixel-noise standard deviation (SD), contrast-to-noise ratio (CNR), and profiles of selected nodules.

We further compared the image quality using the downstream lung nodule detection task. We trained a 3D nodule detection model^{13,14} on the LUNA16 dataset,¹⁵ which contains LDCT scans derived from the LIDC-IDRI database.¹⁶ Our implementation followed the protocol outlined at the Nvidia's MONAI Lung Nodule CT Detection model repository*. During training, the image size of the model input was $192 \times 192 \times 80$ with a pixel size of $0.7 \times 0.7 \times 1.25$ mm³. The HU range was set to $[-1024, 300]$. The model was optimized to minimize a combined binary cross-entropy and L1 loss using the ADAM optimizer with a learning rate of 0.01. Due to the limited number of simulated CT scans in our preliminary work, the lung nodule detection model was not re-trained on the simulated lung CT scans but directly applied to them to detect lung nodules. Since the model was trained on 3D lung CT volumes, to match the image dimensionality of the model training data, each simulated single-slice lung CT was replicated and stacked to form a 3D volume with dimensions of $512 \times 512 \times 64$, with the HU range clipped into $[-1024, 300]$ and then normalized to $[0, 1]$. Case-based sensitivity and specificity were used as endpoints to summarize the performance of NDCT and LDCT with and without the DDPM denoising. A diseased case was considered correctly detected provided that the true nodule was detected at the right location, even with false nodule identifications. Uncertainty estimation of the sensitivity and specificity values were based on the Bootstrap method.

We acknowledge that the model performance is likely degraded due to the distribution shift between the trained 3D CT volume dataset and the pseudo 3D volume generated by stacking the same slice in the current implementation. However, it is the goal of this project to create a realistic virtual workflow with a truly 3D AI-based lung patient generative network such that the lung nodule detection model trained on real patient LDCT data does not need to be retrained to perform nodule detection on the virtually generated lung CT scans. In the future, the nodule detection model will be re-trained when many simulated CT images are generated in our virtual CT workflow and compared to the nodule detection model without retraining. The performance gap between the nodule detection models with and without retraining can be used as an indicator of the clinical realism of the CT images generated by our virtual workflow.

3. RESULTS

A total of 40 healthy and 140 diseased virtual lung patient cases were created by the DDPM-based generative model. Each diseased case contained only one nodule. Fig. 1 shows some of the images generated with our DDPM lung patient model. It is observed that the generated lesions realistically attach to the lung walls or surrounding tissues when in close proximity; this mimics the behavior of tumors in these regions. We also evaluated the realism of our generated images through a Visual Turing Test¹⁷ (<https://visual-turing-test.vercel.app>) with three radiologists from the Icahn School of Medicine at Mount Sinai; where they were asked to label a randomly selected lung CT image as "Real" or "Fake" from a set of 30 real and generated images (half of each type). An accuracy of about 46% indicated that our generative framework passed the Test and can generate

*https://catalog.ngc.nvidia.com/orgs/nvidia/teams/monaitoolkit/models/monai_lung_nodule_ct_detection

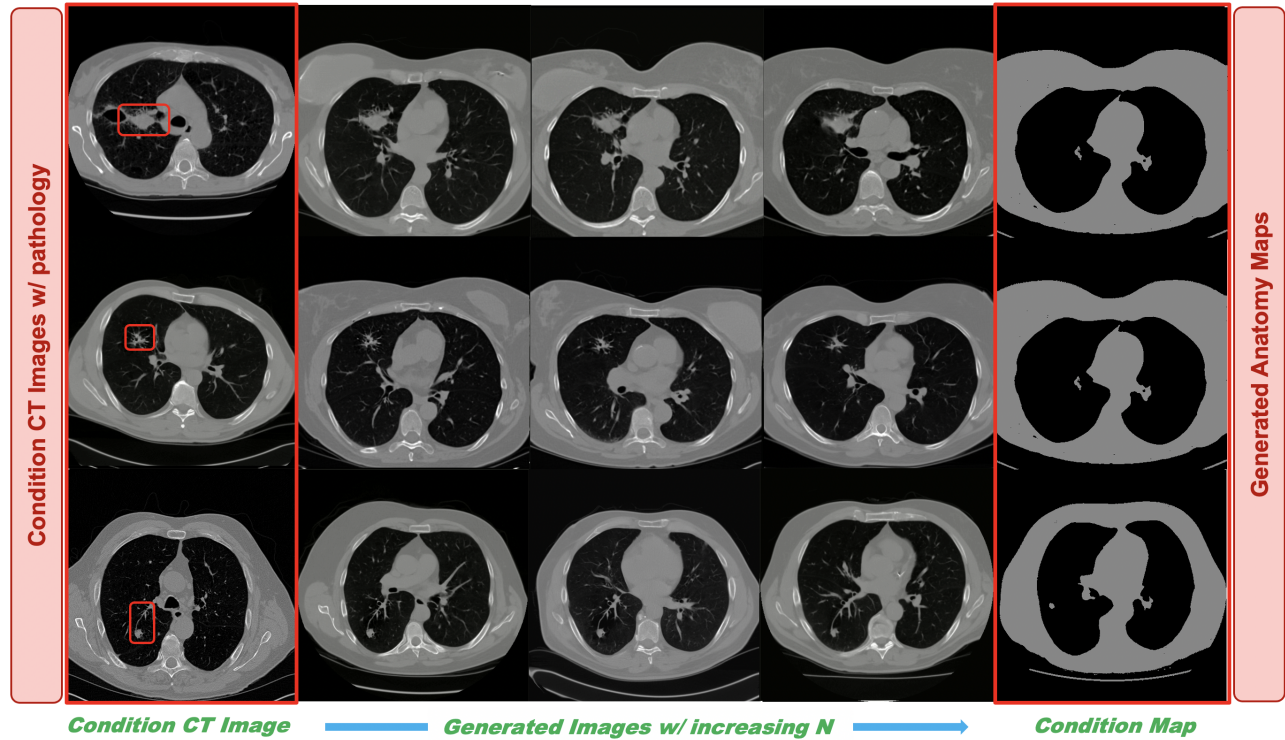


Figure 1. **AI-generated lung patient results.** Examples of generated lung-CT images with pathology using multi-condition sampling; which follows the anatomy of the conditional image (generated using B-Spline¹⁷ curves) on the right and pathology of the real CT image on the left. By varying N (the downsampling factor of a low-pass linear filter) for the conditional pathology image, we can generate different versions of the same cancer/tumor/nodules having varying spiculation patterns.

perceptually realistic images. However, quantitative and statistical feature analysis will be conducted in the future to further validate their clinical realism.

Fig. 2 presents the NDCT and LDCT images before and after DDPM denoising of two cases, along with the profiles of the nodules highlighted in red in the NDCT images. The SD of the selected region of interest (ROI) in LDCT was 184 HU, about double of the SD of 86 HU in the NDCT, matching the quarter-dose setting we used in the LDCT simulation. Using the simulated NDCT data (6 mSv) as a reference, the average PSNR improved from 18.31 to 24.64, and the average SSIM increased from 0.2116 to 0.4089. The CNR of the nodule area improved from 0.9919 to 1.1716 with DDPM denoising, approaching the NDCT value of 1.1756. Additionally, the nodule profiles demonstrate that the denoised nodules closely match the NDCT images, effectively removing the noise in the LDCT image.

Table 1 summarizes the nodule detection accuracies in the NDCT and LDCT images with and without DDPM denoising. The mean and standard deviation of the sensitivity and specificity values were based on the Bootstrap method with 1000 times of sampling with replacement. The results show that the sensitivity was improved in LDCT but the specificity stayed almost the same after the DDPM denoising. A decrease in dose significantly

Table 1. **Nodule detection accuracy.** Mean (Standard deviation) of sensitivity and specificity in the NDCT, LDCT and DDPM denoised LDCT images.

Data type	Sensitivity	Specificity	(Sensitivity + Specificity)/2
NDCT	0.626 (0.039)	0.900 (0.048)	0.763 (0.030)
LDCT	0.569 (0.042)	0.774 (0.066)	0.672 (0.039)
DDPM denoised LDCT	0.687 (0.039)	0.774 (0.065)	0.730 (0.038)

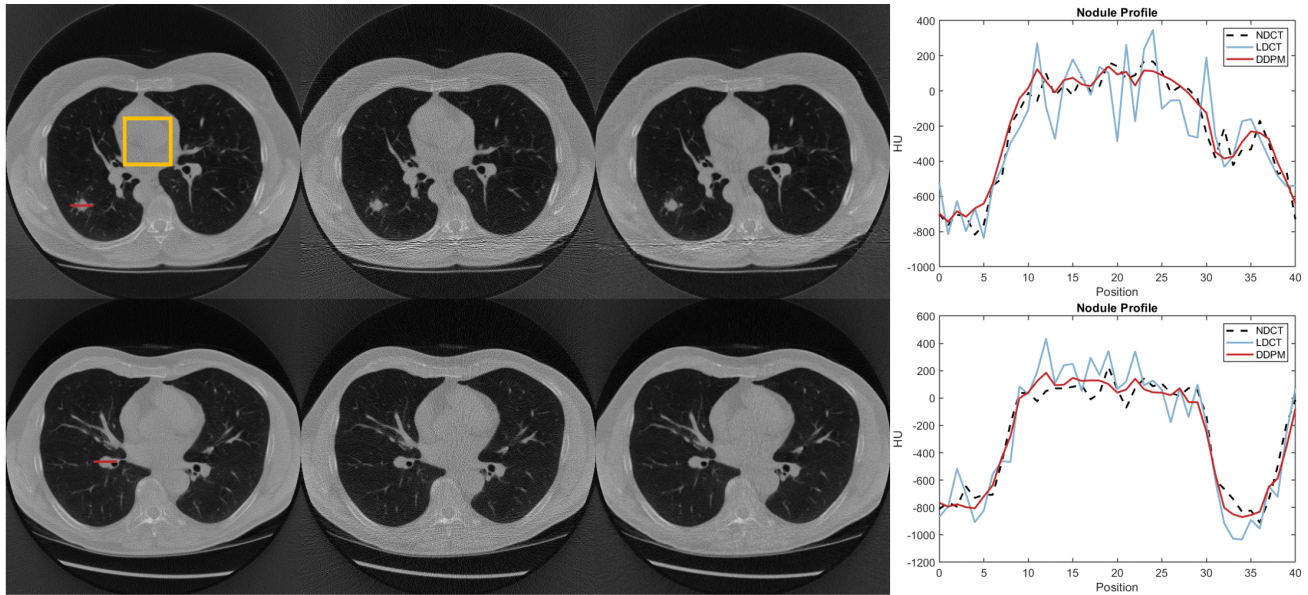


Figure 2. **Reconstruction and Denoising Results.** The figure presents two representative cases, with images displayed from left to right in the following order: NDCT, LDCT, DDPM images, and nodule profiles. The display window is set to $[-1000, 1000]$ HU. The rectangle marks the ROI that was used to calculate standard deviation.

reduced the specificity from 0.90 in ND to about 0.77 in LD regardless of whether DDPM was applied or not, indicating that greater CT noise may have caused more false nodule detection. However, the sensitivity appears to be low, less than 0.7 even in ND.

A manual check on the detection results revealed that the nodule detection model missed some visually obvious nodules. We believe that a data distribution shift between the pseudo-3D CT volumes simulated in the study and the model's training CT volumes may be the dominant factor on the sensitivity performance. As an illustration, see Fig. 3 for the nodule detection for one patient case selected from the 180 cases simulated in this work. In the LDCT image, the detection module was unable to detect the nodule, although it was relatively obvious to human eyes. However, after the image was denoised, the nodule was successfully detected. In the future, truly 3D generative lung patient models will be developed to incorporate lung nodules of various sizes, density, and shapes to allow more diverse task-based performance evaluations such as classification and size estimation in the workflow.

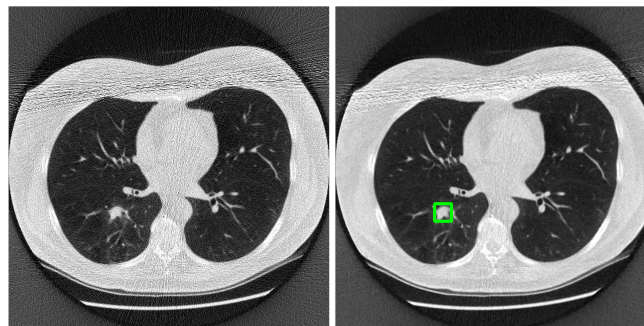


Figure 3. **Nodule Detection Results.** Left: A lung nodule was not detected in the low-dose CT image; Right: The same nodule was detected in the DDPM reconstructed low-dose CT image. Note that all our workflow-generated images were not used in training the detection model. The display window is $[-1,024, 300]$ HU.

4. CONCLUSIONS

This work presents our initial effort to integrate the four major components developed by multiple research institutes into a fully functional virtual CT workflow. The four major components are: AI-generated lung patients with and without pathology, MC-based CT simulator incorporating x-ray physics and flexible scanner geometric settings, state-of-the-art deep image reconstruction and denoising, and image quality evaluation tailored for LDCT-LCS. The preliminary results demonstrate the feasibility of using this workflow to create data for evaluating the performance of DLR/D in LDCT-LCS. However, substantial work is needed to further validate the realism of the lung patient models in downstream clinical tasks, speed up the CT simulator, refine the lung detection and classification models to avoid out-of-distribution bias, and enrich the performance evaluation with subgroup performance analysis and other classification tasks. It is the authors' plan to share the software tools publicly after the virtual workflow is fully developed and validated in the specified context of evaluating DLR/D methods in LDCT-LCS.

5. ACKNOWLEDGEMENTS

The work was supported by the NIH grant R01EB032716. Md Selim acknowledges funding by appointment to the research participant program at the Center for Devices and Radiological Health administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and the U.S. Food and Drug Administration. The mention of commercial products, their sources, or their use in connection with material reported herein is not to be construed as either an actual or implied endorsement of such products by the U.S. Department of Health and Human Services.

REFERENCES

- [1] The National Lung Screening Trial Research Team, "Reduced lung-cancer mortality with low-dose computed tomographic screening," *New England Journal of Medicine* **365**(5), 395–409 (2011).
- [2] de Koning, H. J., van der Aalst, C. M., de Jong, P. A., Scholten, E. T., Nackaerts, K., Heuvelmans, M. A., Lammers, J.-W. J., Weenink, C., Yousaf-Khan, U., Horeweg, N., van 't Westeinde, S., Prokop, M., Mali, W. P., Hoesein, F. A. M., van Ooijen, P. M., Aerts, J. G., den Bakker, M. A., Thunnissen, E., Verschakelen, J., Vliegenthart, R., Walter, J. E., ten Haaf, K., Groen, H. J., and Oudkerk, M., "Reduced lung-cancer mortality with volume CT screening in a randomized trial," *New England Journal of Medicine* **382**(6), 503–513 (2020).
- [3] Field, J. K., Vulkan, D., Davies, M. P., Baldwin, D. R., Brain, K. E., Devaraj, A., Eisen, T., Gosney, J., Green, B. A., Holemans, J. A., Kavanagh, T., Kerr, K. M., Ledson, M., Lifford, K. J., McDonald, F. E., Nair, A., Page, R. D., Parmar, M. K., Rassl, D. M., Rintoul, R. C., Screatton, N. J., Wald, N. J., Weller, D., Whynes, D. K., Williamson, P. R., Yadegarfar, G., Gabe, R., and Duffy, S. W., "Lung cancer mortality reduction by LDCT screening: UKLS randomised trial results and international meta-analysis," *The Lancet Regional Health - Europe* **10**, 100179 (2021).
- [4] Nichol, A. Q. and Dhariwal, P., "Improved denoising diffusion probabilistic models," in [*Proceedings of the 38th International Conference on Machine Learning*], Meila, M. and Zhang, T., eds., *Proceedings of Machine Learning Research* **139**, 8162–8171, PMLR (18–24 Jul 2021).
- [5] Choi, J., Kim, S., Jeong, Y., Gwon, Y., and Yoon, S., "Ilvr: Conditioning method for denoising diffusion probabilistic models," in [*2021 IEEE/CVF International Conference on Computer Vision (ICCV)*], 14347–14356 (2021).
- [6] Schneider, W., Bortfeld, T., and Schlegel, W., "Correlation between CT numbers and tissue parameters needed for Monte Carlo simulations of clinical dose distributions," *Physics in Medicine & Biology* **45**(2), 459 (2000).
- [7] Jia, X., Yan, H., Cerviño, L., Folkerts, M., and Jiang, S. B., "A GPU tool for efficient, accurate, and realistic simulation of cone beam ct projections," *Medical physics* **39**(12), 7368–7378 (2012).
- [8] Mettivier, G., Lai, Y., Jia, X., and Russo, P., "Virtual dosimetry study with three cone-beam breast computed tomography scanners using a fast GPU-based Monte Carlo code," *Physics in Medicine & Biology* **69**(4), 045028 (2024).

- [9] Fu, L., Lee, T.-C., Kim, S. M., Alessio, A. M., Kinahan, P. E., Chang, Z., Sauer, K., Kalra, M. K., and De Man, B., “Comparison between pre-log and post-log statistical models in ultra-low-dose CT reconstruction,” *IEEE transactions on medical imaging* **36**(3), 707–720 (2016).
- [10] Xu, J. and Tsui, B. M., “Electronic noise modeling in statistical iterative reconstruction,” *IEEE Transactions on Image Processing* **18**(6), 1228–1238 (2009).
- [11] Larke, F. J., Kruger, R. L., Cagnon, C. H., Flynn, M. J., McNitt-Gray, M. M., Wu, X., Judy, P. F., and Cody, D. D., “Estimated radiation dose associated with low-dose chest CT of average-size participants in the national lung screening trial,” *American Journal of Roentgenology* **197**(5), 1165–1169 (2011). PMID: 22021510.
- [12] McCollough, C., “TU-FG-207A-04: overview of the low dose CT grand challenge,” *Medical physics* **43**(6Part35), 3759–3760 (2016).
- [13] Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P., “Focal loss for dense object detection,” in [*Proceedings of the IEEE international conference on computer vision*], 2980–2988 (2017).
- [14] Baumgartner, M., Jäger, P. F., Isensee, F., and Maier-Hein, K. H., “nnDetection: a self-configuring method for medical object detection,” in [*Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24*], 530–539, Springer (2021).
- [15] Setio, A. A. A., Traverso, A., De Bel, T., Berens, M. S., Van Den Bogaard, C., Cerello, P., Chen, H., Dou, Q., Fantacci, M. E., Geurts, B., et al., “Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the luna16 challenge,” *Medical image analysis* **42**, 1–13 (2017).
- [16] Armato III, S. G., McLennan, G., Bidaut, L., McNitt-Gray, M. F., Meyer, C. R., Reeves, A. P., Zhao, B., Aberle, D. R., Henschke, C. I., Hoffman, E. A., 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* **38**(2), 915–931 (2011).
- [17] Krishna, A., Yenneti, S., Wang, G., and Mueller, K., “Image factory: A method for synthesizing novel CT images with anatomical guidance,” *Medical Physics* **51**(5), 3464–3479 (2024).