
Article:
Astley, J.R., Marshall, H., Smith, L.J. et al. (2026) Uncertainty-Aware, End-to-End Deep Learning for Functional Lung MRI Quantification Using 129Xe and 1H MRI. Radiology: Cardiothoracic Imaging, 8 (3). e250371. ISSN: 2638-6135

https://doi.org/10.1148/ryct.250371

Uncertainty-Aware, End-to-End Deep Learning for Functional Lung MRI Quantification Using ^{129}Xe and ^{1}H MRI

Keywords: Functional imaging, lung, multi-modal, MRI, uncertainty-aware. **Article Type:**

Original research article

Summary Statement:

An end-to-end, uncertainty-aware deep learning approach enabled direct quantification of functional lung MRI without multi-modal segmentation, performing accurately across a range of pathologies, hospitals, and test-time augmentations.

Key Points:

- In this retrospective study of 574 participants, an uncertainty-aware end-to-end deep learning framework was developed to predict the ventilation defect percentage from paired hyperpolarized ^{129}Xe-MRI and ^{1}H-MRI data.
- The model achieved a median absolute error of 1.01% and 91% clinical accuracy (95% CI: 68%, 94%; n = 32/35) in classifying normal *versus* abnormal ventilation defect percentage, with no evidence of difference compared with state-of-the-art segmentation-based workflows (P = .70).
- Prediction uncertainty was quantified using Monte Carlo dropout-derived distributions, enabling confidence stratification that may support clinical interpretation and workflow efficiency.

Abbreviations:

VDP = ventilation defect percentage, DL = deep learning, ^{129}Xe = hyperpolarized Xenon-129, MC = Monte-Carlo , CNN = convolutional neural network, MedAE = median absolute error

Abstract

Purpose: To develop an automatic end-to-end deep learning pipeline for predicting the ventilation defect percentage (VDP) from co-registered functional hyperpolarized Xenon-129 (^{129}Xe)-MRI and structural proton (^{1}H)-MRI scans without manual intervention.

Materials and Methods: In this retrospective study (2015–2024), ^{129}Xe-MRI and ^{1}H-MRI scans from healthy participants and patients with a range of pulmonary diseases were used to predict VDP and its associated prediction confidence via an uncertainty-aware convolutional neural network framework. Monte-Carlo (MC) dropout was used to quantify model uncertainty. Model robustness was assessed using test-time augmentation to simulate test-retest repeatability. We evaluated the proposed approach on a stratified testing set via the median absolute error (MedAE).

Results: The dataset comprised 574 paired ^{129}Xe-MRI and ^{1}H-MRI scans from 47 healthy participants (mean±SD age = 28.3±17.3; 28F) and 527 patients with a range of pulmonary pathologies (mean±SD age = 44.9±21.9; 295F). The proposed framework produced a MedAE (interquartile range) of 1.01% (0.49%–2.47%) VDP compared to manually corrected, segmentation-derived VDPs; no evidence of difference was found (P = .70). 20 MC dropout iterations were completed, producing VDP prediction distributions that were subsequently clustered into confidence groupings. The proposed approach demonstrated clinical classification accuracy of 91% (95% CI: 68%, 94%; n = 32/35).

Conclusion: An uncertainty-aware, end-to-end deep learning approach enabled accurate prediction of VDP without manual segmentation, with performance comparable to segmentation-based methods and quantification of prediction uncertainty.

1. Introduction

Pulmonary imaging represents a key component in the clinical workflow for many patients with respiratory diseases; a plethora of imaging modalities are available to clinicians and researchers, which provide structural or functional information, aiding in diagnosis, monitoring, and treatment. Hyperpolarized gas MRI is a functional lung imaging modality that can provide regional ventilation information with comparatively high spatial and temporal resolution (1). In this case, Xenon-129 (^{129}Xe) is magnetically polarized and inhaled, providing a map of ventilated regions of the lungs (2). Spatially co-registered functional hyperpolarized ^{129}Xe-MRI and structural proton (^{1}H)-MRI scans facilitate the computation of metrics, such as the ventilation defect percentage (VDP), which gives a percentage of defective lung in comparison to the overall lung parenchymal volume (1). The VDP is a highly-sensitive metric for quantifying lung function that has shown improved sensitivity when compared to traditional lung function tests (3,4). To compute the VDP, the segmentations of the ventilated lung and lung cavity are required. Several approaches have been used to segment functional and structural regions required for the computation of the VDP from paired ^{129}Xe-MRI and ^{1}H-MRI (5-12). In our center, VDPs are currently calculated using the convolutional neural network (CNN)-based workflow depicted in Figure 1.

Segmentation-based approaches to derive the VDP have traditionally been used for VDP calculation; however, manual intervention is still required to compute an accurate VDP (1). This can lead to inter-observer variability between experts and constitutes a time-consuming repetitive process that can cause delays in delivering results to clinicians.

End-to-end deep learning (DL) directly predicts the final target value, frequently by skipping over intermediary steps and has previously demonstrated improved performance on computer vision tasks (13,14). Directly predicting the VDP can remove the need for manual inputs and the inter-observer variability associated with multiple observers while producing accurate VDP values. However, end-to-end DL approaches inherently lack interpretability; they function as a black-box predictor, where users can no longer ensure accuracy by interrogating the intermediary steps of the calculation. To address this and improve trust in end-to-end DL, uncertainty quantification techniques can characterize prediction variability and provide additional information beyond the black-box model. In this work, we propose an automatic, end-to-end pipeline capable

of deriving a key functional lung imaging metric, namely, the VDP, from multi-modal ^{129}Xe-MRI and ^{1}H-MRI volumetric scans without manual intervention. This is achieved via a novel, uncertainty-aware CNN-based framework, validated across a range of scanners and pulmonary pathologies, with the aim of increasing the robustness and repeatability of VDP calculations.

2. Materials and Methods

The dataset contained ^{1}H-MRI and ^{129}Xe-MRI scans from healthy participants and patients with a range of pulmonary diseases. Patient data used in this work were pooled retrospectively from clinical imaging studies and clinical referral cases. All studies received ethical approval by the national research ethics committee with participants (or their guardians) providing informed written consent (REC numbers: 16/YH/0339, 17/YH/0266, 15/SW/0289, 9/LO/1115, 15/NW/0750, 17/NW/0631, 16/EM/0439, 20/YH/0069, 17/LO/0725 and 21/SC/0398). Appropriate consent and permissions have been granted by the sponsors to utilize this data for retrospective analysis. No exclusion criteria were used; the dataset contains all paired ^{129}Xe-MRI and ^{1}H-MRI scans from our center during the period 2015–2024, where permission was granted by the sponsors. These data have, in part, been published previously (9,10,12) for image segmentation applications, but not for end-to-end VDP prediction via regression. Data were pseudo-anonymized before research development. The Checklist for Artificial Intelligence in Medical Imaging (CLAIM) (15) has been completed by the authors and is available as supplementary information.

2.1 *Imaging Protocol*

All participants underwent three-dimensional volumetric ^{129}Xe-MRI and ^{1}H-MRI acquired in the coronal plane at approximately functional residual capacity + inhaled bag volume (400 mL to 1L) with full lung coverage at 1.5T. Patients were scanned across two hospitals and on one of three scanners: an HDx, Artist, or 450w GE scanner (GE Healthcare, Milwaukee, WI, USA) (16). Full image acquisition details are provided in the supplementary material.

Reference standard VDP values are calculated from the segmentation of the ventilated lung in ^{129}Xe-MRI scans and the lung cavity estimations (LCEs) derived from the similar-breath structural ^{1}H-MRI scans. LCEs, which segment the lung cavity in the spatial domain of ^{129}Xe-MRI, were semi-automatically generated using several dual-channel approaches, including CNN- and spatial fuzzy c-means-based methods (5,7,9). LCEs were

subsequently manually corrected to remove any errors. Manual editing was conducted by a wide range of expert observers; all LCEs underwent quality control to ensure consistency by one of the following pulmonary imaging scientists: H.M (15 years of experience), B.A.T (14 years of experience), G.J.C (12 years of experience), L.J.S (10 years of experience), P.J.C.H (10 years of experience), or A.M.B (8 years of experience). ^{129}Xe-MRI ventilated lung segmentations were generated from one of three methods, namely, spatial fuzzy c-means, linear-binning (8), or CNN-based segmentation (12). For linear binning, bias field inhomogeneity was corrected before normalization was applied, consisting of histogram scaling by the mean signal inside the LCE. Histograms were split into six bins and merged into four ventilation categories: the ventilation defect region (VDR) (1st bin), low ventilation region (LVR) (2nd bin), normal ventilation region (NVR) (3rd+4th bins) and high ventilation region (HVR) (5th+6th bins) (8). The six bins were mean-centered with a width of one standard deviation. For the CNN-based approach, the ventilated volume (VV) of the lung was automatically segmented via a nnUNet (12); this can then be subtracted from the total lung volume (TLV) defined by the LCE to give the VDR. Thus, in both approaches, the VDP can be defined in terms of VDR or VV as shown in Equations 1, 2 and 3:

$$VDP^{SFCM}\ (\%) = \left(1 - \frac{VV}{TLV}\right) \times 100 \tag{1}$$

$$VDP^{CNN}\ (\%) = \left(1 - \frac{VV^{DL}}{TLV^{DL}}\right) \times 100 \tag{2}$$

$$VDP^{Binning}\ (\%) = \left(\frac{VDR}{TLV^{DL}}\right) \times 100 \tag{3}$$

Hughes *et al.* indicated that manual editing of segmentations takes between 1 and 2 h (5). Similarly, Astley *et al.* provided an estimate of approximately 1.5 h per patient (9). In our center, we estimate that during our clinical workflow, it now takes approximately 45 min to edit functional and structural lung segmentations due to the improvements gained via CNN-based and binning approaches.

2.3 *Data Split*

For each participant, paired ^{1}H-MRI and ^{129}Xe-MRI scans with corresponding VDP values were collected. A data split of 80:10:10% was used, resulting in the following allocations: training (n = 458), validation (n = 58)

and testing (n = 58). The network was trained for 500 epochs, and validation was performed every 10 epochs. The epoch with the lowest mean absolute error on validation data was used to infer testing cases. Stratified randomization was performed to select validation and testing set cases; participants were first stratified by pulmonary pathology, and then, cases from each disease group were randomly specified as either validation or testing cases. Both validation and testing sets were designed to be representative of the larger dataset in terms of participant pulmonary pathology.

2.2 End-to-end DL VDP Prediction

A dual-channel, multi-modality three-dimensional CNN was developed to predict a VDP value from paired ^{1}H-MRI and ^{129}Xe-MRI scans (17). The network used two channels, representing the structural and functional imaging modalities, which were then concatenated in a channel-wise fashion. The network utilizes dense blocks of increasing size [16, 32, 64, 128] and three residual units. Network parameters were initialized based on previous DL workflows (9,10,12), and a parameter search was conducted on the validation set, resulting in a pReLU activation (18), AdamW (19) optimization ($\alpha = 5 \times 10^{-5}$), and a smooth L1 loss ($\beta = 0.7$) function. A batch size of 1 was used for network training, and instance normalization was applied in each block. A dropout of 0.15 was present in both Dense and Transformation blocks. A network diagram with example inputs is provided in Figure 2.

All scans were subject to preprocessing before network input; scans were resized to 256×256×20 voxels, and image intensities were normalized to the range [0, 1]. During network training, data augmentation was employed to minimize overfitting and improve robustness. Affine rotation and scaling augmentations with limits -12° to 12° and -15% to 15% were employed, respectively, at a probability of 0.1. Elastic three-dimensional rotation and scaling augmentations with limits -12° to 12° and -15% to 15% were employed, respectively, at a probability of 0.5.

Network development, training, inference and evaluation were completed using the open-source medical imaging and deep learning framework monai 1.3.0 built on top of PyTorch 2.2.2. Training and inference were performed across 12 CPUs with 12GB of RAM each, resulting in a training time of approximately 12 h. The code, trained model, and example images are available on GitHub (https://github.com/POLARIS-Sheffield/Direct-VDP-Prediction).

2.4 Quantifying Uncertainty and Robustness

Model uncertainty can be represented as the combination of aleatoric and epistemic uncertainty (20), where epistemic uncertainty represents the noise in the weights and parameters of the predictive model. Monte-Carlo (MC) dropout has been proposed as a method to determine the epistemic uncertainty of DL models by generating parameter space distributions of prediction uncertainty (21); Gal *et al*. (21) have demonstrated that test-time dropout represents a Bayesian approximation of the Gaussian process. Equation 4 shows how the approximate probability distribution, $P(y|x,\Theta)$, can be determined for inputs $\mathbf{x}$, outputs $\mathbf{y}$, and a network with weights $\boldsymbol{\Theta}$ and is equal to a Normal (N) or a Gaussian approximation of the mean $f(x,\Theta)$ and variance $s^2(x,\Theta)$ of the weight parameters:

$$P(y|x,\Theta) = N\big(f(x,\Theta), s^2(x,\Theta)\big) \tag{4}$$

In practice, the distribution is constructed by conducting several forward passes through the network, with random network connections removed on each pass, resulting in an approximate predictive distribution of the target prediction (21). Based on a study by Lemay *et al*. (22), it was determined that 20 forward passes, representing 20 MC dropout iterations, were sufficient to generate an accurate predictive distribution.

Subsequently, K-means clustering was performed to classify these distributions into relative confidence classes (23). To do this, the uncertainty interval width, defined as the central 95% prediction interval across 20 MC dropout samples, was used as the primary metric. To determine the appropriate number of clusters, the inertia was calculated as the within-cluster sum of squares and plotted against K clusters, for values of K between 1 and 9. Subsequently, the "elbow" method was employed to determine the number of clusters (24). Once the appropriate number of clusters had been selected, K-means clustering was employed with random initializations for all 58 testing cases. Clustering was performed using the Python package scikit-learn 1.6.0.

To assess the proposed end-to-end framework's robustness, augmentations were applied to testing cases, simulating test-retest repeatability (22). Specifically, affine and elastic rotation and scaling augmentations of ± 12° and ± 15% were applied to all testing cases, respectively. To measure model robustness, the median absolute error (MedAE) was calculated between non-augmented and augmented end-to-end DL VDP predictions, as well as normal *vs* abnormal clinical classification accuracy.

2.5 *Statistical Analysis*

To evaluate VDP predictions, the MedAE and interquartile range were used as the primary metrics, where the MedAE is defined as the median of absolute errors and the interquartile range refers to the 25th and 75th percentiles of the distribution. Statistical analyses were conducted using Prism v9.2.0 (GraphPad Software, San Diego, CA, USA). Normality was determined via Shapiro-Wilks tests, and the appropriate parametric or non-parametric test was conducted to compare groups. Two-sided Wilcoxon signed-rank tests were used to assess significant differences between segmentation-derived and end-to-end DL VDP values and VDP predictions with and without MC dropout. In addition, Mann-Whitney tests were used to compare performance between testing and validation sets and between confident and unconfident uncertainty meta-categories. Equivalency tests were conducted using two one-sided Wilcoxon tests between segmentation-derived and end-to-end DL VDPs; a ± 1.6% VDP zone-of-indifference, based on within-visit repeatability values of the VDP described by Smith *et al.*, was used (3,25). Bland-Altman analysis was conducted to assess the agreement between segmentation-derived and end-to-end DL VDPs and to understand whether there were systematic biases present in the predictions and what range captures these differences. Spearman correlation was conducted to assess the correlation between segmentation-derived and end-to-end DL VDPs. Spearman correlations were also used to compare forced expiratory volume in 1 second percentage predicted and VDP predictions for the segmentation-derived and proposed end-to-end VDP predictions. We further assessed the Spearman correlation between the width of uncertainty and VDP value. Confusion matrix metrics were calculated for clinical classifications. A *P*-value of $< .05$ was considered statistically significant.

3. Results

3.1 *Dataset Characteristics*

The dataset comprised 574 paired ^{129}Xe-MRI and ^{1}H-MRI scans from 47 healthy participants (mean±SD age = 28.3±17.3; 28F) and 527 patients with a range of pulmonary pathologies (mean±SD age = 44.9±21.9; 295F). A summary of participant pathologies, demographic features, and VDP values for the training, validation, and testing sets is presented in Table 1. VDP (%) Median (IQR) were matched between training (2.19% (0.81-8.53)), validation (2.91% (1.12-6.72)) and testing (2.98% (1.15-6.81)) sets; A flowchart describing the dataset and data split strategy is visualized in Figure 3.

3.2 *End-to-end VDP prediction*

To ensure that the model was not under- or overtrained, MedAE (interquartile range) VDP values were calculated for the validation [1.18% (0.48% – 2.81%)] and testing [1.01% (0.49%–2.47%)] datasets, demonstrating no evidence of difference between groups (P = .70). Statistical comparison between end-to-end DL predicted VDP values and those derived from a conventional segmentation-based workflow indicated no evidence of difference (P = .37) (see Supplementary Figure S1). Furthermore, equivalency testing indicated that the confidence intervals fall entirely within the accepted tolerance of ± 1.6% (3), with observed 90% confidence intervals of -0.18% to 0.74% (confidence interval width = 0.92%).

A minimal bias (standard deviation) of –0.18% (2.77%) was observed in the end-to-end DL predicted VDP values; however, a few patients exhibited large VDP differences, resulting in limits of agreement of -5.60% to 5.24% (Figure 4a). A Spearman correlation of r = .81 (P < .0001) with 95% CIs of 0.69 to 0.88 between segmentation-derived and end-to-end DL VDPs was observed and is shown in Figure 4b.

3.3 *Uncertainty Quantification*

Using predicted VDP values derived from the 20 MC dropout iterations, approximate predictive distributions were constructed representing the epistemic uncertainty of the network. MC dropout reduced the MedAE (interquartile range) VDP from 1.15% (0.47%–2.57%) to 1.01% (0.49%–2.47%); however, this reduction was not significant (p=0.220). Based on 20 MC dropout repeats, the elbow method was employed to determine that four clusters were optimal for K-means clustering, resulting in confidence categories of very confident (n = 11/58, 19.0%), confident (n = 25/58, 43.1%), unconfident (n = 13/58, 22.4%), and very unconfident (n = 9/58, 15.5%), as shown in Supplementary Figure S2. Most cases were classed as confident or very confident (n = 36/58, 62%), with differences in the MedAE observed between confident and unconfident meta-categories (Figure 5). Individually, the MedAE (IQR) VDP of each confidence class was calculated as very confident [0.91% (0.82%–2.26%)], confident [0.62% (0.39%–1.41%)], unconfident [1.69% (0.80%–5.03%)], and very unconfident [1.67% (1.22%–4.51%)]. Figure 5a provides a visualization of distributions for each confidence class along with comparisons between confident and unconfident meta-categories. Example readouts for six participants in the dataset are shown in Figure 6. End-to-end VDP predictions with 20 MC dropout iterations are generated in an average of 80 s per participant across the testing set. Spearman correlation between the width of uncertainty and VDP value, finding a correlation of 0.61 (P < .0001) with 95% CIs of 0.42 to 0.76.

3.4 *Evaluation of Clinical Outcomes*

The upper limit of normal for the VDP is 2.16% as measured in our center (26) and, therefore, a VDP above this value represents abnormal ventilation. We assessed the ability of the proposed approach to replicate clinical outcomes compared to segmentation-based workflows. To do this, we constructed a confusion matrix, indicating the approach's ability to replicate normal *vs* abnormal classifications (Figure 7a); 32 of 35 abnormal VDPs were replicated by the end-to-end DL predictions, resulting in an accuracy of 91% (32/35; 95% CI: 68%, 94%). For the majority of wrongly-classified participants, there were small differences in the predicted values, where the two derived VDPs straddle the 2.16% threshold, as shown in Figure 7b. Furthermore, we examined the relationship between the VDP and forced expiratory volume in one second percentage predicted (FEV_1%) value for both approaches. Spearman correlations between FEV_1% and segmentation-derived VDPs (r = –.44, *P* = .005, 95% CI –.66 to –.13) or FEV_1% end-to-end DL (r = –.40, *P* = .05, 95% CI –.64 to –.10) VDP values were observed (see Supplementary Figure S3).

3.5 *Model Robustness*

Model calibration refers to the ability of a model to reject low-confidence predictions caused by variations in acquisition or patient physiology. A well-calibrated model is robust to these variations and, consequently, can produce reliable and repeatable predictions, despite domain shifts in the data (22,27). CNNs are inherently vulnerable to poorly calibrated predictions, where small interventions can cause vastly different predictions in both classification and regression tasks (28,29). Wilcoxon tests between non-augmented and augmented VDP predictions showed no evidence of differences in predicted VDP values across the applied augmentations. Robustness results are displayed in Table 2.

4. Discussion

The VDP is a highly-sensitive metric for quantifying lung function that has shown improved sensitivity when compared to traditional lung function tests; however, manual intervention in segmentation-based workflows is still required to compute an accurate VDP, leading to inter-observer variability delays in delivering results to clinicians. We propose an end-to-end pipeline utilizing an uncertainty-aware CNN-based framework, capable of predicting VDP directly without manual intervention, with the aim of increasing the robustness and

repeatability of VDP calculations. The proposed approach was validated across a range of scanners and pulmonary pathologies, producing comparable VDP predictions to segmentation-based pipelines and, additionally, uncertainty ranges which can inform confidence in the prediction.

To our knowledge, we utilized the hitherto largest dataset comprising paired ^{129}Xe- and ^{1}H-MRI scans to directly predict the ventilation defect percentage (VDP), without the requirement of image segmentation. This novel, dual-channel, end-to-end DL network exhibited no evidence of differences when compared to conventional segmentation-based workflows and a minimal bias in Bland-Altman analysis between approaches; however, on an individual basis, some discrepancies between methods persist. The proposed approach was trained and tested on a diverse set of healthy participants and patients with pulmonary pathologies, showing robustness to participant disease and demographics. Alongside the diversity of pathologies, the dataset also contains a diversity of expert observers. Scans were pooled retrospectively from a 9-year timeframe, meaning that three separate, semi-automated methods and a range of imaging scientists have been employed to determine expert VDP values. This diversity of methodology and observer acts as inherent noise in the data and provides a unique opportunity to train our model to be invariant to inter-observer variability.

Confidence groupings were established via an uncertainty-aware framework that leverages MC dropout at test-time to generate VDP distributions for each case. The utility of uncertainty quantification in DL is clearly illustrated by the larger MedAE values in unconfident groupings compared to confident groupings (P = .04). Moreover, the four cases with the largest MedAEs were classified as unconfident or very unconfident. By producing not only a predicted VDP but also a confidence value, clinicians and imaging scientists are provided with predictions that are more interpretable and imbue greater trust.

The impact of the proposed approach on clinical outcomes was assessed via normal *vs* abnormal VDP classifications. End-to-end predicted VDPs showed good agreement with current segmentation-based methods; however, for 11 patients, their classification was altered using the proposed method. For the majority of these cases, the error in the VDP was less than the repeatability of the VDP (2.0%) (30); the two cases with the largest VDP difference were classed as very unconfident and unconfident. End-to-end DL-based VDP prediction has several advantages for clinical care, including reducing inter-observer variability inherent within segmentation and facilitating greater consistency and repeatability between visits and/or centers (31). The proposed end-to-end DL approach showed high levels of robustness to elastic and inelastic augmentations,

with no evidence of differences found in prediction accuracy observed between augmented and non-augmented scans (22). Moreover, an end-to-end approach to VDP calculation has the potential to drastically reduce the time required to compute the VDP; however, this was not directly assessed in this work.

Our study had important limitations to consider. A fundamental limitation of predicting the VDP via an end-to-end approach, is that the spatial location and size of defects are not identified as part of the calculation, and thus, this information cannot be used in later analyses or for quality control. A move toward an end-to-end approach, where uncertainty estimations could triage the need for segmentation editing, may be beneficial; for example, in cases where the VDP predicted is low and uncertainty is low, segmentation may not be required. However, in a case with high uncertainty and a large VDP, a segmentation-based workflow could be employed to resolve spatial defects specifically. A further limitation relates to the imbalance in pulmonary pathologies within the dataset; due to the small number of participants with certain diseases, such as lung cancer, and the overrepresentation of certain diseases, such as asthma, further validation is required to ensure suitable application to these pathologies. One limitation of the uncertainty quantification technique deployed is that some proportion of the quantified uncertainty is potentially due to differences in the raw VDP value; for example, patients with large VDPs have larger uncertainty. This indicates that the quantified uncertainty can, at least in part, be explained by the raw VDP value; however, only a moderate correlation was observed, indicating that the uncertainty distribution also captures real network uncertainty, regardless of the VDP value. Furthermore, it should also be noted that the distinction between 'very confident' and 'confident' classification groups does not translate to increased accuracy.

In conclusion, the proposed end-to-end deep learning approach facilitates a more streamlined ventilation analysis workflow, using uncertainty-based confidence groupings to stratify patients requiring manual intervention and increase trust in outcomes. The direct prediction of the VDP facilitates the rapid computation of VDPs and can potentially provide clinicians with important clinical data faster than segmentation-based methods, enabling higher clinical throughput. Future work should assess model robustness across a broader range of augmentations, including synthesized artifacts and noise, and evaluate generalizability using external testing datasets from independent centers. In addition, a similar supervised end-to-end DL approach could be used to predict the ventilation heterogeneity index and other regional functional lung metrics derived from ^{129}Xe-MRI, providing a more comprehensive assessment of lung function beyond VDP. Furthermore, a

prospective analysis of time savings between segmentation-based and end-to-end VDP predictions during our clinical workflow could be undertaken in future research.

References

1. Stewart NJ, Smith LJ, Chan HF, et al. Lung MRI with hyperpolarised gases: current & future clinical perspectives. *Br J Radiol*. June 9 2021:20210207. doi:10.1259/bjr.20210207. Published June 10 2021. Accessed April 28 2026.
2. Fain SB, Korosec FR, Holmes JH, O'Halloran R, Sorkness RL, Grist TM. Functional lung imaging using hyperpolarized gas MRI. *J Magn Reson Imaging*. April 6 2007;25(5):910-23. doi:10.1002/jmri.20876. Published April 6 2007. Accessed April 28 2026.
3. Smith LJ, Horsley A, Bray J, et al. The assessment of short and long term changes in lung function in CF using (129)Xe MRI. *Eur Respir J*. July 6 2020;doi:10.1183/13993003.00441-2020. Published July 8 2020. Accessed April 28 2026.
4. Marshall H, Horsley A, Taylor CJ, et al. Detection of early subclinical lung disease in children with cystic fibrosis by lung ventilation imaging with hyperpolarised gas MRI. *Thorax*. March 6 2017;72(8):760. doi:10.1136/thoraxjnl-2016-208948. Published March 6 2017. Accessed April 28 2026.
5. Hughes PJC, Horn FC, Collier GJ, Biancardi A, Marshall H, Wild JM. Spatial fuzzy c-means thresholding for semiautomated calculation of percentage lung ventilated volume from hyperpolarized gas and (1) H MRI. *J Magn Reson Imaging*. July 7 2017;47(3):640-646. doi:10.1002/jmri.25804. Published July 7 2017. Accessed April 28 2026.
6. Kirby M, Svenningsen S, Owrangi A, et al. Hyperpolarized 3He and 129Xe MR Imaging in Healthy Volunteers and Patients with Chronic Obstructive Pulmonary Disease. *Radiology*. November 1 2012;265(2):600-610. doi:10.1148/radiol.12120485. Published November 1 2012. Accessed April 28 2022.
7. Biancardi A, Acunzo L, Marshall H, et al. A paired approach to the segmentation of proton and hyperpolarized gas MR images of the lungs. 2018:Published June 16-21 2018. Accessed April 28 2026.
8. Collier GJ, Acunzo L, Smith LJ, et al. Linear binning maps for image analysis of pulmonary ventilation with hyperpolarized gas MRI: transferability and clinical applications. 2018:Published June 16-21 2018. Accessed April 28 2026.
9. Astley JR, Biancardi AM, Marshall H, et al. A Dual-Channel Deep Learning Approach for Lung Cavity Estimation From Hyperpolarized Gas and Proton MRI. *J Magn Reson Imaging*. Jun;57(6):1878-1890. doi:10.1002/jmri.28519. Published November 15 2022. Accessed April 28 2026.
10. Astley JR, Biancardi AM, Hughes PJC, et al. 3D Deep Convolutional Neural Network-Based Ventilated Lung Segmentation Using Multi-nuclear Hyperpolarized Gas MRI. Springer International Publishing; 2020:24-35. Published November 4 2020. Accessed April 28 2026.
11. He M, Driehuys B, Que LG, Huang YT. Using Hyperpolarized (129)Xe MRI to Quantify the Pulmonary Ventilation Distribution. *Acad Radiol*. September 13 2016;23(12):1521-1531. doi:10.1016/j.acra.2016.07.014. Published September 13 2016. Accessed April 28 2026.
12. Astley JR, Biancardi AM, Hughes PJC, et al. Large-scale investigation of deep learning approaches for ventilated lung segmentation using multi-nuclear hyperpolarized gas MRI. *Scientific Reports*. June 22 2022;12(1):10566. doi:10.1038/s41598-022-14672-2. Published June 22 2022. Accessed April 28 2026.
13. Oh JH, Kim HG, Lee KM, et al. Effective End-to-End Deep Learning Process in Medical Imaging Using Independent Task Learning: Application for Diagnosis of Maxillary Sinusitis. *Yonsei Med J*. November 25 2021;62(12):1125-1135. doi:10.3349/ymj.2021.62.12.1125. Published November 25 2021. Accessed April 28 2026.
14. Lee Mj, Ha Yg. Autonomous Driving Control Using End-to-End Deep Learning. 2020:470-473. Published Feburary 19-22 2020. Accessed April 28 2026.
15. Tejani AS, Klontzas ME, Gatti AA, et al. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 Update. *Radiology: Artificial Intelligence*. July 1 2024;6(4):e240300. doi:10.1148/ryai.240300. Published Accessed July 22 2025.
16. Stewart NJ, Chan HF, Hughes PJC, et al. Comparison of (3) He and (129) Xe MRI for evaluation of lung microstructure and ventilation at 1.5T. *J Magn Reson Imaging*. March 5 2018;48(3):632-42. doi:10.1002/jmri.25992. Published March 6 2018. Accessed April 28 2026.
17. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. 2017:4700-4708. Published August 27 2017. Accessed April 28 2026.
18. He K, Zhang X, Ren S, Sun J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. *IEEE International Conference on Computer Vision (ICCV 2015)*. February 6 2015;1502doi:10.1109/ICCV.2015.123. Published February 6 2015. Accessed April 28 2026.
19. Loshchilov I. Decoupled weight decay regularization. *arXiv preprint* doi:https://doi.org/10.48550/arXiv.1711.05101. Published November 24 2017. Accessed April 28 2026.

20. Hüllermeier E, Waegeman W. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning*. March 1 2021;110(3):457-506. doi:10.1007/s10994-021-05946-3. Published March 1 2021. Accessed April 28 2026.
21. Gal Y, Ghahramani Z. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. presented at: Proceedings of The 33rd International Conference on Machine Learning; February 8 2023 2016; Proceedings of Machine Learning Research. Accessed April 28 2026. https://proceedings.mlr.press/v48/gal16.html
22. Lemay A, Hoebel K, Bridge CP, et al. Improving the repeatability of deep learning models with Monte Carlo dropout. *NPJ Digit Med*. November 18 2022;5(1):174. doi:10.1038/s41746-022-00709-3. Published November 18 2022. Accessed April 28 2026.
23. Arthur D, Vassilvitskii S. k-means++: the advantages of careful seeding. presented at: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms; January 7 2007 2007; New Orleans, Louisiana. Accessed April 28 2026.
24. Marutho D, Handaka SH, Wijaya E, Muljono. The Determination of Cluster Number at k-Mean Using Elbow Method and Purity Evaluation on Headline News. 2018:533-538. Published September 21-22 2018. Accessed April 28 2026.
25. Lakens D. Equivalence Tests: A Practical Primer for t Tests, Correlations, and Meta-Analyses. *Soc Psychol Personal Sci*. June 25 2017;8(4):355-362. doi:10.1177/1948550617697177. Published June 25 2017. Accessed April 28 2026.
26. Collier GJ, Biancardi AM, Chan HF, et al. Dependence of diffusion and ventilation metrics on hyperpolarized (129)Xe lung MRI with demographics in healthy volunteers. 2025:Published Accessed
27. Kuleshov V, Fenner N, Ermon S. Accurate Uncertainties for Deep Learning Using Calibrated Regression. presented at: Proceedings of the 35th International Conference on Machine Learning; July 03 2018 2018; Proceedings of Machine Learning Research. Accessed April 28 2026. https://proceedings.mlr.press/v80/kuleshov18a.html
28. Guo C, Pleiss G, Sun Y, Weinberger KQ. On Calibration of Modern Neural Networks. presented at: Proceedings of the 34th International Conference on Machine Learning; June 14 2017 2017; Proceedings of Machine Learning Research. Accessed April 28 2026. https://proceedings.mlr.press/v70/guo17a.html
29. Laves M-H, Ihler S, Fast JF, Kahrs LA, Ortmaier T. Well-Calibrated Regression Uncertainty in Medical Imaging with Deep Learning. presented at: Proceedings of the Third Conference on Medical Imaging with Deep Learning; Februrary 8 2023 2020; Proceedings of Machine Learning Research. Accessed April 28 2026. https://proceedings.mlr.press/v121/laves20a.html
30. Marshall H, Kenworthy JC, Horn FC, et al. Peripheral and proximal lung ventilation in asthma: Short-term variation and response to bronchodilator inhalation. *J Allergy Clin Immunol*. Jun;147(6):2154-2161.e6. doi:10.1016/j.jaci.2020.11.035. Published 2020/12/15. Accessed
31. Mukesh M, Benson R, Jena R, et al. Interobserver variation in clinical target volume and organs at risk segmentation in post-parotidectomy radiotherapy: can segmentation protocols help? *Br J Radiol*. June 21 2012;85(1016):e530-6. doi:10.1259/bjr/66693547. Published June 21 2012. Accessed April 28 2026.

Table 1. Summary of participant demographics for the training, validation and testing sets

Data Split	Disease	Number of participants	Age (years)* Median (range)	Sex (F:M) Frequency (%)	VDP (%) Median (IQR)
Training set	Asthma	110	50 (11, 78)	66F (60):44M (40)	2.02 (1.15-4.91)
	Asthma + COPD	36	62 (40, 82)	15F (42):21M (58)	8.97 (5.63-16.31)
	COPD	28	69 (42, 81)	17F (61):11M (39)	19.13 (8.28-24.19)
	Cystic fibrosis	59	15 (6, 62)	35F (59):24M (41)	3.95 (1.16-20.66)
	Healthy	34	28 (9, 71)	18F (53):16M (47)	0.43 (0.11-0.67)
	Interstitial lung disease	32	68 (39, 83)	14F (44):18M (56)	8.00 (5.95-11.07)
	Investigation for possible airways disease	25	49 (11, 74)	19F (76):6M (24)	7.68 (1.31-15.89)
	Lung cancer	9	72 (63, 83)	6F (67):3M (33)	21.47 (18.78-25.23)
	Other**	11	14 (8, 29)	9F (82):2M (18)	1.49 (0.90-7.55)
	Post covid-19	80	49 (19, 79)	37F (46):43M (54)	1.10 (0.66-1.71)
	Preterm birth	34	12 (9, 14)	20F (59):14M (41)	0.44 (0.19-0.88)
	Total training set	458	47 (6, 83)	256F (56):202M (44)	2.19 (0.81-8.53)
Validation set	Asthma	19	50 (31, 73)	8F (42):11M (58)	3.29 (1.87-5.53)
	Asthma + COPD	3	67 (59, 73)	1F (33):2M (67)	12.75 (12.58-18.32)
	Cystic fibrosis	7	14 (5, 35)	5F (71):2M (29)	5.29 (3.32-13.50)
	Healthy	6	35 (11, 60)	5F (83):1M (17)	0.22 (0.16-0.83)
	Interstitial lung disease	4	67 (39, 83)	18F (45):22M (55)	5.49 (4.08-7.07)
	Investigation for possible airways disease	2	52 (46, 58)	2F (100):0M (0)	10.12 (6.31-13.94)
	Lung cancer	1	76 (76, 76)	0F (0):1M (100)	39.34
	Other**	2	22 (10, 34)	2F (100):0M (0)	11.81 (11.42-12.19)
	Post covid-19	9	51 (30, 77)	5F (56):4M (44)	1.41 (0.58-1.87)
	Preterm birth	5	11 (10, 13)	4F (80):1M (20)	0.69 (0.58-0.97)
	Total validation set	58	47 (5, 77)	34F (59):24M (41)	2.91 (1.12-6.72)
Testing set	Asthma	13	58 (25, 79)	7F (54):6M (46)	3.09 (2.02-4.76)
	Asthma + COPD	7	58 (37, 72)	4F (57):3M (43)	3.82 (2.79-5.77)
	COPD	2	62 (61, 64)	2F (100):0M (0)	18.52 (13.54-23.51)
	Cystic fibrosis	6	14 (6, 29)	3F (50):3M (50)	9.94 (4.35-17.05)
	Healthy	7	34 (11, 51)	6F (86):1M (14)	0.79 (0.60-1.19)
	Interstitial lung disease	4	67 (57, 83)	2F (50):2M (50)	12.56 (9.83-14.53)
	Investigation for possible airways disease	4	55 (11, 57)	2F (50):2M (50)	4.40 (2.16-7.12)
	Other**	2	32 (16, 48)	1F (50):1M (50)	2.99 (2.76-3.21)
	Post covid-19	9	50 (36, 69)	5F (56):4M (44)	1.10 (0.48-3.02)
	Preterm birth	4	12 (11, 13)	2F (50):2M (50)	0.68 (0.57-1.43)
	Total testing set	58	50 (6, 83)	34F (59):24M (41)	2.98 (1.15-6.81)
Total		574	47 (5, 83)	323F (56):251M (44)	2.36 (0.86-8.16)

*The age of the patient at scan time was unavailable for seven patients.
**'Other' category includes diseases that affected < 10 participants in the whole dataset.

Abbreviations: F = female, M = male, VDP = ventilation defect percentage, IQR = interquartile range, COPD = chronic obstructive pulmonary disease

Table 2. Summary of model robustness investigations indicating the MedAE (IQR) and accuracy percentages for four different augmentation techniques on all testing set cases.

Deformation	**MedAE (IQR) VDP%**	**Normal *vs* Abnormal classification Accuracy (n, (%))**
None	-	-
Affine scale	0.12 (0.05-0.17) *(ns)*	56/58 (96.6)
Elastic scale	0.12 (0.05-0.22) *(ns)*	57/58 (98.3)
Affine rotate	0.06 (0.02-0.12) *(ns)*	55/58 (94.8)
Elastic rotate	0.08 (0.04-0.16) *(ns)*	56/58 (96.6)

Abbreviations: MedAE = Median absolute error, VDP = ventilation defect percentage, *ns* = not significant, IQR = interquartile range

Figure Legends

Figure 1: Segmentation-based workflow currently in use to compute ventilation defect percentages (VDPs). N4 bias correction is applied to both ^{129}Xe-MRI and ^{1}H-MRI scans. Manual steps are denoted by green boxes, automated steps by blue boxes and deep learning aided steps with a red border.

Figure 2: Network architecture diagram showing the paired processing of ^{1}H-MRI and ^{129}Xe-MRI scans to predict ventilation defect percentage (VDP) in an end-to-end framework.

Figure 3: Study design flowchart. ventilation defect percentage = VDP, mean absolute error = MAE, Monte-Carlo = MC, interquartile range = IQR.

Figure 4: a) Bland-Altman analysis of segmentation-generated and end-to-end deep learning (DL) ventilation defect percentage (VDP) values. Horizontal uneven dashed lines at 1.6% and -1.6% denote zone-of-indifference. b) Spearman correlation and 95% confidence intervals of segmentation-generated and end-to-end DL VDP values. Horizontal and vertical uneven red dashed lines at 2.16% denote upper limit of normal for normal vs abnormal clinical classification.

Figure 5: a) Example probability distributions for one case in each confidence grouping based on 20 Monte-Carlo dropout repeats. UIW = Uncertainty Interval Width. b) Comparison of confident and unconfident meta-categories via Mann-Whitney test, indicating that the median absolute error (MedAE) is significantly larger in the unconfident grouping (P = .04). Red=very unconfident, orange=unconfident, blue=confident, green=very confident.

Figure 6: Six example readouts with structural and functional MRI scans, segmentation-derived and end-to-end deep learning ventilation defect percentage (VDP) values, uncertainty interval widths and confidence classes provided. Patient demographics: (top left) 49-year-old male patient; (top right) 57-year-old female patient; (middle left) 67-year-old female patient; (middle right) 66-year-old male patient; (bottom left) 8-year-old female patient; (bottom right) 71-year-old female patient.

Figure 7: a) Confusion matrix of normal vs abnormal clinical classification. b) Comparison of segmentation-derived and end-to-end deep learning (DL) ventilation defect percentage (VDP) values for wrongly classified cases. Red=very unconfident, orange=unconfident, blue=confident, green=very confident.