Scientific Data

https://doi.org/10.1038/s41597-026-07006-8

Article in Press

BPD-Neo: An MRI Dataset for Lung-Trachea Segmentation with Clinical Data for Neonatal Bronchopulmonary Dysplasia

Received: 23 July 2025

Accepted: 27 February 2026

Cite this article as: Saluja, R., Kovanlikaya, A., Chien, C. *et al.* BPD-Neo: An MRI Dataset for Lung-Trachea Segmentation with Clinical Data for Neonatal Bronchopulmonary Dysplasia. *Sci Data* (2026). https://doi.org/10.1038/s41597-026-07006-8

Rachit Saluja, Arzu Kovanlikaya, Candace Chien, Lauren Kathryn Blatt, Jeffrey M. Perlman, Stefan Worgall, Mert R. Sabuncu & Jonathan P. Dyke

SCIENTIFIC DATA

Tracking no: SDATA-25-04026B

BPD-Neo: MRI Dataset for Lung-Trachea Segmentation with Clinical Data for Neonatal Bronchopulmonary Dysplasia

Authors: Rachit Saluja (Cornell University, Cornell Tech and Weill Cornell Medicine), Arzu Kovanlikaya (Weill Cornell Medicine), Candace Chien (Weill Cornell Medicine), Lauren Blatt (Weill Cornell Medicine), Jeffrey Perlman (Weill Cornell Medicine), Stefan Worgall, Mert Sabuncu, and Jonathan Dyke (Weill Cornell Medicine)

Abstract:
Bronchopulmonary dysplasia (BPD) is a common complication among preterm neonates, with portable X-ray imaging serving as the standard diagnostic modality in neonatal intensive care units (NICUs). However, lung magnetic resonance imaging (MRI) offers a non-invasive alternative that avoids sedation and radiation while providing detailed insights into the underlying mechanisms of BPD. Leveraging high-resolution 3D MRI data, advanced image processing and semantic segmentation algorithms can be developed to assist clinicians in identifying the etiology of BPD. In this dataset, we present MRI scans paired with corresponding semantic segmentations of the lungs and trachea for 40 neonates, the majority of whom are diagnosed with BPD. The imaging data consist of free-breathing 3D stack-of-stars radial gradient echo acquisitions, known as the StarVIBE series. Additionally, we provide comprehensive clinical data and baseline segmentation models, validated against clinical assessments, to support further research and development in neonatal lung imaging.

Datasets:

Repository Name	Dataset Title	Accession Number or DOI	URL to data record	Private reviewer access URL/code
BPD-Neo: An MRI Dataset for Lung-Trachea Segmentation with Clinical Data for Neonatal Bronchopulmonary Dysplasia	BPD-Neo	10.5281/zenodo.15768091	https://zenodo.org/records/15768091	

BPD-Neo: An MRI Dataset for Lung-Trachea Segmentation with Clinical Data for Neonatal Bronchopulmonary Dysplasia

Rachit Saluja *rs2492@cornell.edu*
Cornell University & Cornell Tech
Weill Cornell Medicine

Arzu Kovanlikaya
Weill Cornell Medicine

Candace Chien
Weill Cornell Medicine

Lauren Kathryn Blatt
Weill Cornell Medicine

Jeffrey M. Perlman
Weill Cornell Medicine

Stefan Worgall
Weill Cornell Medicine

Mert R. Sabuncu*
Cornell University & Cornell Tech
Weill Cornell Medicine

Jonathan P. Dyke* † *jpd2001@med.cornell.edu*
Weill Cornell Medicine

Abstract

Bronchopulmonary dysplasia (BPD) is a common complication among preterm neonates, with portable X-ray imaging serving as the standard diagnostic modality in neonatal intensive care units (NICUs). However, lung magnetic resonance imaging (MRI) offers a non-invasive alternative that avoids sedation and radiation while providing detailed insights into the underlying mechanisms of BPD. Leveraging high-resolution 3D MRI data, advanced image processing and semantic segmentation algorithms can be developed to assist clinicians in identifying the etiology of BPD. In this dataset, we present MRI scans paired with corresponding semantic segmentations of the lungs and trachea for 40 neonates, the majority of whom are diagnosed with BPD. The imaging data consist of free-breathing 3D stack-of-stars radial gradient echo acquisitions, known as the StarVIBE series. Additionally, we provide comprehensive clinical data and baseline segmentation models, validated against clinical assessments, to support further research and development in neonatal lung imaging.

* Contributed equally as senior co-authors.
† Corresponding Author.

Background & Summary

Automated segmentation of the neonatal respiratory system is particularly relevant for preterm newborns at risk of developing bronchopulmonary dysplasia (BPD). The etiology of BPD manifests through multiple mechanisms, including tracheobronchomalacia, parenchymal lung disease, pulmonary hypertension, or a combination of these factors (1). While portable X-ray remains the standard imaging modality in neonatal intensive care units (NICUs), it is inherently limited by its 2D planar nature and inability to provide 3D volumetric information on lung and tracheal structures. Although CT imaging can offer such volumetric detail, it is not standard-of-care in neonates due to the higher radiation exposure compared to plain film X-rays. In contrast, lung MRI offers a non-invasive alternative that provides detailed insights into these pathological mechanisms without the need for sedation or ionizing radiation (2; 3). MRI is classified as minimal risk by institutional review boards (IRBs) and additionally offers superior soft tissue contrast and 3D volumetric capabilities. Automated segmentation of the neonatal trachea and lung volume could enhance clinicians' ability to identify the underlying causes of BPD, facilitating improved diagnosis and management.

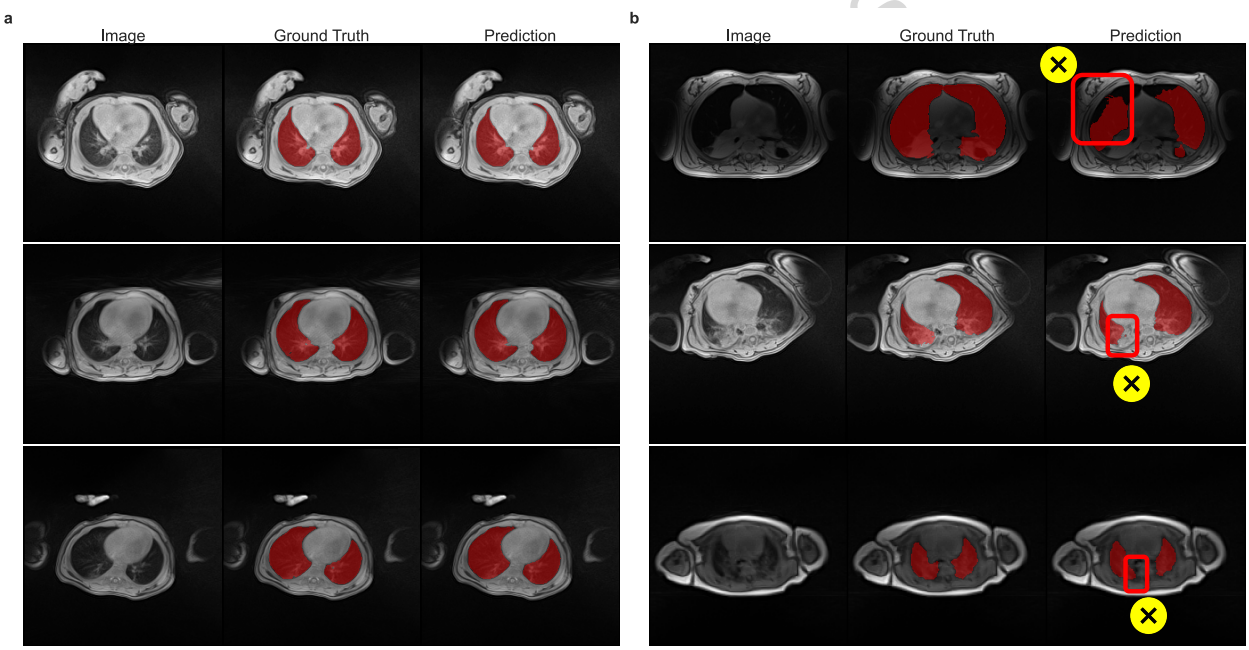


Figure 1: Examples of T1-weighted StarVIBE images (axial view) and their corresponding ground truth and predicted segmentations of lungs (a) illustrates outputs for high-performing cases, (b) highlights cases with lower performance, including instances of both over-segmentation and under-segmentation

The compliant airway in preterm infants is particularly susceptible to injury due to prolonged endotracheal intubation and exposure to positive pressure ventilation. Airway malacia refers to an abnormally compliant airway, leading to excessive collapse (>50% reduction in airway area). Notably, large airway disease is observed in approximately one-third of infants diagnosed with BPD, yet it often remains undiagnosed (4). The current gold-standard diagnostic method, bronchoscopy, requires sedation and carries inherent procedural risks. Tracheobronchomalacia is further associated with increased morbidity, including prolonged hospital stays, a higher incidence of pneumonia, and an increased likelihood of requiring tracheostomy placement. The use of MRI to quantify tracheal airway area presents a promising non-invasive alternative, enabling the automatic segmentation of the airway through deep learning models for more accessible and risk-free diagnosis. Historically, parenchymal lung disease has been the defining feature of BPD. Parenchymal lung disease as a cause in BPD is made evident in both structural and functional lung MRI (2; 5). Preterm birth results in the arrest of lung development resulting in ineffective gas exchange and need for respiratory support and oxygen. Parenchymal injury is complex, resulting from multiple antenatal and postnatal exposures which

further disrupt alveolarization and lead to abnormal repair. Since the introduction of antenatal steroids and surfactant, BPD is mostly characterized by a large simplified alveolar structure. More severe BPD patients have heterogeneous parenchymal disease characterized by atelectasis, hyperinflation, edema, and fibrosis. MRI may be used to quantitate the degree of parenchymal disease in the lung.

The application of semantic segmentation models for quantifying lung and tracheal volumes in MRI images remains uncommon. Recently, a study developed a model to quantify lung volumes in BPD patients; however, while that study had a larger sample size, the dataset was not publicly available, and no clinical data were included (6). Additionally, their analysis did not utilize StarVIBE MRI series, which we believe may offer greater utility for this application. (6) also utilizes the BPD grading system proposed by (7). In contrast, our study adopts the more contemporary 2019 Jensen criteria ((8)), which are also used in current clinical practice and are expected to provide a more accurate and clinically relevant categorization of BPD severity.

Our dataset provides clinicians and researchers with the resources to develop semantic segmentation models capable of automatically segmenting lung and tracheal volumes. The binary image masks produced by the semantic segmentation models may be multiplied by a structural UTE or StarVIBE MRI sequence to produce a parenchymal signal intensity histogram (5; 9). The degree of hyperinflation (ratio of total-lung-volume [TLV] to body-surface-area [BSA]) may also be measured using the Mosteller formula and the derived lung volume. Automatically segmenting the neonatal respiratory system using deep learning-based semantic segmentation methods offers a rapid and objective approach for clinicians to assess tracheal and lung health in relation to the various etiologies of BPD. Our dataset is the first open-source resource to provide paired imaging and segmentation data, facilitating the development of advanced computational models for neonatal respiratory assessment.

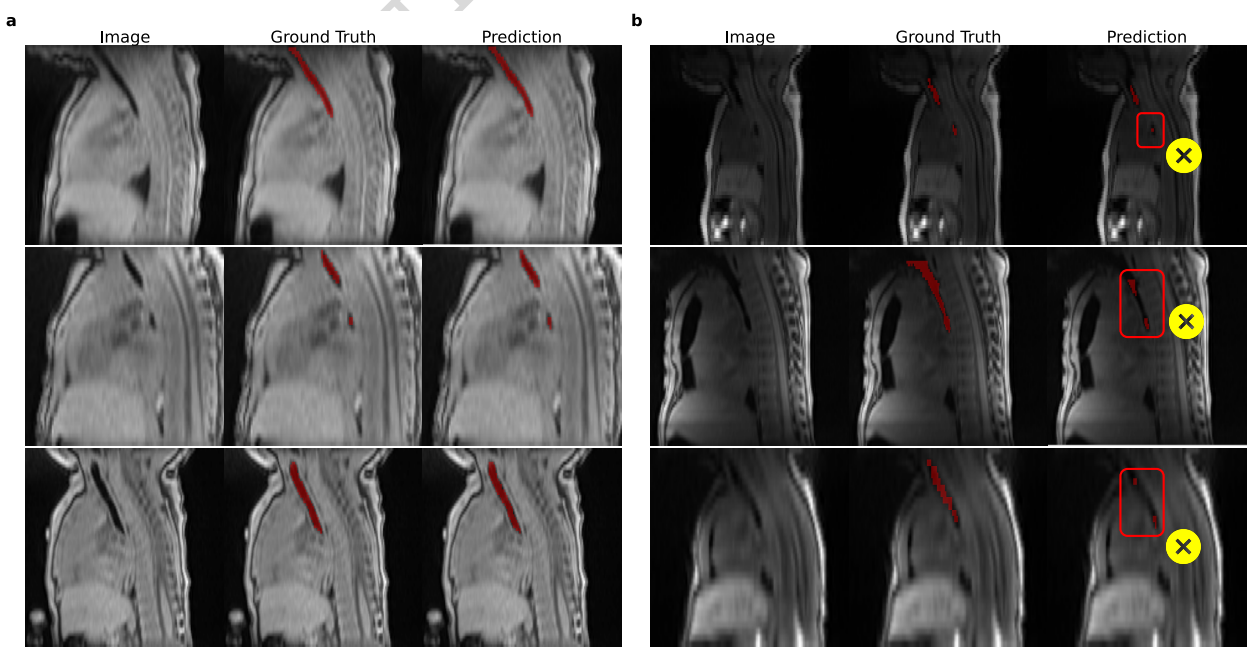


Figure 2: Examples of T1-weighted StarVIBE images (sagittal view) and their corresponding ground truth and predicted segmentations of trachea (a) illustrates outputs for high-performing cases, (b) highlights cases with lower performance, including instances of both over-segmentation and under-segmentation

Methods

Data Collection

Neonatal lung volume was assessed from 40 neonates (18 M/22 F) enrolled in a prospective study assessing physiologic phenotyping of chronic lung disease of prematurity using MRI (NHLBI; R01-HL167003). All parents signed informed consent as part of an approved protocol of our Institutional Review Board (IRB protocol number 20-02022616 (Expiration July 26, 2026)). Participants were recruited by co-investigating pediatric neonatologists on the IRB protocol in the neonatal clinic. Inclusion criteria were infants less than 6 month of age in the NICU already receiving a clinically indicated brain MRI as part of standard of care. An additional 15 minutes was added at the completion of the brain MRI to perform the research lung imaging sequences without the administration of any contrast agents.

All participants provided consent to share fully de-identified data in a research repository or database. This consent was documented through a checkbox option allowing individuals to authorize the institution to retain their protected health information for research purposes. Participants were informed of their right to withdraw this authorization at any time and were provided with a complete copy of the consent form. Consent was obtained by physicians listed on the study protocol who approached the parents of infants already scheduled for a clinical brain MRI. To minimize potential coercion, these physicians were not the primary care providers for the enrolled subjects. Recruitment was conducted in a private setting within the NICU at our institution. The research study and consent form were explained in a quiet environment, allowing parents ample opportunity to ask questions and discuss the study with the physician prior to providing consent.

The MRI data were fully de-identified by the Imaging Data Evaluation & Analysis Laboratory (IDEAL) at our institution. A study-specific identifier was then assigned to each subject's data prior to making it available for investigator access and download. The IDEAL lab is the institutionally approved facility for clinical trial data storage and de-identification, ensuring compliance with privacy standards and protection of patient confidentiality.

MRI Data Acquisition

MRI data was acquired on a 1.5 Tesla Siemens Amira Scanner (Siemens Healthineers; Erlangen, Germany) located in the Neonatal Intensive Care Unit of the New York-Presbyterian Alexandra Cohen Hospital for Women and Newborns. While the availability of MRI within the NICU is a unique feature of our institution which enhances accessibility and comfort for both the infant and family, similar scans can feasibly be conducted outside the NICU, potentially increasing their clinical applicability and broader availability. Infants were fed, swaddled and transported within the NICU to the MRI scanner. Multiple layers of hearing protection were employed to minimize the acoustic noise reaching the infant. Earplugs were used which in general can reduce the noise in the MRI by between 20 and 30 decibels (dB) and are always the first line of defense. In addition to standard foam earplugs, we used MRI safe disposable MiniMuff neonatal noise guards (Natus Medical, San Carlos, CA, USA) which have a gentle hydrogel adhesive to provide a secure fit and to reduce the noise by an additional 7dB. Noise canceling infant MRI headphones were lastly used (Ima-X; Luxembourg) which reduced the noise by an average of 22 dB with up to 30 dB @ 1kHz. During the MRI study, the infant's vitals were monitored continuously by a neonatal nurse using a Philips MR400 patient monitoring system with infant accessories. The infant was also audibly monitored for any signs of distress and the scan immediately stopped should the infant experience any discomfort, and the scan not restarted until they were calmed.

A pair of 8-channel NORAS VARIETY flex coils (20 cm x 22 cm) (NORAS MRI products, Höchberg, Germany), were used to provide one coil anterior and one coil posterior on the infant. A free-breathing 3D stack-of-stars radial gradient echo technique known as StarVIBE was acquired axially for segmentation of both lung and trachea in the neonates (10; 11). StarVIBE is optimally used in pediatric patients and is robust in resisting motion artifacts. Specific acquisition parameters included a 20 cm field of view (FOV) and a 224 x 224 matrix size yielding a 0.9 mm x 0.9 mm x 2 mm (1.6 ml) voxel resolution. A repetition time

Clinical Data	Description
BW (grams)	Weight of premature infant at the time of birth in grams
Length (cm)	Length of premature infant at birth in cms
Sex	Sex of premature infant
Weight (grams)	Weight of premature infant at MRI scan in grams
GA (weeks)	Gestational age in weeks
PMA at Study (weeks)	Postmenstrual age in weeks
Jensen 2019 BPD Definition	Premature infant's Jensen 2019 BPD Classification

Table 1: Full list of clinical data and their description

(TR) of 4.2 ms, an echo time (TE) of 2.0 ms, a flip angle of 4° and a receive bandwidth of 603 Hz/pixel were used.

This acquisition can be replicated at other institutions using different MRI platforms, as the radial gradient echo sequence is a standard clinical MRI technique available across vendors. For example, General Electric (GE) offers a comparable fast radial GRE sequence under the name LAVA or LAVA-STAR, while Philips provides a similar implementation known as THRIVE.

Expert Image Annotation

Segmentation of the lungs and trachea was conducted using 3D Slicer `http://slicer.org` and its segmentation editor (12; 13). Fiducial markers (seed points) were manually placed at the center of each lung and tracheal slice, followed by application of a region-growing algorithm to delineate the structures. A smoothing kernel of 3 mm × 3 mm × 1 mm was applied to the lung segmentations to refine boundaries, whereas no smoothing was applied to the tracheal segmentations due to the limited voxel count in those regions of interest. Final segmentations were reviewed, manually corrected for any errors, and validated by a domain expert before being exported as NIfTI files. Additionally, we provide a small subset of expert segmentations from a second reviewer to facilitate inter-observer variability analysis, as detailed in Section . This reviewer followed the exact same segmentation methodology to ensure consistency in the annotation process.

Clinical Data

In addition to the imaging data, key clinical variables were collected, including birth weight, weight at the time of MRI, gestational age, and postmenstrual age, along with BPD classification based on the Jensen 2019 criteria. The availability of these clinical data alongside imaging data facilitates future research into the identification of biomarkers derived from segmentation-based volumetrics, enabling their integration with clinical variables to enhance understanding of disease progression and outcomes. A comprehensive list of the clinical variables collected and included in the dataset is presented in Table 1.

Data Records

All data records, including the DICOM series, NIfTI files, and clinical data, are available at `https://zenodo.org/records/15768091`, under the CC BY 4.0 license (14). The dataset is accompanied by two MD5 checksums for integrity verification: [e3f1c0b8a9b0ccd8f60190d935ceb715, fe682ecccd363801fe63db76831501a5]. These can be used to ensure the completeness and authenticity of the downloaded files. The dataset includes a XLSX file containing the clinical data, matched to each study by study identifier. The DICOM data comprise all imaging series acquired during the MRI sessions, have been fully anonymized, and are suitable for future research applications. Additionally, the dataset contains NIfTI files for the lung and trachea segmentations corresponding to each study along with some of the multi-rater segmentations. Below is the directory structure of the data record:

```
.
```

```
|-- clinical_data.xlsx                          ## Clinical Data
|-- DICOM-data/                                 ## Dicom Data
|   |-- BPD-Neo-01/                             ## Study
|   |   |-- SER0001/                            ## Dicom Series
|   |   |   |-- IMG00001.dcm
|   |   |   |-- IMG00002.dcm
|   |   |   |-- ...
|   |   |   `-- IMG00018.dcm
|   |   |-- SER0002/
|   |   |-- ...
|   |   `-- SER0007/
|   |-- BPD-Neo-02/
|   |-- ...
|   `-- BPD-Neo-40/
`-- Nifti-data/
    |-- BPD-Neo-01/                             ## Nifti Data
    |   |-- image.nii.gz                        ## Image
    |   |-- lung_seg.nii.gz                     ## Lung Segmentation
    |   `-- trachea_seg.nii.gz                  ## Trachea Segmentation
    |-- ...
    `-- BPD-Neo-40/
```

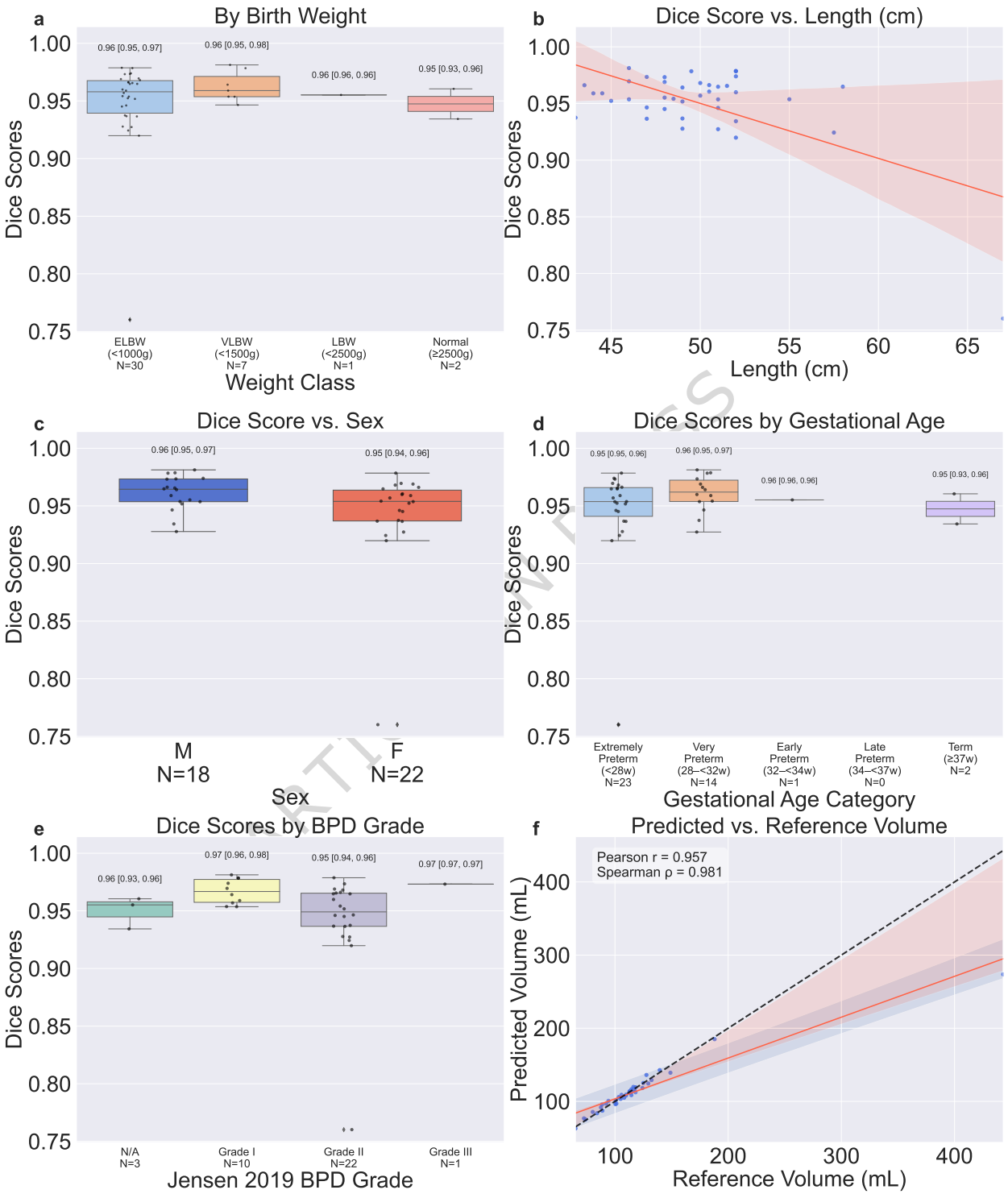


Figure 3: Evaluation of lung segmentation model against clinical variables.

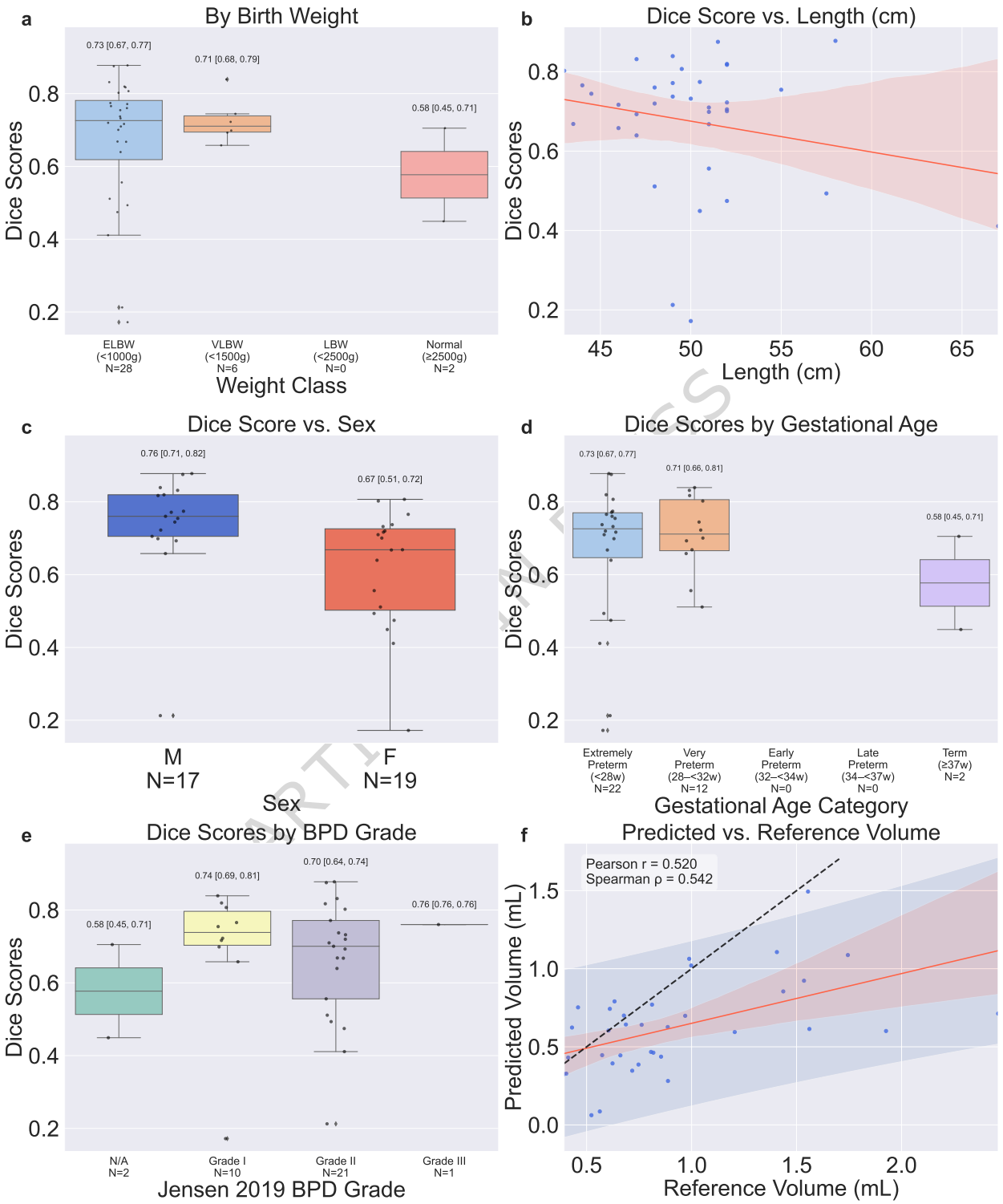


Figure 4: Evaluation of trachea segmentation model against clinical variables.

Technical Validation

Segmentation Model

To support further scientific research, we benchmark our dataset by providing pretrained segmentation models. Specifically, we train two U-Net models (15) for lung and trachea segmentation using the nnUNetV2 framework (16) on T1-weighted StarVIBE MRI data. The models are based on the 3D full-resolution U-Net architecture, optimized to accurately delineate the lungs and trachea through a combination of soft Dice loss and cross-entropy loss with deep supervision.

The lung segmentation model was trained on a dataset comprising 40 samples, partitioned into five folds using an 80/20 split for training and validation. The trachea segmentation model was trained on 36 samples using the same 5-fold cross-validation strategy to ensure robustness and generalizability. Additionally, we conducted an ablation study by replacing the U-Net with residual connections in the encoder (17) to evaluate the impact of architectural modifications on segmentation performance.

		(a) BPD-Neo-Lung (N=40)			**(b) BPD-Neo-Trach (N=36)**		
Method	**Fold**	**Dice ↑**	**HD-95 ↓**	**NSD ↑**	**Dice ↑**	**HD-95 ↓**	**NSD ↑**
nnUNet	0	0.9612	1.9431	0.9136	0.7052	9.2736	0.8240
	1	0.9553	1.9370	0.8881	0.6352	8.7515	0.7780
	2	0.9519	1.9663	0.8887	0.6646	7.4188	0.7681
	3	0.9320	4.4122	0.8740	0.7648	9.2399	0.8673
	4	0.9556	1.9791	0.9141	0.5992	11.3058	0.7139
	Mean	0.9512	2.4475	0.8957	0.6747	9.2000	0.7912
	Std	0.0344	2.5300	0.0772	0.1644	5.6625	0.1618
nnUNet (ResEnc)	0	0.9568	2.7370	0.9045	0.6722	8.4637	0.7975
	1	0.9541	1.9842	0.8877	0.6623	8.4812	0.8029
	2	0.9520	1.8448	0.8905	0.5978	43.7024	0.6957
	3	0.9279	4.2206	0.8716	0.7402	6.6449	0.8716
	4	0.9500	2.6074	0.9062	0.6258	8.8327	0.7478
	Mean	0.9482	2.6788	0.8921	0.6600	15.0372	0.7835
	Std	0.0391	2.8532	0.0802	0.1719	40.8201	0.1821

Table 2: Cross-validation performance of nnUNet and nnUNet (ResEnc) models across three metrics: Dice score, 95th percentile Hausdorff Distance (HD95), and Normalized Surface Distance (NSD).

The cross-validation performance results are presented in Table 2. The lung segmentation model demonstrates strong delineation capabilities, achieving a high mean cross-validation Dice score of 95.1%, indicating reliable performance which is reflected in Figure 1. In contrast, the trachea segmentation model yielded a lower peak Dice score of 67.3%, suggesting that further refinement is needed to improve performance in this task, as shown in Figure 2. One potential factor contributing to the reduced accuracy is the placement of NORAS Flex coils over the lungs, which may have led to a decline in signal intensity, in turn reducing the performance when segmenting the trachea, particularly as the distance from the coils increased.

Segmentation Performance Versus Clinical Data

To assess the clinical relevance of our segmentation models, we evaluated the relationship between Dice scores, segmentation-derived volumes, and corresponding clinical variables. For this analysis, we utilized the nnUNet models rather than the nnUNet (ResEnc) variants, due to their superior performance. As shown in Figure 3 panels (a), (c), (d), and (e), the models achieved consistently high Dice scores across different birth weight classes, sex, gestational age categories, and Jensen 2019 BPD grades, indicating robust and generalizable performance across clinically relevant subgroups.

As shown in Figure 3 panel (f), we also observe a strong correlation between the predicted and reference lung volumes, with a Pearson correlation coefficient of 0.957 and a Spearman correlation coefficient of 0.981,

indicating high agreement between model outputs and expert annotations. Additionally, we note a slight decrease in Dice score, dropping to approximately 0.92, as infant length increases, suggesting that anatomical variability associated with body size may modestly affect segmentation performance.

Target	**N**	**R1 vs R2** (Dice)	$\mathbf{R}_{staple}$ **vs Best Model** (Dice)
Lung	11	0.955 ± 0.014	0.940 ± 0.050
Trachea	07	0.734 ± 0.099	0.676 ± 0.120

Table 3: Inter-reviewer and model performance comparison for lung and trachea segmentation. The mean Dice similarity coefficient (± standard deviation) is reported for both (i) inter-rater agreement between Reviewer 1 (R1) and Reviewer 2 (R2), and (ii) agreement between the STAPLE-generated consensus segmentation (R_{staple}) and the best-performing model. Agreement is notably higher for lung segmentation compared to trachea, both between human raters and between model and consensus.

For tracheal segmentation, the overall Dice performance was lower, and greater variability was observed in relation to clinical variables. As illustrated in Figure 4, Dice scores varied across different birth weight classes, sex, Jensen 2019 BPD grades, and gestational age categories, indicating sensitivity to clinical and anatomical heterogeneity. Additionally, a lower correlation was observed between predicted and reference tracheal volumes, along with increased variation in Dice scores as a function of infant length, suggesting that tracheal segmentation may be more susceptible to anatomical and imaging variability. Future studies may focus on developing higher-performing models for tracheal segmentation, aiming to improve robustness and accuracy in this challenging subset of the dataset.

Inter-Observer Variability

We also assess inter-observer variability for both segmentation tasks, as it provides an important upper bound for model performance and highlights the inherent variability among human reviewers. This benchmark helps contextualize the model's accuracy relative to expert-level agreement and establishes a practical ceiling for achievable performance.

To evaluate inter-observer variability, a second expert independently segmented a subset of the dataset, Lungs (N = 11) and Trachea (N = 7), following the same segmentation protocol as the first observer. This included identical preprocessing steps, fiducial placement, region-growing procedures, and post-processing corrections, ensuring consistency in the annotation methodology for comparative analysis.

We first compare the segmentation performance metrics between Reviewer 1 and Reviewer 2. The inter-observer Dice score for lung segmentation was 0.955 ± 0.014 (from Table 3), indicating very high agreement and serving as a pseudo-upper bound for model performance. Notably, our best model achieved a mean 5-fold cross-validation Dice score of 0.9512, demonstrating that the model performs comparably to expert-level consistency and is well-optimized for the task.

From Table 3, we also observe that inter-observer Dice scores for trachea segmentation are substantially lower compared to lung segmentation, with a mean score of 0.734 ± 0.099. This indicates greater variability and disagreement between reviewers for tracheal annotations, likely due to the smaller structure size and lower signal quality. This inter-observer variability is reflected in the model's performance, with the best-performing trachea segmentation model achieving a 5-fold cross-validation Dice score of 0.6738.

To further analyze inter-observer variability, we computed Bland-Altman plots and measured the volume correlation coefficient between the two reviewers. As shown in Figure 5(a), the majority of differences in lung volumes fall within a narrow range of approximately -10 to +10 cm^3, indicating strong agreement for lung segmentation. In contrast, Figure 5(c) reveals substantially greater variability in tracheal volume estimates relative to their mean, which is expected given the smaller absolute volumes and increased difficulty in delineating the trachea.

Figure 5 panels (b) and (d) illustrate the volume correlations between reviewers. As shown in panel (b), there is strong agreement for lung segmentation, with the reviewers' measurements closely aligned. However, panel (d) highlights substantially greater variability in trachea segmentation, underscoring the challenges

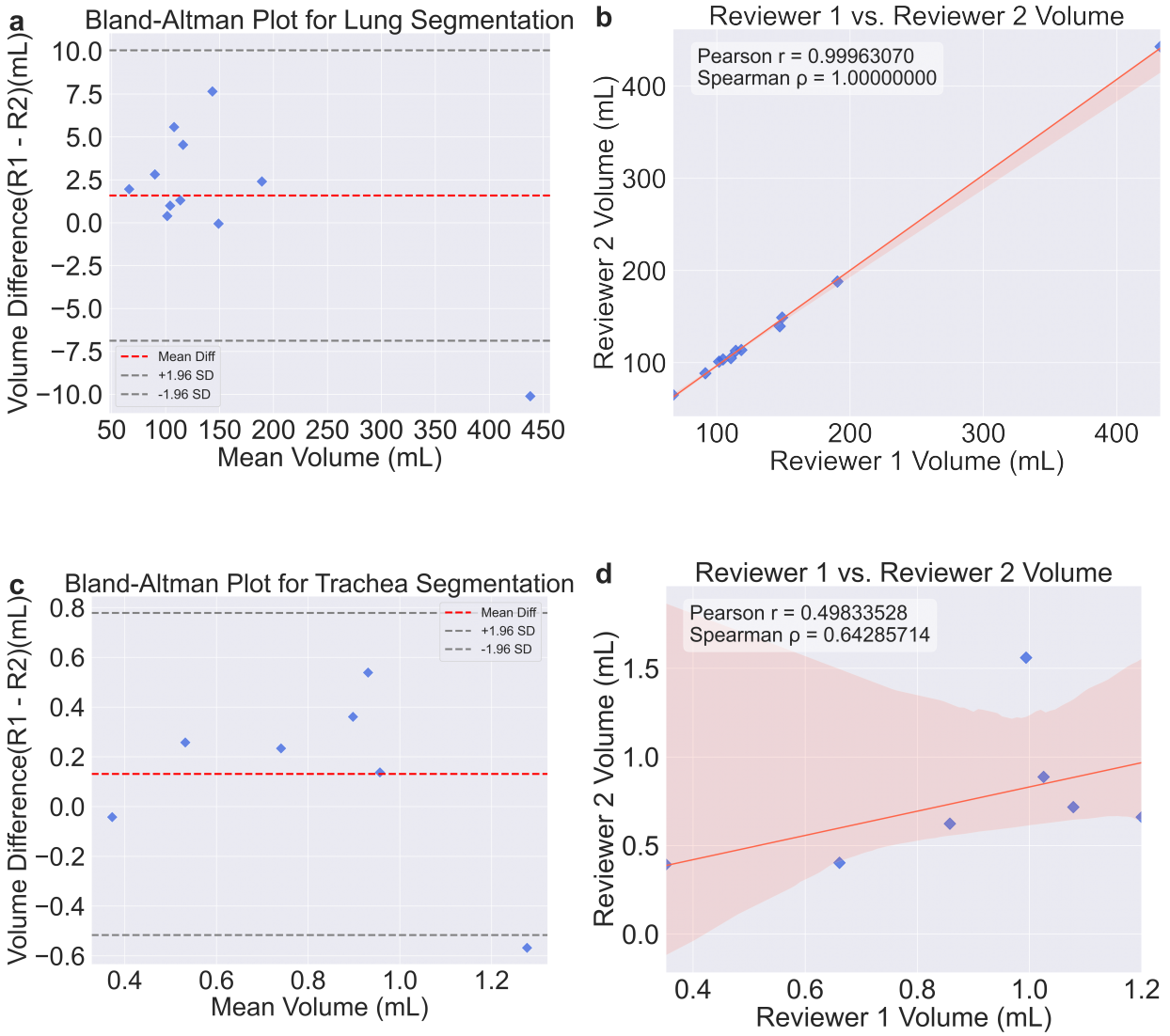


Figure 5: Inter-rater agreement for lung and trachea segmentation volumes. (a, c) Bland–Altman plots comparing volume differences between Reviewer 1 and Reviewer 2 for lung (a) and trachea (c) segmentations. The red dashed line indicates the mean difference, while the gray dashed lines denote the ±1.96 standard deviation limits of agreement. (b, d) Scatter plots with linear regression comparing segmentation volumes between the two reviewers for lung (b) and trachea (d). Pearson correlation coefficients and Spearman correlation coefficients are reported. Agreement is high for lung segmentation (Pearson r ≈ 1.0), whereas trachea segmentation shows weaker correlation and greater variability.

associated with accurately delineating smaller airway structures. This analysis indicates that there remains considerable room for improvement and further research in trachea segmentation for neonatal StarVIBE MR images.

Finally, we generated a consensus segmentation for both lung and trachea using the STAPLE algorithm (18), integrating the annotations from both reviewers. We then computed the Dice scores between this consensus and the predictions from the best-performing model. As shown in Table 3, the trend remains consistent, high Dice scores were achieved for lung segmentation, while lower scores were observed for trachea segmentation.

This further supports the notion that tracheal segmentation remains a more challenging task and highlights the need for continued research in this area.

Data availability

All data records, including the DICOM series, NIfTI files, and clinical data, are available at `https://zenodo.org/records/15768091`, under the CC BY 4.0 license (14).

Code availability

The code repository for the segmentation models can be accessed via `https://github.com/rachitsaluja/BPD-Neo`.

Author Contributions

Conceptualization: R.S, M.S., A.K., J.D., Methodology: R.S., M.S., J.D., Formal Analysis: R.S., M.S., Investigation: R.S, M.S., A.K., J.D., Data Curation: R.S., M.S., J.D., Software: R.S., M.S., Validation: R.S., M.S., Visualization: R.S., M.S., J.D., Writing: R.S., M.S., A.K., J.D., Original Draft: R.S., M.S., A.K., J.D., Writing - Review & Editing: R.S., A.K., C.C., L.B., J.P., S.W., M.S., J.D., Project Administration: J.D., M.S., Supervision: J.D., M.S., Funding Acquisition: A.K., J.D.

Acknowledgments

The authors would like to acknowledge the assistance of the pediatric and NICU nursing staff at NYP who were invaluable in the success of this study. Funding for this work is provided under NHLBI R01HL167003.

References

[1] Katherine Y Wu, Erik A Jensen, Ammie M White, Yan Wang, David M Biko, Kathleen Nilan, María V Fraga, Laura Mercer-Rosa, Huayan Zhang, and Haresh Kirpalani. Characterization of disease phenotype in very preterm infants with severe bronchopulmonary dysplasia. *American journal of respiratory and critical care medicine*, 201(11):1398–1406, 2020.

[2] Jonathan P Dyke, Andreas Voskrebenzev, Lauren K Blatt, Jens Vogel-Claussen, Robert Grimm, Stefan Worgall, Jeffrey M Perlman, and Arzu Kovanlikaya. Assessment of lung ventilation of premature infants with bronchopulmonary dysplasia at 1.5 tesla using phase-resolved functional lung magnetic resonance imaging. *Pediatric Radiology*, 53(6):1076–1084, 2023.

[3] Neil J Stewart, Nara S Higano, Lena Wucherpfennig, Simon MF Triphan, Amy Simmons, Laurie J Smith, Mark O Wielpütz, Jason C Woods, and Jim M Wild. Pulmonary mri in newborns and children. *Journal of Magnetic Resonance Imaging*, 61(5):2094–2115, 2025.

[4] Erik B Hysinger, Nicholas L Friedman, Michael A Padula, Russell T Shinohara, Huayan Zhang, Howard B Panitch, and Steven M Kawut. Tracheobronchomalacia is associated with increased morbidity in bronchopulmonary dysplasia. *Annals of the American Thoracic Society*, 14(9):1428–1435, 2017.

[5] Nara S Higano, Robert J Fleck, David R Spielberg, Laura L Walkup, Andrew D Hahn, Robert P Thomen, Stephanie L Merhar, Paul S Kingma, Jean A Tkach, Sean B Fain, et al. Quantification of neonatal lung parenchymal density via ultrashort echo time mri with comparison to ct. *Journal of Magnetic Resonance Imaging*, 46(4):992–1000, 2017.

[6] Benedikt Mairhörmann, Alejandra Castelblanco, Friederike Häfner, Vanessa Koliogiannis, Lena Haist, Dominik Winter, Andreas Flemmer, Harald Ehrhardt, Sophia Stöcklein, Olaf Dietrich, et al. Automated mri lung segmentation and 3d morphologic features for quantification of neonatal lung disease. *Radiology: Artificial Intelligence*, 5(6):e220239, 2023.

[7] Alan H Jobe and Eduardo Bancalari. Bronchopulmonary dysplasia. *American journal of respiratory and critical care medicine*, 163(7):1723–1729, 2001.

[8] Erik A Jensen, Kevin Dysart, Marie G Gantz, Scott McDonald, Nicolas A Bamat, Martin Keszler, Haresh Kirpalani, Matthew M Laughon, Brenda B Poindexter, Andrea F Duncan, et al. The diagnosis of bronchopulmonary dysplasia in very preterm infants. an evidence-based approach. *American journal of respiratory and critical care medicine*, 200(6):751–759, 2019.

[9] K Vanhaverbeke, A Van Eyck, K Van Hoorenbeeck, B De Winter, A Snoeckx, T Mulder, and S Verhulst. Lung imaging in bronchopulmonary dysplasia: a systematic review. *Respiratory Medicine*, 171:106101, 2020.

[10] Kai Tobias Block, Hersh Chandarana, Girish Fatterpekar, Mari Hagiwara, Sarah Milla, Thomas Mulholland, Mary Bruno, Christian Geppert, and Daniel K Sodickson. Improving the robustness of clinical t1-weighted mri using radial vibe. *Magnetom Flash*, 5:6–11, 2013.

[11] Rafael M Azevedo, Rafael OP de Campos, Miguel Ramalho, Vasco Herédia, Brian M Dale, and Richard C Semelka. Free-breathing 3d t1-weighted gradient-echo sequence with radial data sampling in abdominal mri: preliminary observations. *American journal of roentgenology*, 197(3):650–657, 2011.

[12] Andriy Fedorov, Reinhard Beichel, Jayashree Kalpathy-Cramer, Julien Finet, Jean-Christophe Fillion-Robin, Sonia Pujol, Christian Bauer, Dominique Jennings, Fiona Fennessy, Milan Sonka, et al. 3d slicer as an image computing platform for the quantitative imaging network. *Magnetic resonance imaging*, 30 (9):1323–1341, 2012.

[13] Csaba Pinter, Andras Lasso, and Gabor Fichtinger. Polymorph segmentation representation for medical image computing. *Computer methods and programs in biomedicine*, 171:19–26, 2019.

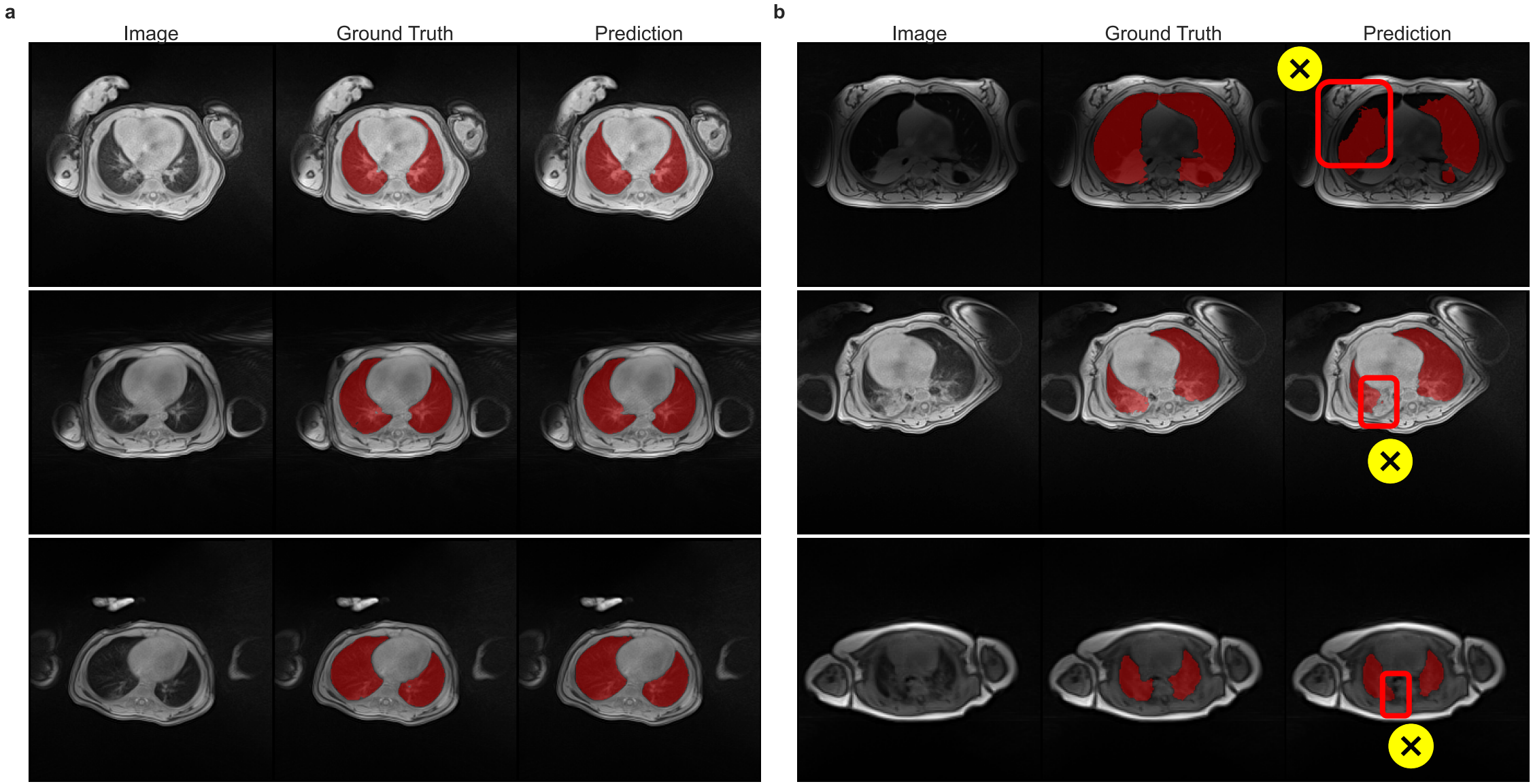
[14] Rachit Saluja. Bpd-neo: An mri dataset for lung-trachea segmentation with clinical data for neonatal bronchopulmonary dysplasia, June 2025. URL `https://doi.org/10.5281/zenodo.15768091`.

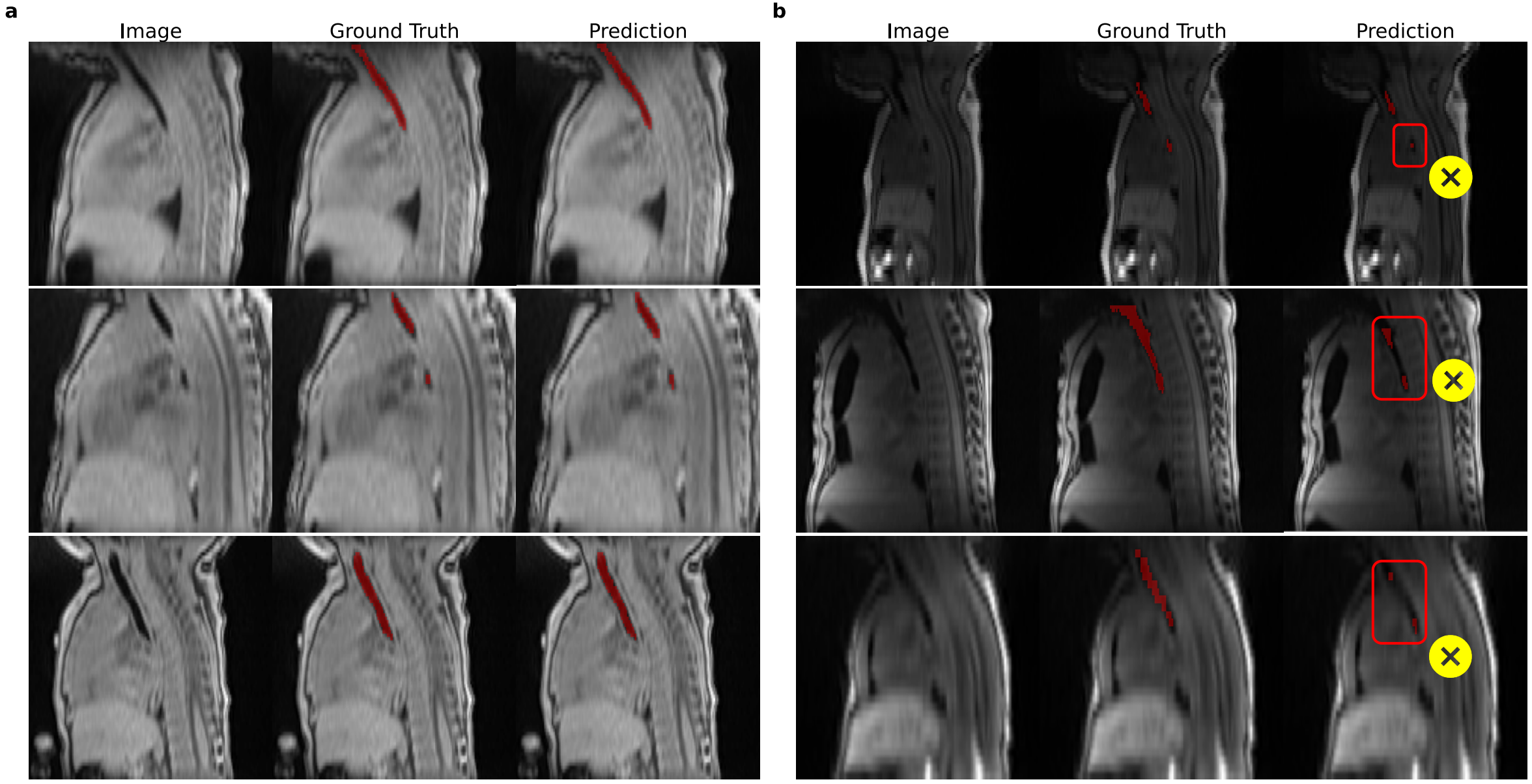
[15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

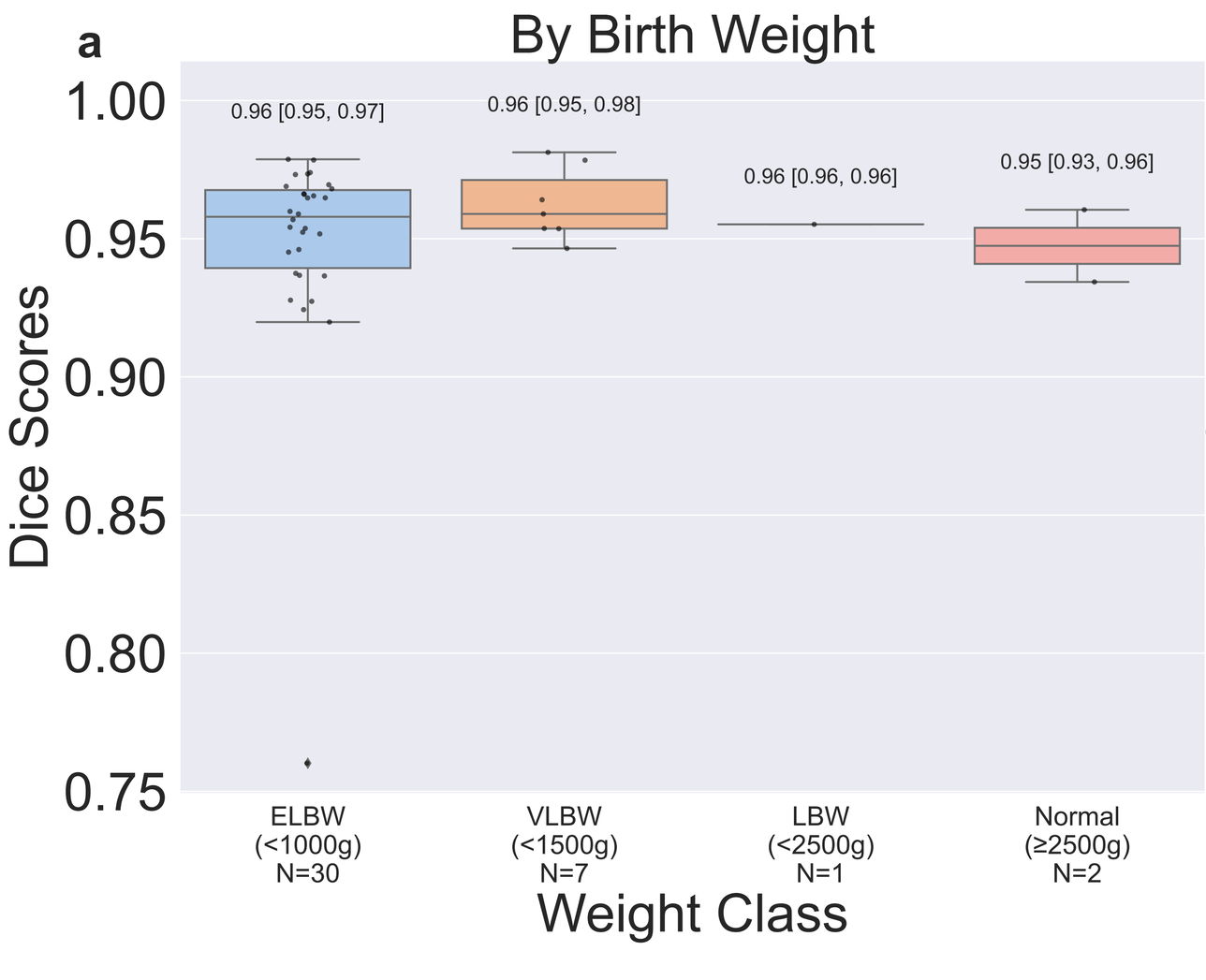
[16] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2): 203–211, 2021.

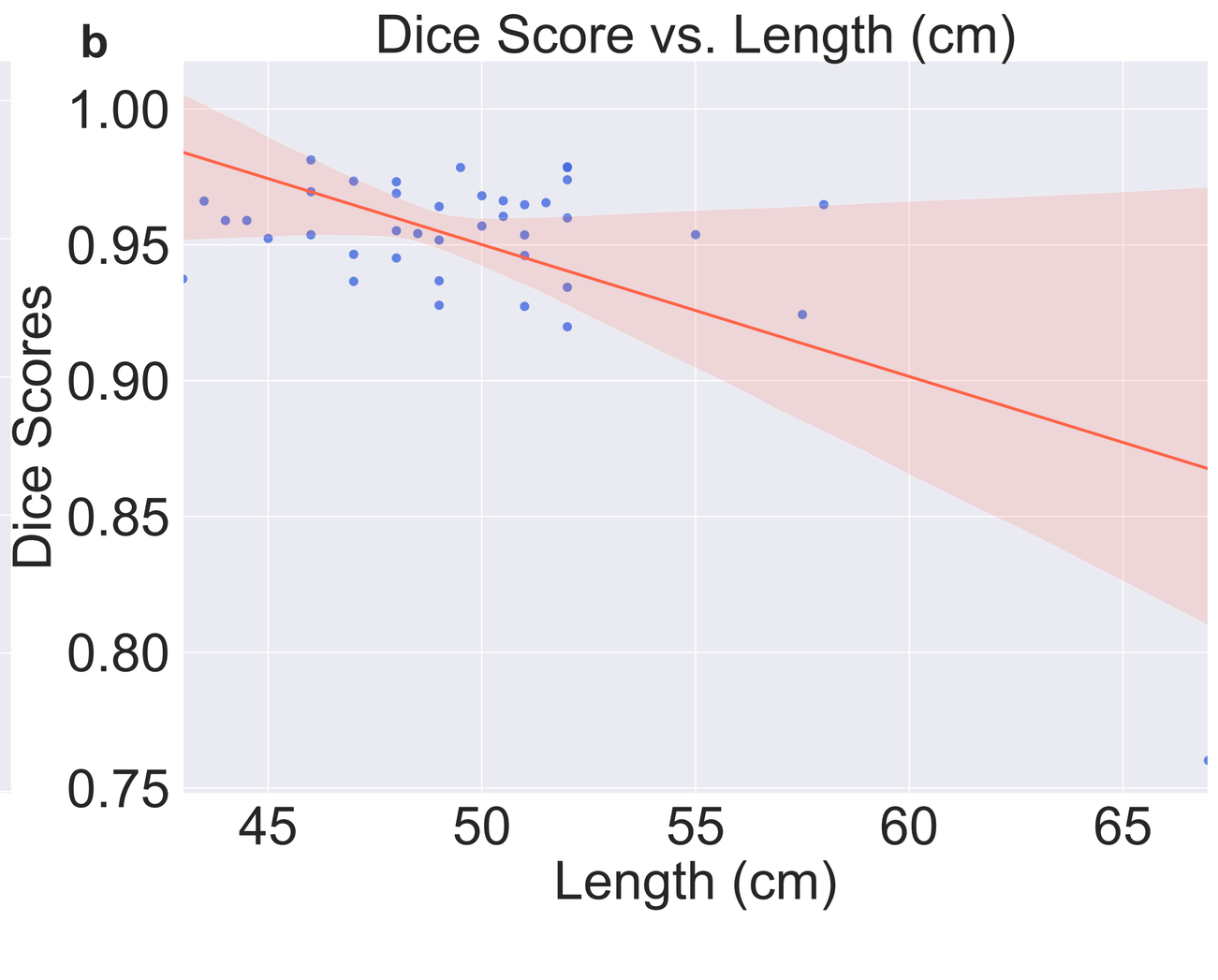
[17] Fabian Isensee, Tassilo Wald, Constantin Ulrich, Michael Baumgartner, Saikat Roy, Klaus Maier-Hein, and Paul F Jaeger. nnu-net revisited: A call for rigorous validation in 3d medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 488–498. Springer, 2024.

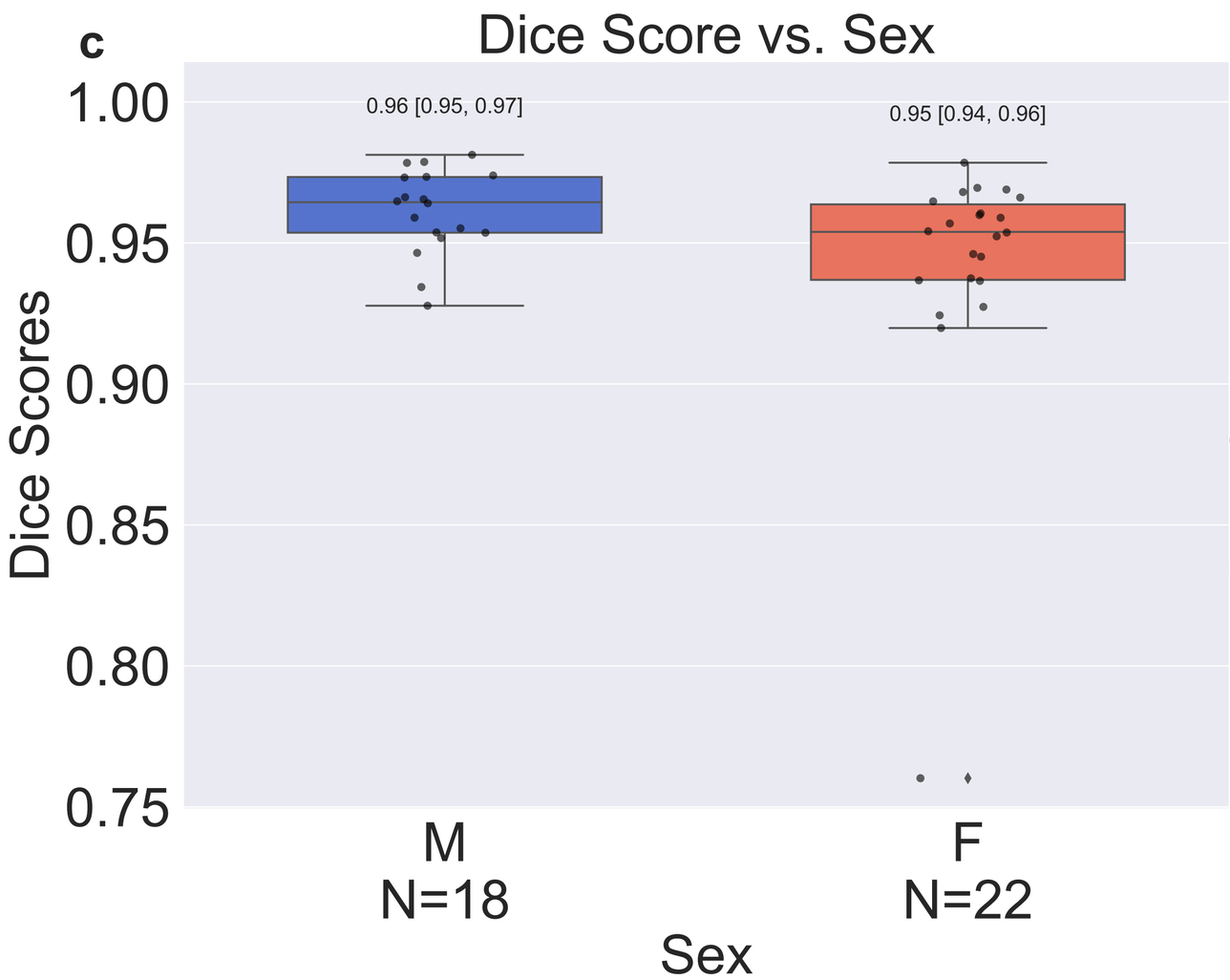
[18] Simon K Warfield, Kelly H Zou, and William M Wells. Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation. *IEEE transactions on medical imaging*, 23(7):903–921, 2004.

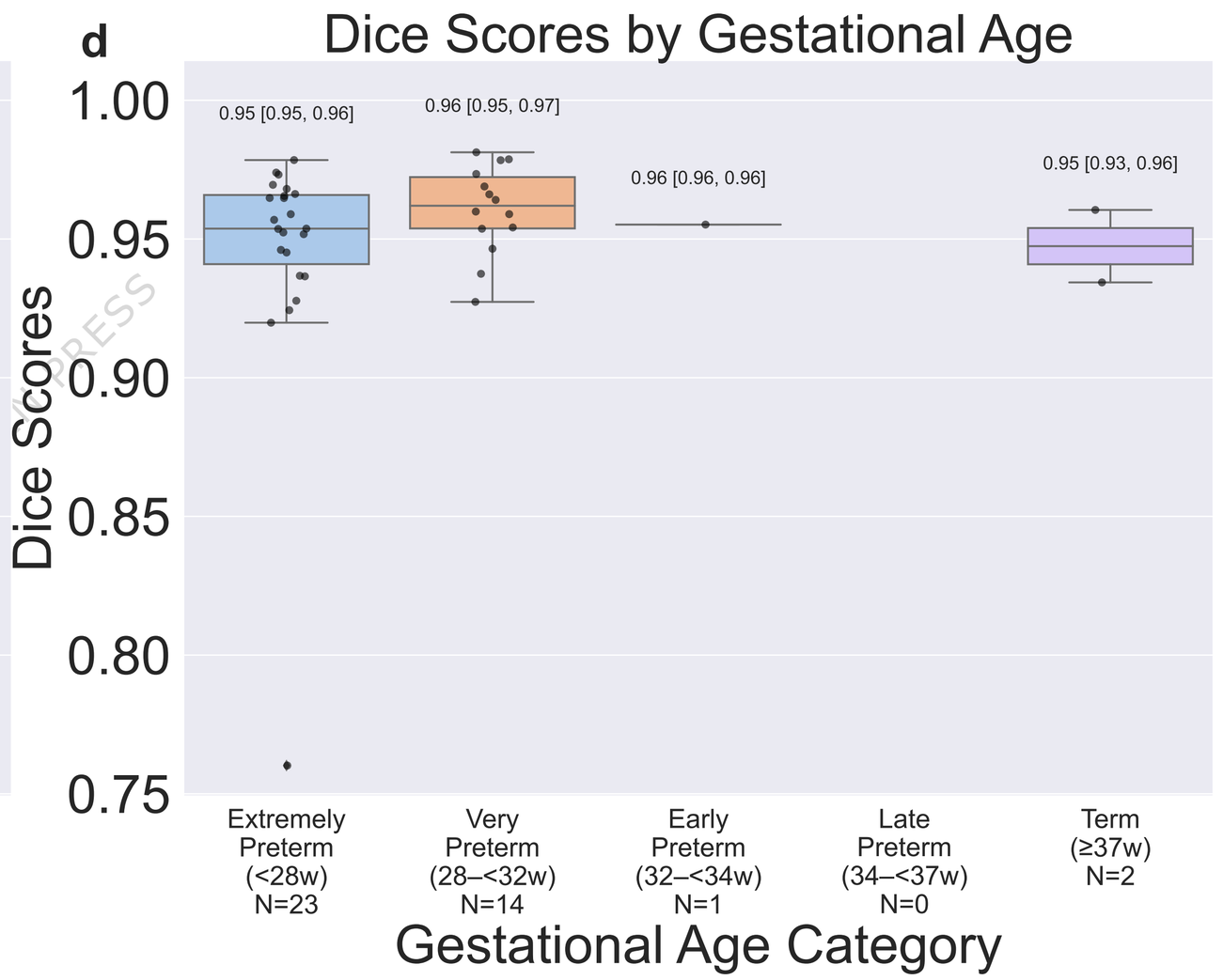

a
Image
Ground Truth
Prediction
b
Image
Ground Truth
Prediction


a
Image
Ground Truth
Prediction
b
Image
Ground Truth
Prediction


a
By Birth Weight
0.96 [0.95, 0.97]
0.96 [0.95, 0.98]
0.96 [0.96, 0.96]
0.95 [0.93, 0.96]
Dice Scores
1.00
0.95
0.90
0.85
0.80
0.75
ELBW (<1000g) N=30
VLBW (<1500g) N=7
LBW (<2500g) N=1
Normal (≥2500g) N=2
Weight Class

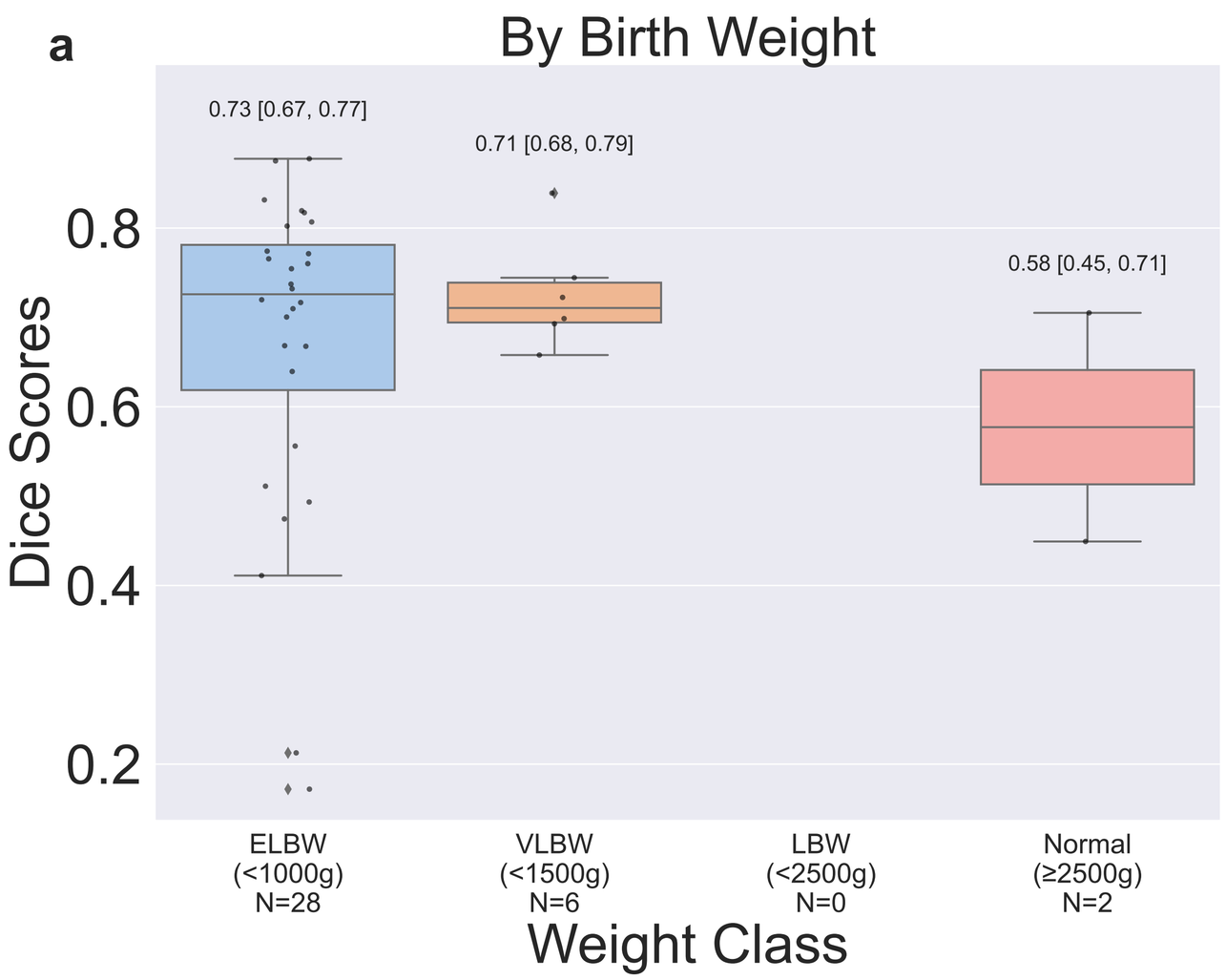

b
Dice Score vs. Length (cm)
Dice Scores
1.00
0.95
0.90
0.85
0.80
0.75
45
50
55
60
65
Length (cm)

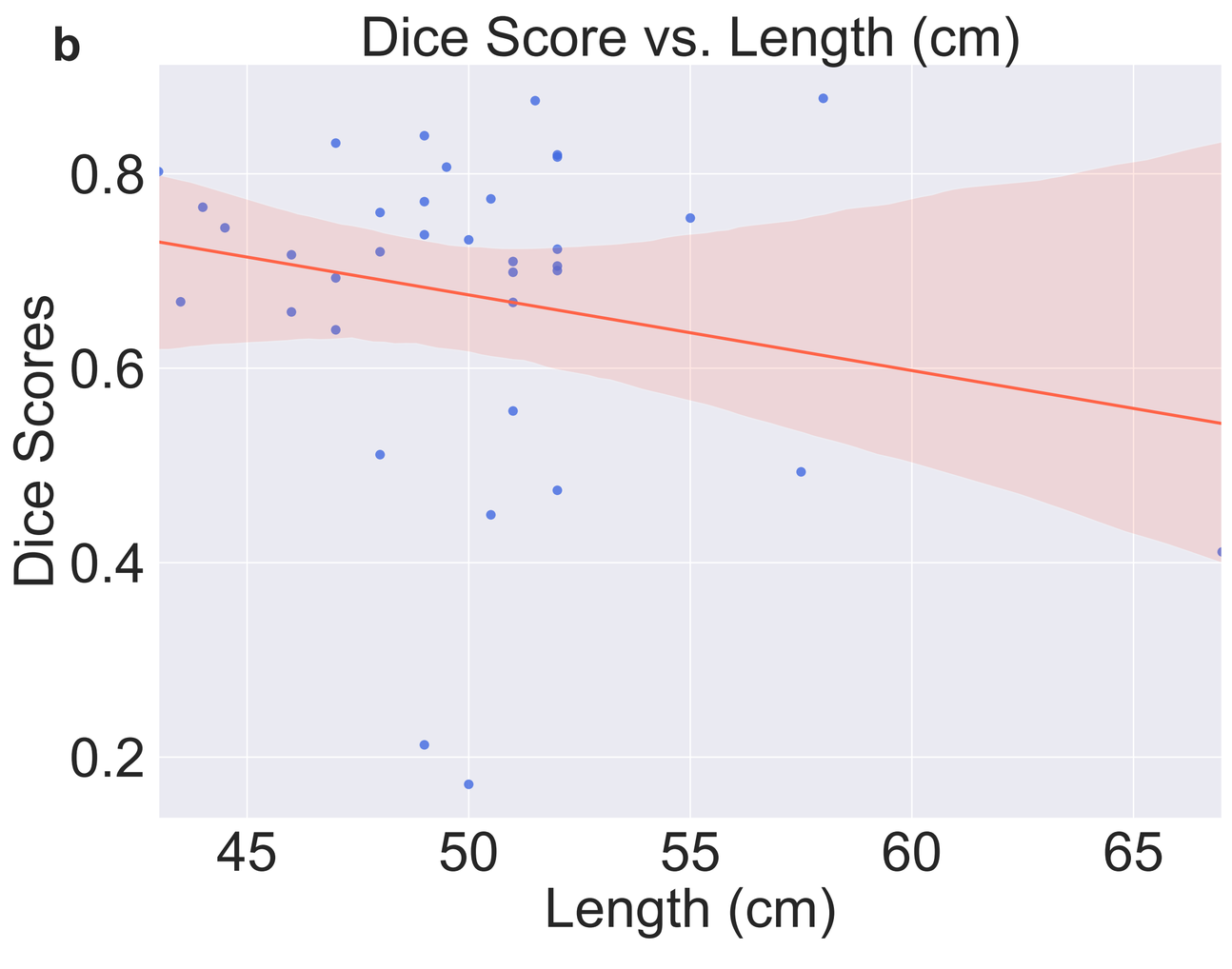

c
Dice Score vs. Sex
0.96 [0.95, 0.97]
0.95 [0.94, 0.96]
Dice Scores
1.00
0.95
0.90
0.85
0.80
0.75
M N=18
F N=22
Sex

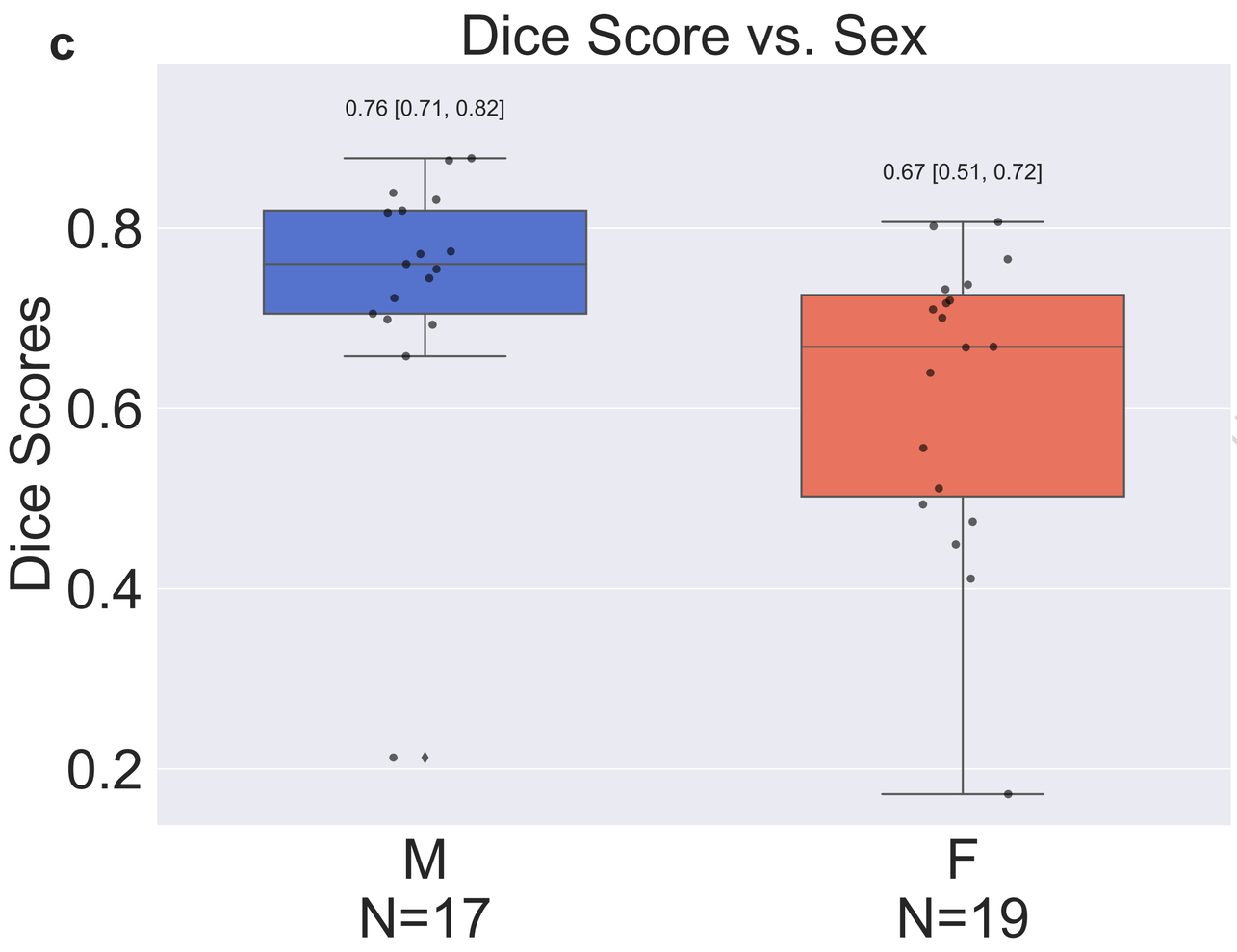

d
Dice Scores by Gestational Age
0.95 [0.95, 0.96]
0.96 [0.95, 0.97]
0.96 [0.96, 0.96]
0.95 [0.93, 0.96]
Dice Scores
1.00
0.95
0.90
0.85
0.80
0.75
Extremely Preterm (<28w) N=23
Very Preterm (28–<32w) N=14
Early Preterm (32–<34w) N=1
Late Preterm (34–<37w) N=0
Term (≥37w) N=2
Gestational Age Category

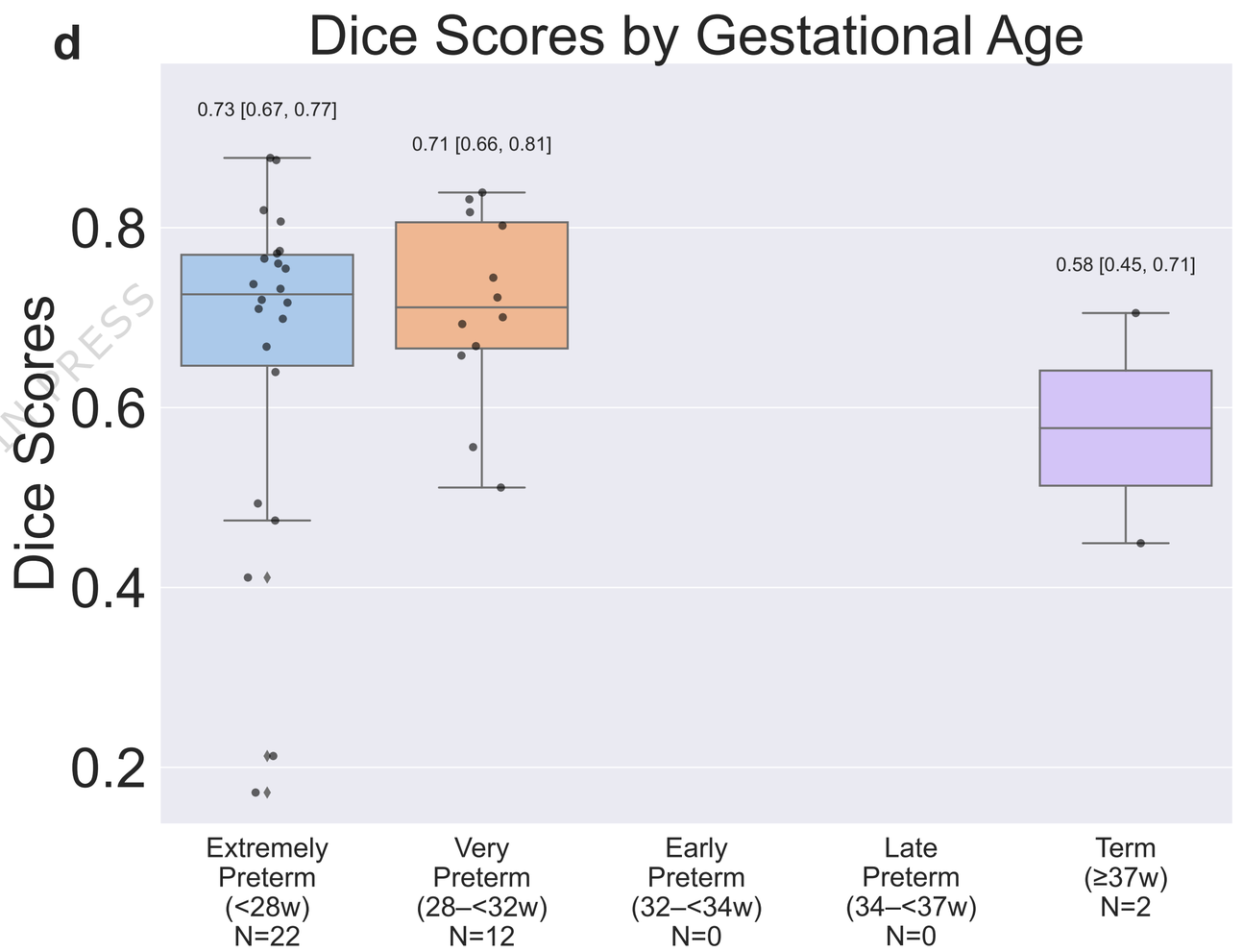
e
Dice Scores by BPD Grade
0.96 [0.93, 0.96]
0.97 [0.96, 0.98]
0.95 [0.94, 0.96]
0.97 [0.97, 0.97]
Dice Scores
1.00
0.95
0.90
0.85
0.80
0.75
N/A N=3
Grade I N=10
Grade II N=22
Grade III N=1
Jensen 2019 BPD Grade

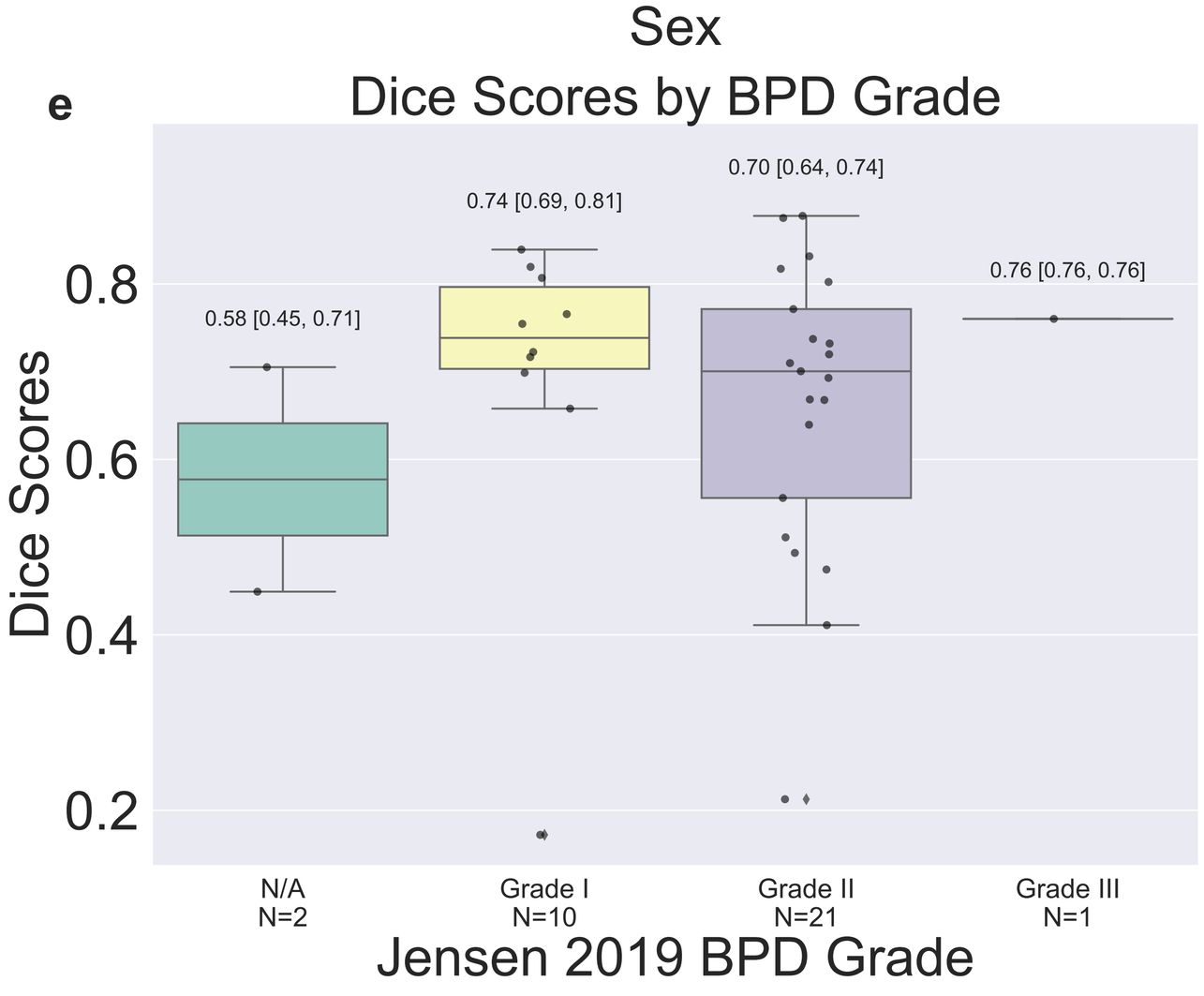
f
Predicted vs. Reference Volume
Pearson r = 0.957
Spearman ρ = 0.981
Predicted Volume (mL)
400
300
200
100
100
200
300
400
Reference Volume (mL)

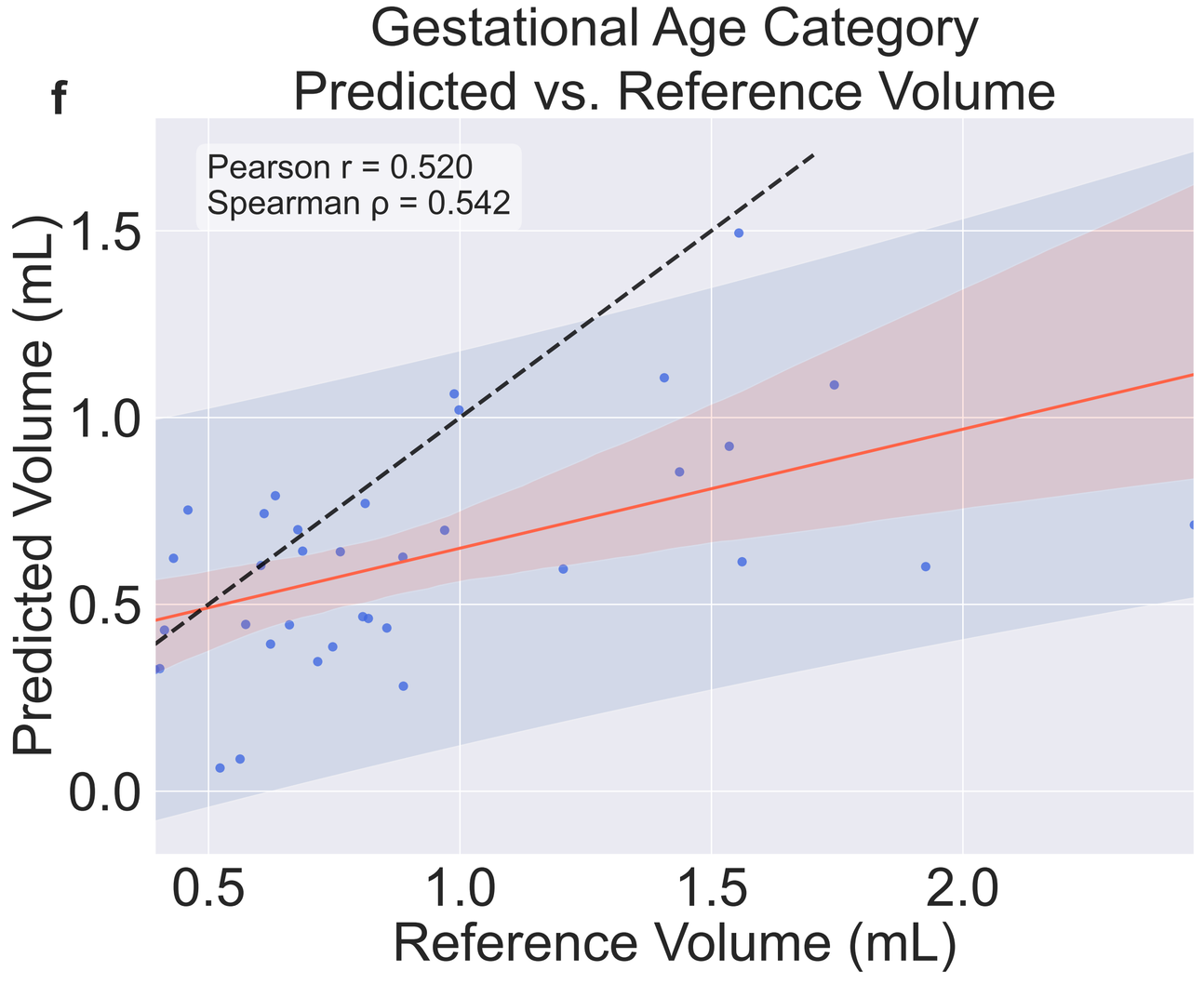


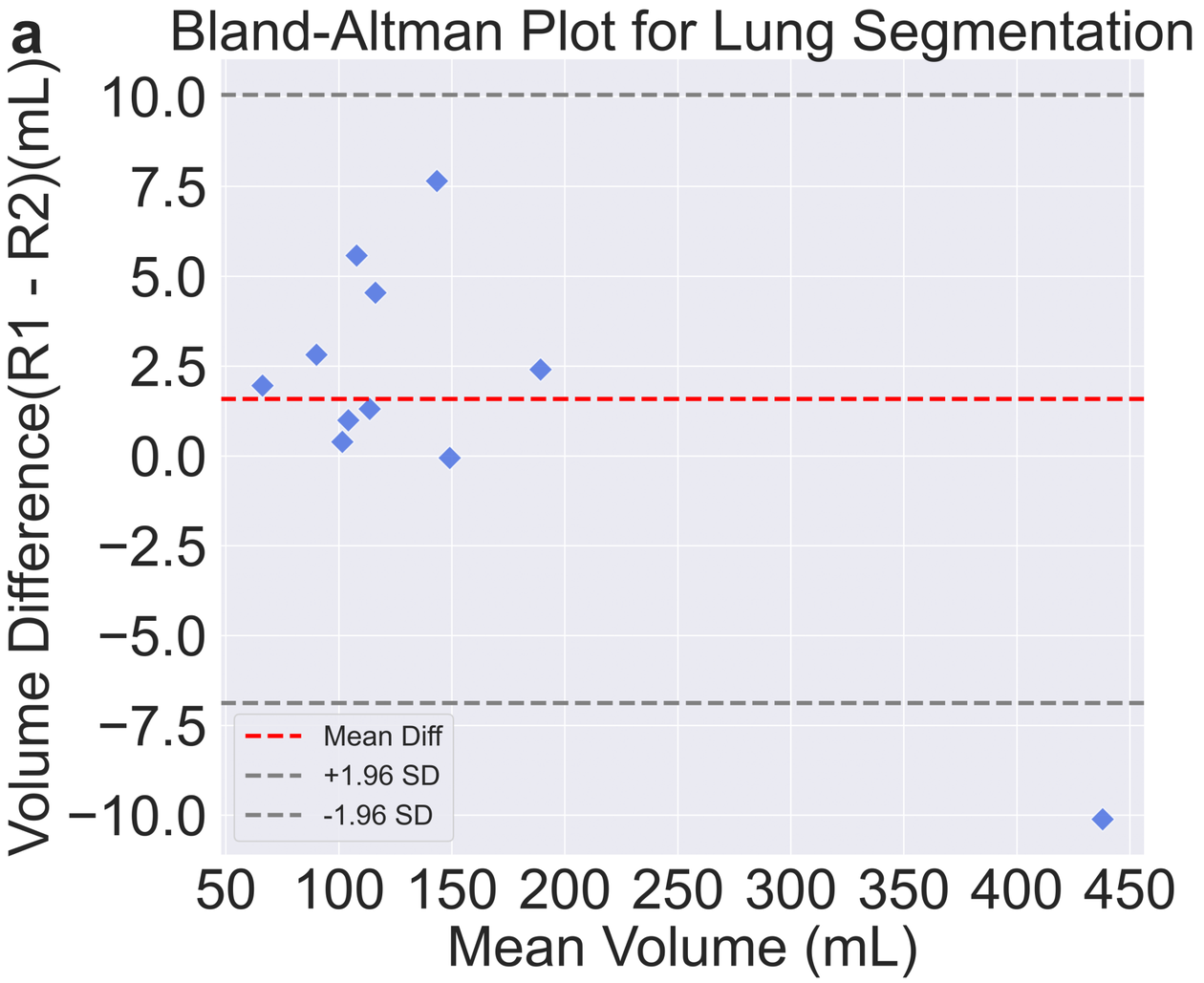


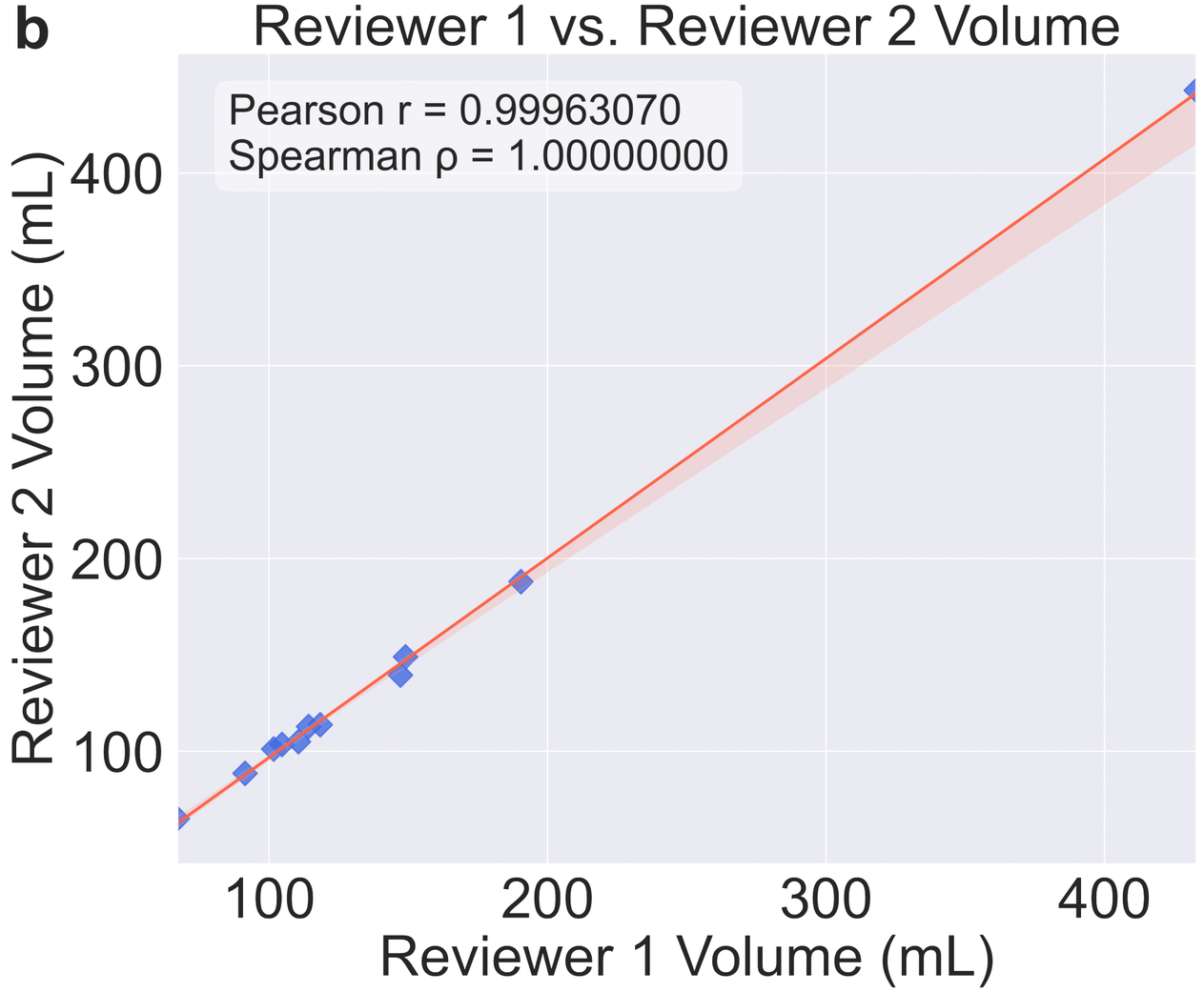


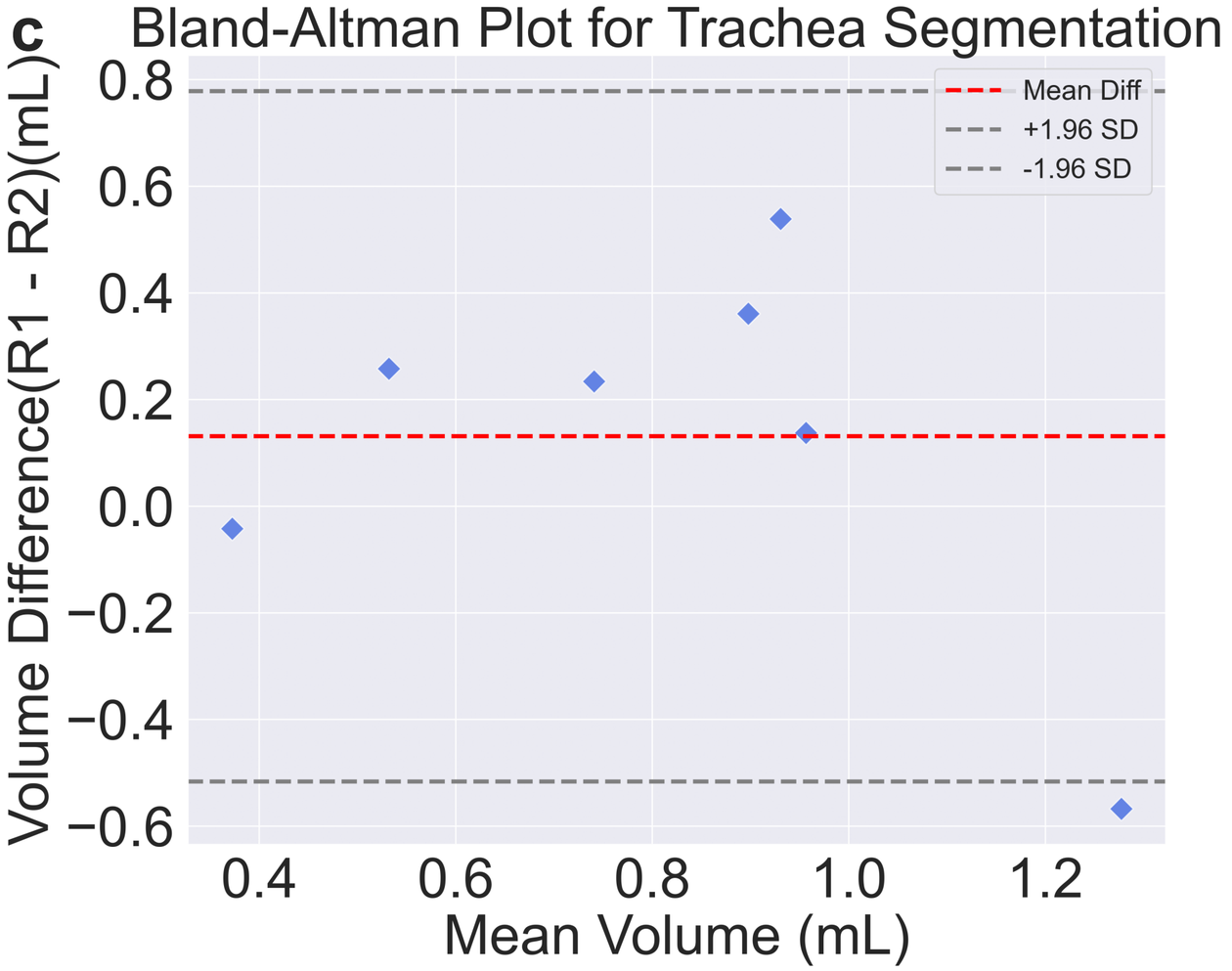


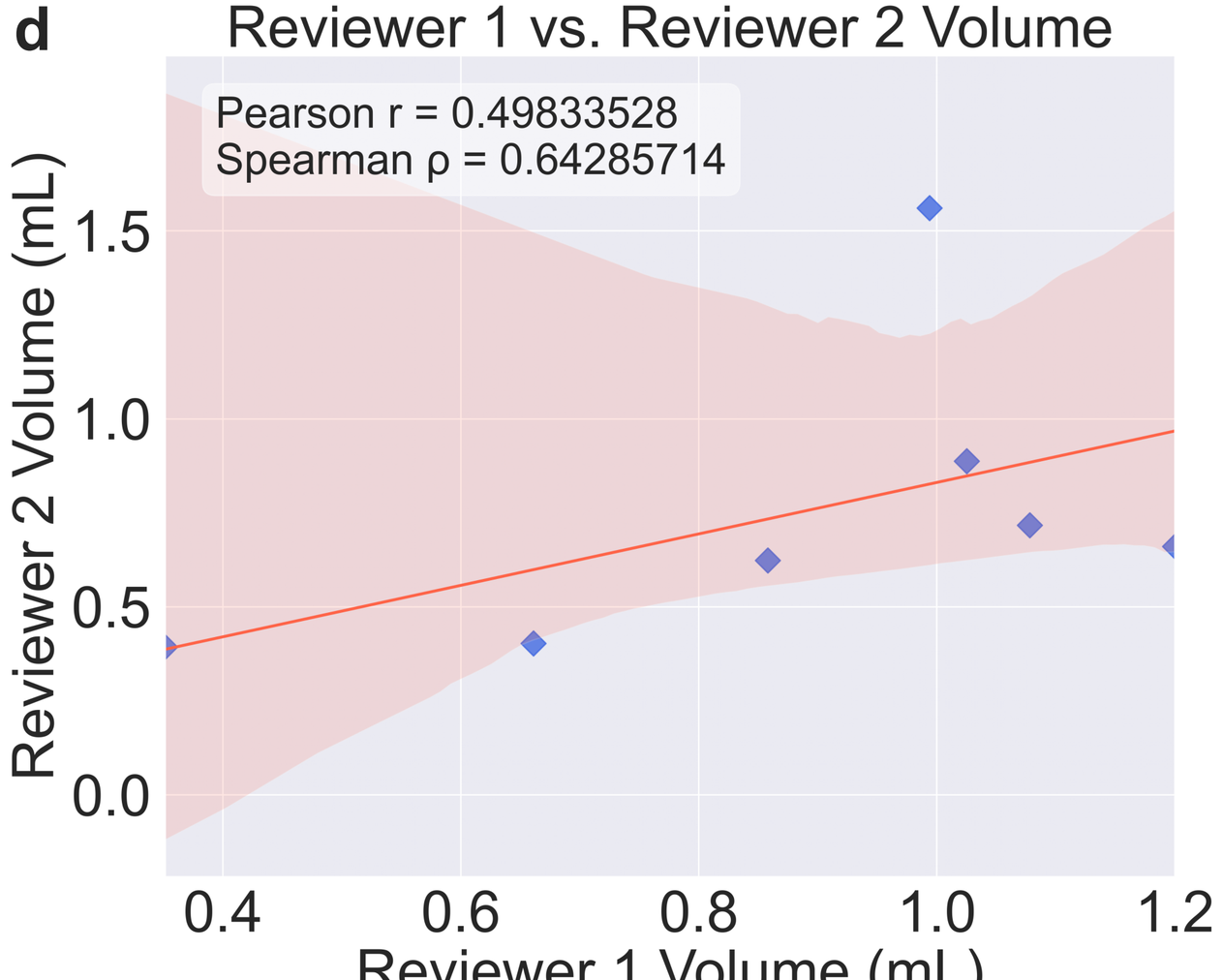





a
Bland-Altman Plot for Lung Segmentation
Volume Difference(R1 - R2)(mL)
Mean Volume (mL)
Mean Diff
+1.96 SD
-1.96 SD
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
50
100
150
200
250
300
350
400
450


b
Reviewer 1 vs. Reviewer 2 Volume
Pearson r = 0.99963070
Spearman ρ = 1.00000000
Reviewer 2 Volume (mL)
Reviewer 1 Volume (mL)
400
300
200
100
100
200
300
400


c
Bland-Altman Plot for Trachea Segmentation
Volume Difference(R1 - R2)(mL)
Mean Volume (mL)
Mean Diff
+1.96 SD
-1.96 SD
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
0.4
0.6
0.8
1.0
1.2


d
Reviewer 1 vs. Reviewer 2 Volume
Pearson r = 0.49833528
Spearman ρ = 0.64285714
Reviewer 2 Volume (mL)
Reviewer 1 Volume (mL)
1.5
1.0
0.5
0.0
0.4
0.6
0.8
1.0
1.2