

# 1           **Deep Learning-Based Magnetic Resonance Imaging Lung** 2           **Segmentation and Volumetric Marker Extraction in Preterm Infants**

## 3 4           **Authors**

5           Benedikt Mairhörn<sup>1,§</sup>, Alejandra Castelblanco<sup>1,§</sup>, Friederike Häfner<sup>2,4,§</sup>, Vanessa Pfahler<sup>5</sup>,  
6           Lena Haist<sup>2</sup>, Dominik Waibel<sup>1</sup>, Andreas Flemmer<sup>4</sup>, Harald Ehrhardt<sup>7</sup>, Sophia Stoecklein<sup>5</sup>, Olaf  
7           Dietrich<sup>5</sup>, Kai Foerster<sup>4</sup>, Anne Hilgendorff<sup>2,3,†</sup>, Benjamin Schubert<sup>1,6, †,\*</sup>

8  
9           <sup>1</sup>Institute of Computational Biology, Helmholtz Zentrum München, 85764 Neuherberg, Germany

10           <sup>2</sup>Institute for Lung Biology and Disease and Comprehensive Pneumology Center, Helmholtz Zentrum München,  
11           Germany, Member of the German Center for Lung Research (DZL)

12           <sup>3</sup>Center for Comprehensive Developmental Care (CDeC<sup>LMU</sup>) Hospital of the Ludwig-Maximilians University,  
13           Munich, Germany

14           <sup>4</sup>Department of Neonatology, Perinatal Center, Hospital of the Ludwig-Maximilians University, Munich, Germany

15           <sup>5</sup>Department of Radiology, Hospital of the Ludwig-Maximilians University, Munich, Germany

16           <sup>6</sup>Department of Mathematics, Technical University of Munich, 85748 Garching bei München, Germany

17           <sup>7</sup>Department of General Pediatrics & Neonatology, Justus-Liebig-University, Giessen, Germany, Member of the  
18           German Center for Lung Research (DZL)

19           <sup>§</sup>The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First  
20           Authors.

21           <sup>†</sup> The authors wish it to be known that, in their opinion, the last two authors should be regarded as Joint Last  
22           Authors.

23           \*Corresponding author: [benjamin.schubert@helmholtz-muenchen.de](mailto:benjamin.schubert@helmholtz-muenchen.de)

34 **Abstract**

35 *Objective and Impact Statement:* We apply a deep learning (DL) segmentation method and  
36 automate the extraction of imaging markers for neonatal lung structure using magnetic resonance  
37 imaging (MRI) in order to inform clinical care with robust and quantifiable information about the  
38 neonatal lung.

39 *Introduction:* Quantification of lung structural information in a standardized fashion is crucial to  
40 inform diagnostic processes that enable personalized treatment and monitoring strategies.  
41 Increased efficiency and accuracy in image quantification is especially needed in prematurely born  
42 infants, for whom long-term survival is critically determined by acute and chronic pulmonary  
43 complications, currently diagnosed based on clinical criteria due to the lack of routinely applicable  
44 diagnostic tools.

45 *Methods:* We prospectively enrolled 107 premature infants in two clinical centers with and without  
46 chronic lung disease, i.e., Bronchopulmonary Dysplasia (BPD) to perform quiet-breathing lung  
47 MRI. An ensemble of deep convolutional neural networks was developed to perform lung  
48 segmentation, with a subsequent reconstruction of the 3-dimensional lung and computation of MRI  
49 volumetric measurements and compared to the standard manual segmentation.

50 *Results:* The DL model successfully annotates lung segments with a volumetric dice score of 0.908  
51 (Site 1) and 0.880 (Site 2), thereby reaching expert-level performance while demonstrating high  
52 transferability between study sites and robustness towards technical (low spatial resolution,  
53 movement artifacts) and disease conditions. Estimated lung volumes correlated with infant lung  
54 function testing measures and enabled the separation of neonates with and without BPD.

55 *Conclusion:* Our work demonstrates the potential of AI-supported MRI measures to perform  
56 monitoring of neonatal lung development and characterization of respiratory diseases in this high-  
57 risk patient cohort.

58

59 **Keywords**

60 Bronchopulmonary Dysplasia, Chronic Lung Disease, Preterm Infant, Lung Segmentation, Lung  
61 Magnetic Resonance Imaging, BPD Severity Prediction, Deep Learning.

62

63

64 **MAIN TEXT**

65

66 **1. Introduction**

67 Clinical decision-making in lung disease is mostly based on clinical observations and lung function  
68 measurements, ideally complemented by structural information from imaging strategies. Although  
69 quantification of lung structural information in a standardized fashion can critically inform the  
70 diagnostic process and help to implement personalized treatment and monitoring strategies, the  
71 much needed increase in efficiency, accuracy and comparability in image quantification most often  
72 remains an unmastered challenge in critical areas of today's clinical care.

73 These unmet diagnostic clinical needs become especially apparent in the most vulnerable, yet  
74 diagnostically challenging patient cohort. Born extremely immature, the preterm infant postnatally  
75 faces the inevitable development of acute lung injury, subsequently evolving into a chronic  
76 condition in the majority of cases [1]. While the diagnostic process still solely relies on clinical  
77 indicators of late-stage pulmonary function [2,3], the standardized assessment of radiation-free,  
78 sensitive imaging strategies would allow for diagnosing and monitoring respiratory disease from  
79 neonatal life into adulthood.

80 Given the low sensitivity and diagnostic value of conventional chest radiography to sensitively  
81 discriminate disease conditions and the limitations of Computed Tomography (CT) due to  
82 radiation exposure [4,5], alternative imaging techniques such as Magnetic Resonance Imaging  
83 (MRI) are being explored to provide quantitative information with prognostic relevance when  
84 assessing the diseased neonatal lung [6,7,8]. MRI in the neonatal lung is technically challenged by  
85 small subject sizes, lower spatial resolution, and sensitivity to infant motion, resulting in blurring,  
86 ghosting, and other image artifacts [9]. These conditions demand expert knowledge to obtain  
87 measurements from the acquired pulmonary images [7,10] and affect inter-rater concordances  
88 resulting in low standardization and limitations in high-throughput of MRI-based monitoring in  
89 neonatal lung disease.

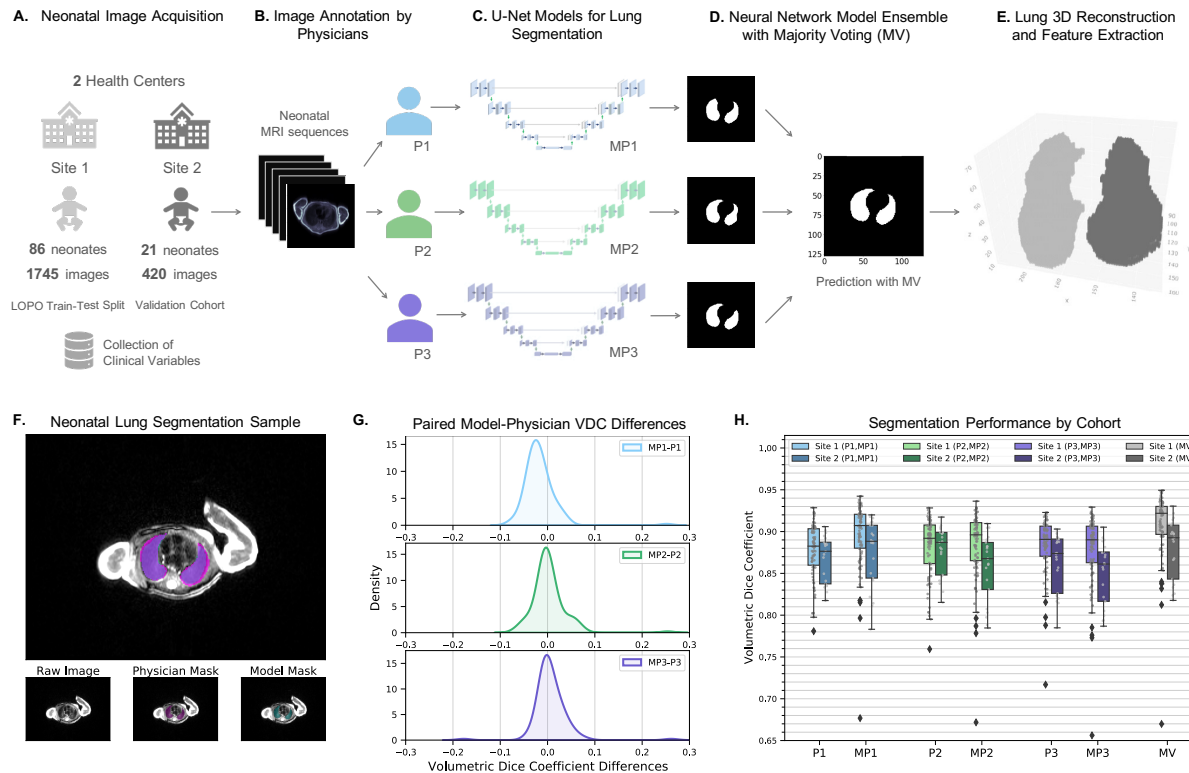
90 We therefore developed a deep learning-based system to support robust and standardized MRI  
91 analysis in the lungs of preterm neonates. To this end, we prospectively enrolled 107 cases with  
92 and without chronic lung disease, also known as Bronchopulmonary Dysplasia (BPD), from two  
93 study sites undergoing quiet-breathing lung MRI near term age. BPD is a chronic lung injury

94 syndrome resulting from a structurally immature lung condition that comprises a range of  
95 functional abnormalities (e.g., alveolar septation, airway injuries, decreased microvascular  
96 development) leading to insufficient gas exchange, with severity of the disease determined by the  
97 need of supplemental oxygen [1]. We combined recent advances in computational methods, i.e.,  
98 convolutional neural networks (CNN), which have been integrated in various biomedical imaging  
99 applications [11,12,13], to test the applicability and robustness of deep learning (DL) methods for  
100 performing MRI lung segmentation in preterm infants with and without BPD. Subsequently, the  
101 obtained lung segmentations were used to compute MRI-based 3-dimensional (3D) lung  
102 volumetric features, that enabled the accurate separation of healthy cases from premature infants  
103 with different BPD severity grades.

## 104 **2. Results**

105 The proposed automated pipeline was developed and validated in a cohort of 107 preterm infants  
106 near term during quiet sleep. Quiet-breathing T2-weighted single-shot fast-spin-echo 3T-MRI lung  
107 sequences and clinical information were acquired at two study sites (Fig. 1A, Table 1).

108 We trained a set of U-Net CNN models to perform 2D lung segmentations on each slice of the  
109 collected neonatal MRI scans, with each model based on manual annotations of three different  
110 physicians (Fig. 1B-C) and combined them through pixel-wise majority voting (MV) to an  
111 ensemble model (Fig. 1D). A 3D representation of the lung and a clustering method to separate  
112 left and right lobes were used to calculate volumetric features (Fig. 1E).



113

114 **Figure 1. MRI-based Neonatal Lung Volume Analysis Pipeline.** (A) MRI neonatal image acquisition and data collection. (B)  
 115 Manual image annotation performed by three trained physicians. (C-D) CNN model training and model prediction with majority  
 116 voting. (E) Lung 3D reconstruction and volume-based feature extraction. (F) Lung segmentation example. (G) Paired Volumetric  
 117 Dice Coefficient (VDC) differences between the reference CNN model and the corresponding manual annotation. (H) Manual  
 118 annotation and model performances by cohort.

119 *2.1 Deep Learning Enables Human-Level Neonatal Lung Segmentation in MRI*

120 To investigate whether human-level performance could be achieved, the segmentations generated  
 121 by one of the DL models (e.g., MP1), which was trained with the manual annotations from one  
 122 physician (e.g., P1), were compared with each ground-truth of the remaining manual  
 123 segmentations (i.e., P2 and P3), we report the average model performance across these  
 124 comparisons. As evaluation metric, we used the volumetric dice coefficient (VDC) integrating the  
 125 segmentation performance of all pixels and slices from one MRI sequence. The average inter-rater  
 126 VDC concordance between physicians (e.g., P1 vs each of the manual annotations P2 and P3) was  
 127 also calculated as a reference for the model performance.

128 For study Site 1, models were trained in a leave-one-patient-out (LOPO) cross-validation scheme  
129 to generate unbiased performance estimates, while samples from study Site 2 were solely used for  
130 validation, performance for both cohorts was analyzed.

131 The proposed DL models achieved high segmentation performances comparable to the reference  
132 manual performance, while demonstrating low average VDC differences between the model and  
133 the reference manual segmentation when comparing identical MRI sequences ( $-0.017 \pm 0.035$ ,  
134  $0.003 \pm 0.036$ ,  $0.006 \pm 0.037$ , for MP1, MP2, MP3; Fig. 1G). One case even exceeded average  
135 performance (one-sided Wilcoxon signed-rank test,  $p$ -value= $0.0565$  for MP3). Small differences  
136 between the paired performance of the model and the manual segmentation indicates an adequate  
137 representational power of the proposed models to abstract the segmentation knowledge from the  
138 provided training dataset.

139 The average VDC of the manual annotations were  $0.875 \pm 0.032$ ,  $0.881 \pm 0.034$ , and  $0.879 \pm 0.035$   
140 (P1, P2, P3), whereas the DL models had VDCs of  $0.89 \pm 0.041$ ,  $0.878 \pm 0.042$ , and  $0.872 \pm 0.043$   
141 (MP1, MP2, MP3), demonstrating differences of less than 0.016 points in the overall performance  
142 thus indicating high comparability between models and manual annotations.

143 The MV ensemble model prediction was evaluated by comparing its performance against a ground  
144 truth generated with all manual annotations (P1, P2, P3) also aggregated through pixel-wise  
145 majority voting. The MV ensemble model showed the highest segmentation performance (Fig.  
146 1H), with an average VDC of  $0.902 \pm 0.039$  improving by 2.1% over the highest manual  
147 segmentation's VDC average, indicating human-level accuracy of the AI based segmentation  
148 method for quiet-breathing neonatal lung MRI.

149

## 150 *2.2 Robust Automatic Segmentations Across Clinical Sites and Diseased Lungs*

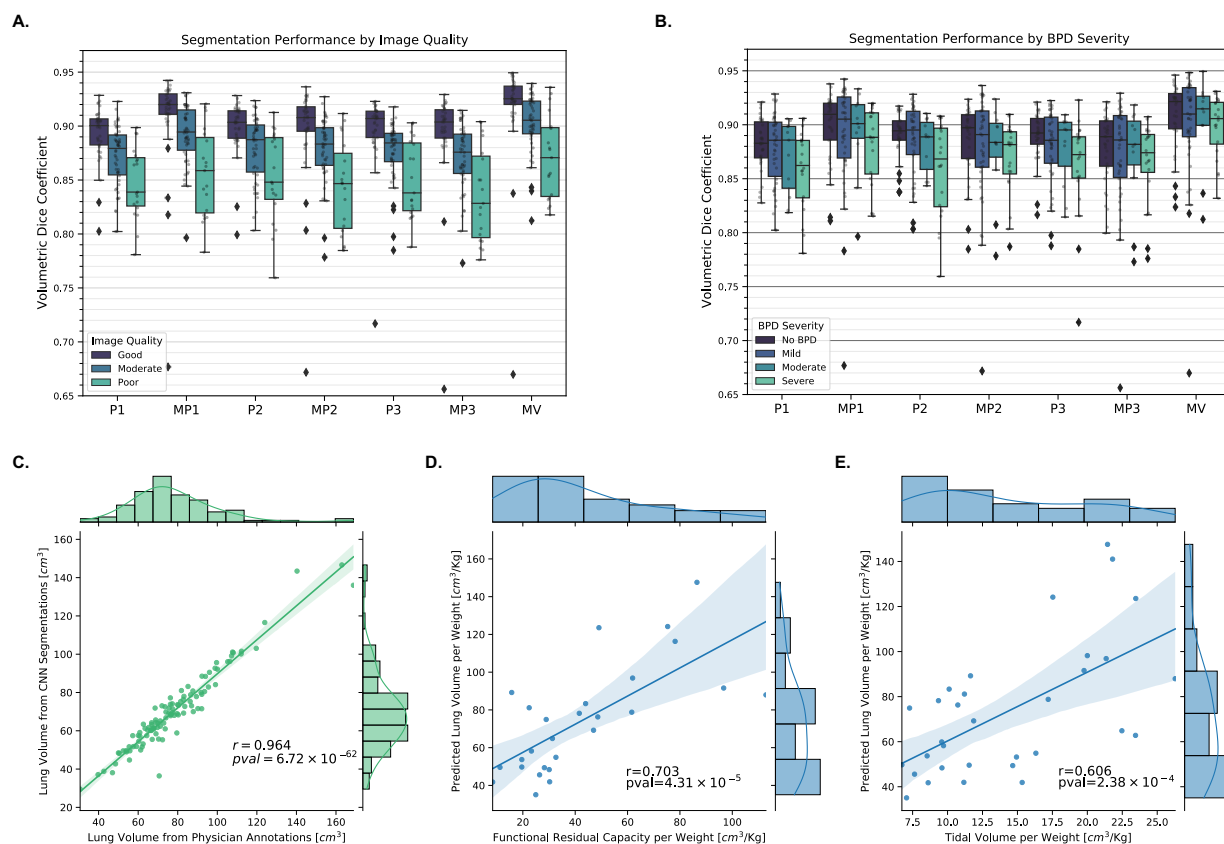
151 Performance across different study sites was assessed to confirm the models generalizability (Fig.  
152 1H). The average VDC of the MV ensemble model was  $0.908 \pm 0.039$  for study Site 1 and  
153  $0.880 \pm 0.036$  for study Site 2. Differences in the average VDC between cohorts were below 0.0286  
154 points for all trained models, indicating transferability of the models between cohorts while facing  
155 minor changes in the segmentation performance. Minor differences in performance across sites  
156 were also observed in the manual segmentations, potentially originating from imbalances in image

157 quality demonstrated by the average image quality scores of 1.7 (study Site 1) and 2.3 (study Site  
 158 2), respectively (1= best, 3=worst quality score).

159 We then investigated image quality as a confounding factor for segmentation performance (Fig.  
 160 2A) showing a significant effect for both manual segmentations (Kruskal-Wallis P1, P2, P3, p-  
 161 values= $[1.24 \times 10^{-6}, 1.14 \times 10^{-5}, 4.21 \times 10^{-8}]$ , n=107) and the models accordingly (Kruskal-Wallis  
 162 MP1, MP2, MP3, MV, p-values= $[1.53 \times 10^{-7}, 7.14 \times 10^{-7}, 4.67 \times 10^{-8}, 2.53 \times 10^{-7}]$ , n=107), indicating  
 163 lower MRI quality to result in lower segmentation performance.

164 Model robustness was furthermore tested for lung structural differences, i.e., presence of BPD-  
 165 characteristic changes (Fig 2B), showing no significant differences for segmentation performance  
 166 between disease severity grades (Kruskal-Wallis MP1, MP2, MP3, MV, p-values= $[0.30, 0.20,$   
 167  $0.55, 0.48]$ , n=107).

168



169  
 170 **Figure 2. Lung Segmentation and Lung Volume Analysis.** (A) Image quality vs segmentation performance for manual  
 171 annotations and models (n=107). (B) Manual annotations and model segmentation performance by BPD Severity (mild, moderate,

172 and severe) (n=107). (C) Estimated lung volume from the CNN lung segmentations vs estimated lung volume from manual  
173 segmentations (n=107). (D) Correlation of functional residual capacity per weight vs predicted lung volume per weight (n=27). E)  
174 Correlation of tidal volume per weight vs lung volume per weight (n=32).

175

### 176 *2.3 MRI-based Lung Volume Estimates Correlate Well with Lung Function*

177 In order to approximate lung volume, a 3D reconstruction of the lungs was performed using the  
178 fully segmented lung sequences (*Methods - MRI-Lung Volumetric Features*). We classified the  
179 lung volume further by addressing each lung lobe separately using an automatic clustering  
180 technique allowing for the differentiation of the lung lobe volume ratio.

181 Predicted lung volumes showed a significant correlation (Pearson,  $r=0.964$ ,  $p\text{-value}=6.72\times 10^{-62}$ ,  
182  $n=107$ ) when compared to the lung volumes generated by manual segmentations (Fig. 2C),  
183 indicating that the high segmentation accuracy of the DL ensemble model enabled a robust  
184 downstream estimation of the lung volumes, including sequences with low image quality (Pearson,  
185  $r=0.963$ ,  $p\text{-value}=3.69\times 10^{-11}$ ,  $n=19$ ).

186 A significant correlation was also observed between the estimated lung volume normalized to  
187 bodyweight and parameters derived from infant lung function testing (ILFT by  
188 bodyplethysmography), such as functional residual capacity normalized to bodyweight (Fig. 2D,  
189  $r=0.703$ ,  $p\text{-value}=4.31\times 10^{-5}$ ,  $n=27$ ), and tidal volume normalized to bodyweight (Fig. 2E,  $r=0.606$ ,  
190  $p\text{-value}=2.38\times 10^{-4}$ ,  $n=32$ ), thereby validating the accuracy of the automated volume prediction by  
191 an MRI-independent measure.

192

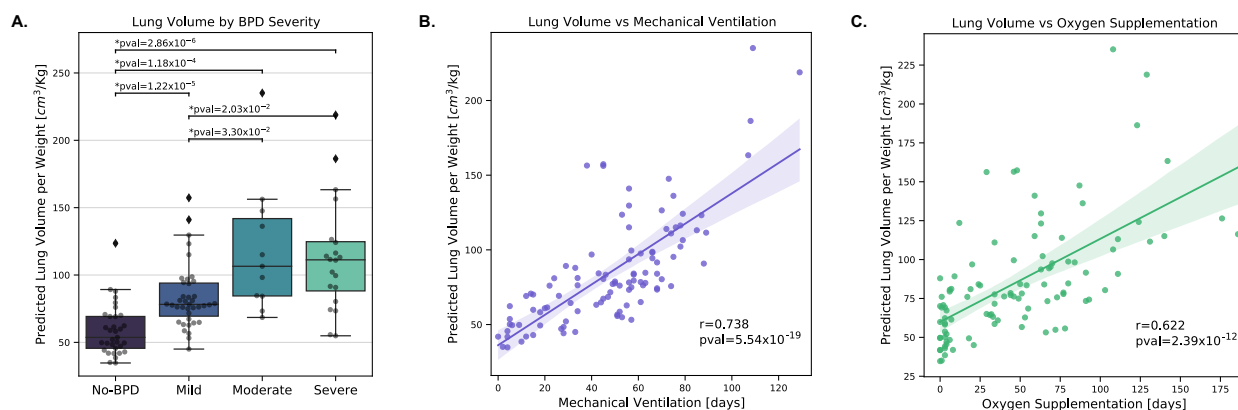
### 193 *2.4 MRI-based Lung Features Demonstrate Predictive Performance for BPD Severity* 194 *Classification*

195 To investigate the clinical value of the MRI-based lung volume in the preterm neonate, we tested  
196 the performance of ML models for BPD severity prediction.

197 First, an exploratory analysis of the relation between the predicted lung volumes and indicators of  
198 BPD severity was performed. We demonstrated significant differences between the lung volume  
199 per bodyweight distributions for different BPD severity grades (Kruskal-Wallis,  $k=43.86$ ,  $p\text{-}$   
200  $\text{value}=1.61\times 10^{-9}$ ,  $n=103$ ), with higher lung volumes corresponding to increased BPD severity  
201 levels (Wilcoxon–Mann–Whitney U test with Bonferroni correction, no-BPD vs mild, no-BPD vs



202 moderate, no-BPD vs severe, mild vs moderate and mild vs severe,  $k=[226, 25, 57, 98, 210]$ ,  $p$ -  
203 values= $[1.22 \times 10^{-5}, 1.18 \times 10^{-4}, 2.86 \times 10^{-6}, 3.30 \times 10^{-2}, 2.03 \times 10^{-2}]$ ,  $n=103$ ; Fig. 3A). Moreover,  
204 significant positive correlations were observed between the predicted lung volume normalized to  
205 bodyweight and the duration of mechanical ventilation (invasive and non-invasive) ( $r=0.738$ ,  $p$ -  
206 value= $5.54 \times 10^{-19}$ ,  $n=103$ ; Fig. 3B), as well as, the duration of oxygen supplementation ( $r=0.622$ ,  
207  $p$ -value= $2.39 \times 10^{-12}$ ,  $n=103$ ; Fig. 3C). These findings thereby confirm the positive correlation  
208 demonstrated for formerly preterm infants at school age showing an increase in functional residual  
209 capacity in higher BPD severity grades and a prolonged history of respiratory support [14], in line  
210 with previous observations reporting elevated lung volumes in severe BPD cases [7,10].  
211

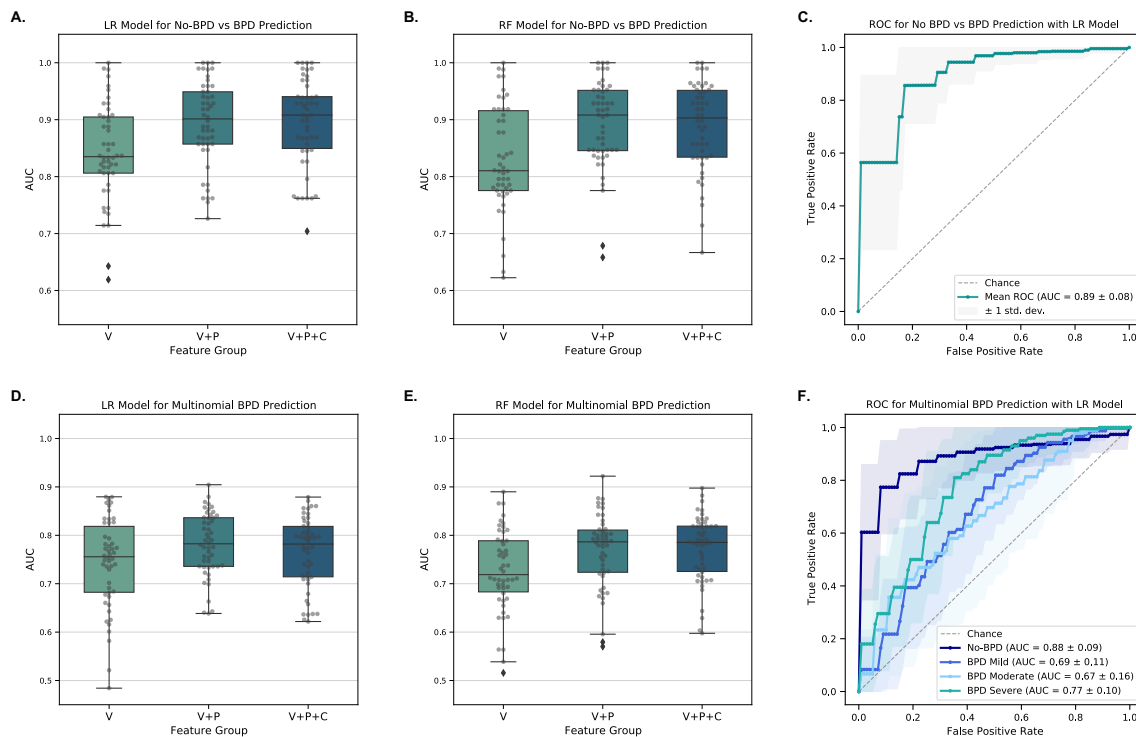


212  
213 **Figure 3. Lung Volume vs BPD Severity Indicators.** (A) Distribution of the predicted lung volume normalized by bodyweight  
214 against BPD severity grades. ( $n=103$ , \* $p$ -values for Wilcoxon–Mann–Whitney U-test with Bonferroni correction). (B) Correlation  
215 of lung volume normalized by bodyweight vs duration of mechanical ventilation ( $n=103$ ). (C) Correlation of lung volume  
216 normalized by bodyweight vs duration of oxygen supplementation ( $n=103$ ).

217  
218 Next, MRI-based lung volumes normalized to bodyweight as well as lung-lobe ratios were used  
219 as explanatory features to predict BPD severity and days of mechanical ventilation using Logistic  
220 Regression (LR), Random Forest (RF), and Poisson regression. Additional clinical variables were  
221 sequentially added as features to investigate the overall predictive performance of the model. Ten  
222 repeated nested cross-validations were performed with five outer and five inner-folds for  
223 performance estimation and hyperparameter tuning. The dataset used for the predictions consisted  
224 of 103 patients with a complete set of explanatory variables. Three prediction scenarios were  
225 explored with our models: i) a binary classification model comparing infants with BPD and  
226 without BPD diagnosis including all severity levels, ii) a scenario with multinomial classification

227 comparing the three different BPD severity levels to infants without BPD, and iii) a regression  
228 scenario predicting the number of days with mechanical ventilation as a continuous measure  
229 indicating BPD severity.

230 For the binary classification, a high classification performance was found with lung-volumetric  
231 features (V) (Table 1) indicating an average AUC performance of 84.31%±8.66% for the LR  
232 Model and 82.66%±10.64% for the RF Model. The inclusion of patient information (P) increased  
233 the AUC classification performance to 89.40%±7.95% for the LR model and 89.17%±7.70% for  
234 the RF model (Fig. 4A-B). Further inclusion of clinical variables (C) showed similar AUC  
235 performances of 89.51%±8.20% for the LR model and 88.89%±7.95% for the RF model (Fig. 4A-  
236 B). F1 scores were higher for the RF models with 85.10% ±6.11% for the highest F1 score using  
237 all feature groups (Table 1). The ROC curves for the binary classification LR model with all feature  
238 sets (V,P,C) also showed an overall stable model for the different train-test splits (Fig. 4C). The  
239 high performance of the classification models that exclusively used volumetric features indicates  
240 their potential as descriptors of BPD severity, which can be improved to an overall high  
241 performance when adding known clinical risk factors, such as gestational age (GA) and growth  
242 [15,16].



243

244 **Figure 4. BPD Prediction Performance.** V= Volumetric features (lung volume per bodyweight, lung lobe volume ratio), P= Patient information (gender, gestational age, birth weight, body Size), C= Clinical Parameters (APGAR 5 min score, early onset infection, steroids treatment). (A) AUC performance of logistic regression models for BPD prediction (BPD vs. no BPD) by feature groups. (B) AUC performance of random forest models for BPD prediction (BPD vs. no BPD) by feature groups. (C) ROC for BPD binomial classification with logistic regression and all feature groups (V+P+C). (D) AUC performance of logistic regression models for multinomial BPD prediction by feature groups. (E) AUC performance of random forest models for multinomial BPD prediction by feature groups. (F) ROC for BPD multinomial classification with logistic regression using all feature groups (V+P+C).

252

253 The multiclass prediction of BPD that only used volumetric features as explanatory variables

254 resulted in an pairwise average AUC of 74.33%±9.30% with the LR model, and a pairwise average

255 AUC of 72.65%±8.49% for the RF model (Fig. 4 D-E, Fig. S1). The best performing model for

256 multinomial classification was the LR model using features from volumetric measurements and

257 patient-information with a pairwise average AUC of 78.06%±6.51% (Table 2). The ROC of the

258 multinomial LR classification using all features showed that the best one class vs all AUC

259 performances can be reached for no-BPD and severe BPD classes (Fig. 4F).

260

261 **Table 2 - Binary and Multinomial BPD Severity Classification Performance**

	Feature Group → ↓ Score	Logistic Regression Model			Random Forest Model		
		V	V + P	V+P+C	V	V + P	V+P+C
<b>Binary: No BPD vs BPD (all severity levels)</b>	AUC [%]	84.31 ±8.66	89.40 ±7.95	89.51 ±8.20	82.66 ±10.64	89.17 ±7.70	88.89 ±7.95
	Weighted F1 Score [%]	81.54 ±8.29	83.71 ±7.35	83.32 ±8.20	80.22 ±8.25	84.71 ±5.78	85.10 ±6.11
<b>Multinomial: No BPD, BPD Mild, Moderate and Severe.</b>	AUC [%]	74.33 ±9.30	78.06 ±6.51	76.44 ±7.35	72.65 ±8.49	76.84 ±7.69	77.22 ±6.92
	Weighted F1 Score [%]	46.75 ±8.60	50.71 ±7.22	52.31 ±8.13	48.97 ±10.37	52.94 ±9.49	53.87 ±8.67

262 V= Volumetric features (lung volume per bodyweight, lung lobe volume ratio), P= Patient information (gender, gestational age,

263 birth weight, body Size), C= Clinical Parameters (APGAR 5 min score, early onset infection, steroids treatment).

264

265 The prediction of days of mechanical respiratory support was evaluated with RF regression, and

266 Poisson regression, next to an analysis by feature groups (Table 3). The RF regression model

267 achieved the lowest mean average prediction error (MAE) with 14.15 ±2.13 days using only

268 volumetric features,  $10.87 \pm 1.61$  days with the (V+P) feature groups, and  $10.78 \pm 1.76$  days for the  
269 (V+P+C) feature groups.

270

271 **Table 3 - Mean Average Prediction Error of Days with Mechanical Respiratory Support**

	Mean Average Error (MAE) in days by Feature Group		
	V	V + P	V+P+C
Random Forest Regression	14.15 $\pm 2.13$	10.87 $\pm 1.61$	10.78 $\pm 1.76$
Poisson Regression	17.40 $\pm 2.93$	13.11 $\pm 2.75$	12.67 $\pm 1.99$

272 V= Volumetric features (lung volume per bodyweight, lung lobe volume ratio), P= Patient information (gender, gestational age,  
273 birth weight, body Size), C= Clinical Parameters (APGAR 5 min score, early onset infection, steroids treatment).

274

### 275 **3. Discussion**

276 In order to improve standardized image assessment and thus diagnostic accuracy for lung disease  
277 in high-risk patient cohorts, we successfully applied DL models to demonstrate the viability for  
278 accurate segmentation of neonatal lung in MRI sequences thereby addressing the most challenging  
279 conditions. The high comparability and low variability of the CNN models in comparison to the  
280 manual annotations, implies the significant potential of the models to overcome the technical  
281 challenges of newborn quiet-breathing MRI, including small volumes, motion artifacts, blurring,  
282 and low image resolutions. Previous studies that performed MRI lung segmentation in neonates  
283 faced limitations in scalability and sensitivity due to smaller cohorts and the use of shape-based  
284 image-segmentation methods. For instance, Heimann et al. [17] used lung shape-appearance  
285 models to perform free-breathing MRI lung segmentation in a cohort of 32 children reporting an  
286 average volumetric overlap of 85% with the annotated ground truth. In adult subjects, MRI  
287 acquisition protocols are improved to meet the need of such automated approaches by the use of  
288 breath holding maneuvers, impossible in the spontaneously breathing infant. By the use of this  
289 technique, Kohlmann et al. achieved a ground truth segmentation overlap of 94% using  
290 thresholding and 3D lung region-growing-based methods with 14 patients [18], whereas other  
291 adult MRI lung segmentation methods reported VDCs in the range of (82%-86%) [19, 20]. In  
292 comparison, our ensemble DL model applied in the most challenging of conditions achieved a  
293 significant performance with an average VDC of 90.2%, while using a multi-center approach.

294 Furthermore, the equivalent segmentation performance in healthy and diseased lungs indicates  
295 persistent accuracy in different lung structure-related image conditions for the proposed 2D  
296 segmentation method even when structural differences apply.

297 As a result of the high segmentation performance of the proposed ensemble model, our automated  
298 pipeline also enabled an accurate downstream estimation of the neonatal lung volumes,  
299 significantly correlated to the corresponding volumes abstracted from manual annotations  
300 ( $r=0.964$ ). Despite the significant technical challenges faced in neonatal lung MRI, our results  
301 thereby reach comparable performance levels reported for adult cohorts, where MRI lung  
302 volumetric estimations were reported with Pearson correlations above 0.98 when comparing  
303 manual vs automatic volume predictions [18]. Furthermore, our MRI-based automatic volume  
304 estimations demonstrate high consistency when compared with lung function measurements,  
305 providing a valuable validation independent from image-based annotations.

306 Next, we show that the predicted lung volume and lung-lobe volume ratio hold potential to reflect  
307 lung health and disease, i.e., BPD. Here, the significant performance of the BPD classification and  
308 regression models that include volumetric features indicates their significant value for disease  
309 prediction (AUC  $0.895\pm 0.082$ ), thereby exceeding previous imaging-based BPD prediction  
310 models that reported AUC binary prediction performances of 0.834-0.858 using ultrasound, and  
311 an AUC of 0.8 when using MRI time-relaxation periods [21]. Hence, our results motivate further  
312 research in the application of automated segmentation and extraction of lung-volumetric features  
313 for monitoring infant lung diseases, potentially while integrating additional lung structural  
314 information [21].

315 Further improvements could be achieved by future studies through the collection of larger and  
316 more diverse datasets of manual annotations to strengthen the generalizability and performance of  
317 the ensemble model, next to the inclusion of different lung pathological conditions. Investigation  
318 of additional volumetric and spatial features that relate to the characterization of the pulmonary  
319 condition of the lung, together with the analysis of MRI-lung volumetric features in longitudinal  
320 approaches could become crucial to inform medical decision making through early prediction of  
321 outcome.

322 Our work herewith contributes to the generation of AI-driven scientific evidence required to  
323 integrate MRI volume-based features as a biomarker to monitor neonatal lung development in  
324 health and disease in daily clinical practice while avoiding radiation exposure. The proposed DL  
325 segmentation method and automated extraction of structural measurements from neonatal lung  
326 MRI, enables the translation of medical expertise to larger scale applications, including its  
327 transferability to health centers that face different expertise levels as well as longitudinal  
328 measurements over prolonged periods of time. Therefore significantly contributing to the  
329 standardization and comparability of critical features in respiratory disease monitoring in  
330 newborns and infants.

331

## 332 4. Materials and Methods

### 333 4.1 Cohort Characteristics

334 A total of 107 preterm infants, gestational age (GA)  $27 \pm 2.13$  weeks, with and without BPD, were  
335 prospectively included in the study from two medical centers after informed parental consent: the  
336 Perinatal Centre LMU Munich (Site 1,  $n=86$ ; EC #195–07) and the Perinatal Centre UKGM  
337 Giessen (Site 2,  $n=21$ ; EC #135–12). In total, 73 of the participants were diagnosed with BPD and  
338 classified into three severity levels: mild ( $n=42$ ), moderate ( $n=11$ ) and severe ( $n=20$ ), according to  
339 the definition by Jobe et. al. [1], 34 participants did not develop BPD. Clinical information on  
340 neonatal health conditions and treatments was also collected from both cohorts, all the clinical  
341 variables were available for 103 patients of the complete cohort (Table 1). Pulmonary function  
342 tests including tidal breathing analysis and bodyplethysmographic functional residual capacity,  
343 were performed for a subgroup of neonates ( $n=32$ ) at 36 weeks GA, according to the guidelines of  
344 the American Thoracic and European Respiratory Society.

345

346 **Table 1 - Clinical Information of the Preterm Neonatal Cohort (N=103)**

Clinical Variable	All (N=103) Average $\pm$ SD	No-BPD (N=33) Average $\pm$ SD	BPD-Mild (N=39) Average $\pm$ SD	BPD-Moderate (N=11) Average $\pm$ SD	BPD-Severe (N=20) Average $\pm$ SD
Gestational Age (weeks)	26.96 $\pm 2.12$	29.09 $\pm 1.43$	26.20 $\pm 1.48$	25.69 $\pm 2.06$	25.62 $\pm 1.43$

Birth Weight (g)	908.25 ± 304.58	1206.21 ±292.39	829.74 ± 182.77	641.82 ± 177.85	716.25 ± 154.31
*Body Size (cm)	34.38 ± 4.02	38.38 ±3.23	33.26 ± 2.92	31.06 ±2.43	31.76 ±2.25
APGAR Score - 5 min	7.71 ± 1.40	8.06 ±1.00	7.87 ±1.10	7.36 ±2.38	7.00 ±1.59
†Early Onset Infection	No (N=80), Yes (N=23)	No (N=29), Yes (N=4)	No (N=30), Yes (N=9)	No (N=7), Yes (N=4)	No (N=14), Yes (N=6)
Administration of post-natal corticosteroids	No (N=61), Yes (N=42)	No (N=28), Yes (N=5)	No (N=22), Yes (N=17)	No (N=6), Yes (N=5)	No (N=15), Yes (N=5)
Oxygen Supplementation (days)	47.30 ± 43.18	5.18 ±7.72	45.56 ±21.23	81.55 ±30.96	101.35 ±40.78
Mechanical Ventilation (days; invasive and non-invasive)	48.22 ±26.93	19.91 ±15.62	52.51 ±13.74	66.55 ±19.20	76.50 ±21.08

347 \*Linear BMI imputation performed for missing body sizes. †Early Onset Infection as defined by [22].

348

#### 349 4.2 Imaging and Segmentation Protocols

350 Preterm infants underwent MRI near term age, i.e., at approximately 36 weeks GA. T2-weighted  
351 lung MRI sequences were acquired in unsedated spontaneous sleep for the Perinatal Centre LMU  
352 Munich cohort, and under light sedation with chloral hydrate (30-40 mg/kg administered orally)  
353 for Perinatal Centre UKGM Giessen.

354 Axial images were acquired with a T2-weighted half-Fourier-acquired single-shot turbo spin echo  
355 (HASTE) protocol for lung structural assessment. An ECG-triggered 2D multi-slice single-shot  
356 fast spin-echo sequence with an echo time (TE) of 57 ms was used; the repetition time was set to  
357 2 RR intervals. The spatial resolution was 1.3×1.9 mm<sup>2</sup> in plane with a slice thickness of 4 mm  
358 and 0.4 mm slice gap. Parallel imaging with an acceleration factor of 2 was applied and 2 averages  
359 were acquired for each slice.

360 In sum, a total of 107 MRI sequences with 2,165 axial images, with a resolution of 256×192 pixels,  
361 were acquired using 3T MRI scanners (Siemens Skyra for the Perinatal Centre LMU Munich and  
362 Siemens Verio for Perinatal Centre UKGM Giessen). Pseudonymization of image and clinical  
363 information was performed to guarantee a blinded analysis.

364 Manual lung segmentation of the MRI sequences was performed independently by three  
365 physicians, with different training levels (one radiologist and two late-stage image analysis trained

366 medical students). The software ITK-SNAP [23] was used to collect the manual segmentations.  
367 Image quality of the sequences was rated by a fourth independent radiologist. To remove  
368 unnecessary background, we first identify the centroid of all pixels that are above the 5% intensity  
369 quantile threshold across all slices and then crop all slices to a square of 128×128 pixels centered  
370 at the centroid.

371

### 372 *4.3 Deep-Learning MRI Lung Segmentation Model*

373 The proposed 2D lung segmentation DL models are based on the U-Net neural network (NN)  
374 architecture [24]. U-Net models produce a latent representation of the image by processing it  
375 through a set of convolutional layers in a contracting path and then processing the features through  
376 an expansion path of up-convolutional layers, with skipped connections at each level, returning a  
377 high-resolution binary pixel-wise segmentation map of the image. Our U-Net architecture has 4  
378 down and 4 up convolutional blocks and a fifth intermediate convolutional block, batch  
379 normalization was included after every building block of the U-Net and a Dropout Layer (dropout  
380 rate=0.1) after each encoding block was added. Detailed architecture parameters are available in  
381 Table S3 and in the code-repository (*Section 4.6*). The Instant-DL framework, which is designed  
382 to efficiently train U-Net segmentation models for medical imaging applications, was adapted for  
383 our study [25]. Hyperparameters were optimized using grid search (Table S2) for three randomly  
384 selected leave-one-patient-out models in a 4-fold cross-validation scheme, the best performances  
385 were achieved with 300 training epochs, 0.001 learning rate, using a binary cross-entropy loss for  
386 the NN optimization, and applying image augmentations with 0.1 random zoom, 0.1 translations  
387 and up to 22.5° random rotations.

388 The dice-coefficient, defined as  $DC = \frac{2pg+1}{p^2+g^2+1}$ , with  $p$  being the predicted positive-class pixels  
389 and  $g$  being the ground truth pixels [26], was used as the metric for evaluation of the model  
390 performance. Ground truths were generated with the annotations from a single physician, for the  
391 physician-based models, and with majority voting from all the physicians, for the integrated model.  
392 Optimization was performed with Adam [27].

393 Using a leave-one-patient-out cross-validation scheme, a set of  $k$  models were trained exclusively  
394 with the sequences from Site 1, that is, for each  $k^{\text{th}}$  model, the data of the  $k^{\text{th}}$  participant is used



395 only for validation. In addition, MRI sequences from study Site 2 were used exclusively for  
396 validation of a MV model trained with all the sequences from study Site 1. Lung segmentation  
397 accuracy was measured using the volumetric dice coefficient (VDC), where pixels from all the  
398 slices in the MRI sequence are aggregated and evaluated for segmentation validity. Resulting VDC  
399 scores of the models and manual annotations for each sequence can be found on Table S1.

400

#### 401 *4.4 MRI-Lung Volumetric Features*

402 A 3D volumetric representation of the lung was created using the segmented lung regions and the  
403 DICOM pixel-spacing and patient orientation metadata from the MRI 2D sequences. The volume  
404 was generated with voxels that considered both the distance to the neighboring pixels in the 2D  
405 slice (dx, dy), as well as the slice thickness and the space between the slices (dz). The total lung  
406 volume was calculated by adding the individual voxel volumes of all the segmented pixels in the  
407 MRI sequence.

408 In addition, the lung was divided between the left and right lobes using a two-step algorithm  
409 involving a K-means (K=2) clustering with further refinement of the class labels using a soft-  
410 margin SVM classifier (penalty=0.001). For the K-means algorithm, we initialized the centroids  
411 using the *k-means++* initialization [28] and used a weighted Euclidean distance favoring the x and  
412 y dimensions for improved left-right lung lobe clustering with weights of (1, 1, 0.1) for x, y, and  
413 z coordinates. Using k-means annotations as labels, the SVM algorithm was applied iteratively  
414 updating the voxel labels until convergence of the resulting silhouette score; methods are available  
415 in the code-repository (*Section 4.6*). Once the voxels were classified in the left or right lung lobes,  
416 the lung lobe volume ratio was calculated by finding the volume of each lobe and then by dividing  
417 the larger over the smaller lobe volume as an asymmetry indicator. The resulting MRI-based lung  
418 volume features are available in Supplementary Table S1.

419

#### 420 *4.5 BPD Severity Prediction Models*

421 A regression analysis to predict the severity of BPD was performed for which explanatory  
422 variables were grouped in three categories, MRI-based lung volume features (V) (lung volume per  
423 birth weight and the lung lobe volume ratio), patient-related (P) features (gestational age, birth

424 weight, body size, gender) and also clinical parameters (C) (APGAR score - 5 min, early onset  
425 infection, steroid treatment).

426 Random Forest (RF) [29] and Logistic regression models [30] were trained to perform binomial  
427 (BPD vs. No BPD) and multinomial (no BPD, mild, moderate, and severe BPD) classification. A  
428 nested cross-validation scheme was implemented to find the best hyperparameters using grid-  
429 search (Table S2), the average performance of the model was estimated with multiple repetitions  
430 of the nested cross-validation scheme (10 times with different random seeds). A stratified 5-fold  
431 train-test split was used for both the inner and outer loops of each nested cross-validation. For the  
432 Logistic regression model, features were standardized removing the mean and scaling to unit  
433 variance of the training set. Additional regression models, Poisson and RF, to predict the number  
434 of days with required respiratory support (adding invasive and non-invasive days) were trained  
435 using the same nested cross-validation scheme.

436

#### 437 *4.6 Data and Source Code Availability*

438 Source code of models for lung segmentation, 3D volume-feature estimations, and regression  
439 models can be found at <https://github.com/SchubertLab/NeoLUNet>.

440 Resulting weights of the U-Net models used for BPD prediction will be made available at  
441 (<https://zenodo.org>).

442

#### 4435. **Acknowledgments**

444 We sincerely thank the patients and their families of the AIRR study cohort for their significant  
445 contribution to the study by providing the samples. The present study was supported by the Young  
446 Investigator Grant NWG VH-NG-829 by the Helmholtz Foundation and the Helmholtz Zentrum  
447 Muenchen, Germany and the German Center for Lung Research (DZL) - German Ministry of  
448 Education and Health (BMBF).

449 B.M. and A.C. are supported by the Helmholtz Association under the joint research school Munich  
450 School for Data Science - MUDS. B.S. acknowledges financial support by the Postdoctoral  
451 Fellowship Program of the Helmholtz Zentrum München.

452

453 **REFERENCES**

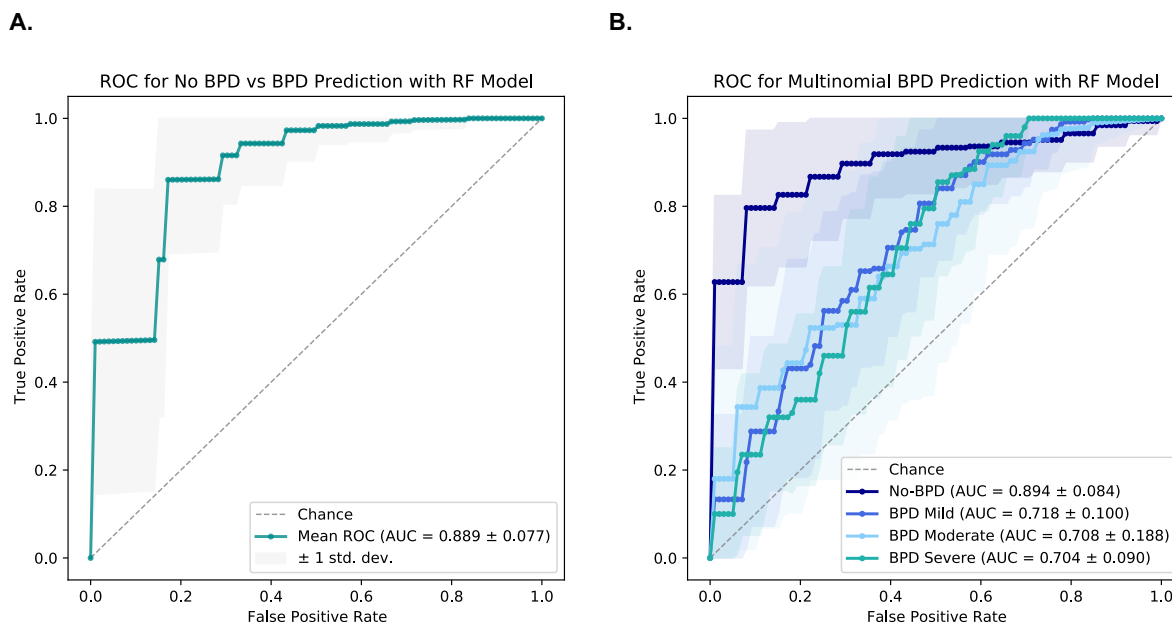
- 454 [1] A. H. Jobe and E. Bancalari, “Bronchopulmonary dysplasia,” *Am. J. Respir. Crit. Care*  
455 *Med.*, vol. 163, no. 7, pp. 1723–1729, Jun. 2001.
- 456 [2] A. H. Jobe, “Mechanisms of Lung Injury and Bronchopulmonary Dysplasia,” *Am. J.*  
457 *Perinatol.*, vol. 33, no. 11, pp. 1076–1078, Sep. 2016.
- 458 [3] T. R. Kalikkot, M. C. Guaman, and B. Shivanna, “Bronchopulmonary dysplasia: A review  
459 of pathogenesis and pathophysiology,” *Respir. Med.*, vol. 132, Nov. 2017, doi:  
460 10.1016/j.rmed.2017.10.014.
- 461 [4] G. R. Washko, “Diagnostic imaging in COPD,” *Semin. Respir. Crit. Care Med.*, vol. 31,  
462 no. 3, pp. 276–285, Jun. 2010.
- 463 [5] C. May, M. Prendergast, S. Salman, G. F. Rafferty, and A. Greenough, “Chest radiograph  
464 thoracic areas and lung volumes in infants developing bronchopulmonary dysplasia,” *Pediatr.*  
465 *Pulmonol.*, vol. 44, no. 1, pp. 80–85, Jan. 2009.
- 466 [6] B. Loi et al., “Lung Ultrasound to Monitor Extremely Preterm Infants and Predict  
467 Bronchopulmonary Dysplasia. A Multicenter Longitudinal Cohort Study,” *Am. J. Respir. Crit.*  
468 *Care Med.*, vol. 203, no. 11, pp. 1398–1409, Jun. 2021.
- 469 [7] L. M. Yoder et al., “Elevated lung volumes in neonates with bronchopulmonary dysplasia  
470 measured via MRI,” *Pediatric Pulmonology*. 2019, doi: 10.1002/ppul.24378.
- 471 [8] P. J. Critser et al., “Cardiac Magnetic Resonance Imaging Evaluation of Neonatal  
472 Bronchopulmonary Dysplasia-associated Pulmonary Hypertension,” *Am. J. Respir. Crit. Care*  
473 *Med.*, vol. 201, no. 1, pp. 73–82, Jan. 2020.
- 474 [9] D. C. Dean 3rd et al., “Pediatric neuroimaging using magnetic resonance imaging during  
475 non-sedated sleep,” *Pediatr. Radiol.*, vol. 44, no. 1, pp. 64–72, Jan. 2014.
- 476 [10] L. L. Walkup et al., “Quantitative Magnetic Resonance Imaging of Bronchopulmonary  
477 Dysplasia in the Neonatal Intensive Care Unit Environment,” *Am. J. Respir. Crit. Care Med.*, vol.  
478 192, no. 10, pp. 1215–1222, Nov. 2015.

- 479 [11] M. H. Hesamian, W. Jia, X. He, and P. Kennedy, “Deep Learning Techniques for Medical  
480 Image Segmentation: Achievements and Challenges,” *J. Digit. Imaging*, vol. 32, no. 4, pp. 582–  
481 596, Aug. 2019.
- 482 [12] J. Islam and Y. Zhang, “Towards Robust Lung Segmentation in Chest Radiographs with  
483 Deep Learning,” *arXiv [cs.CV]*, Nov. 30, 2018.
- 484 [13] B. Ait Skourt, A. El Hassani, and A. Majda, “Lung CT Image Segmentation Using Deep  
485 Neural Networks,” *Procedia Comput. Sci.*, vol. 127, pp. 109–113, Jan. 2018.
- 486 [14] J. S. Landry, T. Chan, L. Lands, and D. Menzies, “Long-term impact of bronchopulmonary  
487 dysplasia on pulmonary function,” *Can. Respir. J.*, vol. 18, no. 5, pp. 265–270, Sep. 2011.
- 488 [15] A. P. Popova, “Mechanisms of bronchopulmonary dysplasia,” *J. Cell Commun. Signal.*,  
489 vol. 7, no. 2, p. 119, Jun. 2013, Accessed: May 10, 2021. [Online].
- 490 [16] B. Thébaud et al., “Bronchopulmonary dysplasia,” *Nature reviews. Disease primers*, vol.  
491 5, no. 1, Nov. 2019, doi: 10.1038/s41572-019-0127-7.
- 492 [17] T. Heimann, M. Eichinger, G. Bauman, A. Bischoff, M. Puderbach, and H.-P. Meinzer,  
493 “Automated scoring of regional lung perfusion in children from contrast enhanced 3D MRI,”  
494 *Medical Imaging 2012: Computer-Aided Diagnosis*. 2012, doi: 10.1117/12.911946.
- 495 [18] P. Kohlmann et al., “Automatic lung segmentation method for MRI-based lung perfusion  
496 studies of patients with chronic obstructive pulmonary disease,” *Int. J. Comput. Assist. Radiol.*  
497 *Surg.*, vol. 10, no. 4, pp. 403–417, Apr. 2015.
- 498 [19] W. F. Sensakovic, S. G. Armato 3rd, A. Starkey, and P. Caligiuri, “Automated lung  
499 segmentation of diseased and artifact-corrupted magnetic resonance sections,” *Med. Phys.*, vol.  
500 33, no. 9, pp. 3085–3093, Sep. 2006.
- 501 [20] T. Böttger et al., “Implementation and evaluation of a new workflow for registration and  
502 segmentation of pulmonary MRI data for regional lung perfusion assessment,” *Phys. Med. Biol.*,  
503 vol. 52, no. 5, pp. 1261–1275, Mar. 2007.
- 504 [21] K. Förster et al., “Altered relaxation times in MRI indicate bronchopulmonary dysplasia,”  
505 *Thorax*, vol. 75, no. 2, pp. 184–187, Feb. 2020.

- 506 [22] M. P. Sherman, B. W. Goetzman, C. E. Ahlfors, and R. P. Wennberg, “Tracheal  
507 Aspiration and Its Clinical Correlates in the Diagnosis of Congenital Pneumonia,” *Pediatrics*, vol.  
508 65, no. 2. pp. 258–263, 1980. doi: 10.1542/peds.65.2.258.
- 509 [23] P. A. Yushkevich et al., “User-guided 3D active contour segmentation of anatomical  
510 structures: significantly improved efficiency and reliability,” *Neuroimage*, vol. 31, no. 3, pp. 1116–  
511 1128, Jul. 2006.
- 512 [24] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional Networks for Biomedical  
513 Image Segmentation,” *Lecture Notes in Computer Science*. pp. 234–241, 2015, doi: 10.1007/978-  
514 3-319-24574-4\_28.
- 515 [25] D. Waibel, S. S. Boushehri, and C. Marr, “InstantDL - An easy-to-use deep learning  
516 pipeline for image segmentation and classification.” doi: 10.1101/2020.06.22.164103.
- 517 [26] F. Milletari, N. Navab, and S.-A. Ahmadi, “V-Net: Fully Convolutional Neural Networks  
518 for Volumetric Medical Image Segmentation,” 2016 Fourth International Conference on 3D  
519 Vision (3DV). 2016, doi: 10.1109/3dv.2016.79.
- 520 [27] Diederik P. Kingma and Jimmy Ba, “ADAM: A method for stochastic optimization”. 3rd  
521 International Conference for Learning Representations, 2015.
- 522 [28] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” *J. Mach. Learn. Res.*, vol.  
523 12, no. 85, pp. 2825–2830, 2011.
- 524 [29] P. Geurts, D. Ernst, and L. Wehenkel, “Extremely randomized trees,” *Machine Learning*,  
525 vol. 63, no. 1. pp. 3–42, 2006. doi: 10.1007/s10994-006-6226-1.
- 526 [30] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer Verlag, 2006.  
527  
528

529 **Supplementary Material**

530 *Figure S1: Receiver operator characteristics for Random Forest models.*



531  
 532 **Figure S1. ROC for Binary and Multinomial RF Classification Models.** (A) ROC for BPD binomial classification with random  
 533 forest models and all feature groups (V+P+C). (B) ROC for BPD multinomial classification with random forest models using all  
 534 feature groups (V+P+C).

535  
 536 *Table S1: De-identified resulting physician and model VDC segmentation performances,*  
 537 *volumetric predictions and lung-lobe volume ratio.*

538 *Table S2: Grid Search Parameters.*

	<b>Hyperparameters Evaluated with Grid Search</b>
<b>U-Net for Lung Segmentation</b>	Epochs = [100, 200, 300, 400] Loss Functions = Mean squared error, Binary cross-entropy, Dice-loss. Learning Rate = [0.001, 0.0001] Augmentation = with augmentations (0.1 random zoom, 0.1 translations and up to 22.5° random rotations) or without augmentations.

<b>Logistic Regression</b>	Penalty= L1, L2  C = log space sampled array (n=10 points). <i>np.logspace(-4, 1, 10, endpoint=True)</i>  Grid search scoring = F1
<b>Random Forest</b>	Max. Depth = [3,4,5,6,7,8,9,10]  Grid search scoring = F1

539

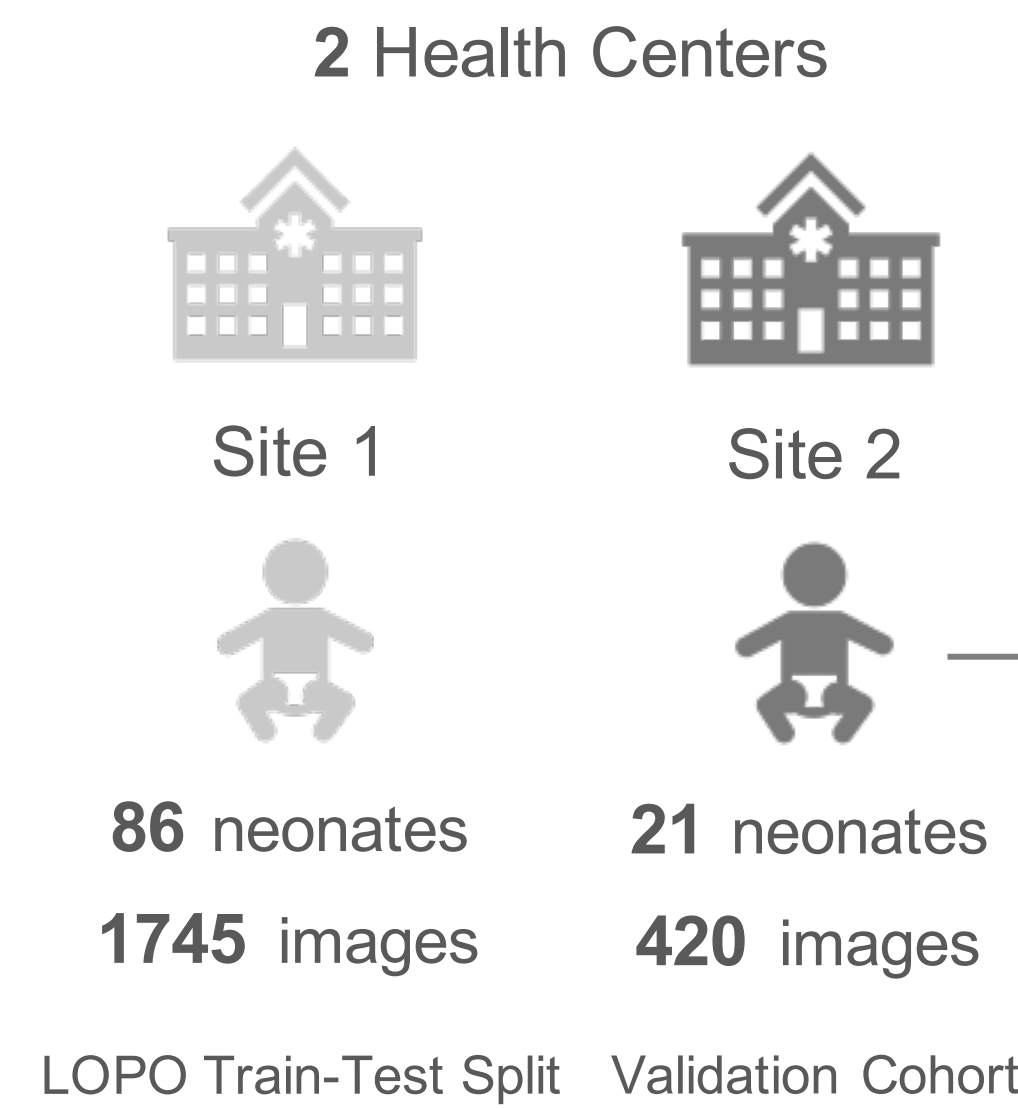
540 *Table S3: U-Net Architecture Parameters*

	<b>Block Description</b>
Convolutional Block 1 (CNN-1)	CNN Filters = 64, Kernel = 3×3 Activation function = LeakyReLU (alpha=0.2) Batch Normalization  CNN Filters = 64, Kernel = 3×3 Activation function = LeakyReLU (alpha=0.2) Batch Normalization
Dropout and Max Pooling	Dropout fraction = 0.1 Max. Pooling Kernel Size = 2×2
Convolutional Block 2 (CNN-2)	CNN Filters = 128, Kernel = 3×3 Activation function = LeakyReLU (alpha=0.2) Batch Normalization  CNN Filters = 128, Kernel = 3×3 Activation function = LeakyReLU (alpha=0.2) Batch Normalization
Dropout and Max Pooling	Dropout fraction = 0.1 Max. Pooling Kernel Size = 2×2
Convolutional Block 3 (CNN-3)	CNN Filters = 256, Kernel = 3×3 Activation function = LeakyReLU (alpha=0.2) Batch Normalization  CNN Filters = 256, Kernel = 3×3 Activation function = LeakyReLU (alpha=0.2) Batch Normalization
Dropout and Max Pooling	Dropout fraction = 0.1 Max. Pooling Kernel Size = 2×2
Convolutional Block 4 (CNN-4)	CNN Filters = 512, Kernel = 3×3 Activation function = LeakyReLU (alpha=0.2) Batch Normalization  CNN Filters = 512, Kernel = 3×3 Activation function = LeakyReLU (alpha=0.2) Batch Normalization
Dropout and Max Pooling	Dropout fraction = 0.1 Max. Pooling Kernel Size = 2×2

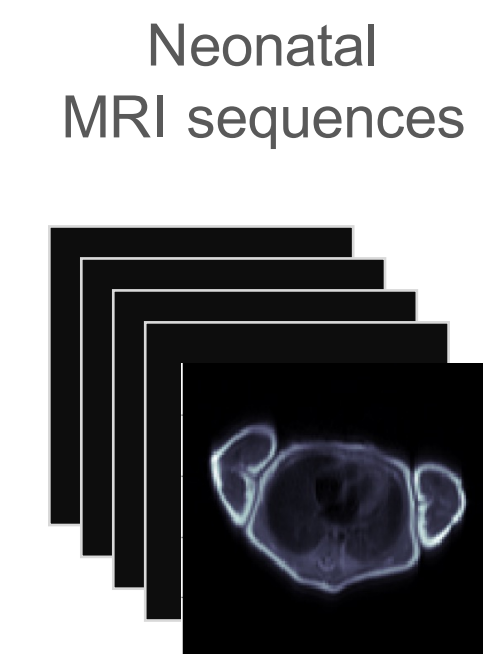
Convolutional Block 5 (CNN-5)	<p>CNN Filters = 1024, Kernel = 3×3 Activation function = LeakyReLU (alpha=0.2) Batch Normalization</p> <p>CNN Filters = 1024, Kernel = 3×3 Activation function = LeakyReLU (alpha=0.2) Batch Normalization</p>
Dropout and Max Pooling	<p>Dropout fraction = 0.1 Max. Pooling Kernel Size = 2×2</p>
Convolutional Block 6 (CNN-6)	<p>Up-6 Features: feature size= 512, up-sampling-kernel = 2×2 Concatenation: CNN-4 Features + Up-6 Features</p> <p>CNN Filters = 512, Kernel = 3×3 Activation function = LeakyReLU (alpha=0.1) Batch Normalization</p> <p>CNN Filters = 512, Kernel = 3×3 Activation function = LeakyReLU (alpha=0.1) Batch Normalization</p>
Convolutional Block 7 (CNN-7)	<p>Up-7 Features: feature size= 256, up-sampling-kernel = 2×2 Concatenation: CNN-3 Features + Up-7 Features</p> <p>CNN Filters = 256, Kernel = 3×3 Activation function = LeakyReLU (alpha=0.1) Batch Normalization</p> <p>CNN Filters = 256, Kernel = 3×3 Activation function = LeakyReLU (alpha=0.1) Batch Normalization</p>
Convolutional Block 8 (CNN-8)	<p>Up-8 Features: feature size= 128, up-sampling-kernel = 2×2 Concatenation: CNN-2 Features + Up-8 Features</p> <p>CNN Filters = 128, Kernel = 3×3 Activation function = LeakyReLU (alpha=0.1) Batch Normalization</p> <p>CNN Filters = 128, Kernel = 3×3 Activation function = LeakyReLU (alpha=0.1) Batch Normalization</p>
Convolutional Block 9 (CNN-9)	<p>Up-9 Features: feature size= 64, up-sampling-kernel = 2×2 Concatenation: CNN-1 Features + Up-9 Features</p> <p>CNN Filters = 64, Kernel = 3×3 Activation function = LeakyReLU (alpha=0.1) Batch Normalization</p> <p>CNN Filters = 64, Kernel = 3×3 Activation function = LeakyReLU (alpha=0.1) Batch Normalization</p>
Output Layer	Activation function = Sigmoid



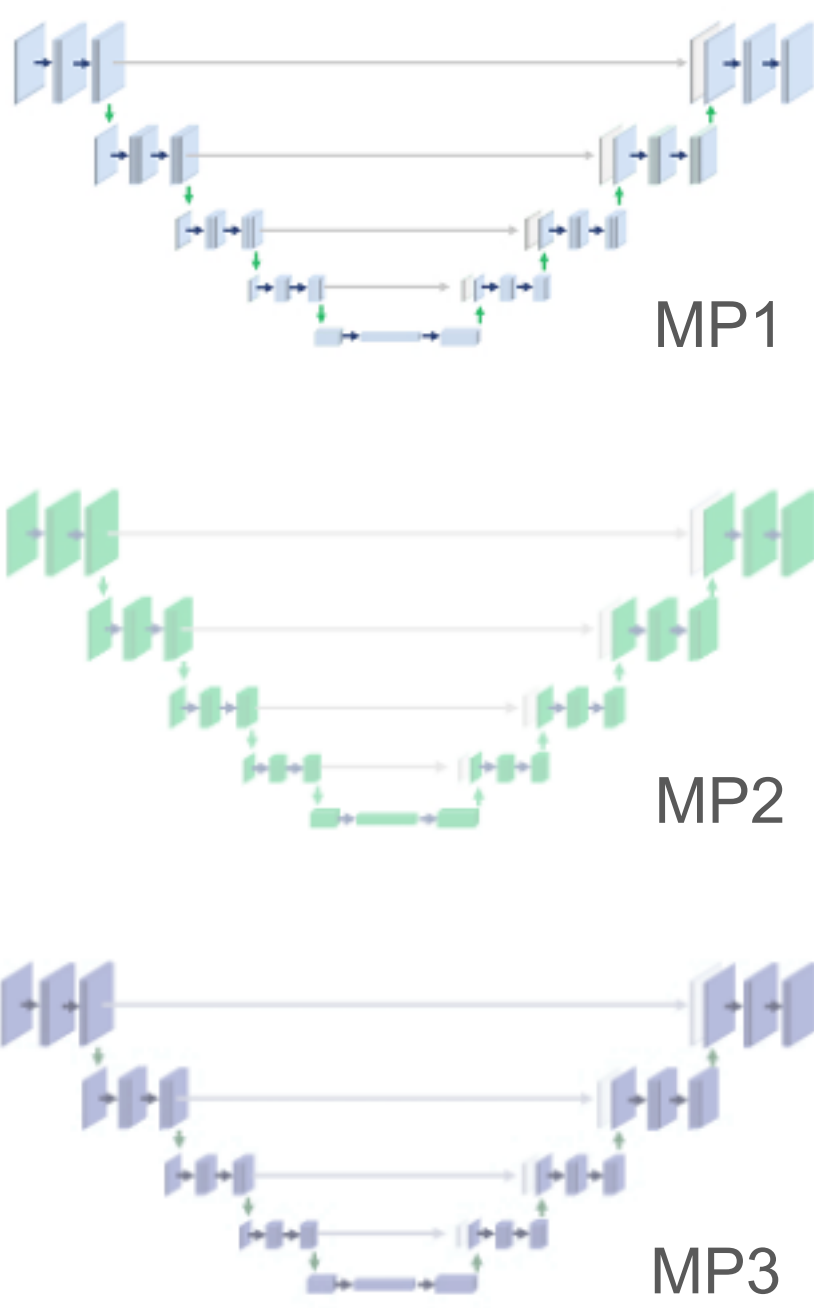
### A. Neonatal Image Acquisition



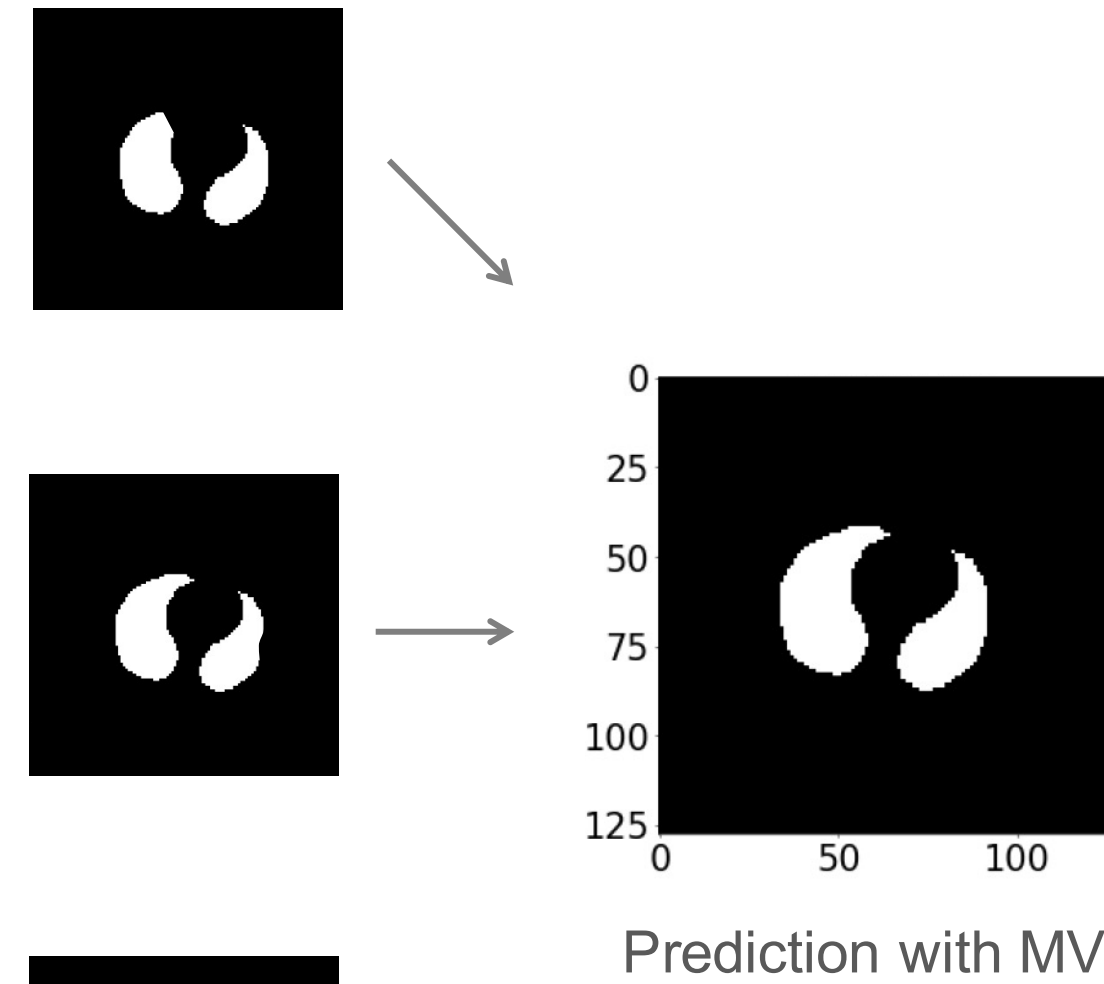
### B. Image Annotation by Physicians



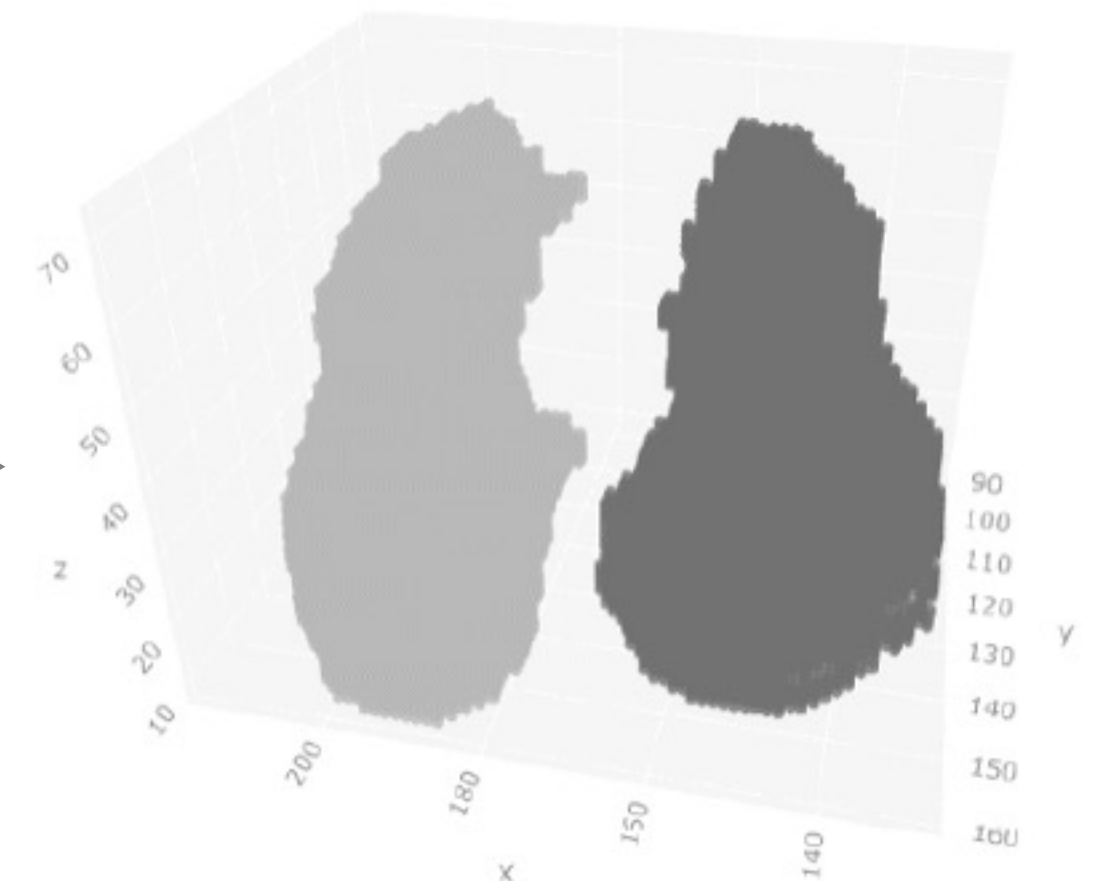
### C. U-Net Models for Lung Segmentation



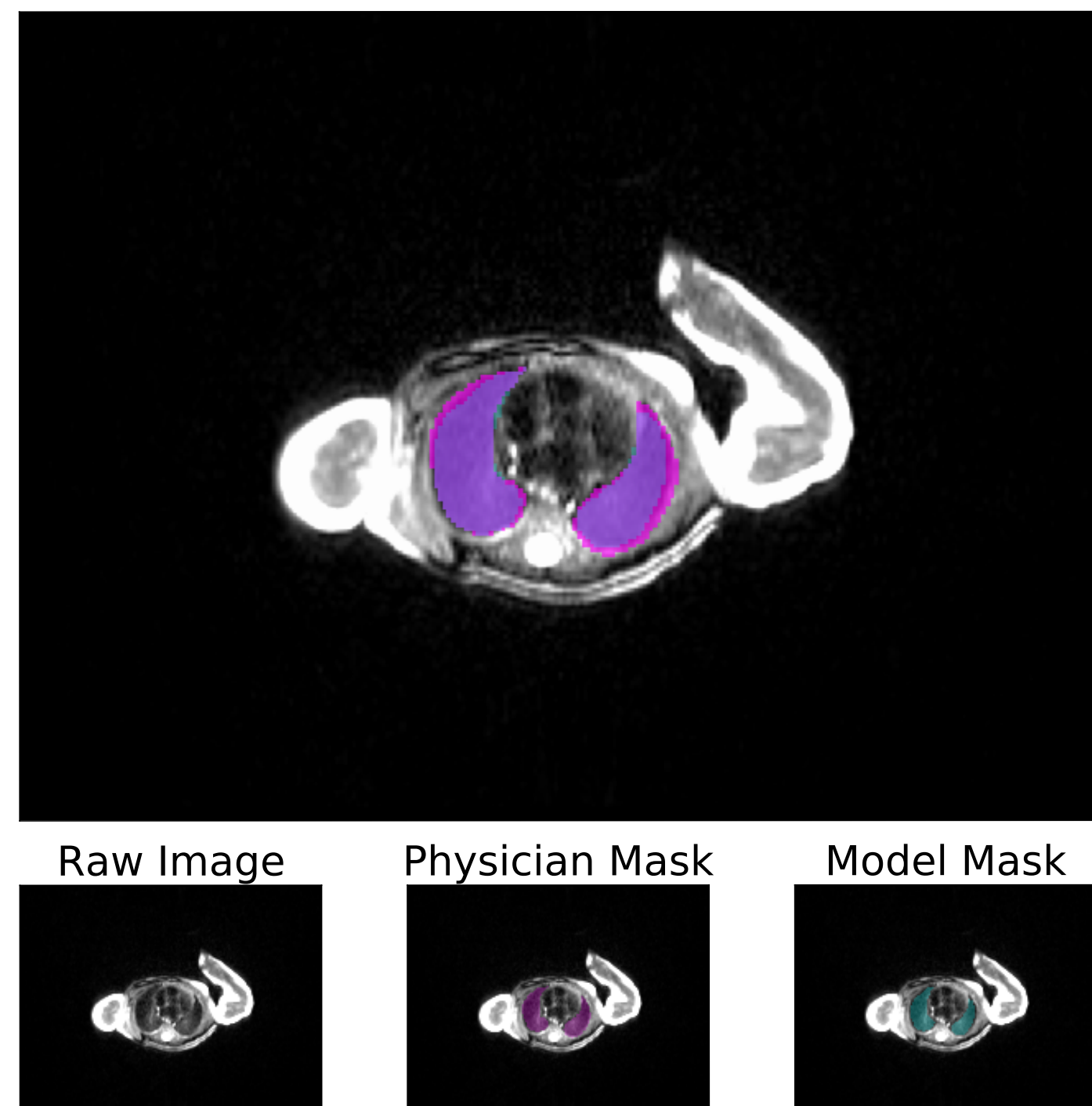
### D. Neural Network Model Ensemble with Majority Voting (MV)



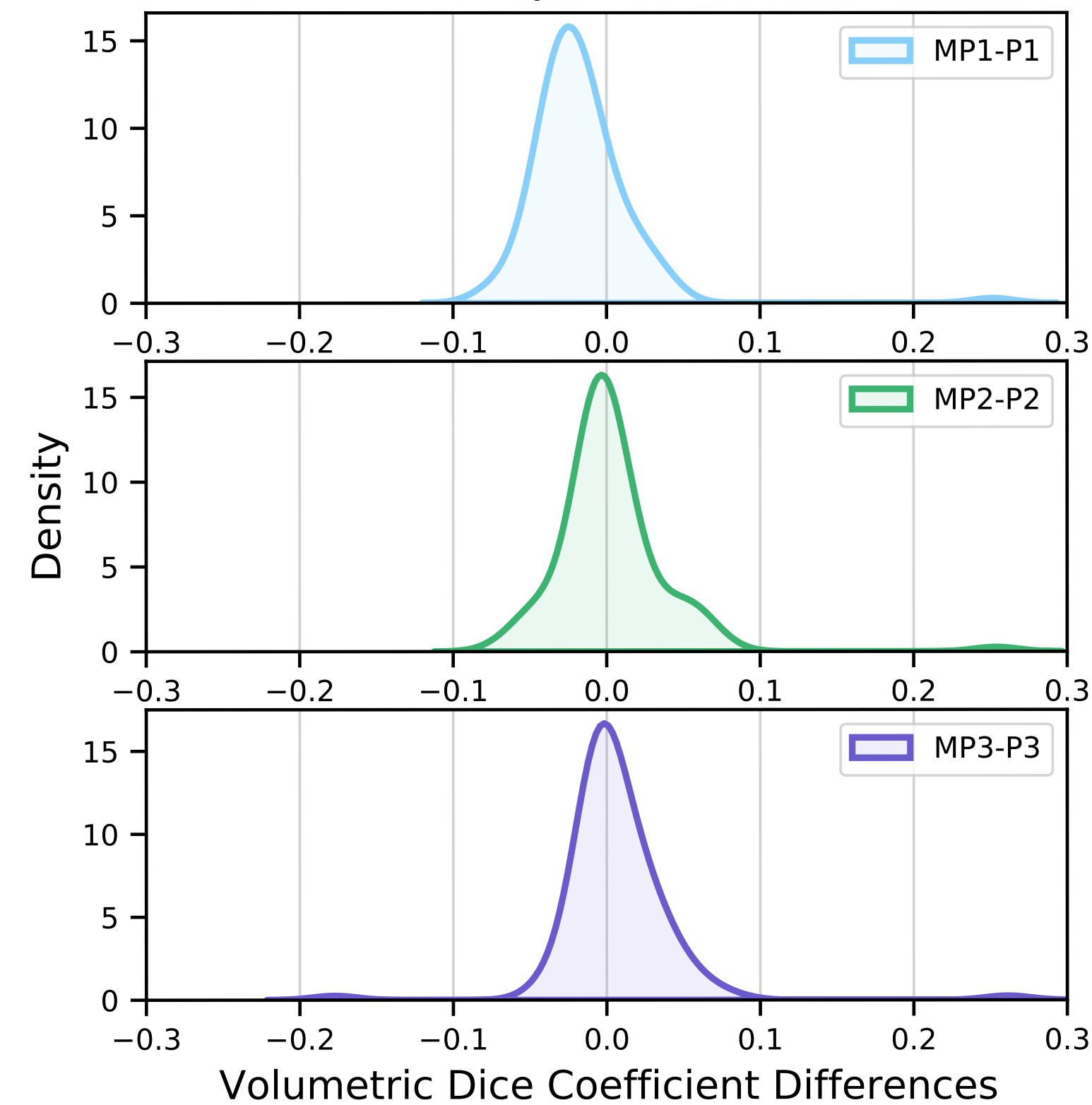
### E. Lung 3D Reconstruction and Feature Extraction



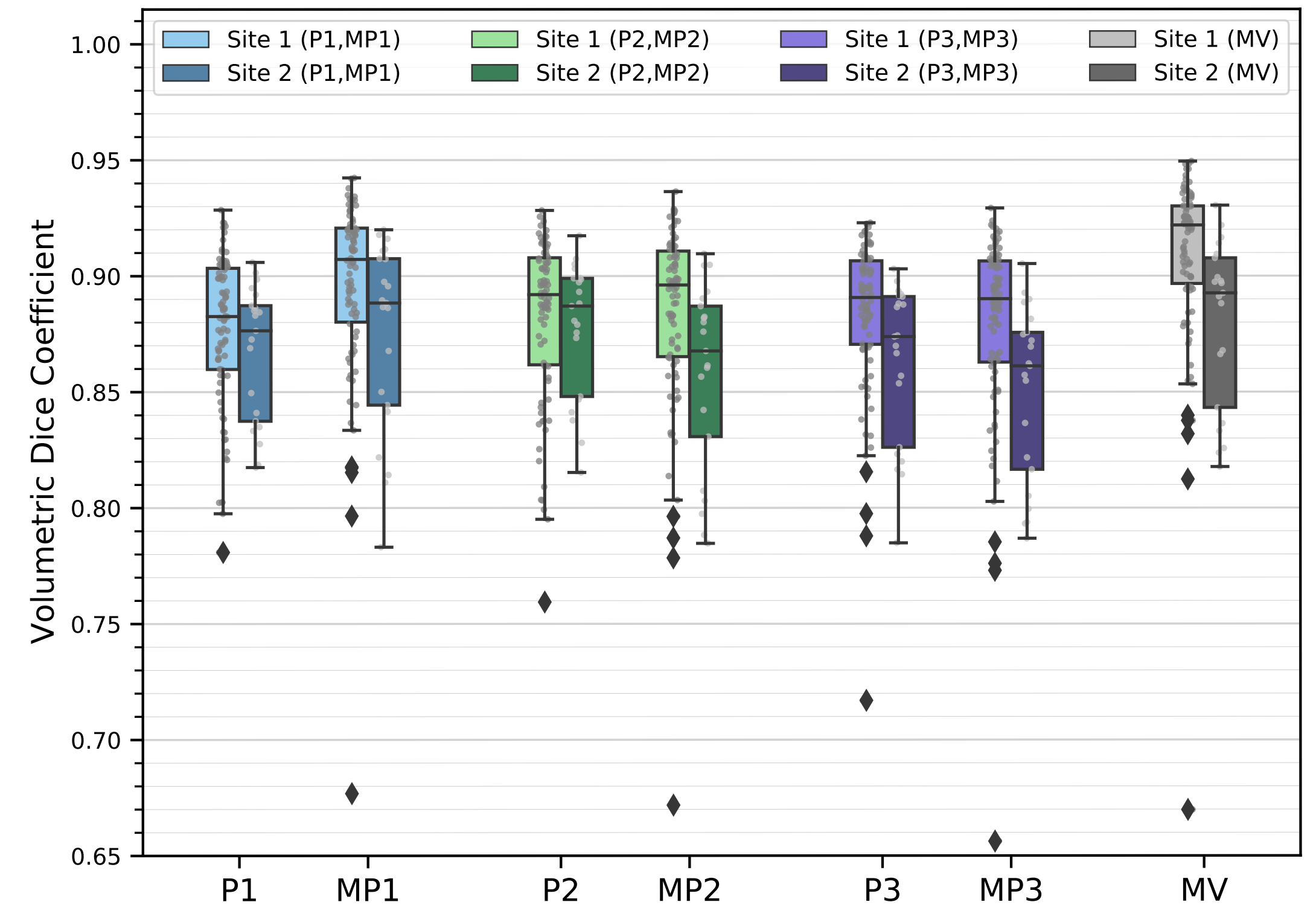
### F. Neonatal Lung Segmentation Sample



### G. Paired Model-Physician VDC Differences

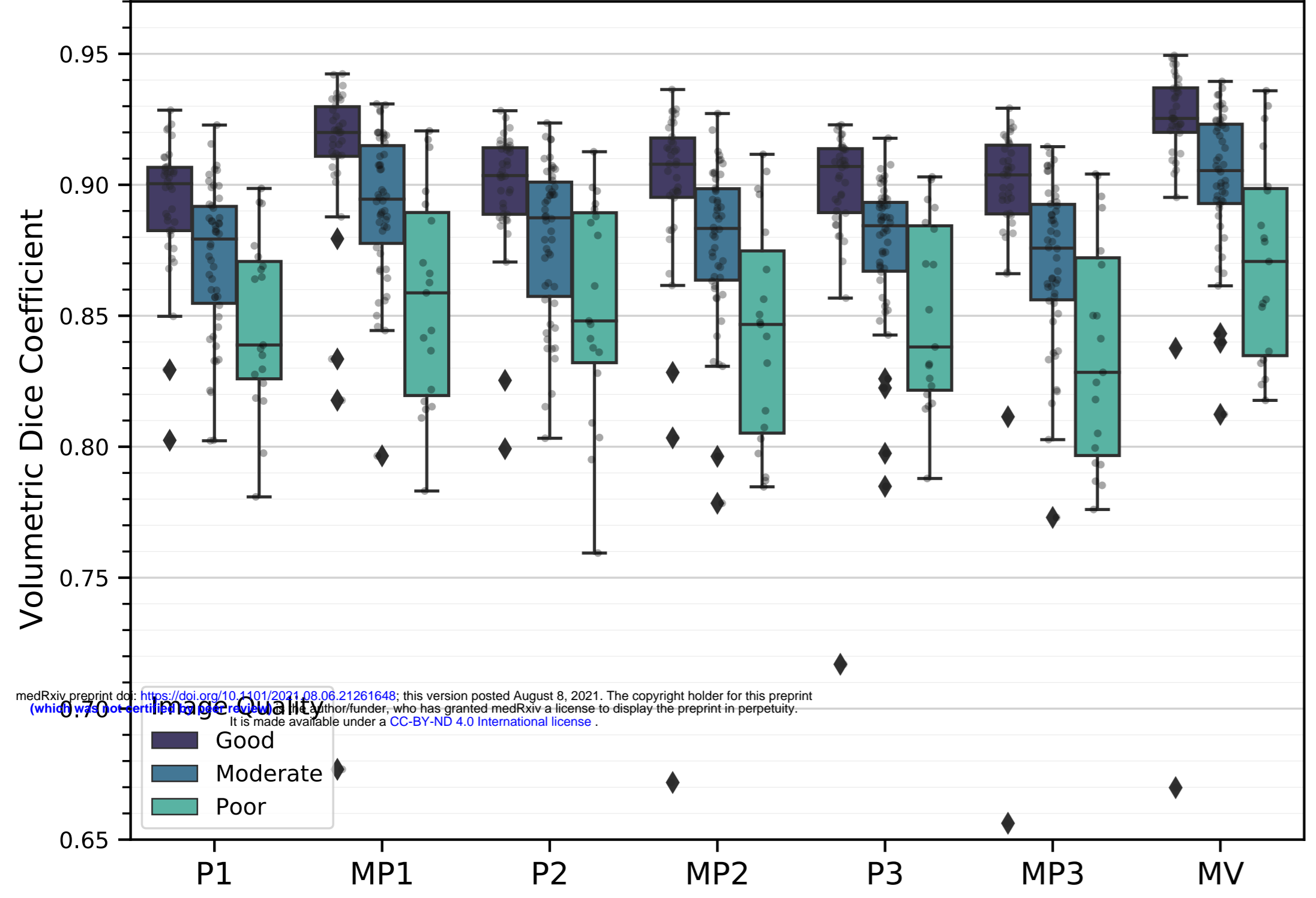


### H. Segmentation Performance by Cohort

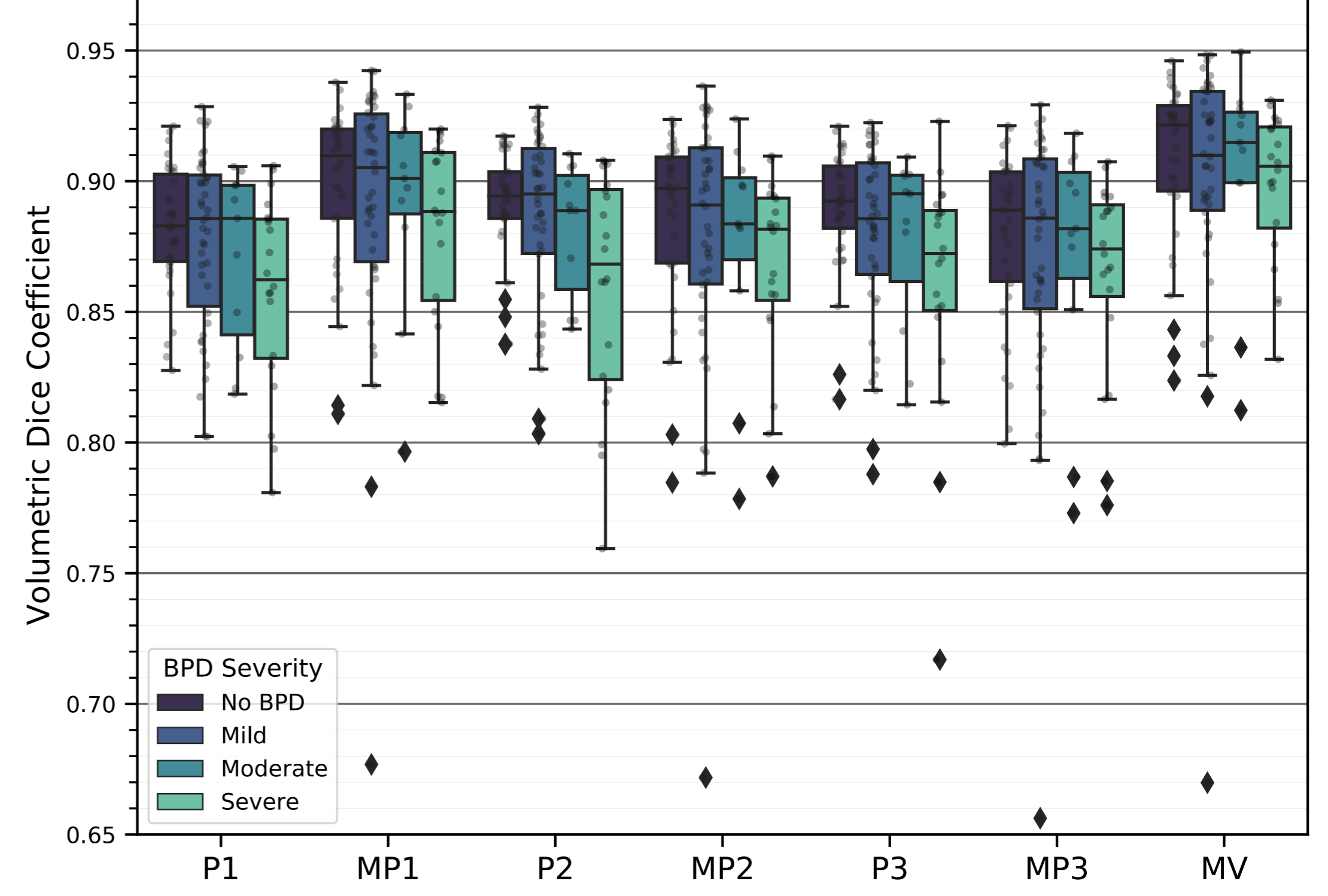
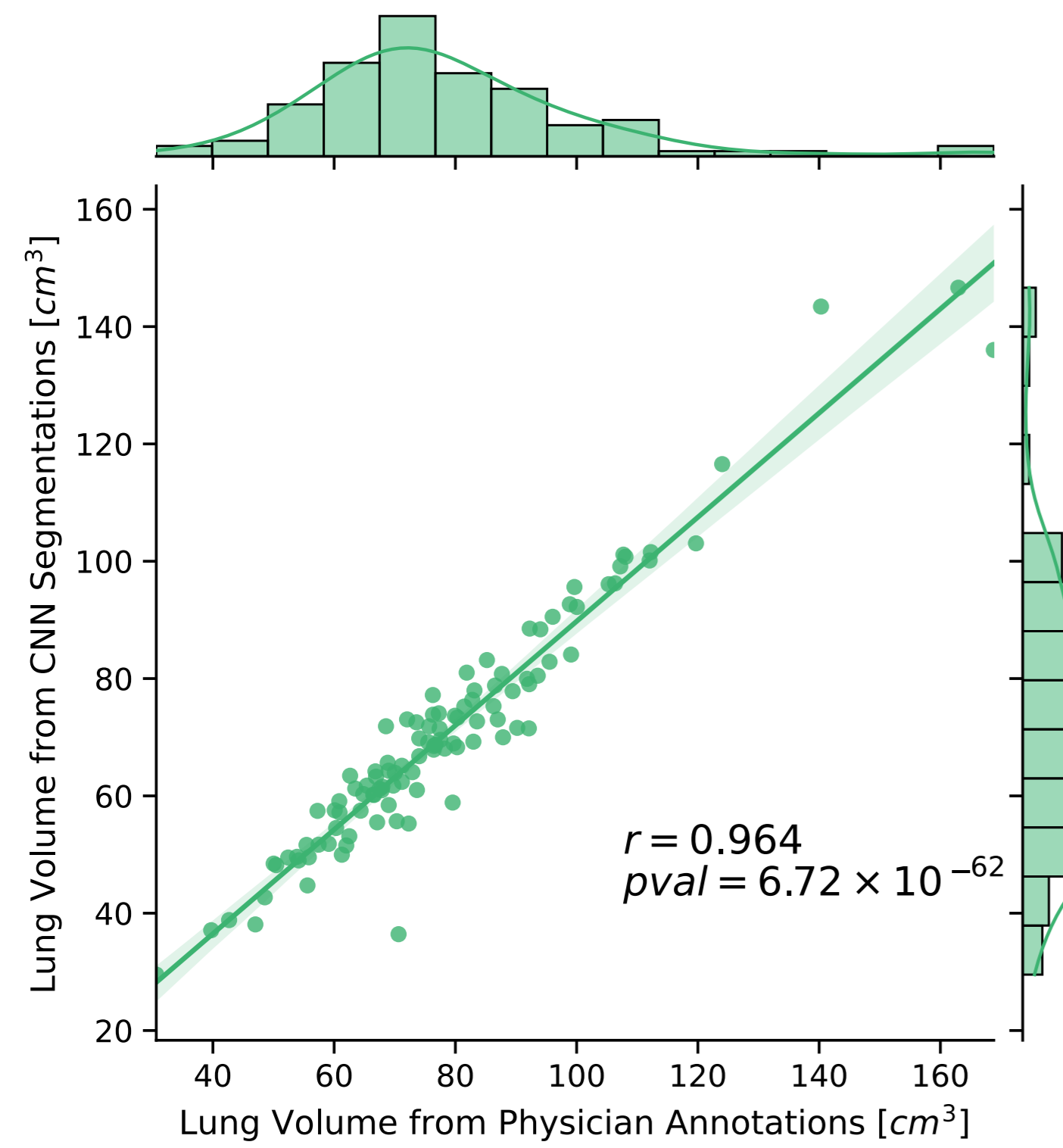
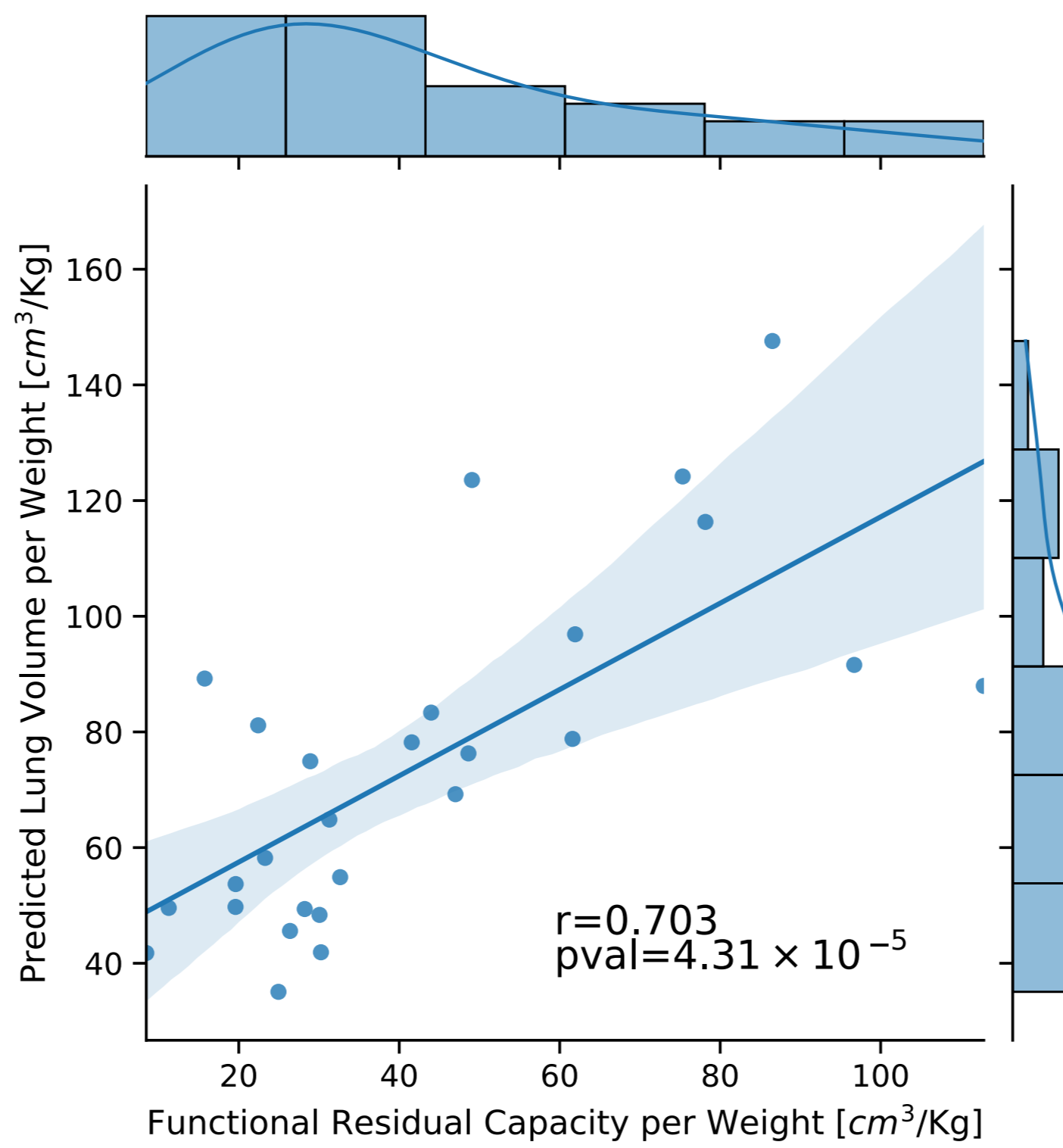


**A.**

Segmentation Performance by Image Quality

**B.**

Segmentation Performance by BPD Severity

**C.****D.****E.**