Table 3. Quantitative comparison between model prediction and rater segmentation on a subset of 10 patients in internal test data using STAPLE algorithm-generated results as the reference standard.

	Rater 1	**Rater 2**	**Rater 3**	**Rater 4**	**Proposed Model**
DC (%)	82.9 (39.9-94.1)	54.4 (49.4-70.7)	88.3 (58.6-93.5)	57.5 (49.1-64.6)	64.7 (45.7-73.7)
ASSD (mm)	0.31 (0.08-1.15)	0.72 (0.47-1.01)	0.16 (0.08-0.6)	1.26 (0.62-1.6)	0.67 (0.55-1.07)
HD95 (mm)	2.07 (0.62-5.86)	1.96 (1.71-3.99)	1.24 (0.62-1.7)	7.2 (1.8-10.09)	3.6 (1.11-6.65)
Reference Length (mm)	**20.53 (18.05-42.17)**				
Prediction Length (mm) †	28.72 (13.86-39.71)	23.13 (13.64-50.68)	22.8 (13.49-28.37)	29.66 (24.71-52.09)	17.37 (4.86-31.46)
$\boldsymbol{\delta L_{diff}}$ **(mm)**	-0.67 (-6.72-6.57)	-3.32 (-15.75-7.12)	-2.3 (-6.55-0.83)	7.56 (-6.2-12.38)	-5.64 (-20.04--1.99)
$\boldsymbol{\lvert\delta L_{diff}\rvert}$ **(mm)**	6.65 (4.19-10.28)	12.09 (7.12-16.84)	4.1 (1.87-6.57)	9.29 (6.36-12.38)	5.64 (2.01-20.04)
Reference Volume (mm^3)	**52.49 (22.99-166.42)**				

Prediction Volume (mm^3) †	95.3 (83.87-247.37)	80.42 (64.88-275.53)	79.04 (62.94-213.48)	69.7 (66.47-210.12)	82.4 (55.37-240.16)
δV_{diff} (mm^3)	1.9 (0.93-7.18)	2.35 (1.02-4.19)	1.9 (0.43-4.3)	3.65 (0.97-4.77)	0.3 (-1.92-6.23)
$\lvert\delta V_{diff}\rvert$ (mm^3)	4.02 (1.28-7.18)	2.35 (1.63-4.19)	1.9 (0.43-4.3)	4.21 (1.99-5.05)	5.05 (1.85-7.59)
Volume correlation	0.93 (0.84-0.98)	0.87 (0.81-0.94)	0.97 (0.89-0.99)	0.84 (0.79-0.9)	0.91 (0.82-0.96)

Note. Values are shown as medians with the IQR in parentheses. $\delta()$ and $|\delta()|$ represent relative error and absolute, respectively.

† It represents the human segmentation result for the raters and the prediction result for the proposed model.