## Appendix A.

## A.1 Patients characteristics

*A.1. 1 Patient characteristics of the internal and external datasets*

| Variables | | Training (n=303) n (%) | Internal Test (n=104) n (%) | External Test (n=519) n (%) |
|---|---|---|---|---|
| Gender | Male | 209 (69) | 70 (67) | 424 (82) |
| | Female | 94 (31) | 34 (33) | 95 (18) |
| T Stage | T1 | 44 (15) | 12 (12) | 21 (4) |
| | T2 | 73 (24) | 29 (28) | 91 (18) |
| | T3 | 34 (11) | 19 (18) | 80 (15) |
| | T4 | 152 (50) | 44 (42) | 44 (8) |
| | Unknown | - | - | 283 (55) |
| N Stage | N0 | 54 (18) | 18 (17) | 35 (7) |
| | N1 | 36 (12) | 4 (4) | 37 (7) |
| | N2a | 11 (4) | 4 (4) | 111 (21) |
| | N2b | 97 (32) | 39 (38) | 17 (3) |
| | N2c | 91 (30) | 37 (36) | 15 (3) |
| | N3 | 14 (4) | 2 (1) | 21 (4) |
| | Unknown | - | - | 283 (55) |
| HPV status | Negative | 137 (45) | 57 (55) | 60 (12) |
| | Positive | 133 (44) | 41 (39) | 277 (53) |
| | Unknown | 33 (11) | 6 (6) | 182 (35) |

## A.2 Data acquisition and image reconstruction

*A.2. 1 Details about internal data acquisition and reconstruction*

| **CT scan** | ***Scanner Manufacturer, Model*** | *count (%)* |
|---|---|---|
| | Siemens, Biograph64 | 208 (51.1) |
| | Siemens, SOMATOM Definition AS | 184 (45.2) |
| | Siemens, Sensation Open | 10 (2.5) |
| | Siemens, SOMATOM go.Open Pro | 5 (1.2) |
| | **Image dimensions** | *value (min – max)* |
| | Matrix size | 512×512 |

| | | |
|---|---|---|
| | Number of slices | 151 - 374 |
| | Pixel Size (mm) | 0.94 - 1.52 |
| | Slice thickness (mm) | 1.95 – 2.00 |
| **PET scan** | **Scanner Manufacturer, Model** | *count (%)* |
| | Siemens, Biograph 64_mCT | 370 (91) |
| | Siemens, Biograph128_Vision 600 Edge | 35 (8.6) |
| | Biograph40_mCT | 2 (0.4) |
| | **Reconstruction Method** | *count (%)* |
| | PSF+TOF 3i21s | 350 (86) |
| | PSF+TOF 4i5s | 37 (9) |
| | OSEM3D 3i24s | 20 (5) |
| | **Image dimensions** | *value (min – max)* |
| | Matrix size | 200×200 - 512×512 |
| | Number of slices | 111 - 730 |
| | Pixel Size (mm) | 0.98 - 4.07 |
| | Slice thickness (mm) | 1.5 – 2 |

*A.2. 2. Details about external PET-CT data*

| **CT scan details** | Pixel Size (mm) *(min – max)* | Slice thickness (mm) *(min – max)* |
|---|---|---|
| CHUM | 0.98 - 1.37 | 1.50 – 3.27 |
| CHUP | 0.98 - 0.98 | 1.50 – 2 |
| CHUS | 0.68 - 1.17 | 2 – 5 |
| CHUV | 0.98 - 1.37 | 1 – 4.25 |
| HGJ | 0.983 - 0.98 | 3.27 - 3.27 |
| HMR | 0.98 - 1.37 | 3.27 - 3.27 |
| MDA | 0.49 - 2.73 | 2 - 5 |

Details about the scanners used to acquire the CT and PET images of the external test set can be found at https://hecktor.grand-challenge.org/Data/ .

## A.3 Training details

### Single-label segmentation

As the volume of the positive class increased, sequence selection required some optimization. In the training/validation sets, 80% of the sequences contained tumour pixels for more than 2.5% of the total volume, 15% for less than 2.5%, and 5% negative sequences (without tumour).

### Multi-label segmentation

Primary tumour and lymph nodes structures can either exist independently or appear together in a slice of each 3D volume. In order to ensure convergence and the model's ability to effectively differentiate and accurately segment both GTVp and GTVln, proper sequence selection was necessary. Sequences were always considered for inclusion if 1) either the GTVp or GTVln volume equalled or exceeded 2.5% of the total sequence volume, and/or if 2) both structures appeared in the same sequence. These chosen sequences accounted for 85% of the total sequences utilized for training and validation. For the remaining sequences: 10% were selected from sequences where only one structure was present and with less than 2.5% of the sequence volume, and 5% were selected from the negative sequences. When training using sequences extracted from the axial view, for two out of the 5 folds the last two percentages were adjusted to 0.05, 0.08 and 0.04, 0.08, respectively, to ensure model convergence.

## Appendix B.

### B.1 Patients characteristics per centre

In the external test set, three (CHUP, CHUV, and MDA) out of the seven centres had some patients without primary tumour (2/72 for CHUP, 1/53 for CHUV, and 13/194 for MDA). Patients without metastatic lymph nodes were: 4/55 for CHUM, 8/72 for CHUP, 17/72 for CHUS, 7/53 for CHUV, 4/55 for HGJ, 1/18 for HMR, and 18/194 for MDA.

Compared to the internal test set, the centres CHUP and CHUS showed significant differences in their GTVp ($p<0.001$) , GTVln, ($p<0.001$, $p=0.001$) and GTVtot ($p<0.001$) volume distribution. CHUM and MDA showed differences ($p<0.001$) in GTVp and GTVtot, but not in GTVln. CHUV and HGJ were only found to be different ($p<0.001$) in terms of GTVp. HMR did not show any statistically significant differences with the internal test set.

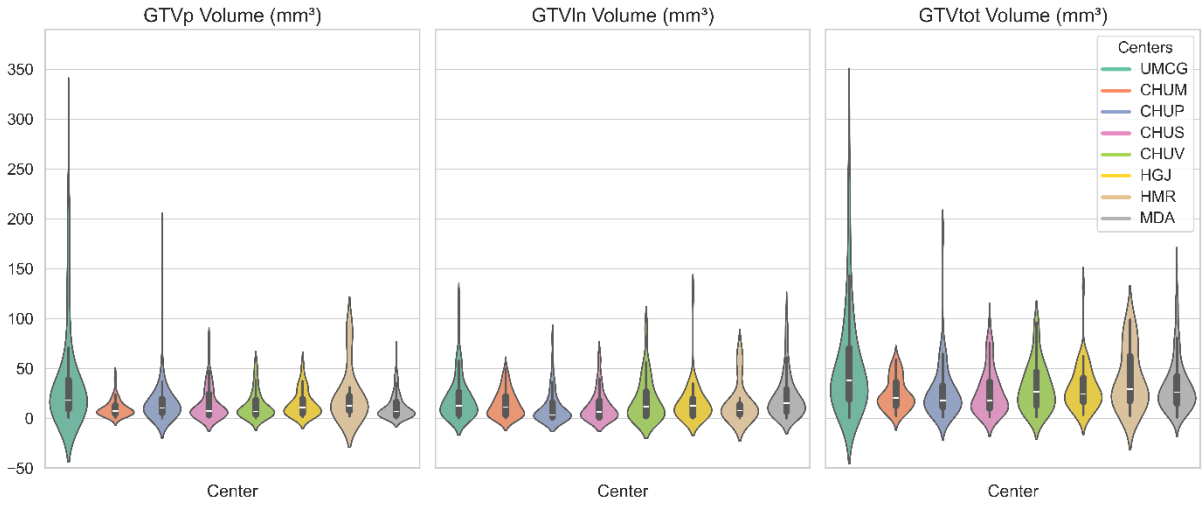



*Figure B. 1 Distribution of tumour volumes across different centres in the test sets.*

## B.2 Segmentation analysis

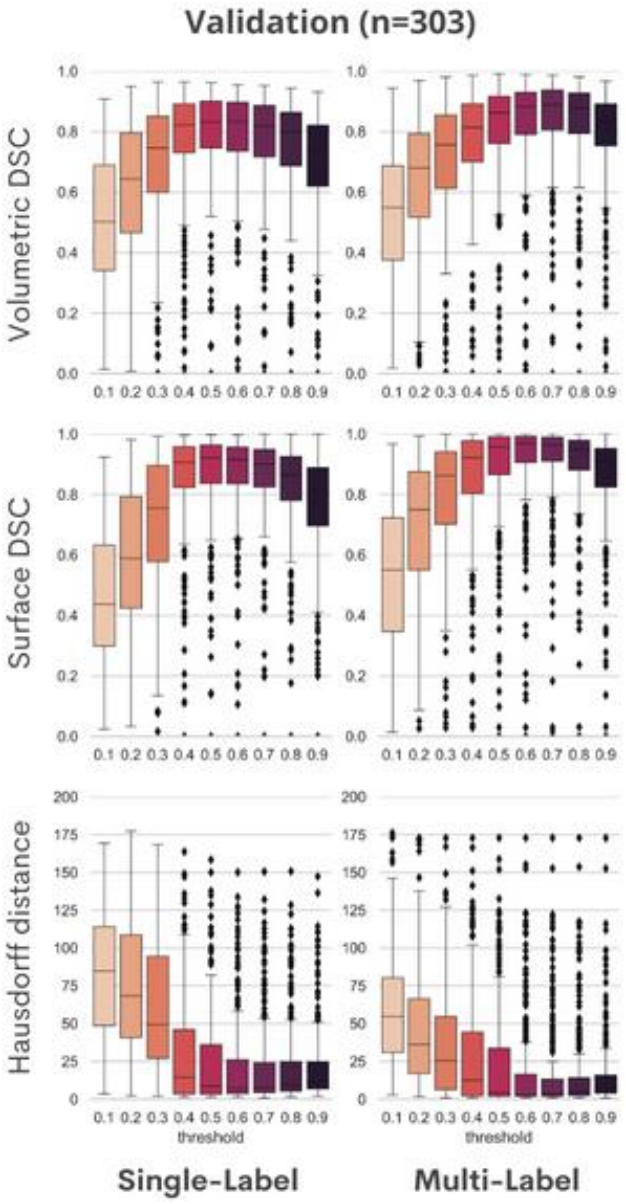



*Figure B. 2 Validation results for single- and multi-label segmentation.*

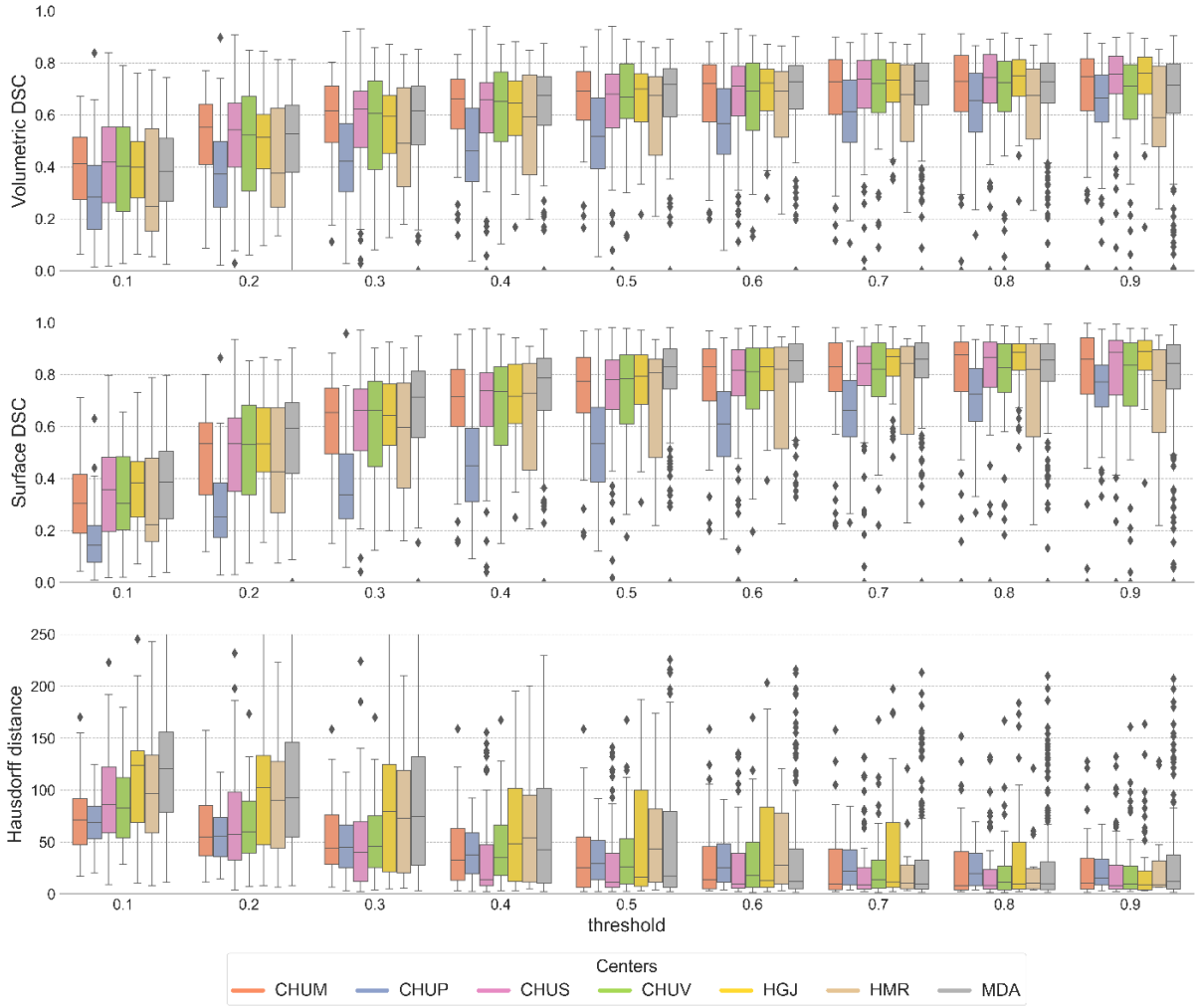



*Figure B. 3 External test results per centre.*

| | CHUM $DSC_{agg}$ | | | CHUP $DSC_{agg}$ | | | CHUS $DSC_{agg}$ | | | CHUV $DSC_{agg}$ | | | HGJ $DSC_{agg}$ | | | HMR $DSC_{agg}$ | | | MDA $DSC_{agg}$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $threshold$ | $GTV_p$ | $GTV_{ln}$ | $mean$ | $GTV_p$ | $GTV_{ln}$ | $mean$ | $GTV_p$ | $GTV_{ln}$ | $mean$ | $GTV_p$ | $GTV_{ln}$ | $mean$ | $GTV_p$ | $GTV_{ln}$ | $mean$ | $GTV_p$ | $GTV_{ln}$ | $mean$ | $GTV_p$ | $GTV_{ln}$ | $mean$ |
| 0.1 | 0.32 | 0.53 | 0.42 | 0.33 | 0.26 | 0.30 | 0.42 | 0.42 | 0.42 | 0.37 | 0.55 | 0.46 | 0.40 | 0.43 | 0.42 | 0.30 | 0.25 | 0.28 | 0.31 | 0.46 | 0.39 |
| 0.2 | 0.44 | 0.66 | 0.55 | 0.44 | 0.35 | 0.40 | 0.53 | 0.54 | 0.54 | 0.50 | 0.66 | 0.58 | 0.51 | 0.55 | 0.53 | 0.43 | 0.39 | 0.41 | 0.44 | 0.59 | 0.52 |
| 0.3 | 0.53 | 0.72 | 0.62 | 0.52 | 0.42 | 0.47 | 0.62 | 0.61 | 0.61 | 0.59 | 0.72 | 0.65 | 0.59 | 0.62 | 0.61 | 0.53 | 0.51 | 0.52 | 0.55 | 0.66 | 0.61 |
| 0.4 | 0.59 | 0.76 | 0.67 | 0.59 | 0.47 | 0.53 | 0.67 | 0.66 | 0.67 | 0.65 | 0.75 | 0.70 | 0.65 | 0.67 | 0.66 | 0.61 | 0.60 | 0.60 | 0.64 | 0.70 | 0.67 |
| 0.5 | 0.64 | 0.78 | 0.71 | 0.65 | 0.50 | 0.58 | 0.72 | 0.69 | 0.71 | 0.71 | **0.77** | 0.74 | 0.70 | 0.70 | 0.70 | 0.66 | 0.65 | 0.66 | 0.70 | 0.72 | 0.71 |
| 0.6 | 0.67 | **0.79** | 0.73 | 0.70 | 0.53 | 0.62 | 0.75 | 0.72 | 0.73 | 0.75 | **0.77** | 0.76 | 0.74 | 0.72 | 0.73 | 0.71 | **0.67** | **0.69** | 0.74 | **0.73** | 0.73 |
| 0.7 | 0.71 | **0.79** | 0.75 | 0.74 | 0.55 | 0.64 | 0.77 | **0.73** | 0.75 | 0.79 | 0.76 | **0.78** | 0.77 | **0.73** | 0.75 | 0.74 | 0.63 | **0.69** | 0.76 | **0.73** | **0.74** |
| 0.8 | 0.74 | 0.78 | 0.76 | **0.76** | **0.57** | **0.66** | 0.79 | 0.72 | **0.76** | **0.82** | 0.74 | **0.78** | 0.80 | **0.73** | 0.76 | **0.76** | 0.59 | 0.67 | **0.77** | 0.72 | **0.74** |
| 0.9 | **0.77** | 0.76 | **0.77** | 0.74 | 0.54 | 0.64 | **0.80** | 0.70 | 0.75 | **0.82** | 0.68 | 0.75 | **0.83** | 0.70 | **0.77** | 0.75 | 0.51 | 0.63 | 0.76 | 0.68 | 0.72 |

*Table B. 1 Aggregate DSC results per centre in the external test.*

## B.4 Correlation analysis

In the internal test set, the correlation coefficients of the CV versus volumetric DSC ranged between r = -0.70 and r = -0.77. The correlation between the CV and surface DSC ranged between r = -0.54 and r = -0.72. In the external test, the results were in the intervals of (r = -0.33, r = -0.55) for volumetric DSC and (r = -0.20, r = -0.49) for surface DSC. For the internal set results, correlation coefficients for volumetric DSC and surface DSC were not significantly different (p>=0.01) across threshold values. The highest correlation coefficient values for the external set were obtained at a probability threshold of 0.9. A statistical difference (p<0.01) was observed between correlation coefficients calculated at probability thresholds of 0.3,0.4,0.5,0.6 and 0.8,0.9 for both volumetric and surface DSC.

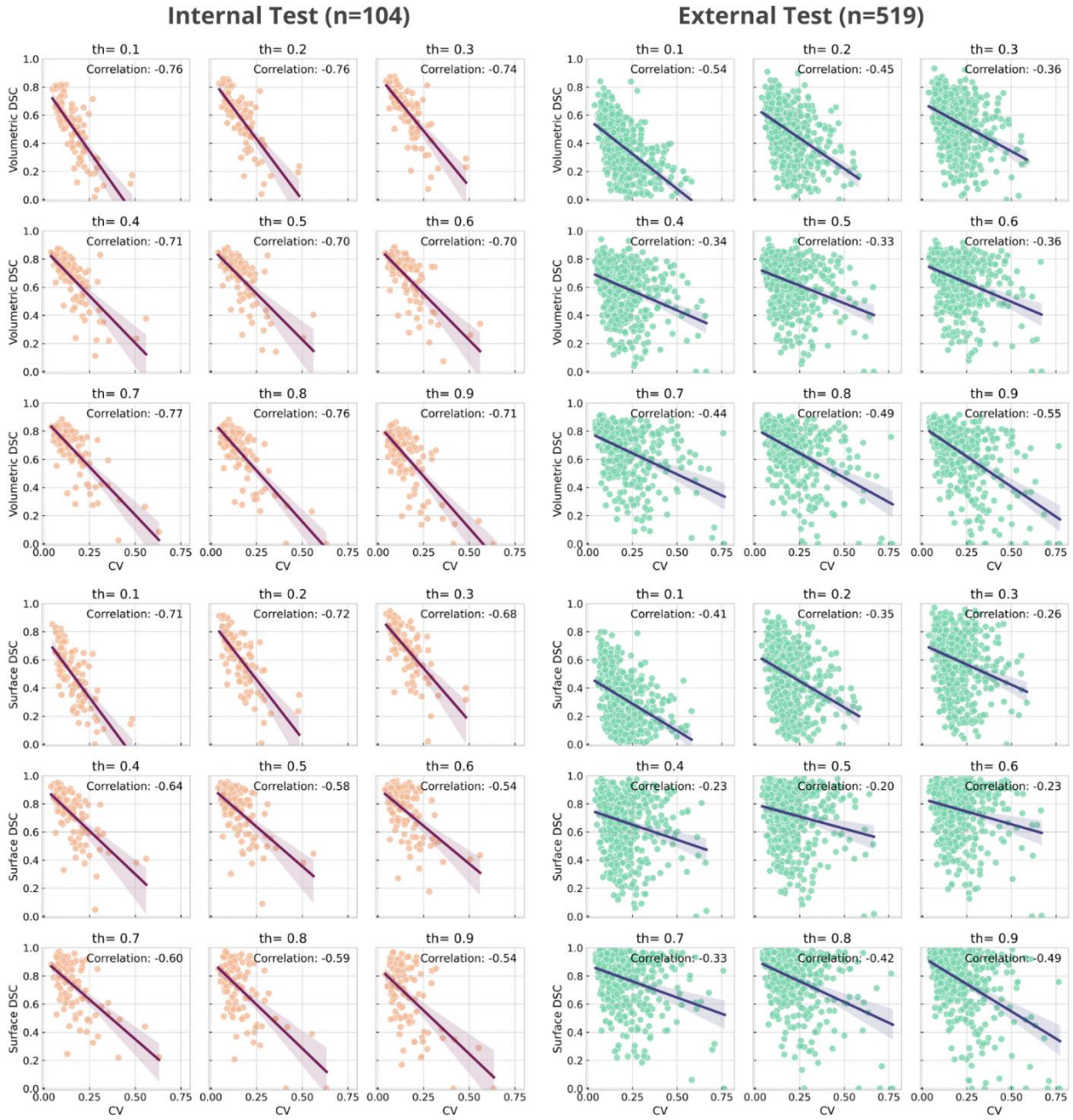



*Figure B. 4 Correlation plots of CV versus Volumetric DSC (upper three rows) and CV versus Surface DSC (lower three rows) calculated at different probability threshold values for the GTVtot in the multi-label segmentation approach for the internal (first three columns) and external (last three columns) test sets.*

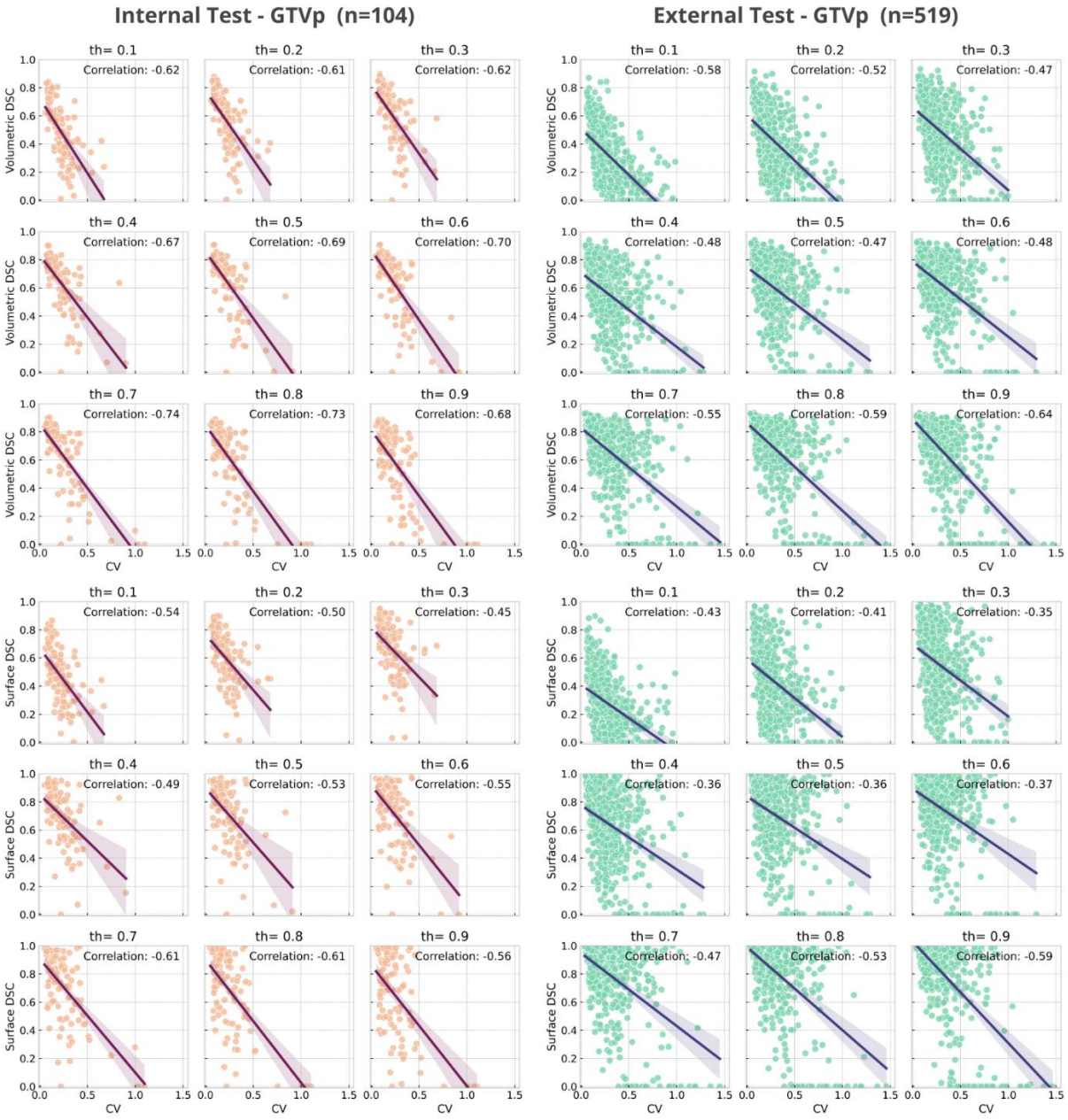



*Figure B. 5 Correlation plots of CV versus Volumetric DSC (first three rows) and CV versus Surface DSC (last three rows) calculated at different probability threshold values for the GTVp in the multi-label segmentation approach for the internal and external test sets.*

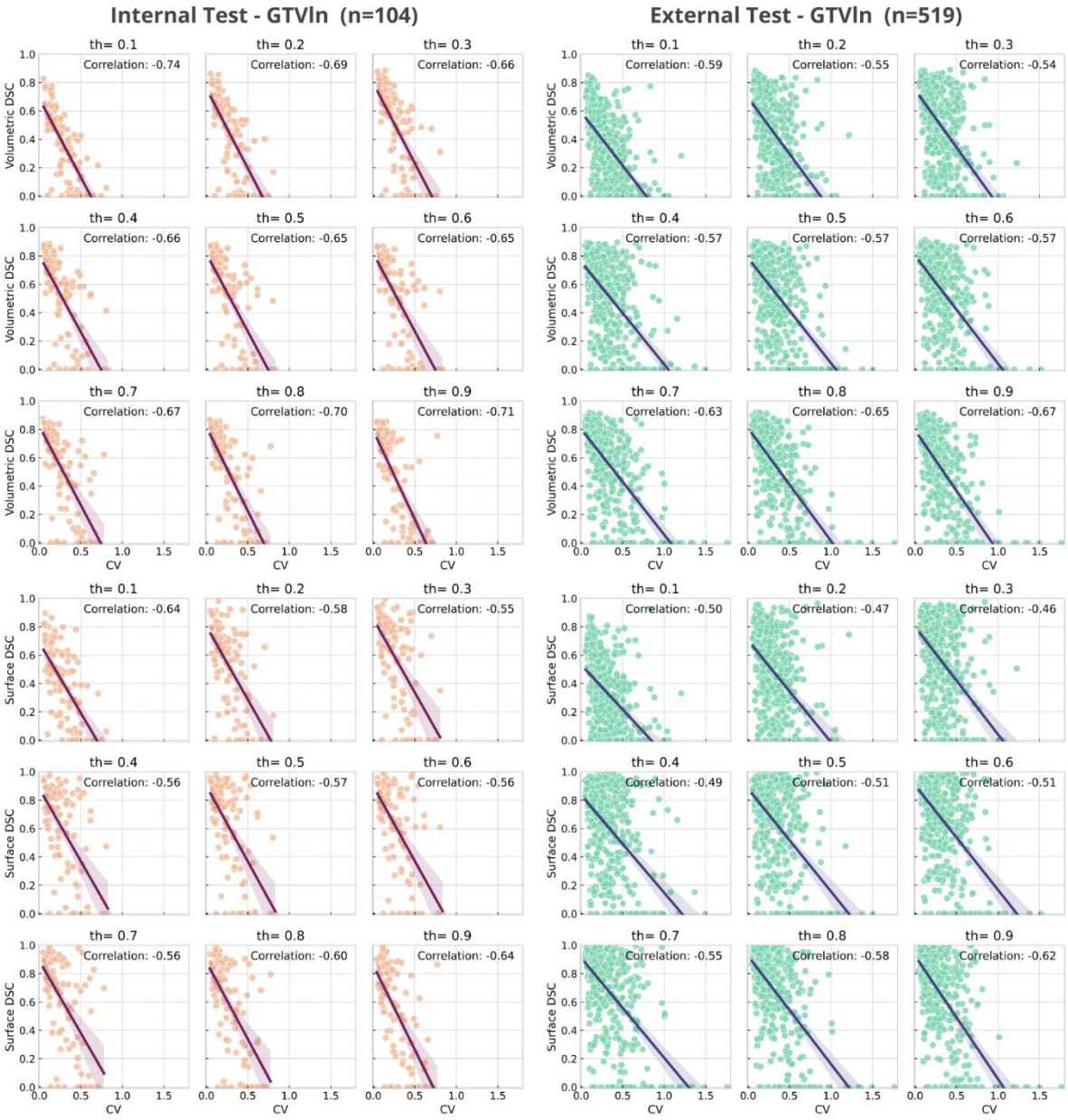



*Figure B. 6 Correlation plots of CV versus Volumetric DSC (first three rows) and CV versus Surface DSC (last three rows) calculated at different probability threshold values for the GTVln in the multi-label segmentation approach for the internal and external test sets.*

## B.5 Patient-specific analysis

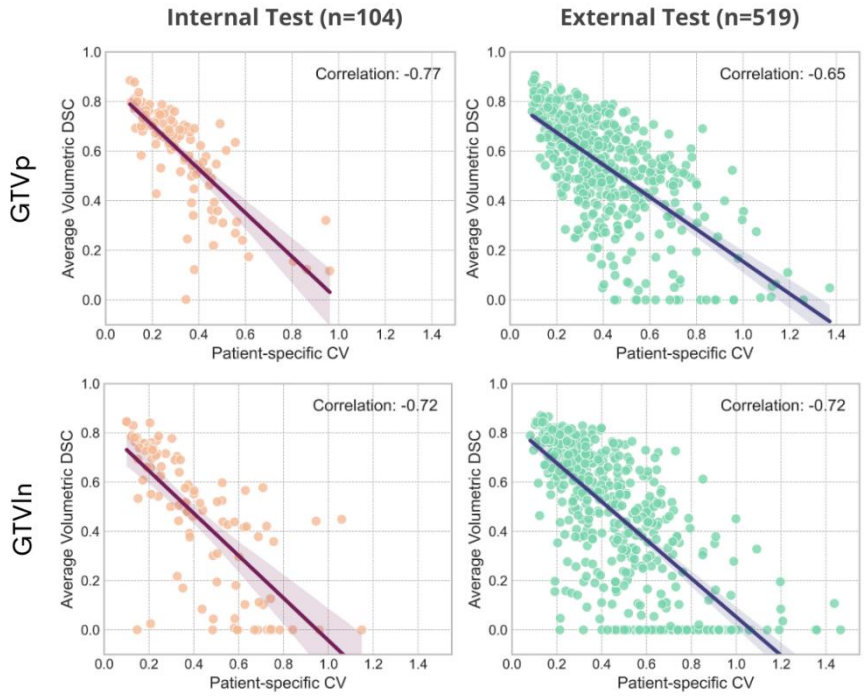
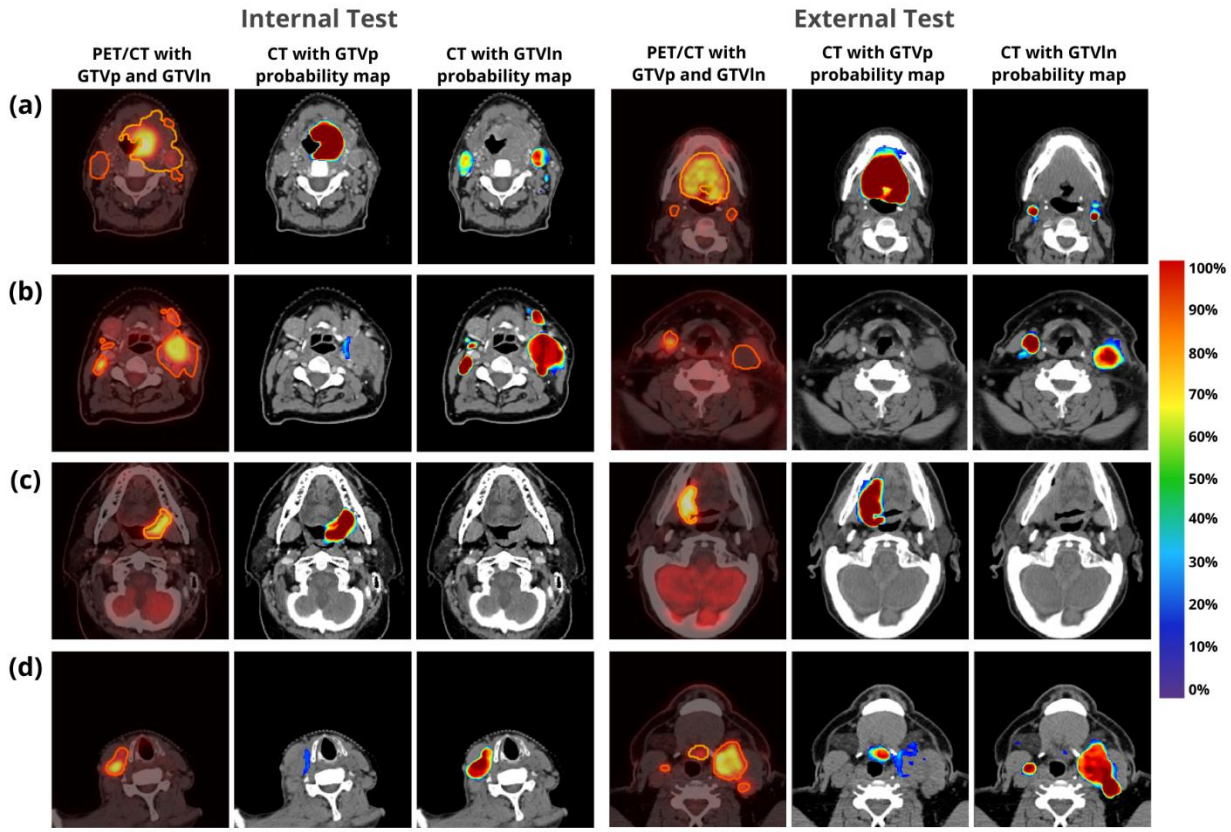


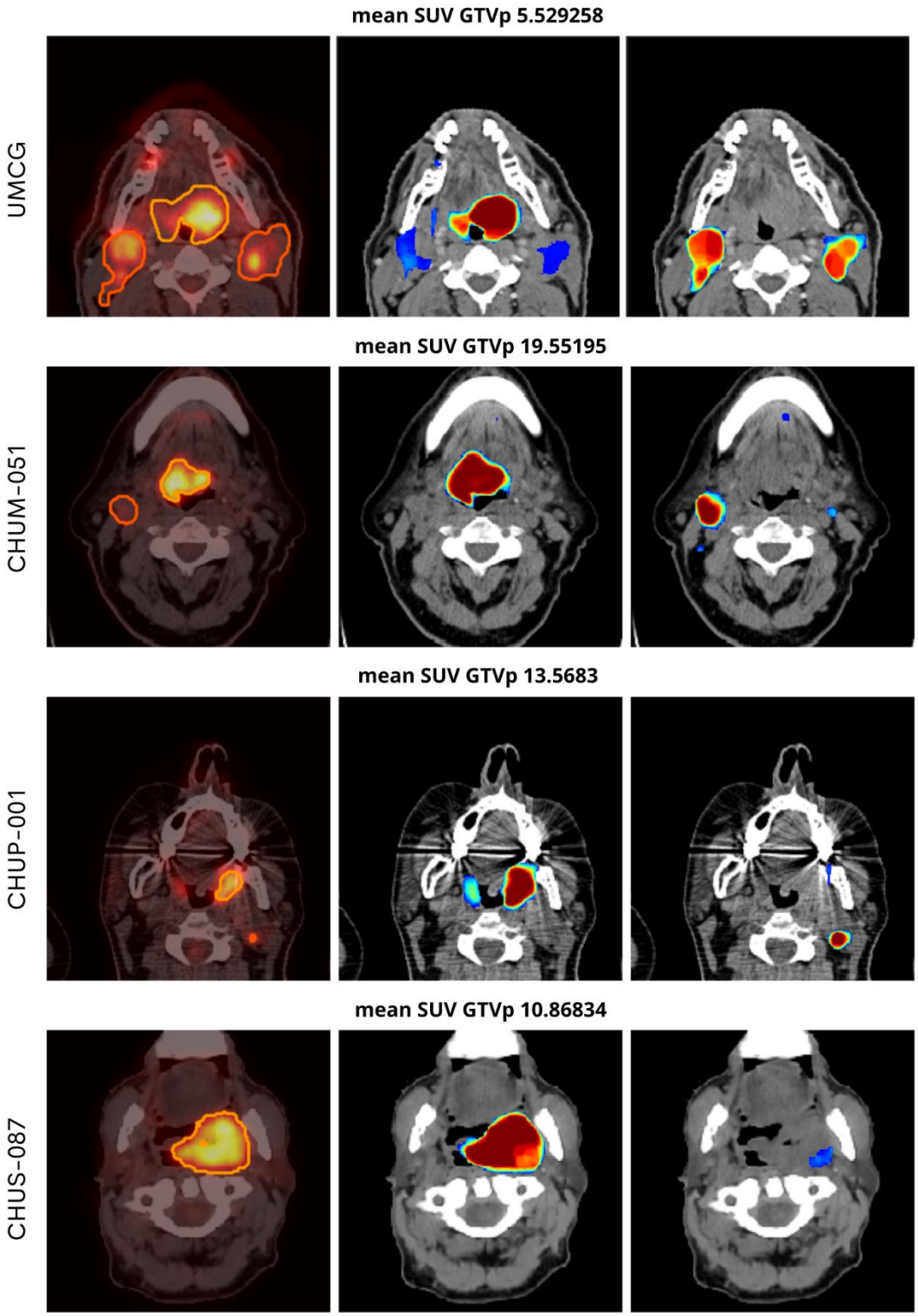

*Figure B. 7. Correlation plots between patient-specific CV and mean volumetric DSC across probability threshold (Average Volumetric DSC) calculated per patient. The correlation plots are reported per structure.*

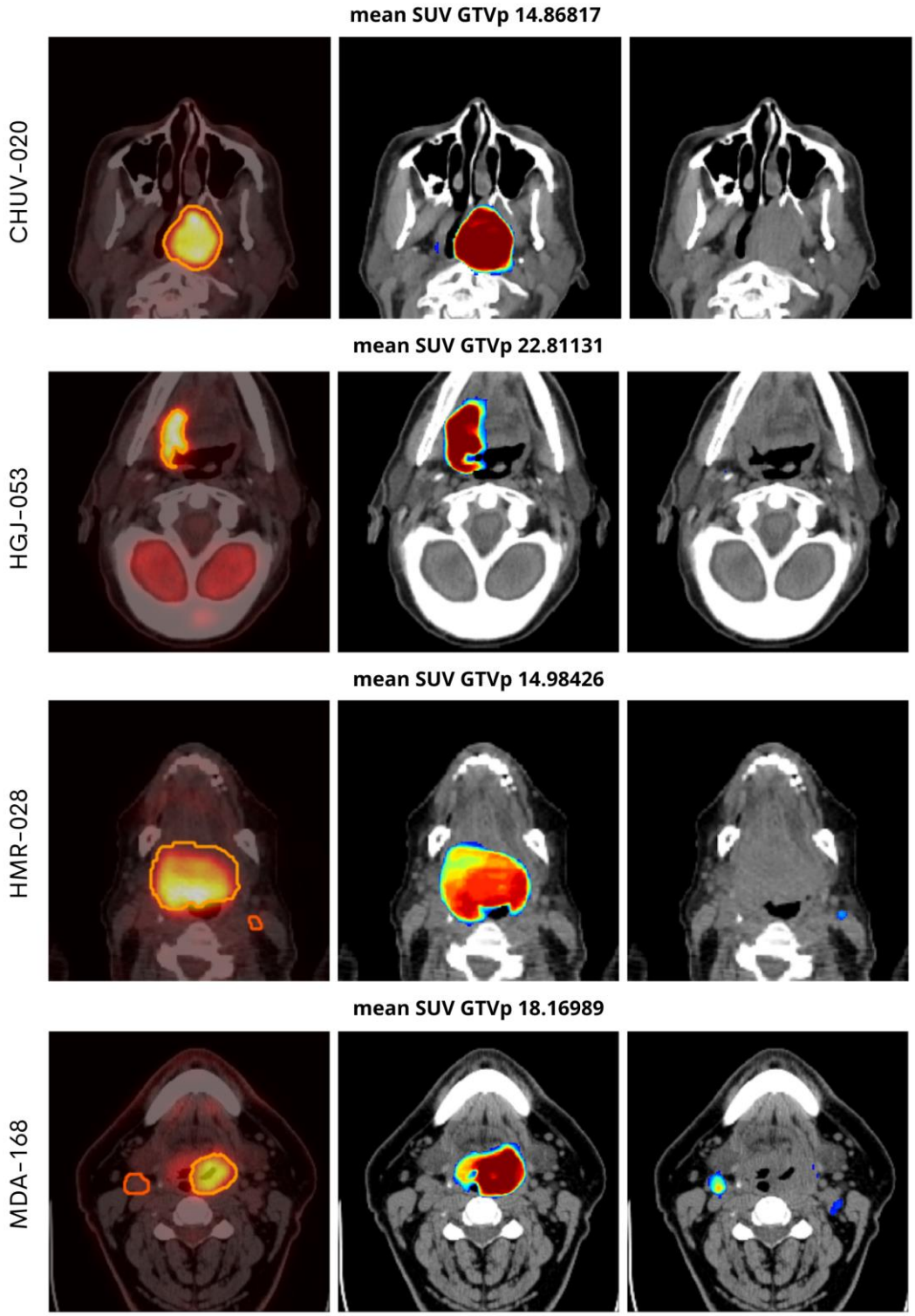
## B.6 Qualitative results

In rows (a) and (b), cases of patients with large volumes of GTVp and GTVln are presented, highlighting differences in tumour presentation on PET/CT and in delineation. In rows (c) and (d), cases of high Mean SUV GTVp and GTVln are presented for both internal and external test cases.



Examples of predicted tumour probability maps of GTVp and GTVln in the multi-label segmentation approach for patients with high Mean SUV GTVp in each center.




mean SUV GTVp 14.86817
CHUV-020
mean SUV GTVp 22.81131
HGJ-053
mean SUV GTVp 14.98426
HMR-028
mean SUV GTVp 18.16989
MDA-168