

Teamwork Makes the AI Work: Multi-Institutions Lung SBRT GTV Autosegmentation Using Federated Learning

Ben Coull-Neveu¹, Mathieu Goulet², Nathalie Panov³, Tanner Connell³, Éric Bonenfant⁴, Cédric Charrois-Durand², Marie Duclos³, James Man Git Tsui³

¹McGill University, ²Santé Québec Chaudière-Appalaches (SQCA), ³McGill University Health Centre (MUHC),
⁴Santé Québec Estrie – Centre Hospitalier Universitaire de Sherbrooke (SQE-CHUS)

ABSTRACT

An nnU-Net-based model was trained using a federated learning framework on datasets from the McGill University Health Centre (MUHC), the Santé Québec Chaudière-Appalaches (SQCA), and the Santé Québec Estrie – Centre Hospitalier Universitaire de Sherbrooke (SQE-CHUS). Performance was evaluated on cross-validation and independent test sets using the DICE similarity coefficient and compared with institution-specific models and the public TotalSegmentator and MedSAM2 models.

The federated model achieved one of the highest overall DICE scores while demonstrating improved generalizability across institutions. It consistently outperformed TotalSegmentator and achieved substantially fewer segmentation failures. Performance was comparable to institution-specific models and MedSAM2 on the test sets, indicating that federated learning can maintain competitive accuracy without the need for centralized data sharing. These findings support federated learning as a practical and scalable framework for collaborative AI development in radiotherapy.

REFERENCES

- Yang T, Yu X, McKeown MJ, Wang ZJ (2024), "When Federated Learning Meets Medical Image Analysis: A Systematic Review with Challenges and Solutions". *APSIPA Transactions on Signal and Information Processing*, Vol. 13 No. 1 pp. 1–55
- Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J., & Maier-Hein, K. H. (2021). nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2), 203-211.
- Wasserthal, J., Breit, H.-C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D., Cyriac, J., Yang, S., Bach, M., Segeroth, M., 2023. TotalSegmentator: Robust Segmentation of 104 Anatomic Structures in CT Images. *Radiology: Artificial Intelligence*.
- Ma, J., Yang, Z., Kim, S., Chen, B., Baharoon, M., Fallahpour, A., Asakereh, R., Lyu, H., & Wang, B. (2025). MedSAM2: Segment Anything in 3D medical images and videos. *arXiv preprint arXiv:2504.03600*.
- Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6(2), 65–70.

INTRODUCTION

Developing robust AI models for radiotherapy requires large and diverse datasets. These are difficult to share between institutions due to patient confidentiality, IT restrictions, and limited local machine learning expertise or resources. Federated learning offers a solution by enabling collaborative model training without exchanging raw data [1].

Objective: Evaluate the feasibility of federated training between three institutions by developing a lung GTV segmentation model for SBRT planning.

METHODS

Model: nnU-Net [2], adapted for federated learning

Federation: Shared One-Drive for weight exchange after every epoch

Institutions:

- MUHC (215 scans)
- SQCA (210 scans)
- SQE-CHUS (206 scans)

Training:

- 80:20 train/test split
- 5-fold cross-validation

Comparison:

- Local models
- TotalSegmentator [3]
- MedSAM2 [4]

Statistics:

- Repeated-measures ANOVA, followed by post-hoc, Holm-Bonferroni-corrected paired t-tests.
- Institution-agnostic paired t-tests on patient-level DICE improvement relative to local models.

DATA SELECTION & PREPROCESSING

- Excluded SBRT 4DCT phases with motion artifacts or lesions >2cm into the diaphragm.
- Lung masks generated using **TotalSegmentator v2.11** [3].
- Voxels >2cm outside the lung mask were set to zero (Figure 1).
- Applied ignore labels [2] beyond 2cm from labelled GTV(s).

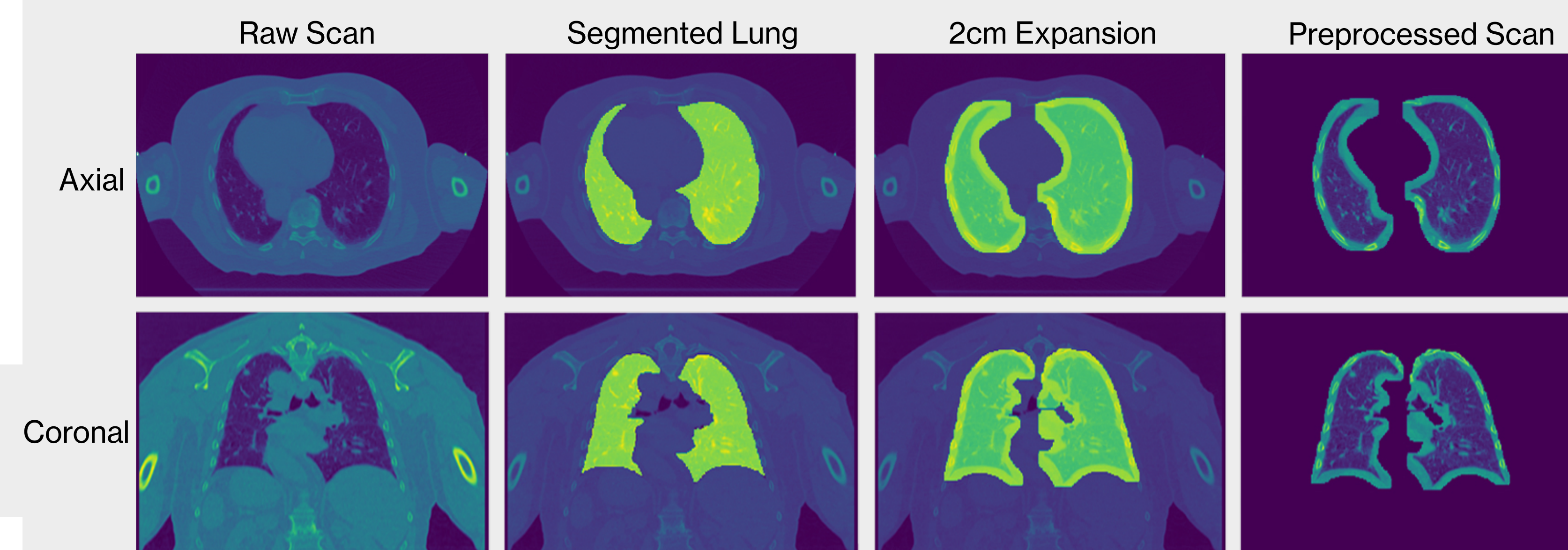


Figure 1. 4DCT scan reprocessing step. Omitted is the ignore-label around the GTV(s) which is applied to the preprocessed scan.

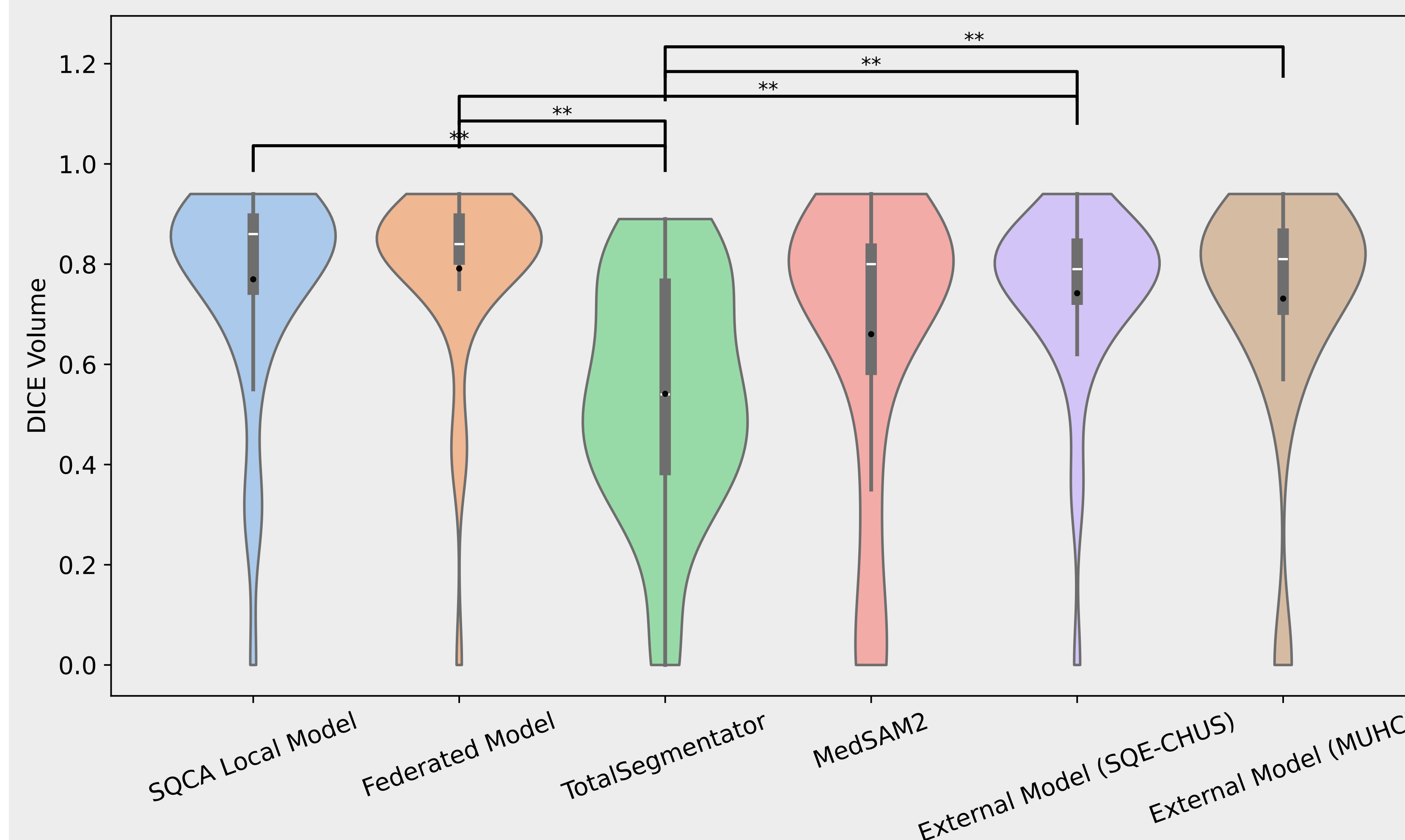


Figure 2. Violin plot visualizing the DICE score distribution of models inferred on the SQCA test set. The black points indicate the mean, while the white lines indicate the median. The boxes show the interquartile range (IQR), while the extended lines are the whisker lines (1.5 x IQR). Models linked with double stars (**) are significantly different ($p < 0.01$).

RESULTS

- **ANOVA analysis** showed significant inter-model differences for all institutions.
- **Post-hoc paired t-tests** showed:
 - Federated model significantly outperformed TotalSegmentator across the board,
 - No significant differences between MedSAM2, federated, and local models on independent test sets.
 - Federated model significantly outperformed MUHC and SQCA models on cross-validation sets.
 - See Figure 2 for SQCA model performance distributions.
- **Institution-agnostic analysis** revealed that the federated model achieved (Tables 1 & 2):
 - one of the highest DICE scores, with significant improvements over locally-inferred models only on the cross-validation set.
 - The lowest segmentation failure rates among the models (excluding local models inferred on local data):
 - Federated: **18.5%**
 - MedSAM2: **19.4%**
 - TotalSegmentator: **51.6%**

DISCUSSION

- Local models generalized poorly to external institutions.
- Federated learning exhibited highest overall mean DICE while improving inter-institution robustness.
- Federated performance matched or exceeded all model, public or local.
- Federated model showed consistent performance with lower failure rates.
- Federated training particularly benefits institutions with limited datasets and relevant resources.

CONCLUSIONS

- Federated learning enables collaborative AI development, while maintaining patient privacy in regulated healthcare environments.
- Federated model yielded comparable to superior performance to local and public models.
- It also showed improved generalizability across institutions.

This work supports the idea that federated platforms provide substantial advantages, particularly for smaller institutions with insufficient data or resources for in-house model training.

CONTACT

Ben Coull-Neveu
McGill University
benjamin.coull-neveu@mail.mcgill.ca

James Man Git Tsui
McGill University Health Center
james.tsui@mcgill.ca

Model	DICE Score ($\mu \pm \sigma$)	DICE Improvement over Local ($\mu \pm \sigma$)	p	Data Source(s)	DICE < 0.5
Federated Model	0.748 $\pm$ 0.236	+ 0.047 $\pm$ 0.186	< 0.001	All	10.7%
External Model (MUHC)	0.683 $\pm$ 0.273	- 0.057 $\pm$ 0.222	< 0.001	SQCA & SQE-CHUS	16.7%
External Model (SQCA)	0.614 $\pm$ 0.315	- 0.080 $\pm$ 0.304	< 0.001	MUHC & SQE-CHUS	25.7%
External Model (SQE-CHUS)	0.642 $\pm$ 0.269	- 0.027 $\pm$ 0.217	0.024	MUHC & SQCA	18.5%
TotalSegmentator	0.546 $\pm$ 0.236	- 0.155 $\pm$ 0.256	< 0.001	All	35.5%
MedSAM2	0.659 $\pm$ 0.278	- 0.042 $\pm$ 0.334	< 0.001	All	20.5%
MUHC Local Model	0.625 $\pm$ 0.311	—	—	MUHC	25.7%
SQCA Local Model	0.714 $\pm$ 0.252	—	—	SQCA	12.4%
SQE-CHUS Local Model	0.766 $\pm$ 0.246	—	—	SQE-CHUS	9.0%

Table 1. DICE scores and improvements over the local models given the *Cross-Validation* set. DICE improvement is calculated for each scan and averaged across all scans. Two-tailed p-values indicate the significance of DICE score difference compared to the local models. DICE < 0.5 indicates the fraction of the scans that resulted in DICE under 0.5 (poor overlap).

Model	DICE Score ($\mu \pm \sigma$)	DICE Improvement over Local ($\mu \pm \sigma$)	p	Data Source(s)	DICE < 0.5
Federated Model	0.695 $\pm$ 0.296	+ 0.011 $\pm$ 0.083	0.16	All	18.5%
External Model (MUHC)	0.663 $\pm$ 0.282	- 0.058 $\pm$ 0.128	< 0.001	SQCA & SQE-CHUS	20.0%
External Model (SQCA)	0.591 $\pm$ 0.308	- 0.050 $\pm$ 0.197	0.022	MUHC & SQE-CHUS	27.7%
External Model (SQE-CHUS)	0.686 $\pm$ 0.233	- 0.005 $\pm$ 0.172	0.80	MUHC & SQCA	15.3%
TotalSegmentator	0.495 $\pm$ 0.241	- 0.190 $\pm$ 0.230	< 0.001	All	51.6%
MedSAM2	0.651 $\pm$ 0.273	- 0.033 $\pm$ 0.322	0.26	All	19.4%
MUHC Local Model	0.617 $\pm$ 0.345	—	—	MUHC	31.8%
SQCA Local Model	0.770 $\pm$ 0.203	—	—	SQCA	9.8%
SQE-CHUS Local Model	0.670 $\pm$ 0.352	—	—	SQE-CHUS	20.5%

Table 2. DICE scores and improvements over the local models, as reported in Table 1, but inferred on the *Test* set.