Towards Informative Uncertainty Measures for MRI Segmentation in Clinical Practice: Application to Multiple Sclerosis

Nataliia Molchanova[1,2,3], Vatsal Raina[3,4], Francesco La Rosa[5], Andrey Malinin[6], Henning Müller[3], Mark Gales[4], Cristina Granziera[7], Mara Graziani[3,8], and Merirxell Bach Cuadra[1,9]
[1]Radiology department, Lausanne University Hospital (CHUV), Lausanne, Switzerland, [2]Doctoral School of the Faculty of Biology and Medicine, University of Lausanne (UNIL), Lausanne, Switzerland, [3]University of Applied Sciences of Western Switzerland, Sierre, Switzerland, [4]University of Cambridge, Cambridge, United Kingdom, [5]Icahn School of Medicine at Mount Sinai, New York, NY, United States, [6]Shifts Project, Helsinki, Finland, [7]University Hospital Basel, Basel, Switzerland, [8]IBM Research Europe, Zurich, Switzerland, [9]Center for Biomedical Imaging (CIBM), University of Lausanne, Lausanne, Switzerland

Synopsis

Keywords: Machine Learning/Artificial Intelligence, Multiple Sclerosis, Brain, Uncertainty estimation, Reliable AI

We approach the problem of quantifying the degree of reliability of supervised deep learning models used by clinicians for automatic multiple sclerosis lesion segmentation on MRI. In particular, we quantify the correspondence of various uncertainty measures to the errors that a deep learning model makes in overall segmentation or lesion detection. The evaluation is done both on in- and out-of- domain datasets (40 and 99 patients respectively), and provides insights about the measures that can point clinicians to potential errors of an automatic algorithm regardless of the distributional shift.

Introduction

MRI plays an important role in diagnosing and monitoring multiple sclerosis (MS)[1]. White matter lesions (WML) identified on T2 and FLAIR brain scans is a hallmark of the disease[1-3]. Over the past years various deep learning (DL) algorithms have been developed to replace a time-consuming skill-demanding procedure of manual WML annotation[4]. On the other hand, WML segmentation with black-box DL models is not necessarily reliable, especially when tested on out-of-domain data, e.g. different scanners, centres, patients, *etc*[6-9]. Thus, automatic predictions should be verified and corrected by clinicians. In this work, we investigate different voxel- and lesion-scale uncertainty measures as a method of pointing clinicians to potential model errors in overall segmentation or lesion detection.

Methods

We evaluate six voxel-scale uncertainty measures[6,9] and seven lesion-scale measures[6-8](full list in Figure 1). Uncertainty is estimated using deep ensembles[5], where the base model is a 3D U-net, which was previously used in uncertainty studies for the WML segmentation task[6-9]. The absolute values of uncertainty are not necessarily meaningful, hence we should only rely on the ranking of the uncertainties for different predictions. Error retention curves (RC) allow quantifying the correspondence between an uncertainty measure and model errors while only looking at the ranking of predictions in terms of uncertainty[5,6]. An RC for a single subject is built by iteratively replacing a fraction of the most uncertain predictions (voxels or lesions) with the ground truth, and recomputing model performance on this subject in terms of overall segmentation or lesion detection (see Figure 2). As a segmentation quality measure at the voxel scale the Dice similarity coefficient (DSC) is used; at the lesion scale the detection quality is evaluated using the lesion positive predictive value (LPPV) (see Figure 2). Average across subjects areas under respective RCs, *i.e* DSC-AUC or LPPV-AUC, quantify for the particular dataset the correspondence between voxel or lesion uncertainty measures and errors made in segmentation or lesion detection.

We employ a dataset provided by the Shifts project[9]. It contains FLAIR scans, which underwent denoising, skull stripping, bias field correction and interpolation to 1 mm^3 space, and their manual WML annotations used as the ground truth. The Shifts dataset embraces four publicly available and one private datasets acquired at six different medical centres with six different scanner models (both 1.5T and 3T field strength). Training and validation sets contain data from four different medical centres with 33 and 7 scans respectively. The Shifts dataset allows to separate the RC analysis between in-domain (same centres as the training data) and out-of-domain (two new centres) sets containing 40 and 99 subjects respectively.

Results and Discussion

Examples of uncertainty maps on voxel and lesion scales are shown in Figure 3. The resulting voxel- and lesion-scale RCs computed separately for in- and out-of-domain data, as well as for the whole dataset are shown in Figure 4. The respective areas under the RCs are ranked and shown in Figure 5.

The entropy based measures (ExE and EoE) have the highest DSC-AUC on the shifted dataset, indicating a superior ability in capturing model segmentation errors compared to other voxel-scale measures. However ExE loses informativeness for the lesion detection, showing the lowest LPPV-AUC. In principle, regions of high voxel uncertainty are often located on lesion borders and should be related to lesion delineation more than detection (Figure 3). The lesion-scale measure DDU$_{true}$is not based on the voxel-scale uncertainty but computes the disagreement in structural predictions between models in an ensemble. DDU$_{true}$ shows the highest LPPV-AUC on both in- and out-of-domain data. Despite that, a visual examination of voxel uncertainty maps sometimes shows non-zero uncertainties inside false negative (FN) lesions, while lesion-scale uncertainties cannot be computed for FN lesions and, thus, cannot be used for FN localisation (see Figure 3).

On the other hand, the ranking of the voxel-scale uncertainty measures in terms of DSC-AUC is different for the in- and out-of-domain datasets. In particular, the DSC-AUC of the negated confidence measure is the highest in the initial domain, but is one of the lowest in the shifted domain. The ranking of the lesion uncertainty measures, however, does not change under the distributional shift.

Conclusions

In this study, we promote the use of uncertainty measures to quantify the degree of reliability of DL models for WML segmentation in MS. We compared different uncertainty measures both on voxel and lesion scales on the in- and out-of-domain data, showing that lesion-scale uncertainty measures in comparison to the voxel-scale ones yield a more consistent ranking of measures in terms of capturing model errors. Additionally, we observe that the lesion uncertainty DDU_{true} has a superior ability to capture model errors related to lesion detection, what withholds for both in- and out-of-domain. We believe that lesion-scale detection uncertainty is needed to support the adoption of automatic DL-based methods for WML segmentation into the clinical practice. Our study guides towards which uncertainty measures are more informative for pinpointing potential errors in voxel- or lesion-scale predictions. It is yet important to verify in practice if the information brought by the uncertainty maps can simplify or speed up a semi-automatic segmentation by pointing clinicians to potential model errors.

Acknowledgements

This work was supported by the Hasler Foundation Responsible AI programme (MSxplain) and the EU Horizon 2020 project AI4Media (grant 951911). We acknowledge access to the facilities and expertise of the CIBM Center for Biomedical Imaging, a Swiss research center of excellence founded and supported by Lausanne University Hospital (CHUV), University of Lausanne (UNIL), École polytechnique fédérale de Lausanne (EPFL), University of Geneva (UNIGE) and Geneva University Hospitals (HUG).

References

1. Hemond CC, Bakshi R. Magnetic Resonance Imaging in Multiple Sclerosis. Cold Spring Harb Perspect Med. 2018;8(5):a028969. doi:10.1101/cshperspect.a028969

2. Thompson AJ, Banwell BL, Barkhof F, et al. Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. Lancet Neurol. 2018;17(2):162-173. doi:10.1016/S1474-4422(17)30470-2

3. Bendfeldt K, Kuster P, Traud S, et al. Association of regional gray matter volume loss and progression of white matter lesions in multiple sclerosis - A longitudinal voxel-based morphometry study. Neuroimage. 2009;45(1):60-67. doi:10.1016/j.neuroimage.2008.10.006

4. Zeng C, Gu L, Liu Z, Zhao S. Review of Deep Learning Approaches for the Segmentation of Multiple Sclerosis Lesions on Brain MRI. Front Neuroinform. 2020;14:610967. Published 2020 Nov 20. doi:10.3389/fninf.2020.610967

5. Malinin A. Uncertainty estimation in deep learning with application to spoken language assessment, Ph.D. thesis, University of Cambridge, United Kingdom, 2019.

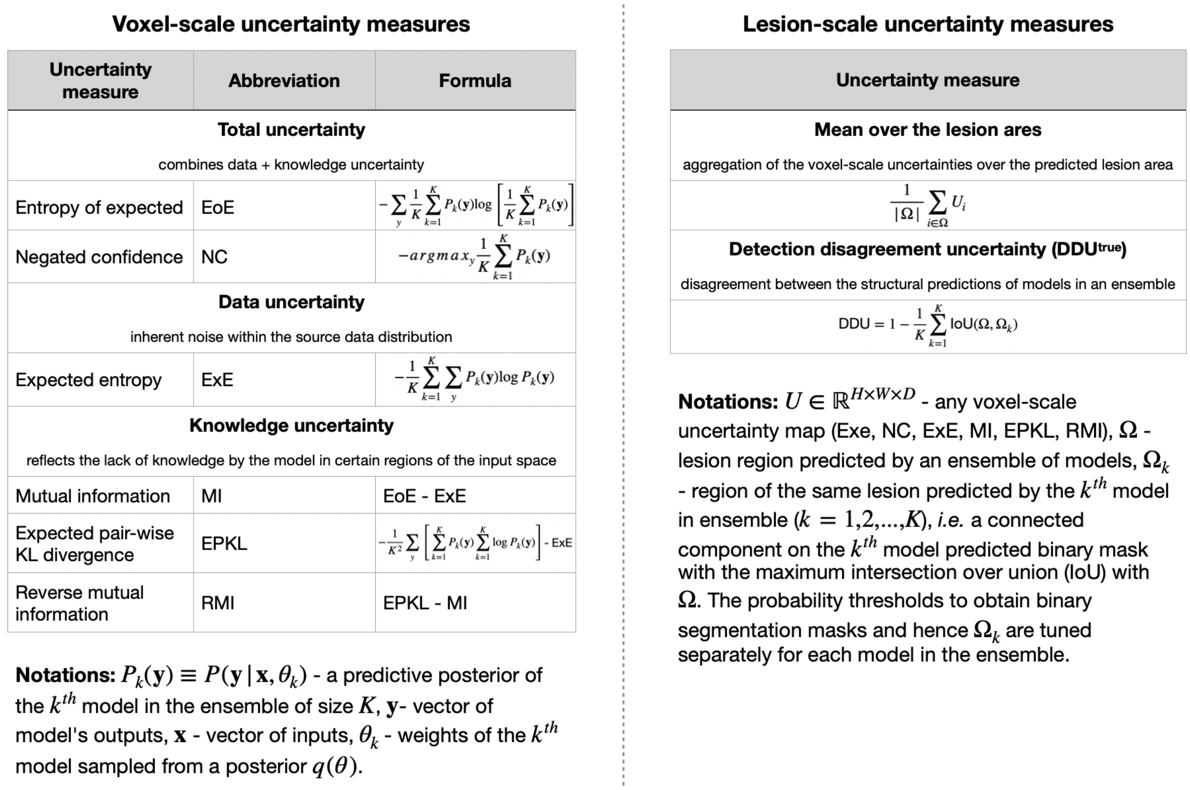
6. Molchanova N, Raina V, Malinin A, et al. Novel structural-scale uncertainty measures and error retention curves: application to multiple sclerosis. ArXiv.

7. Lambert B, Forbes F, Tucholka A, Doyle S, Dojat M. Multi-Scale Evaluation of Uncertainty Quantification Techniques for Deep Learning based MRI Segmentation. In ISMRM-ESMRMB & ISMRT 2022 - 31st Joint Annual Meeting International Society or Magnetic Resonance in Medicine London, United Kingdom, May 2022.

8. Nair T, Precu D, Arnold DL, Arbel T. Exploring Uncertainty Measures in Deep Networks for Multiple Sclerosis Lesion Detection and Segmentation. Medical image analysis. 2018,59:101557. doi:10.1007/978-3-030-00928-1_74

9. Malinin A, Athanasopoulos A, Barakovic M, et al. Shifts 2.0: Extending The Dataset of Real Distributional Shifts. ArXiv. doi:10.48550/arxiv.2206.15407

Figure 1
Definitions of voxel- and lesion- scale uncertainty measures estimated using deep ensembles within this study.



Voxel-scale uncertainty measures

Uncertainty measure	Abbreviation	Formula
Total uncertainty combines data + knowledge uncertainty		
Entropy of expected	EoE	$-\sum_{y}\frac{1}{K}\sum_{k=1}^{K}P_k(\mathbf{y})\log\left[\frac{1}{K}\sum_{k=1}^{K}P_k(\mathbf{y})\right]$
Negated confidence	NC	$-argmax_y\frac{1}{K}\sum_{k=1}^{K}P_k(\mathbf{y})$
Data uncertainty inherent noise within the source data distribution		
Expected entropy	ExE	$-\frac{1}{K}\sum_{k=1}^{K}\sum_{y}P_k(\mathbf{y})\log P_k(\mathbf{y})$
Knowledge uncertainty reflects the lack of knowledge by the model in certain regions of the input space		
Mutual information	MI	EoE - ExE
Expected pair-wise KL divergence	EPKL	$-\frac{1}{K^2}\sum_{y}\left[\sum_{k=1}^{K}P_k(\mathbf{y})\sum_{k=1}^{K}\log P_k(\mathbf{y})\right]$ - ExE
Reverse mutual information	RMI	EPKL - MI

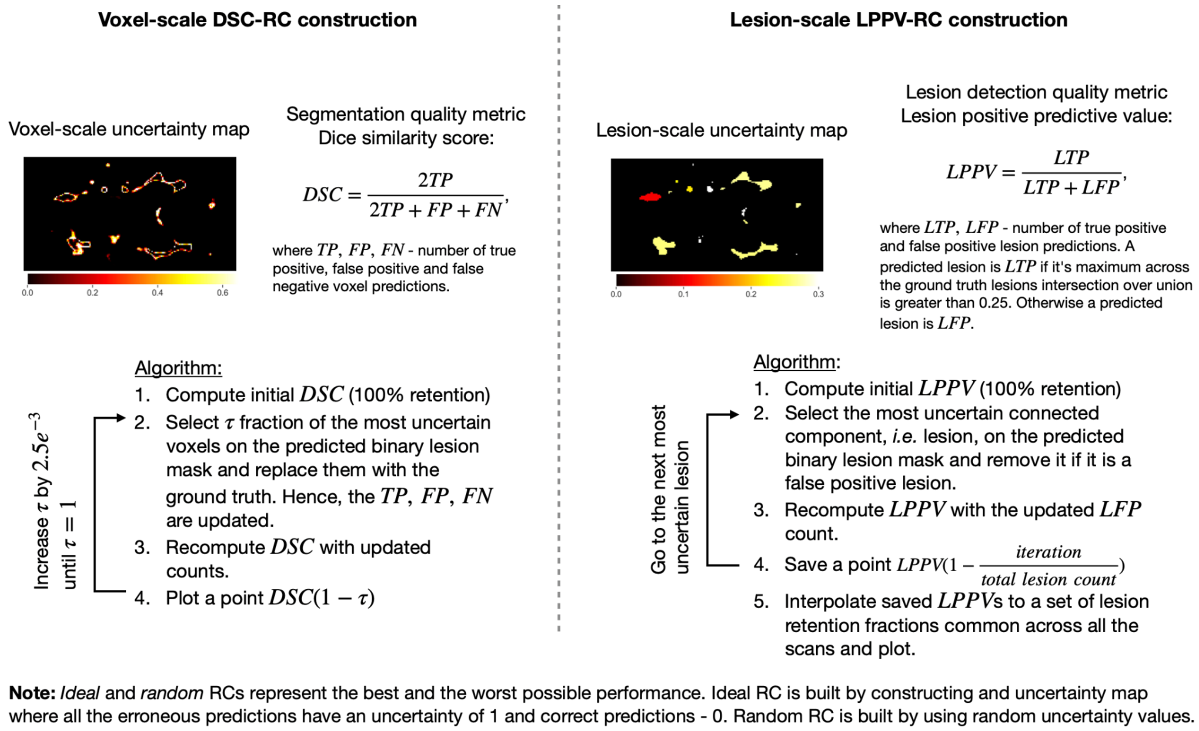
Notations: $P_k(\mathbf{y}) \equiv P(\mathbf{y}\,|\,\mathbf{x},\theta_k)$ - a predictive posterior of the k^{th} model in the ensemble of size K, $\mathbf{y}$- vector of model's outputs, $\mathbf{x}$ - vector of inputs, θ_k - weights of the k^{th} model sampled from a posterior $q(\theta)$.

Lesion-scale uncertainty measures

Uncertainty measure
Mean over the lesion ares aggregation of the voxel-scale uncertainties over the predicted lesion area
$\frac{1}{\vert\Omega\vert}\sum_{i\in\Omega}U_i$
Detection disagreement uncertainty (DDU^{true}) disagreement between the structural predictions of models in an ensemble
DDU $= 1 - \frac{1}{K}\sum_{k=1}^{K}$ IoU(Ω,Ω_k)

Notations: $U \in \mathbb{R}^{H\times W\times D}$ - any voxel-scale uncertainty map (Exe, NC, ExE, MI, EPKL, RMI), $\mathbf{\Omega}$ - lesion region predicted by an ensemble of models, $\mathbf{\Omega}_k$ - region of the same lesion predicted by the k^{th} model in ensemble ($k = 1,2,...,K$), *i.e.* a connected component on the k^{th} model predicted binary mask with the maximum intersection over union (IoU) with Ω. The probability thresholds to obtain binary segmentation masks and hence Ω_k are tuned separately for each model in the ensemble.

Figure 2
Explanation of the retention curves (RC) construction for a single patient: DSC-RC for quantifying the correspondence between voxel-scale uncertainty measures and errors in segmentation, LPPV-RC on the lesion-scale for quantifying the correspondence between lesion-scale measures and errors in lesion detection.



Voxel-scale DSC-RC construction

Voxel-scale uncertainty map

Segmentation quality metric
Dice similarity score:

$$DSC = \frac{2TP}{2TP + FP + FN},$$

where TP, FP, FN - number of true positive, false positive and false negative voxel predictions.

Algorithm:

1. Compute initial DSC (100% retention)
2. Select τ fraction of the most uncertain voxels on the predicted binary lesion mask and replace them with the ground truth. Hence, the TP, FP, FN are updated.
3. Recompute DSC with updated counts.
4. Plot a point $DSC(1-\tau)$

Increase τ by $2.5e^{-3}$ until $\tau = 1$

Lesion-scale LPPV-RC construction

Lesion-scale uncertainty map

Lesion detection quality metric
Lesion positive predictive value:

$$LPPV = \frac{LTP}{LTP + LFP},$$

where LTP, LFP - number of true positive and false positive lesion predictions. A predicted lesion is LTP if it's maximum across the ground truth lesions intersection over union is greater than 0.25. Otherwise a predicted lesion is LFP.

Algorithm:

1. Compute initial $LPPV$ (100% retention)
2. Select the most uncertain connected component, *i.e.* lesion, on the predicted binary lesion mask and remove it if it is a false positive lesion.
3. Recompute $LPPV$ with the updated LFP count.
4. Save a point $LPPV(1 - \frac{iteration}{total\ lesion\ count})$
5. Interpolate saved $LPPV$s to a set of lesion retention fractions common across all the scans and plot.

Go to the next most uncertain lesion

Note: *Ideal* and *random* RCs represent the best and the worst possible performance. Ideal RC is built by constructing and uncertainty map where all the erroneous predictions have an uncertainty of 1 and correct predictions - 0. Random RC is built by using random uncertainty values.

Figure 3

Examples of uncertainty maps on voxel and lesion scales for one patient.

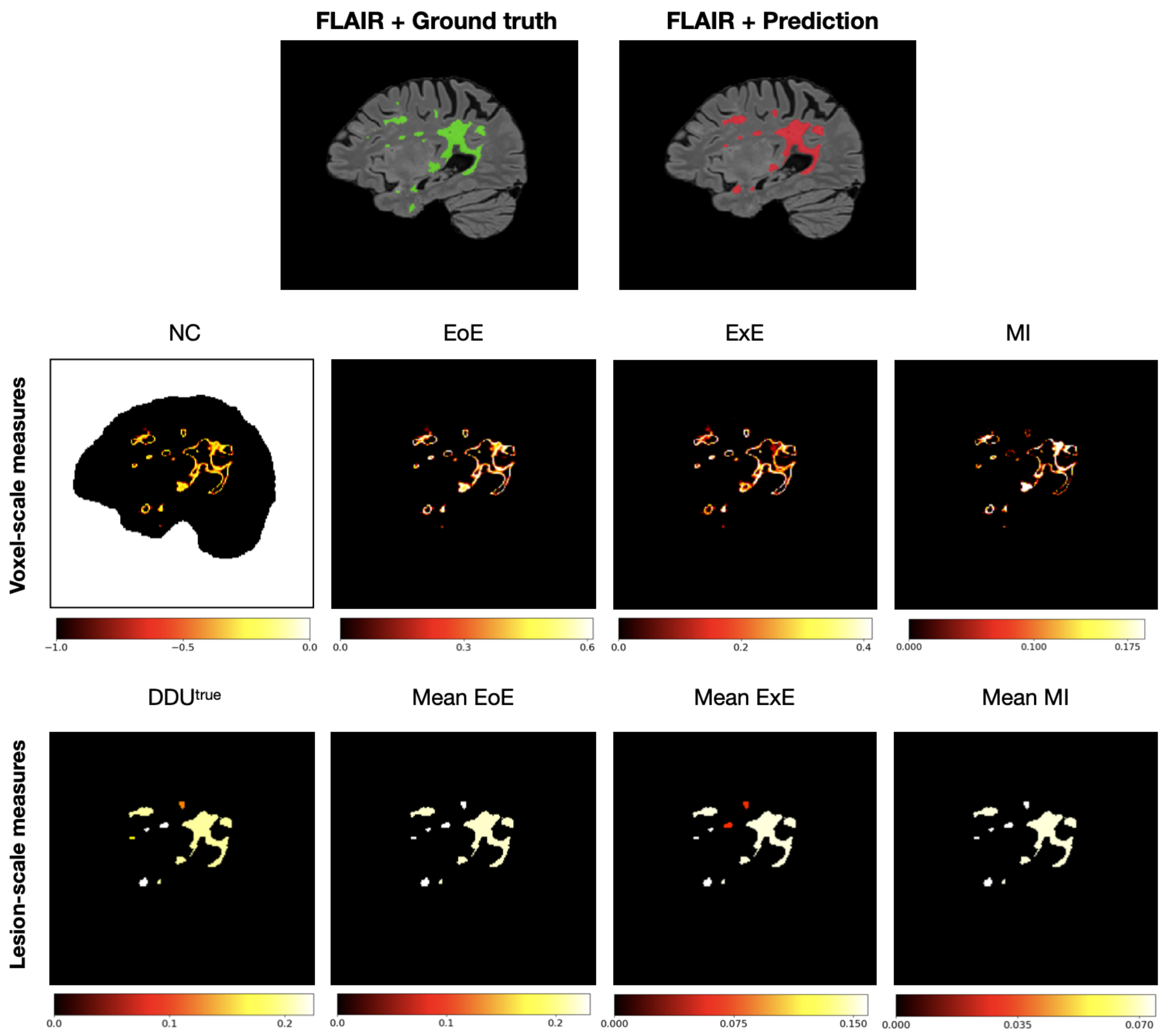


Figure 4

Resulting average across patients DSC-RC and LPPV-RC obtained on different sets of data, i.e in-domain and out-of-domain datasets separately and their joint set.

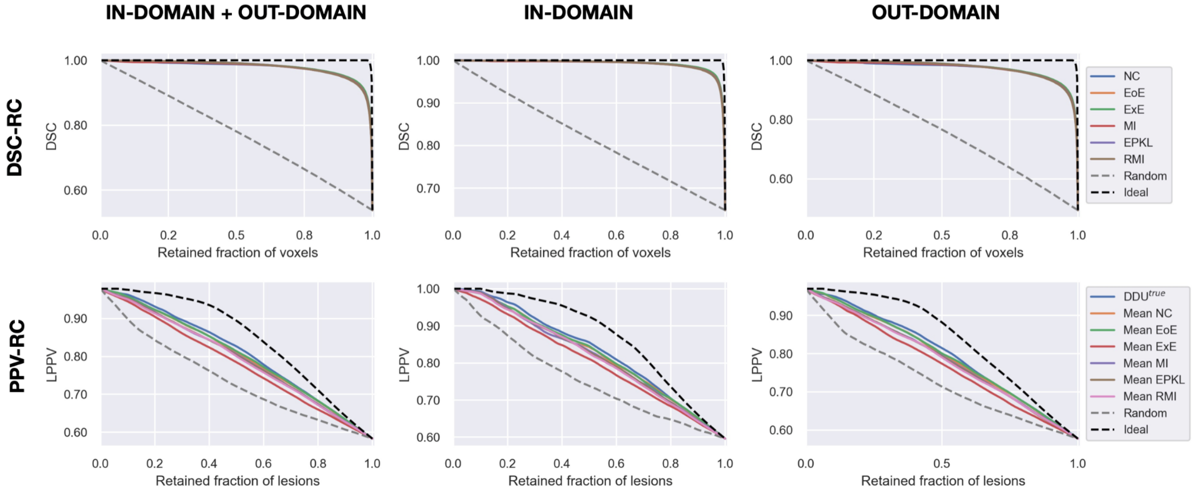
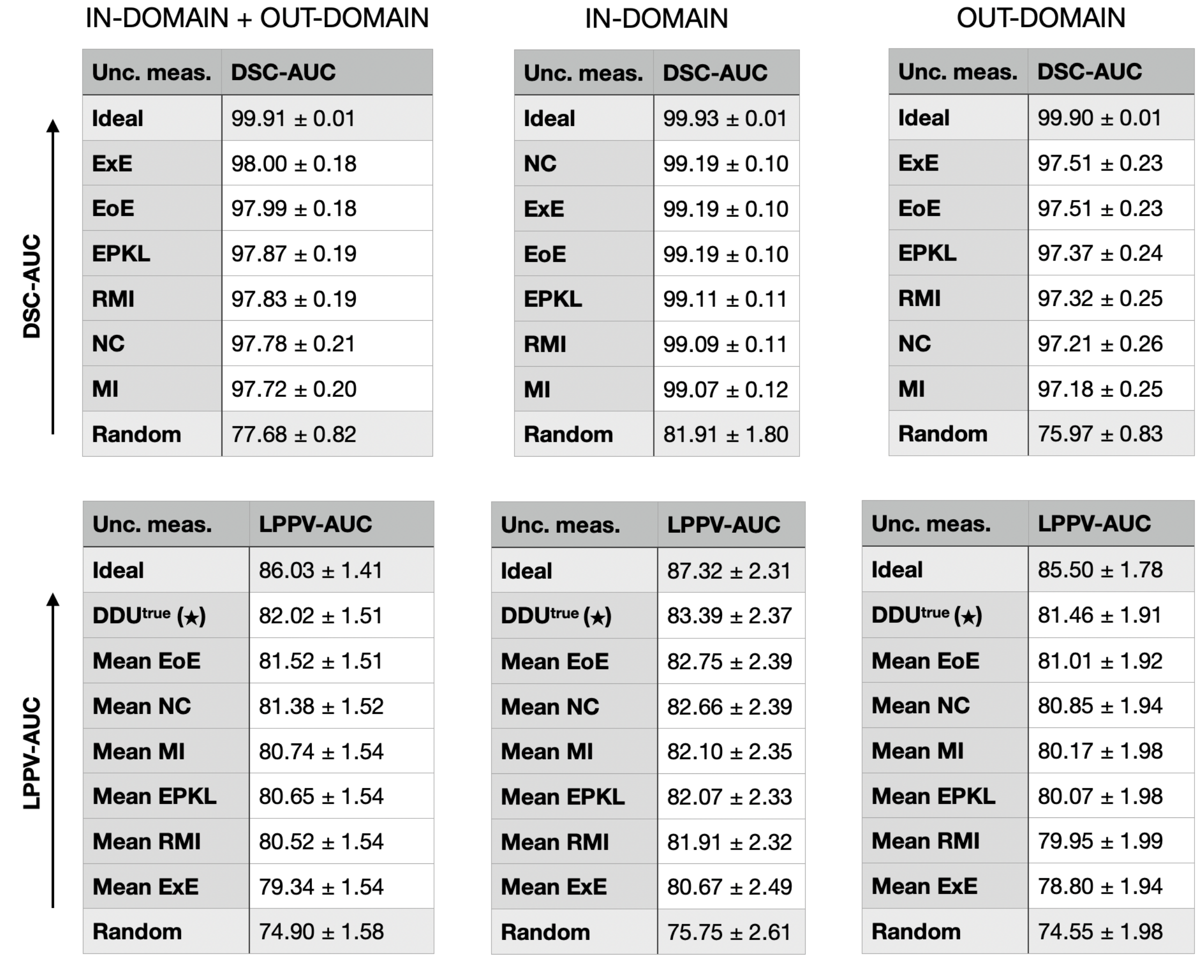


Figure 5

Resulting average across patients areas under the retention curves, i.e. DSC-AUC/LPPV-AUC, measuring the correspondence between voxel-/lesion-scale uncertainty measures and model errors in segmentation/lesion detection. AUCs computation performed on different sets of data:

in-domain and out-of-domain datasets separately and their joint set. Standard errors are computed using bootstrapping with the sample size of 85% of the population size for 10,000 repetitions.


DSC-AUC
LPPV-AUC

IN-DOMAIN + OUT-DOMAIN

Unc. meas.	DSC-AUC
Ideal	99.91 ± 0.01
ExE	98.00 ± 0.18
EoE	97.99 ± 0.18
EPKL	97.87 ± 0.19
RMI	97.83 ± 0.19
NC	97.78 ± 0.21
MI	97.72 ± 0.20
Random	77.68 ± 0.82

IN-DOMAIN

Unc. meas.	DSC-AUC
Ideal	99.93 ± 0.01
NC	99.19 ± 0.10
ExE	99.19 ± 0.10
EoE	99.19 ± 0.10
EPKL	99.11 ± 0.11
RMI	99.09 ± 0.11
MI	99.07 ± 0.12
Random	81.91 ± 1.80

OUT-DOMAIN

Unc. meas.	DSC-AUC
Ideal	99.90 ± 0.01
ExE	97.51 ± 0.23
EoE	97.51 ± 0.23
EPKL	97.37 ± 0.24
RMI	97.32 ± 0.25
NC	97.21 ± 0.26
MI	97.18 ± 0.25
Random	75.97 ± 0.83

Unc. meas.	LPPV-AUC
Ideal	86.03 ± 1.41
DDU^{true} (★)	82.02 ± 1.51
Mean EoE	81.52 ± 1.51
Mean NC	81.38 ± 1.52
Mean MI	80.74 ± 1.54
Mean EPKL	80.65 ± 1.54
Mean RMI	80.52 ± 1.54
Mean ExE	79.34 ± 1.54
Random	74.90 ± 1.58

Unc. meas.	LPPV-AUC
Ideal	87.32 ± 2.31
DDU^{true} (★)	83.39 ± 2.37
Mean EoE	82.75 ± 2.39
Mean NC	82.66 ± 2.39
Mean MI	82.10 ± 2.35
Mean EPKL	82.07 ± 2.33
Mean RMI	81.91 ± 2.32
Mean ExE	80.67 ± 2.49
Random	75.75 ± 2.61

Unc. meas.	LPPV-AUC
Ideal	85.50 ± 1.78
DDU^{true} (★)	81.46 ± 1.91
Mean EoE	81.01 ± 1.92
Mean NC	80.85 ± 1.94
Mean MI	80.17 ± 1.98
Mean EPKL	80.07 ± 1.98
Mean RMI	79.95 ± 1.99
Mean ExE	78.80 ± 1.94
Random	74.55 ± 1.98

(★) - statistically significant difference is detected between DDU^{true} and the rest of the measures according to one-sided Wilcoxon tests (H1: Median DDU^{true} > Median Unc. meas.) at a significance level of 0.01.