Computers in Biology and Medicine 184 (2025) 109336

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Structural-based uncertainty in deep learning across anatomical scales: Analysis in white matter lesion segmentation

Nataliia Molchanova [a,b,c,*], Vatsal Raina [d,1], Andrey Malinin [e,2], Francesco La Rosa [f], Adrien Depeursinge [a,b], Mark Gales [d], Cristina Granziera [g,h,i], Henning Müller [b,j], Mara Graziani [b], Meritxell Bach Cuadra [a,c]

[a] Radiology Department, University of Lausanne and Lausanne University Hospital, Lausanne, Switzerland
[b] MedGIFT, Institute of Informatics, School of Management, HES-SO Valais-Wallis University of Applied Sciences and Arts Western Switzerland, Sierre, Switzerland
[c] CIBM Center for Biomedical Imaging, Lausanne, Switzerland
[d] ALTA Institute, University of Cambridge, Cambridge, United Kingdom
[e] Isomorphic Labs, London, United Kingdom
[f] Icahn School of Medicine at Mount Sinai, New York City, United States of America
[g] Translational Imaging in Neurology (ThINK) Basel, Department of Medicine and Biomedical Engineering, University Hospital Basel and University of Basel, Basel, Switzerland
[h] Department of Neurology, University Hospital Basel, Basel, Switzerland
[i] Research Center for Clinical Neuroimmunology and Neuroscience Basel (RC2NB), University Hospital Basel and University of Basel, Basel, Switzerland
[j] Department of Radiology and Medical Informatics, University of Geneva, Geneva, Switzerland

ARTICLE INFO

Keywords:
Multiple sclerosis
White matter lesion segmentation
Magnetic resonance imaging
Deep learning
Uncertainty quantification
Instancescale uncertainty
Patientscale uncertainty

ABSTRACT

This paper explores uncertainty quantification (UQ) as an indicator of the trustworthiness of automated deep-learning (DL) tools in the context of white matter lesion (WML) segmentation from magnetic resonance imaging (MRI) scans of multiple sclerosis (MS) patients. Our study focuses on two principal aspects of uncertainty in structured output segmentation tasks. First, we postulate that a reliable uncertainty measure should indicate predictions likely to be incorrect with high uncertainty values. Second, we investigate the merit of quantifying uncertainty at different anatomical scales (voxel, lesion, or patient). We hypothesize that uncertainty at each scale is related to specific types of errors. Our study aims to confirm this relationship by conducting separate analyses for in-domain and out-of-domain settings. Our primary methodological contributions are (i) the development of novel measures for quantifying uncertainty at lesion and patient scales, derived from structural prediction discrepancies, and (ii) the extension of an error retention curve analysis framework to facilitate the evaluation of UQ performance at both lesion and patient scales. The results from a multi-centric MRI dataset of 444 patients demonstrate that our proposed measures more effectively capture model errors at the lesion and patient scales compared to measures that average voxel-scale uncertainty values. We provide the UQ protocols code at https://github.com/Medical-Image-Analysis-Laboratory/MS_WML_uncs.

1. Introduction

Multiple sclerosis (MS) is a chronic, progressive autoimmune disorder of the central nervous system affecting approximately 2.8 million people worldwide [1]. The primary characteristics of MS are demyelination, axonal damage, and inflammation due to the breakdown of the blood–brain barrier [2,3]. The diagnostic criteria for MS include both neurological symptoms observation and magnetic resonance imaging (MRI) examination for the presence of lesions disseminated in time and space [3–5]. White matter lesions (WMLs) are a hallmark of MS, indicating the regions of inflammation in the brain, typically assessed through FLAIR or T1-weighted modalities [4,6]. On FLAIR scans, WMLs are visible as hyperintense regions with periventricular area, brainstem, and spinal cord being prevalent lesion sites. The size, shape, and count of WMLs vary markedly across patients. While crucial for diagnosis and monitoring, the manual annotation of new and enlarged lesions is a time-consuming and skill-demanding process.

[*] Corresponding author at: Radiology Department, University of Lausanne and Lausanne University Hospital, Lausanne, Switzerland.
E-mail address: nataliia.molchanova@unil.ch (N. Molchanova).
[1] Work done while at HES-SO Valais, Sierre, Switzerland
[2] Work done while at Yandex Research, Moscow, Russia

https://doi.org/10.1016/j.compbiomed.2024.109336
Received 1 December 2023; Received in revised form 23 October 2024; Accepted 23 October 2024
Available online 15 November 2024

The task of automated WML segmentation has propelled the development of novel image processing techniques for many years [7,8]. More recently, algorithms have been boosted by the success of deep learning (DL) in computer vision. DL methods quickly became state-of-the-art for WML segmentation, providing better performance at faster processing times [9,10]. Various DL models were explored in application to WML segmentation, with U-Net architecture being the most common model at faster processing times [10].

The potential clinical application of DL methods raises safety concerns. These include the black-box nature of such approaches and their susceptibility to variations in test data, known as domain shifts [11]. Additionally, common factors such as limited data availability, imperfect annotations, and ground-truth ambiguity due to inter-rater variability further challenge the reliability of DL model predictions, potentially hindering their seamless integration into clinical practice [12]. The field of uncertainty quantification (UQ) offers a possibility to tackle this issue by estimating the "degree of untrustworthiness" of model predictions [12], focusing on two main uncertainty sources [13]: (i) data noise, captured by data uncertainty, and (ii) training data scarcity or domain shifts, captured by model uncertainty. In the context of high-risk AI applications, the information about the trustworthiness of model predictions is important not only from an engineering perspective, but also for the end-users, *e.g.* clinicians [14].

Consequently, UQ is gaining popularity within the field of medical image analysis not only as a way to assess prediction trustworthiness. However, the usage of uncertainty extends beyond quality control to accommodate such applications as improving prediction quality, domain adaptation, active learning, and other applications [13,15–17]. In medical image segmentation tasks, uncertainty is usually assessed by treating semantic segmentation as pixel or voxel classification, computing uncertainty for each pixel or voxel prediction. Given the structure of a segmentation model output, it is also possible to explore uncertainty values associated with some region of prediction. Several works explore uncertainty associated with a segmented region of interest, *e.g.* structure- or lesion-wise [18–22], or for a whole prediction on a patient [23,24].

1.1. Related works on uncertainty quantification in multiple sclerosis

Prior research on UQ for WML segmentation explored different techniques, including single-network deterministic methods [22,25], Monte Carlo Dropout (MCDP) [21], batch ensembles [22]. Our previous study [26] investigated the deep ensembles [27] and compared them with the MCDP method [28], showing the advantage of the first one. The utility of a specific UQ method depends on a particular application and available resources [13,15–17]. Deep ensembles were subsequently shown to have a higher quality of uncertainty estimates compared to other methods, while being computationally less effective compared to single-shot models or batch ensembles [13,15–17]. The deep ensemble is a deterministic method as the inference of each member is; thus, the reliability of this UQ method can be studied without a concern about the repeatability of the results.

Using ensemble methods or sampling UQ methods, based on obtaining samples from the posterior distribution, allows for the exploration of various uncertainty measures. Several measures of voxel-scale uncertainty have been explored, including variance, entropy, mutual information [21,29]. Our previous study expanded this list by exploring a common negated confidence and more advanced measures of model uncertainty, such as reverse mutual information and expected pairwise Kullback–Leibler divergence [26,30]. Several studies with different UQ methods and measures used, observe that voxel scale uncertainty tends to be the highest at the borders of WMLs, especially larger ones [21,25,26,29,30], resembling partial-volume [31,32] or inter-rater disagreement maps.

In MS, some works explored uncertainty associated with a segmented region of interest, *i.e.* at the lesion scale [21,29,30]. The pioneering study [21] suggested computing a log-sum of voxel-scale uncertainties across a predicted lesion region, using different voxel-scale uncertainty maps. Analogously, mean average voxel uncertainty values across the lesion region were explored [29]. Lambert et al. [29] showed the advantages of structural UQ based on graph neural networks over voxel aggregation methods. Our prior research [29] demonstrated that lesion-scale uncertainty, computed through disagreement in structural predictions, is more effective at identifying false-positive lesions than aggregating voxel-scale uncertainties. Although we explored advanced measures such as expected KL divergence and reverse mutual information [33], they did not exhibit any significant advantage over the more commonly employed entropy and mutual information in medical image analysis. In the context of MS lesion segmentation, the patient-scale uncertainty remains less explored.

Besides these various measures, prior works proposed different ways to compare uncertainty measures. Ideally, a high uncertainty score should highlight the predictions that are most likely to be wrong. Hence, we expect a reliable uncertainty measure to reflect the increased likelihood of an erroneous prediction and thus correlate with model mistakes. For classification tasks, a calibration of uncertainty is measured to assess its quality, similarly the uncertainty quality can be compared at the voxel scale. At the lesion-/ patient- scales the calibration metrics are not explicitly defined. When investigating lesion-scale measures, Nair et al. [21] looked into uncertainty-based prediction filtering as a means to correlate uncertainty and false positive errors, and Lambert et al. [29] used accuracy-confidence curves. Our previous work redefines an error retention curve analysis to quantify the relationship between uncertainty and lesion detection errors [30]. Prior to that the error retention curve analysis has been explored to compare classification or segmentation pixel-/voxel-scale uncertainty measures for various tasks as a way to quantify its relationship with an error/quality metric of a choice [33–35]. This is a necessary analysis for various practical clinical implementations, including a signaling uncertainty-based system to warn medical specialists about the potential errors in automatic predictions, automatic uncertainty-based filtering of errors, or active learning where the hardest, *i.e.* most likely erroneous examples need to be selected.

Various studies on UQ for WML segmentation use similar U-net-like deep learning models [21,26,29,36,37], which have been widely explored in application to the MS lesion segmentation task [7,10,38,39]. While there is an agreement about the DL model, studies were conducted on various datasets, predominantly private ones. There had not been a public benchmark dataset for the UQ methods evaluation within the context of WML segmentation before the Shifts 2.0 Challenge [26].

1.2. Our contributions

This study extends our previous work [30] and introduces advancements in uncertainty quantification (UQ) methods, focusing on MRI segmentation across voxel, lesion, and patient scales. We introduce a novel patient-scale uncertainty measure that leverages ensemble member disagreement to more accurately identify segmentation errors. To compare patient-scale measures, we redefine the error retention curve analysis, enabling a better understanding of their performance in detecting poor segmentation quality. Our quantitative evaluation is conducted in both in-domain and out-of-domain settings using a total of 404 scans to mirror the diversity of MRI data coming from several studies, medical centers, and scanners. Additionally, this research provides a comparison of uncertainty measures across different anatomical scales, highlighting their capacity to detect voxel misclassification, lesion false discovery, and general segmentation inaccuracies, considering clinically relevant applications. The proposed UQ framework is specifically tailored for WML segmentation on FLAIR MRI scans. Through additional evaluation, we confirm the generalizability of a similar task of white matter hyperintensity segmentation on 2D FLAIR MRI scans.

Our contributions include:

- Proposing the error retention curves analysis for instance-detection tasks, enabling an evaluation of lesion-scale UQ methods in their ability to capture lesion false detection errors.
- Proposing a patient-scale uncertainty measure, a novel approach for WML segmentation evaluation, enhancing the understanding of overall segmentation failure.
- Proposing the extension of the error retention curves analysis for patient-scale to compare the ability of different uncertainty measures to capture overall segmentation quality.

2. Materials and methods

2.1. Data

The initial study creating the data was designed as a part of the Shifts 2.0 Challenge [26] specifically for the exploration of uncertainty quantification across shifted domains. This configuration comprises three publicly available datasets and a single private one. Data is separated into in-domain (Train, Val, Test$_{in}$) and out-of-domain (Test$_{out}$) subsets. This enables UQ evaluation both with and without the domain shift. Data split into in- and out-of-domain sets is designed to maximize the drop of model performance in lesion segmentation in the out-of-domain test. From a clinical perspective, the domain shift is provided by the difference in medical center, scanner, annotators, and MS stages (Table 1). The Test$_{in}$ and Test$_{out}$ show a prominent difference in lesion distributions likely brought by the differences of MS stages distributions (see Fig. 1).

We extend this existing public benchmark by including a large in-house dataset (Test$_{private}$, 162 patients) collected in the Basel University Hospital, Switzerland [40]. While Test$_{private}$ should be treated as an out-of-domain, the lesion profiles overlap with both Test$_{in}$ and Test$_{out}$ (see Fig. 1).

For the additional assessment of generalizability and repeatability, we add an evaluation on a similar task of white matter hyperintensity (WMH) segmentation. We use a publicly available test set from the WMH Segmentation Challenge [41] comprising 110 subjects. On MRI FLAIR scans, WMH has a similar WML MS visual representation, but not localization [6]. WMHs come from a different pathology related to vascular abnormalities rather than MS [42]. The WMH Segmentation Challenge dataset contains 2D FLAIR scans with 3 mm thickness, compared to 0.8–2.2 mm slice thickness in the rest of the datasets. The lack of information in the z-axis contributes to the domain shift in addition to differences in study, medical center, underlying pathology, annotation protocol, among others. Additionally, this cohort exhibits higher lesion loads and larger lesion sizes (see Fig. 1).

For WML and WMH segmentation, this study uses FLAIR MRI scans and their manual WML annotations. FLAIR scans from Test$_{private}$ and Test$_{WMH}$ underwent a common pre-processing pipeline similar to the Shifts 2.0 Challenge pre-processing, including skull stripping [43], bias field correction [44], and interpolation to 1 mm isovoxel space. Information about data sources, metadata, and data splits is provided in Table 1. Fig. 1 illustrates some differences between domains brought by variations in MS stage distributions and scanner changes, affecting the lesion characterization and intensity features, respectively. Other factors, such as changes in study design, lesion annotators, scanner operators, may also contribute to the domain shift.

2.2. Uncertainty quantification

This work implements deep ensembles [27] for UQ by training multiple networks with identical architecture but different random seed initializations. The random seed controls several factors, for instance, weights initialization, training sample selection, random augmentations, and stochastic optimization algorithms. Although each ensemble member has distinct model weights, they all stem from the same posterior distribution. This causes varied predictions among ensemble members for the same input example. The spread or variation in these predictions serves as an uncertainty estimate.

2.2.1. Uncertainty quantification at different anatomical scales

In an image segmentation task, a class prediction is not a single value but an image-size map. Thus, the disagreement between the ensemble members can be quantified not only for each voxel of the prediction but also for a subset of its elements. For WML segmentation, the model prediction is a 3D probability map. We can quantify the uncertainty associated with the decision taken in each voxel, thus obtaining another 3D map with voxel-scale uncertainty values. We can also quantify uncertainty associated with a set of predictions within a region of a particular lesion, thus obtaining an uncertainty score for each predicted lesion. Similarly, we can quantify uncertainty for the whole patient. We implement several uncertainty measures at each anatomical scale (voxel, lesion, or patient). The exact mathematical formulation for the previous existing and proposed UQ measures are summarized in Table 2 and described hereafter.

Voxel-scale uncertainty measures. Perceiving segmentation as a classification of each voxel of an image, one could use uncertainty measures available for classification tasks to quantify uncertainty for per-voxel predictions. The common uncertainty measures in this case will be negated confidence and information theory measures such as entropy of expected, expected entropy, or mutual information which respectively depict different *total*, *data*, and *model* uncertainty.

Lesion-scale uncertainty measures. Given a WML segmentation task, we can compute a single uncertainty score for each predicted connected component, *i.e.* lesion. Differently from previous measures that aggregate voxel-scale uncertainties [21,22]. Our previous work [30] proposes a novel lesion-scale uncertainty defined directly through the disagreement between the lesion structural predictions of ensemble members. We hypothesize that looking at the disagreement in structural predictions, *i.e.* predicted lesion regions, might be more beneficial for the discovery of false positive lesions.

To define our proposed measure, we consider the ensemble of M models, where each member model is parametrized by weights θ^m, $m \in \{0, 1, \dots, M-1\}$. The ensemble probability prediction is obtained by computing a mean average across members. Then, the binary lesion segmentation mask is obtained by applying a threshold α to the softmax ensemble prediction, where α is chosen based on the Dice similarity coefficient maximized on the validation dataset. Analogously, by applying the threshold α to the softmax predictions of each of the ensemble models, we can obtain the binary lesion segmentation masks predicted by each model m in the ensemble. Let L be a *predicted lesion* that is a connected component from the binary segmentation map obtained from the ensemble model; and L^m is the *corresponding lesion* predicted by the model m, determined as the connected component on the binary segmentation map predicted by the m-th member with maximum intersection over union (IoU) with L. If the softmax probability threshold is optimized for each member model separately based on the highest Dice score, the resulting thresholds will be different from α and will be member-specific: α^m, $m \in \{0, 1, \dots, M-1\}$, instead of α. Then, the binary segmentation maps obtained with α^m will lead to different corresponding lesion regions, called $L^{m,+}$. Then, the proposed measure, lesion structural uncertainty (LSU), is defined as follows:

$$LSU = 1 - \frac{1}{M}\sum_{m=0}^{M-1} IoU(L, L^m), \tag{1}$$

Table 1
Data splits and meta information. MS stages are clinically isolated syndrome (CIS), relapsing remitting (RR), primary progressive (PP), and secondary progressive (SP). Computed statistics are median (Q2) and interquartile range (IQR). Computed statistics are median (Q2) and interquartile range (IQR).

Domain	In-domain			Out-of-domain MS		Out-of-domain WMH
Source	Carass et al. [45], Commowick et al. [46]			Lesjak et al. [47], Bonnier et al. [48]	Granziera [40]	Kuijf et al. [41]
Medical center location	Rennes, Bordeaux and Lyon (France), Best (Netherlands)			Ljubljana (Slovenia), Lausanne (Switzerland)	Basel (Switzerland)	Utrecht and Amsterdam (Netherlands), Singapore
Scanners	Siemens (Aera 1.5T, Verio 3.0T), GE Disc 3.0T, Philips (Ingenia 3.0T, Medical 3.0T)			Siemens Magnetom Trio 3.0T	Siemens Magnetom Prisma 3.0T	3T Philips Achieva, Siemens TrioTim 3.0T, Philips Achieva 3.0T, Ingenuity 3.0T, GE Signa (1.5T, 3.0T)
M:F ratio range	0.21–0.4			0.23-0.70	0.68	–
MS stages	RR, PP, SP			CIS, RR, SP, PP	RR, PP, SP	–
# raters	2/7			consensus/3	consensus	consensus
Inter-rater agreement (Dice score)	0.63 and 0.71			0.78 and -	-	–
Set name	Train	Val	Test$_{in}$	Test$_{out}$	Test$_{private}$	Test$_{WMH}$
# scans	33	7	33	99	162	110
# lesions per scan, Q2 (IQR)	34 (20–50)	26 (19–61)	30 (15–47)	39 (20-77)	63 (25-88)	60 (37-83)
Total lesion volume per scan, Q2 (IQR) [mL]	12.5 (3.1-27.8)	15.5 (4.0–24.7)	7.2 (3.7–11.3)	2.7 (1.3–7.3)	7.4 (2.4–14.3)	9.4 (3.3–20.3)

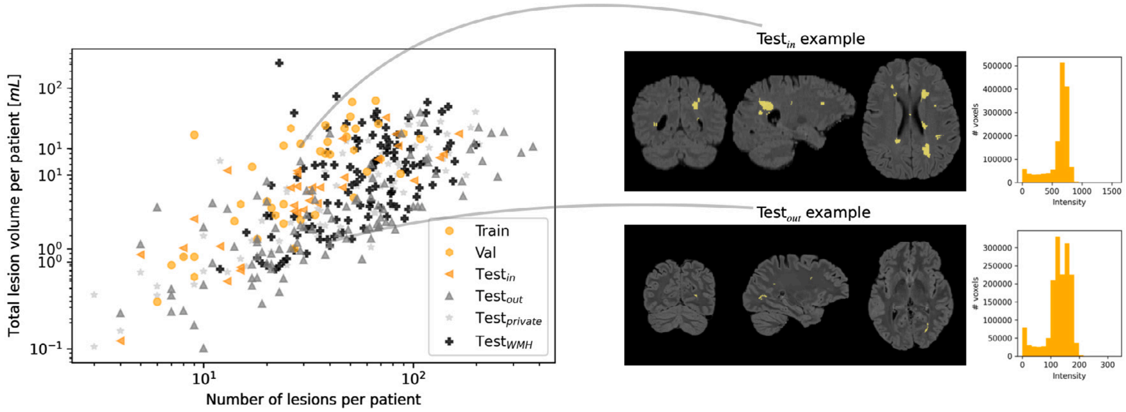


Fig. 1. Illustration of the domain shift between the in-domain datasets (Train, Val, Test$_{in}$) and the out-of-domain dataset (Test$_{out}$, Test$_{private}$, and Test$_{WMH}$) brought by the differences in the MS stages and medical centers. On the left, the plot of the total lesion volume in milliliters versus the number of lesions per scan for in-domain (orange) and out-of-domain (gray and black) sets reveals the difference in the lesion load (as a proxy to an MS stage) between different domains. On the right, typical examples from the Test$_{in}$ and Test$_{out}$ sets illustrate the difference in the lesion load, as well as the intensity differences brought by the change of the medical center (*i.e.* scanner, technicians, annotators, and other parameters contributing to the domain shift) and MS stages (*i.e.* smaller lesion load and size).

and

$$LSU^+ = 1 - \frac{1}{M} \sum_{m=0}^{M-1} IoU(L, L^{m,+}). \tag{2}$$

Patient-scale uncertainty measures. Patient-scale uncertainty offers the most compact way of uncertainty representation considering the clinical practice, that is presenting a single uncertainty score per patient. Analogously to the lesion scale, the patient-scale uncertainty can be computed by averaging voxel or lesion uncertainties. Using similar reasoning as for the lesion scale, we propose a patient-scale measure analogous to LSU (Eq. (1)), where instead of the lesion region L the total segmented lesion region is used. To define these measures, let S be a set of voxels predicted as lesion class by the ensemble model, S^m - set of voxels predicted as lesion class by the m-th member model in the ensemble, and $S^{m,+}$ is the same, but obtained with the member-specific threshold α^m. Then, the proposed patient structural uncertainty measures are defined as:

$$PSU = 1 - \frac{1}{M} \sum_{m=0}^{M-1} IoU(S, S^m), \tag{3}$$

and

$$PSU^+ = 1 - \frac{1}{M} \sum_{m=0}^{M-1} IoU(S, S^{m,+}). \tag{4}$$

2.3. Quantitative evaluation of uncertainty measures

Uncertainty has a relation to errors made by a model: ideally, a higher uncertainty expresses an increased likelihood of erroneous prediction. For each of the anatomical scales: voxel, lesion, and patient, the “error” definition can vary. For example, a voxel-scale error can be simply defined as a voxel misclassification, a lesion-scale error can be defined as a lesion misdetection, and a patient-scale error can be a summary of voxel errors. In this work, we want to compare voxel-, lesion-, and patient-scale uncertainty measures in terms of their ability to capture errors of different kinds. For this, we use an error retention curve analysis [26,34,35], previously introduced only for voxel-scale uncertainty, and extended for lesion and patient scales in this work.

Table 2
Definitions of uncertainty measures at three anatomical scales: voxel, lesion, and patient.

(a) **Voxel-scale uncertainty measures** computed for each pixel $i \in B$ of an input scan $\mathbf{x}$ (B is a set of voxels defining the brain region), $\mathbf{y}$ - targets, $c \in \{0, 1, \dots, C-1\}$ is the class ($C = 2$ for binary WML segmentation), $P(y_i = c\|\mathbf{x}, \theta^m)$ is a softmax probability predicted by the m-th member in the ensemble of M models, and $\hat{P}(y_i = c\|x)_i = \frac{1}{M}\sum_{m=0}^{M-1} P(y_i = c\|\mathbf{x}, \theta^m)$ is the probability predicted by ensemble.	
Negated confidence	$NC_i = -\underset{c\in\{0,1,\dots,C-1\}}{\operatorname{argmax}} \frac{1}{M}\sum_{m=0}^{M-1} P(y_i = c\|\mathbf{x}, \theta^m)$
Entropy of expected	$EoE_i = -\sum_{c=0}^{C-1} \hat{P}(y_i = c\|x) \log \hat{P}(y_i = c\|x)$
Expected entropy	$ExE_i = -\frac{1}{M}\sum_{m=0}^{M-1}\sum_{c=0}^{C-1} P(y_i = c\|\mathbf{x}, \theta^m) \log P(y_i = c\|\mathbf{x}, \theta^m)$
Mutual information	$MI_i = EoE_i - ExE_i$
(b) **Lesion-scale uncertainty measures** computed for each *predicted lesion* L, that is a connected component on the predicted binary segmentation map. The last is obtained by applying a threshold α to the softmax ensemble prediction $\hat{P}(\mathbf{y} = \mathbf{1}\|x) = \frac{1}{M}\sum_{m=0}^{M-1} P(\mathbf{y} = \mathbf{1}\|\mathbf{x}, \theta^m)$, where α is chosen based on the Dice similarity coefficient maximized on the validation dataset. L^m is the corresponding lesion predicted by the m-th member model, determined as the connected component on the binary segmentation map predicted by the m-th member (threshold α applied to $P(\mathbf{y} = \mathbf{1}\|\mathbf{x}, \theta^m), m \in \{0, 1, \dots, M-1\}$) with maximum intersection over union (IoU) with L. If the softmax probability threshold is optimized based on the highest Dice score for each member model separately, the resulting thresholds will be different from α and will be member-specific: $\alpha^m, m \in \{0, 1, \dots, M-1\}$ instead of α. Then, the binary segmentation maps obtained by applying α^m to $P(\mathbf{y} = \mathbf{1}\|\mathbf{x}, \theta^m), m \in \{0, 1, \dots, M-1\}$ will lead to different corresponding lesion regions, called $L^{m,+}$.	
Voxel uncertainties aggregation via mean average	$\overline{EoE}_L = \frac{1}{\|L\|}\sum_{i\in L} EoE_i$. Analogously, $\overline{ExE}_L, \overline{NC}_L, \overline{MI}_L$ are defined.
Proposed lesion structural uncertainty (LSU)	$LSU = 1 - \frac{1}{M}\sum_{m=0}^{M-1} IoU(L, L^m)$ and $LSU^+ = 1 - \frac{1}{M}\sum_{m=0}^{M-1} IoU(L, L^{m,+})$
(c) **Patient-scale uncertainty measures** computed for patient. S is a set of voxels in a scan predicted as lesions by the ensemble model, S^m is a set of voxels predicted as lesions by the model m, and $S^{m,+}$ is the same, but obtained with the member-specific threshold $\alpha^m, m \in \{0, 1, \dots, M-1\}$. W - set of lesions predicted by the ensemble model.	
Voxel uncertainties aggregation via mean average	$\overline{EoE}_B = \frac{1}{\|B\|}\sum_{i\in B} EoE_i$. Analogously, $\overline{ExE}_B, \overline{NC}_B, \overline{MI}_B$ are defined.
Proposed lesion uncertainties aggregation via mean average	$\overline{LSU} = \frac{1}{\|W\|}\sum_{l\in W} LSU_l$. Analogously, $\overline{LSU^+}$ is defined.
Proposed patient structural uncertainty (PSU)	$PSU = 1 - \frac{1}{M}\sum_{m=0}^{M-1} IoU(S, S^m)$ and $PSU^+ = 1 - \frac{1}{M}\sum_{m=0}^{M-1} IoU(S, S^{m,+})$

2.3.1. Error and quality metrics

We start by defining errors on the voxel and lesion scale as well as quality metrics used in this work for model performance characterization and error retention curve analysis.

Voxel-scale errors. Similarly to a classification task, the errors at the voxel scale will include false positives and negatives (FP and FN, respectively). Based on FP, FN, true positives (TP), and true negatives (TN), one derives metrics like true positive rate (TPR) and positive predictive value (PPV), which measure correctly classified voxels against ground truth or predicted lesions, respectively. To evaluate both error types, we use the F_1 score, also known as the Dice similarity score (DSC) in image processing. However, it is well known that the DSC metric suffers from a bias to the occurrence rate of the positive class, *i.e.* lesion load, jeopardizing the comparison of results. We thus additionally utilize the normalized DSC (nDSC) [49] for the model evaluation. In a nutshell, nDSC scales the precision at a fixed recall rate to tackle the lesion load bias.

Lesion-scale errors. Analogously, true positive, false positive, and false negative lesions (TPL, FPL, FNL) can be defined if the criteria for lesion (mis)detection are given. While some studies accept minimal overlap for detection [21,39,45], we apply a 25% intersection over the union threshold for a predicted lesion to be considered a TPL. For the FNL definition, we consider a zero overlap with the prediction. A FNL is a ground truth lesion that has no overlap with predictions. Metrics derived from TPL, FPL, and FNL include Lesion TPR, PPV, and F_1, further referred to as LTPR, LPPV, LF_1. The differences at the voxel scale include: (i) uncertainty cannot be quantified for FNLs, as they are not predicted lesions; (ii) it is not possible to define a true negative lesion. The metrics definitions can be found in Appendix A.

2.3.2. Error retention curve analysis

The error retention curve (RCs) [26,34,35] assess the correspondence between a chosen uncertainty measure and an error or a quality metric. By quantifying this correspondence for various uncertainty measures we can choose a measure that is better at pointing out errors in model predictions. This is relevant for clinical applications, where uncertainty constitutes a signaling system requiring human verification.

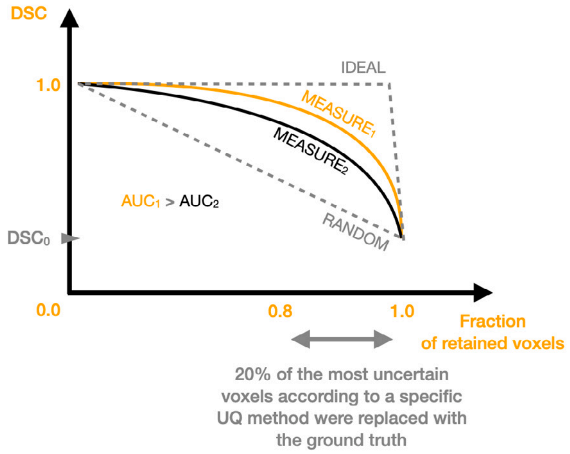


Fig. 2. An illustration of a Dice score retention curve (DSC-RC) for assessing the correspondence between voxel uncertainty ($\mathrm{MEASURE}_1$ and $\mathrm{MEASURE}_2$) and segmentation quality measured by DSC. DSC_0 - quality of the predicted segmentation before voxel replacement. IDEAL and RANDOM RCs are built for the ideal and random uncertainty and are the upper and lower bounds of the uncertainty-robustness performance.

Compared to the uncertainty calibration analysis [13], error RCs only consider the ranking of uncertainty values within a particular scan, thus, avoiding uncertainty values scaling present in the calibration metrics. Additionally, they allow for the choice of a quality metric w.r.t. to which the uncertainty measure is compared. Thus, allowing for extending their definition to different scales, *e.g.* lesion or patient. Moreover, compared to calibration metrics, the RC analysis allows us to estimate the upper and lower bounds of the uncertainty-robustness performance.

Voxel-scale DSC-RC. Similarly to our previous investigation [30], we use voxel-scale RCs to quantify the average across patients correspondence between per-voxel uncertainty and DSC, *i.e.* per-voxel misclassification errors of different kinds: either FP or FN. For one patient, a voxel-scale DSC-RC is built by sequentially replacing a fraction τ of the most uncertain voxel predictions within the brain mask with the ground truth and re-computing the DSC. If one measure has a better ability to capture model errors than another measure, then the most uncertain voxels will be faster replaced with the ground truth and the DSC-RC will grow faster. Thus, the area under the DSC retention curve (DSC-RC), further referred to as DSC-AUC can be used to compare different uncertainty measures in their ability to capture model segmentation errors. It is possible to estimate lower and upper bounds of performance by building *random* and *ideal* RCs. For a random RC, we assign random uncertainty values to each voxel of predictions. For the ideal one, a zero uncertainty is assigned to true positive and negative (TP and TN) voxels while false positive and negative (FP and FN) voxels have an uncertainty of 1. To build the RCs, we use $\tau = 2.5 \cdot 10^{-3}$. An illustrative explanation of a voxel-scale RC can be found in Fig. 2.

Lesion-scale LPPV-RC (proposed). In our previous investigation [30] we proposed an extension of the error RC analysis to the lesion scale through LF1-RC. LF1-RC assesses the correspondence between lesion-scale uncertainty and errors in lesion detection within a patient. As defined in Section 2.3.1, the LF1 is reflective of both FNL and FPL. However, uncertainty cannot be defined for FNLs as they are not predicted, but ground-truth lesions. Thus, LF1-RCs are more suitable for the comparison of different models or uncertainty quantification methods, for which the number of FNL can vary. However, for the comparison of lesion-scale uncertainty measures, where the number of FNLs does not change, the LPPV-RC analysis is sufficient. Thus, we propose the LPPV-RC assesses the correspondence between lesion-scale uncertainty and lesion false positive errors within a patient. Intuitively, this analysis helps to understand which uncertainty measure is the best at pointing to false positive lesions.

Building a LPPV-RC for a patient starts with computing the number of TPL and FPL, *i.e.* $\#_{TPL}$ and $\#_{FPL}$, and uncertainty values for each of these lesions. Further, the most uncertain lesions are sequentially replaced with TPL, and LPPV is recomputed. Analogously to the voxel scale, if a lesion-scale uncertainty measure has a better ability to capture FPL than another measure, then FPL will be replaced faster, and the curve will grow faster. Thus, the area under the LPPV-RC, that is LPPV-AUC, can be used to compare different measures in their ability to capture FPL detection errors. As each patient has a different number of predicted lesions, to obtain an average across the dataset LPPV-AUC, we first need to interpolate all LPPV-RCs to a similar set of retention fractions. For this, we use a piecewise linear interpolation and a set of retention fractions similar to the voxel scale. Additionally, similarly to the voxel scale, the ideal and random RCs are built. The ideal curve is built by considering all TPLs having an uncertainty of 0 and all FPLs having an uncertainty of 1. The random curve is built by using random uncertainties for each of the lesions.

Patient-scale DSC-RC (proposed). In this work, we propose a way to extend an error RCs analysis to the patient scale to assess the correspondence between patient-scale uncertainty measures and overall prediction quality in a patient. We use DSC as a measure of overall segmentation quality. Then, a patient-scale DSC-RC is built by sequentially excluding the most uncertain patients, that is replacing their DSC with 1.0, and recomputing the average across the dataset DSC. Similarly to the voxel and lesion scales, the area under the patient-scale DSC-RC is used to compare the ability of different patient-scale uncertainty measures to capture patients with a greater number of erroneous predictions. In analogy to the voxel and lesion scales, we want to assess the upper and lower bounds of the performance with ideal and random patient-scale DSC-RCs. To build a random curve we assign random uncertainties to each of the patients. To build the ideal curve, we use a negated DSC score as an uncertainty measure, as we want ideal uncertainty to point to the most erroneous examples in terms of DSC.

Statistical testing. For the voxel and lesion scales, the error retention curves analysis, namely DSC-RC and LPPV-RC, are computed per patient. Therefore, when comparing different uncertainty measures across each other, one can assess the differences in AUC distributions across measures, *e.g.* statistics. For the patient scale, DSC-RC is computed per dataset (by iterative replacement of the most uncertain patients). Nevertheless, it is possible to estimate the bootstrap confidence intervals by treating the patient-scale DSC-RC as a statistic itself. Thus, to conduct the measures ranking for the patient-scale uncertainty measures, we compare the mean patient-scale DSC-AUC, paying attention to the corresponding confidence intervals.

2.3.3. *Patient-scale uncertainty as a proxy for segmentation quality*

In addition to the information brought by the error RC, we would like to study if a patient-scale uncertainty can serve as a proxy to the model segmentation quality, measured by DSC. For this, we compute Spearman's correlation coefficient ρ between the DSC and uncertainty values. The Spearman's correlation is computed for different test sets separately, and then jointly. The joint correlation coefficient should show if the uncertainty measure can be used as a proxy for the segmentation quality regardless of the domain shift. This might be particularly useful for the scenario where the domain shift is unknown.

2.4. *WML segmentation model*

For this study, we consider two models based on a 3D U-Net architecture. Similar 3D-U-net-based models have been previously used for WML segmentation and compared to other approaches [7,9,10,39]. Furthermore, our choice is supported by the fact that the same model has been extensively used previously for UQ exploration within the same WML segmentation task in MS [21,22,25,26,29]. The first model is the baseline model from the Shifts 2.0 Challenge [26] dedicated to UQ for WML segmentation. The second model is a self-configuring nnU-Net architecture [43]. Both models are ensembles with 5 members, where each member is a 3D U-Net model [36,37]. There are several crucial differences between the Shifts Baseline (SB) U-Net and the nnU-Net models: (i) architecture, *i.e* SB has the depth reduced by one and, thus, less trainable parameters; (ii) loss function, *i.e.* Focal-Dice loss for SB and cross-entropy and Dice loss for nnU-Net; (iii) deep supervision is utilized by nnU-Net, compared to SB; (iv) input, SB's input are patches of the size 96 × 96 × 96 cropped from the brain using a sequence of transforms, while nnU-Net uses patches 112 × 160 × 128 cropped around the whole brain. Both models represent public benchmarks, and their training and inference code is available online.[3] For the SB model, the only difference, compared to the original model, is an addition of 2 more ensemble members, obtained using the original training code. For the nnU-Net model, we used a "3d_fullres" configuration, we ensured the consistency of training and validation examples across folds (for the model to be comparable to SB) and limited the number of training epochs to 200 (due to the validation loss stagnation, to prevent overfitting). Since the Shifts dataset does not contain lesions less than 10 voxels, we process the outputs of each of the models to remove all the connected components with less than 10 voxels.

[3] The original code including model implementation and weights, training and inference code can be found at the Shifts Challenge GitGub: https://github.com/Shifts-Project/shifts/tree/main/mswml. nnU-Net model code is publicly available at https://github.com/MIC-DKFZ/nnUNet. Model weights can be found on our GitHub: https://github.com/Medical-Image-Analysis-Laboratory/MS_WML_uncs.

Table 3
Mean average model performance in segmentation (DSC and nDSC) and lesion detection (LF1 and LPPV). 90% confidence intervals were computed using bootstrapping. SB - Shifts 2.0 Challenge baseline model.

Set	DSC		nDSC		LF1		LPPV	
	SB	nnU-Net	SB	nnU-Net	SB	nnU-Net	SB	nnU-Net
Train	0.756 [0.737, 0.774]	0.906 [0.892, 0.917]	0.725 [0.699, 0.749]	0.856 [0.826, 0.883]	0.547 [0.493, 0.596]	0.845 [0.787, 0.876]	0.689 [0.627, 0.735]	0.971 [0.957, 0.981]
Val	0.720 [0.602, 0.783]	0.776 [0.701, 0.821]	0.684 [0.625, 0.740]	0.736 [0.669, 0.783]	0.444 [0.345, 0.547]	0.643 [0.555, 0.707]	0.533 [0.425, 0.608]	0.762 [0.624, 0.871]
Test$_{in}$	0.633 [0.582, 0.673]	0.707 [0.671, 0.739]	0.689 [0.662, 0.717]	0.741 [0.715, 0.768]	0.487 [0.439, 0.528]	0.701 [0.666, 0.733]	0.610 [0.552, 0.660]	0.762 [0.721, 0.797]
Test$_{out}$	0.488 [0.457, 0.515]	0.571 [0.538, 0.600]	0.533 [0.501, 0.560]	0.603 [0.570, 0.630]	0.333 [0.308, 0.361]	0.502 [0.477, 0.525]	0.623 [0.586, 0.659]	0.828 [0.799, 0.852]
Test$_{private}$	0.601 [0.578, 0.621]	0.646 [0.626, 0.665]	0.628 [0.608, 0.645]	0.653 [0.635, 0.670]	0.416 [0.396, 0.437]	0.562 [0.543, 0.581]	0.581 [0.556, 0.605]	0.799 [0.779, 0.817]
Test$_{WMH}$	0.591 [0.564, 0.616]	0.648 [0.623, 0.671]	0.599 [0.580, 0.617]	0.651 [0.632, 0.668]	0.373 [0.353, 0.391]	0.555 [0.534, 0.574]	0.488 [0.456, 0.518]	0.696 [0.665, 0.724]

3. Results

3.1. Model performance

The evaluation of the ensemble model performance in terms of average segmentation and lesion detection quality is presented in Table 3 for training, validation, and testing sets. Regardless of the model, SB or nnU-Net, the in-domain performance reaches its upper bound determined by the inter-rater agreement reported in. There is a considerable drop in performance (around 10% depending on the metric) between in- and out-of-domain sets both in terms of segmentation (DSC and nDSC) and lesion detection (LF1). The performance on Test$_{private}$ and Test$_{WMH}$ datasets lies in between Test$_{in}$ and Test$_{out}$ with regards to segmentation and lesion detection quality. Between the two models, nnU-Net shows higher performance in terms of segmentation and lesion detection.

3.2. Quantitative evaluation of uncertainty measures

3.2.1. Error retention curve analysis

The RCs for the assessment of uncertainty measures on each of the anatomical scales (voxel, lesion, and patient) are presented in Fig. 3. The voxel-scale DSC-RCs and lesion-scale LPPV-RCs were obtained by averaging across the respective datasets. The mean areas under the error retention curves and the results of the statistical testing are presented in Table 4.

Regardless of the test set, all **voxel-scale** uncertainty measures outperform random uncertainty and are closer to the ideal uncertainty in terms of mean DSC-AUC, indicating their ability to capture errors in segmentation. However, the marginal difference between DSC-AUCs of different measures is relatively small. On the in-domain Test$_{in}$, there is no agreement between two models in terms of the measures with the highest mean DSC-AUC: while total and data uncertainty (NC_i, EoE_i, ExE_i) have higher DSC-AUC for the SB model, model uncertainty (MI_i) has a higher DSC-AUC for the nnU-Net model. On the out-of-domain Test$_{out}$ and Test$_{private}$ datasets, the entropy-based total and data uncertainty measures (EoE_i and ExE_i) tend to have an advantage compared to other measures, contributing to their overall advantage in the whole evaluation. Nevertheless, the aggregation of data uncertainty ExE_i for the lesion-/ patient- uncertainty computation usually yields the worst results in terms of lesion-scale LPPV-AUC/patient-scale DSC-AUC. This means that a good performance of an uncertainty measure in capturing voxel misclassifications, when aggregated, does not necessarily lead to an optimal uncertainty measure for detecting lesion false positive or overall segmentation failure.

Regardless of the test set, at the **lesion scale**, there is a greater marginal difference between different measures, particularly for the SB model. For the SB model, the proposed measure LSU^{+} has an advantage in the mean LPPV-AUC over other measures, indicating a better ability to capture lesion false positive errors. While LSU and LSU^{+} have similar LPPV-AUCs, there is usually some difference in their performances, benefiting the LSU^{+} measure. Among the measures based on the aggregation of voxel uncertainties, aggregated total uncertainty $\overline{EoE}_L$, generally provides slightly higher mean LPPV-AUC. Despite the differences between the mean LPPV-AUCs among lesion-scale measures, the 90% confidence interval overlap suggests that these differences are limited.

At the **patient scale**, the marginal differences between various measures are prominent compared to the voxel and lesion scales, especially on the out-of-domain sets. The results are aligned for both in- and out-of-domain test sets and for both models, SB and nnU-Net. The proposed PSU and PSU^{+} measures have comparable and the highest patient-scale DSC-AUCs, suggesting their superior ability to capture overall segmentation failure. The aggregation of the best in terms of LPPV-AUC lesion scale uncertainty (*i.e.* LSU and LSU^{+}) yields lower patient DSC-AUC. Averaging voxel uncertainties across the brain generally provides worse-than-random performance in the error retention curve analysis. The last means that an average across-subject voxel-scale uncertainty is not informative of an overall segmentation performance on a particular subject measured by DSC or has an inverse relationship with errors.

3.2.2. Patient-scale uncertainty as a proxy to the segmentation quality

Extending the analysis of the relationship between the patient-scale uncertainty measures and the segmentation quality measures by DSC, Table 5 presents corresponding Spearman's correlation coefficients. Fig. 4 contains plots DSC and patient uncertainty for the measures with the highest (proposed $PSU^{(+)}$), median (proposed $\overline{LSU}^{(+)}$), and worse-than-random ($\overline{NC}_B$ and $\overline{EoE}_B$) patient-scale DSC-AUC values. For the SB model and the rest of the measures, the same analysis and trends can be found in Appendix B.2. The results show the highest correlation between the patient uncertainty and DSC is provided by the proposed $PSU^{(+)}$ measures, with ρ around 0.8 across different test sets. For the aggregation of the lesion-scale uncertainty, the correlation with the segmentation quality drops at least twice. For the measures based on the voxel-scale uncertainty aggregation, the correlation is either weak, *e.g.* $\overline{NC}_B$, or positive. There is a positive correlation between $\overline{EoE}_B$, $\overline{ExE}_B$, and $\overline{MI}_B$, suggesting that high uncertainty can point to examples with high DSC. The absolute value of this correlation is around 0.5, which is higher than for $\overline{LSU}^{(+)}$, yet lower than for the proposed $PSU^{(+)}$.

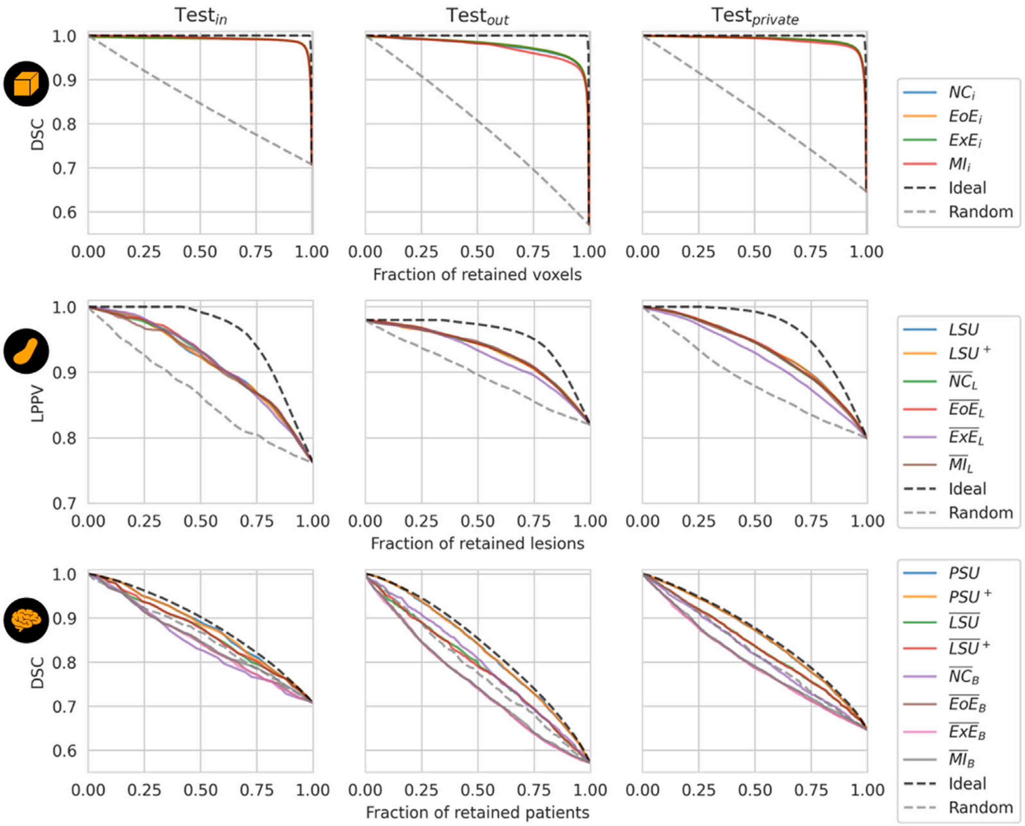


Fig. 3. Error retention curves for the assessment of uncertainty measures at the voxel, lesion, and patient anatomical scales across the in-domain Test$_{in}$ (left column) and the out-of-domain Test$_{out}$ (center column) and Test$_{private}$ (left column) sets for the nnU-Net model. Different rows correspond to different anatomical scales indicated with icons on the left. The voxel-scale DSC-RCs and lesion-scale LPPV-RCs were obtained by averaging across the respective datasets. At each of the scales, the ideal (black dashed) line indicates the upper bound of an uncertainty measure performance in its ability to capture model errors; the random (gray dashed) indicates no relationship between an uncertainty measure and error; a worse-than-random performance indicates an inverse relationship. Analogous results for the SB model are shown in Appendix B.1.

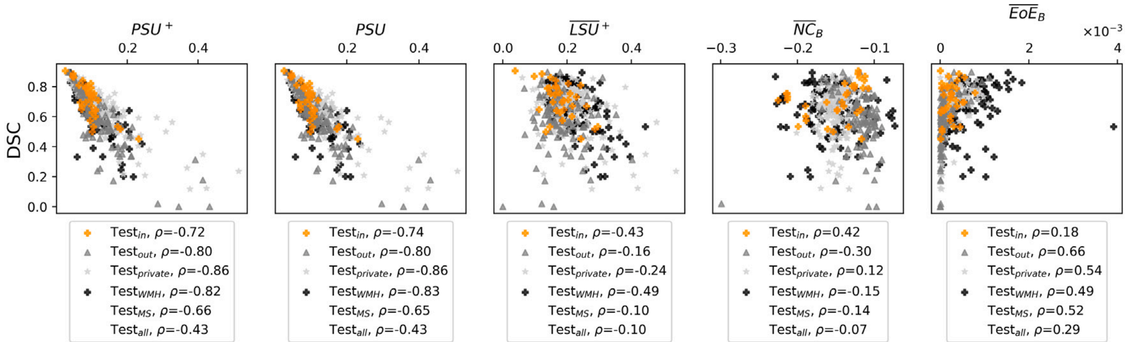


Fig. 4. The relationship between DSC and patient-scale uncertainty is assessed for Test$_{in}$ (orange), Test$_{out}$ (gray), Test$_{private}$ (light gray), and Test$_{WMH}$ (black) separately and jointly for the nnU-Net model. The presented uncertainty measures were chosen based on the results of the error RC analysis (Fig. 3 and Table 4) to illustrate the relationship between DSC and uncertainty brought by measures with the highest (proposed $PSU^{(+)}$), median (proposed $\overline{LSU^{(+)}}$), and worse-than-random ($\overline{NC}_B$ and $\overline{EoE}_B$) DSC-AUC values. Results for other measures and for the SB model can be found in Appendix B.2.

3.2.3. Generalizability of the analysis on white matter hyperintensity

Beyond MS patients, the multi-scale error retention curve and DSC-uncertainty correlation analyses were replicated on a large publicly available cohort of subjects with WMH (Test$_{WMH}$). The full analysis is available in Appendix B.3.

The observed performance of the proposed patient-scale measures discussed in the previous sections is replicated for this new task of WMH segmentation. The results in Fig. 4 and Table 5 confirm that the proposed measures $PSU^{(+)}$ have a stronger relationship segmentation quality compared to the aggregation measures.

3.3. Qualitative evaluation of the uncertainty maps

Our results show that uncertainty quantification mainly at the lesion and patient scales can well depict model error predictions, however, various anatomical scales provide information about different types of errors. In Fig. 5 the uncertainty maps and values are shown for four

Table 4
Mean average areas under error retention curves and 90% bootstrap confidence intervals for the assessment of the uncertainty measures at the voxel, lesion, and patient anatomical scales across the in-domain Test$_{in}$ (left column) and the out-of-domain Test$_{out}$ (center column) and Test$_{private}$ (right column) sets. Results are presented for the Shifts Challenge Baseline (SB) and nnU-Net models. Highest AUC values for each dataset, model, and anatomical scale are highlighted in **bold**, lowest - in *italic*; ideal and random values are in gray color and indicate the upper and lower bounds of performance, respectively.

Measure	Test$_{in}$		Test$_{out}$		Test$_{private}$	
	SB	nnU-Net	SB	nnU-Net	SB	nnU-Net
			Voxel-scale DSC-AUC (↑)			
Ideal	99.93 [99.91, 99.94]	99.94 [99.92, 99.95]	99.90 [99.88, 99.91]	99.93 [99.92, 99.92]	99.93 [99.91, 99.94]	99.93 [99.91, 99.94]
NC_i	**99.17** [98.99, 99.31]	99.17 [98.29, 99.49]	96.74 [96.23, 97.12]	97.59 [97.02, 0.9797]	98.56 [98.36, 98.70]	**99.02** [98.82, 99.16]
EoE_i	**99.16** [98.99, 99.31]	*99.11* [98.10, 99.46]	**97.02** [96.56, 97.37]	**97.72** [97.22, 98.05]	**98.65** [98.46, 98.79]	**99.02** [98.82, 99.17]
ExE_i	**99.16** [98.99, 99.31]	*99.11* [98.09, 99.46]	**97.02** [96.56, 97.38]	**99.71** [97.21, 98.05]	**98.65** [98.46, 98.80]	**99.02** [98.82, 99.16]
MI_i	*99.05* [98.85, 99.21]	**99.27** [98.74, 99.50]	*96.69* [96.19, 97.08]	*97.28* [96.70, 97.68]	*98.46* [98.25, 98.62]	*98.86* [98.63, 99.01]
Random	80.91 [76.77, 83.36]	84.87 [82.79, 86.69]	76.20 [74.88, 77.36]	80.00 [78.72, 81.21]	80.18 [78.99, 81.19]	82.79 [81.85, 83.62]
			Lesion-scale LPPV-AUC (↑)			
Ideal	87.88 [82.60, 90.91]	95.72 [93.89, 96.88]	87.07 [83.40, 89.46]	96.47 [93.13, 97.66]	86.41 [84.54, 87.93]	96.36 [95.51, 96.96]
LSU	83.54 [75.80, 87.04]	91.54 [89.57, 93.15]	83.28 [79.63, 85.91]	**94.06** [90.87, 95.41]	82.63 [80.74, 84.28]	**93.29** [92.15, 94.21]
LSU^+	**83.90** [78.83, 87.31]	91.51 [89.53, 93.12]	**83.89** [80.27, 86.45]	93.97 [90.80, 95.33]	**82.70** [80.83, 84.37]	**93.29** [92.15, 94.20]
$\overline{NC}_L$	83.33 [78.34, 86.77]	91.71 [89.46, 93.92]	83.24 [79.60, 85.86]	**94.06** [90.84, 95.39]	82.34 [80.38, 84.04]	93.14 [92.05, 94.05]
$\overline{EoE}_L$	83.38 [78.41, 86.83]	**91.81** [89.61, 93.93]	83.26 [79.63, 85.88]	**94.07** [90.86, 95.40]	82.28 [80.30, 83.99]	93.22 [92.11, 94.11]
$\overline{ExE}_L$	*81.73* [76.70, 85.24]	91.70 [89.50, 93.27]	*81.55* [77.88, 84.17]	*93.41* [90.32, 94.77]	***78.74*** **[76.80, 80.56]**	*91.99* [90.77, 93.00]
$\overline{MI}_L$	82.63 [77.70, 86.03]	*91.37* [89.22, 92.98]	82.31 [78.64, 85.00]	**94.06** [90.86, 95.40]	81.62 [79.69, 83.34]	93.05 [91.89, 93.96]
Random	76.69 [71.57, 80.48]	86.65 [83.96, 88.94]	76.35 [72.71, 79.19]	90.59 [87.65, 92.10]	73.97 [71.91, 75.81]	88.61 [87.18, 89.88]
			Patient-scale DSC-AUC (↑)			
Ideal	85.74 [84.16, 87.52]	88.72 [87.22, 90.36]	79.21 [77.96, 80.52]	83.55 [82.30, 84.95]	84.48 [83.72, 85.26]	86.23 [85.56, 86.91]
PSU	**84.99** [83.16, 86.81]	**87.90** [86.25, 89.73]	**78.40** [77.11, 79.73]	**82.68** [81.26, 84.18]	**83.63** [82.79, 84.47]	**85.73** [85.02, 86.46]
PSU^+	**84.82** [82.97, 86.68]	**87.84** [86.17, 89.70]	**78.39** [77.10, 79.70]	**82.70** [81.28, 84.20]	**83.60** [82.75, 84.44]	**85.75** [85.04, 86.47]
$\overline{LSU}$	83.77 [81.99, 85.42]	86.80 [84.69, 88.64]	75.48 [74.55, 77.26]	79.90 [77.66, 81.22]	79.91 [80.02, 82.28]	83.55 [82.54, 84.44]
$\overline{LSU^+}$	83.13 [81.04, 84.88]	86.87 [84.80, 88.97]	75.28 [74.36, 77.08]	79.76 [75.73, 81.16]	79.91 [80.02, 82.28]	83.52 [82.51, 84.42]
$\overline{NC}_B$	80.70 [78.27, 82.42]	*84.13* [82.40, 85.57]	74.82 [72.96, 76.57]	79.90 [77.85, 81.73]	79.79 [75.81, 78.79]	82.00 [80.84, 82.97]
$\overline{EoE}_B$	80.19 [76.20, 82.86]	84.71 [82.40, 86.80]	*71.60* [69.62, 73.32]	75.04 [72.69, 76.72]	77.43 [75.81, 78.79]	79.98 [78.49, 81.19]
$\overline{ExE}_B$	*80.19* [76.20, 82.87]	84.64 [82.34, 86.72]	*71.57* [69.53, 73.31]	*75.44* [72.69, 76.72]	*77.37* [75.74, 78.73]	*79.83* [78.33, 81.04]
$\overline{MI}_B$	80.28 [76.25, 83.02]	85.07 [82.69, 87.20]	71.70 [69.76, 73.39]	75.18 [72.93, 77.00]	77.60 [75.97, 78.97]	80.20 [78.72, 81.42]
Random	81.87 [79.09, 83.84]	85.73 [83.58, 87.62]	74.10 [72.53, 75.53]	78.03 [76.20, 79.60]	80.08 [78.85, 81.14]	82.26 [81.14, 83.23]

Table 5
Spearman's correlation coefficients quantifying the relationship between different patient-scale uncertainty values and segmentation quality measured by DSC for different test sets and their combinations. The highest negative correlation values are highlighted in **bold**.

Measures	Test$_{in}$		Test$_{out}$		Test$_{private}$		Test$_{WMH}$	
	SB	nnU-Net	SB	nnU-Net	SB	nnU-Net	SB	nnU-Net
PSU	**−0.81**	**−0.74**	**−0.81**	**−0.84**	**−0.86**	**−0.87**	**−0.83**	**−0.46**
PSU^+	−0.72	−0.72	−0.80	**−0.84**	−0.86	−0.86	**−0.83**	−0.43
$\overline{LSU}$	−0.41	−0.41	−0.22	−0.37	−0.25	−0.36	−0.49	−0.11
$\overline{LSU}^+$	−0.29	−0.43	−0.22	−0.34	−0.42	−0.49	−0.11	−0.11
$\overline{NC}_B$	0.36	0.42	0.11	0.30	0.30	0.14	−0.09	0.07
$\overline{EoE}_B$	0.23	0.18	0.55	0.56	0.54	0.54	0.53	0.31
$\overline{ExE}_B$	0.23	0.21	0.55	0.68	0.56	0.57	0.57	0.31
$\overline{MI}_B$	0.20	0.07	0.53	0.63	0.49	0.47	0.33	0.24

different subjects, corresponding to different scenarios with respect to the quality of lesion segmentation.

Voxel-scale maps provide refined information about the misclassifications in each voxel. Moreover, voxel-scale uncertainty is always high at the borders of lesions. Hypothetically, this is a reflection of the inter-rater variability or the noise in the ground truth, which are also known to be higher at the borders of lesions. The noise in the data-generation process increases the likelihood of mistakes at the borders of lesions. Nevertheless, the voxel-scale uncertainty can be high in the center of the lesion, signaling that the model is uncertain in the whole lesion region, not only at the borders. Sometimes high uncertainty regions can be related to the FNLs.

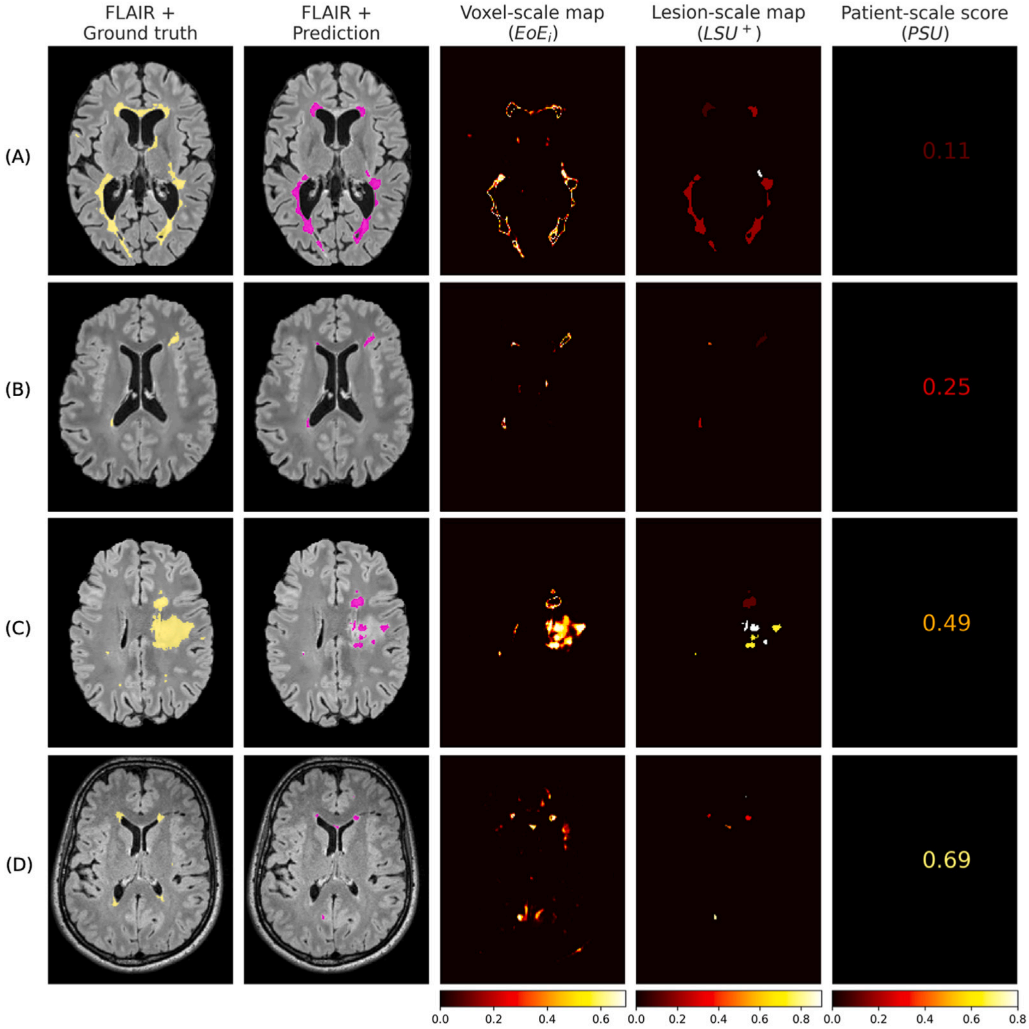


Fig. 5. Examples of uncertainty maps at the voxel and lesion scales and patient uncertainty values. The two left columns illustrate axial slices of a FLAIR scan with the ground truth (in yellow) and predicted (in pink) WML masks; the middle column - voxel-scale uncertainty maps computed with the EoE_i measure; the fourth column - lesion-scale uncertainty maps computed with the proposed LSU^+; the fifth column - the patient-scale uncertainty value computed with the proposed PSU^+. The choice of measures is based on the results of the error retention curves analysis. (A), (B), (C), and (D) represent different scenarios with gradually decreasing DSC. Cases (A) and (B) represent good and mediocre model performance, respectively. Patient (C) has an atypical large lesion, which the algorithm fails as expected. Patient (D) was not correctly preprocessed (the skull is not removed) which led to the algorithm's low performance and high patient uncertainty.

Lesion-scale maps provide a visually more intuitive way to assess the correctness of the predicted lesion regions compared to the voxel-scale maps. Particularly, lesion-scale maps can be used to highlight FPLs. Nonetheless, high lesion uncertainty may be an indicator of wrong delineation rather than detection. Let us note that, compared to the voxel-scale, the lesion-scale maps lose all the information about the FNLs.

Patient-scale values inform about the overall quality of the segmentation without indicating the particular reasons for the segmentation failure. As for the chosen examples (C) and (D), high patient uncertainty reveals the fact of the algorithm failure, however for (C) the problem is in the atypical large lesion and for (D) it is a wrong preprocessing, *i.e.* the absence of skull-stripping.

4. Discussion

Our research offers a detailed framework for the assessment of uncertainty quantification for a clinically relevant task of white matter lesion segmentation in multiple sclerosis. The specificity of the segmentation task allowed for the exploration of UQ at different anatomical scales: voxel, lesion, and patient. We introduced novel structure-based UQ measures at the lesion and patient scales. For each of these scales, we performed a comparative study between different uncertainty measures (among the state-of-the-art and the proposed) to determine the measures that can point to specific model errors: voxel misclassification, lesion false discovery, or overall low quality of segmentation. For this, we use the error retention curves analysis previously introduced for the pixel or voxel scales [26,34,35] and extended it to the structural scales in this and our previous work [30]. Our proposed uncertainty measures ($LSU^{(+)}$ on the lesion scale and $PSU^{(+)}$ on the patient scale from the Eqs. (1)–(4)) quantify the disagreement in the structural predictions between the ensemble model and its members, demonstrating enhanced error detection over state-of-the-art aggregation-based metrics on both in- and out-of-domain datasets. Furthermore, $PSU^{(+)}$ is shown to be a reliable indicator of overall segmentation quality both in- and out-of-domain.

This study compares a variety of voxel-scale measurements adopted from classification tasks, noting their similar capabilities in capturing voxel misclassification errors. A more pronounced difference between these measures is observed after aggregation at other anatomical scales.

Particularly, at the lesion scale, higher areas under the respective RCs are observed for the total uncertainty measures, compared to the measure of model uncertainty, and even more data uncertainty. However, voxel uncertainty aggregation at the patient scale yielded results akin to random uncertainty judging by the error RC analysis. Closer examination of the correlation between patient scale uncertainty measures and the DSC revealed a positive relationship, suggesting that a higher average voxel uncertainty correlates with improved DSC. A high positive correlation of the aggregation-based measures ($\overline{EoE}_B$, $\overline{ExE}_B$, and $\overline{MI}_B$) and the total lesion volume in a subject (see Appendix B.4) also goes against common knowledge about the bias in better segmenting subjects with higher lesion loads [49]. Similar behavior of the measures based on an aggregation of voxel uncertainties has been previously observed for the task of brain tumor segmentation [23], but not for the task of brain structures segmentation [18], where the segmented objects are the same and of similar sizes in each of the images. This supports our hypothesis that voxel-scale uncertainty aggregation is unsuitable for tasks affected by this bias. In such cases, structural disagreement metrics present a viable alternative to aggregation-based methods, showing a strong connection to different error types.

Limitations and future work. The fact that lesion and patient uncertainty measures depend on the choice of the threshold at the model's output, necessary for the instances or segmented region definition, remains a matter of ongoing debate. We proposed to address the issue by introducing two analogs of the same measure corresponding to different strategies of the threshold choice, *i.e.* LSU versus LSU^+ and PSU versus PSU^+. Nevertheless, a more detailed investigation of this aspect might be needed. For instance, exploring model calibration as a way to circumvent threshold tuning or investigating measures of uncertainty where this dependence is mitigated.

This paper is focused on the WML segmentation task. While this is a relevant task in clinical practice, there are several medical image segmentation tasks that could adopt the proposed multi-scale approach for UQ. This includes, for instance, nuclei segmentation on histopathology images [50], bone metastases segmentation on the full-body MRI or CT [51,52], vascularized lymph nodes on CT or MRI [53], or white matter lesions in MRI from non-MS patients [54]. However, finding the multi-center data and benchmarks needed for UQ methods validation under the domain shift in these new tasks remains challenging.

CRediT authorship contribution statement

Nataliia Molchanova: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Vatsal Raina:** Writing – review & editing, Writing – original draft, Validation, Software, Investigation, Data curation, Conceptualization. **Andrey Malinin:** Writing – review & editing, Writing – original draft, Supervision, Software, Data curation, Conceptualization. **Francesco La Rosa:** Writing – review & editing, Validation, Supervision, Software. **Adrien Depeursinge:** Writing – review & editing, Validation, Project administration, Funding acquisition. **Mark Gales:** Writing – review & editing, Supervision, Resources, Funding acquisition. **Cristina Granziera:** Writing – review & editing, Supervision, Resources, Funding acquisition, Data curation. **Henning Müller:** Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Funding acquisition. **Mara Graziani:** Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Investigation, Conceptualization. **Meritxell Bach Cuadra:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT-3.5/-4 and Grammarly for the detection of grammatic and stylistic errors in the manuscript. After using these tools, the authors reviewed and edited the content as needed and all take full responsibility for the content of the publication.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: AM: work done while at Yandex and Isomorphic Labs. HM: mandates with Roche. MG: work done while at IBM Research. CG: The University Hospital Basel (USB), as the employer of C.G., has received the following fees which were used exclusively for research support: (i) advisory boards and consultancy fees from Actelion, Novartis, Genzyme-Sanofi, GeNeuro, Hoffmann La Roche and Siemens; (ii) speaker fees from Biogen, Hoffmann La Roche, Teva, Novartis, Merck, Jannsen Pharmaceuticals and Genzyme-Sanofi; (iii) research grants: Biogen, Genzyme Sanofi, Hoffmann La Roche, GeNeuro. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. NM, VR, FLR, AD, MBC nothing to disclose.

Acknowledgments

This work was supported by the Hasler Foundation Responsible AI program (MSxplain), the EU Horizon 2020 project AI4Media (grant 951911), and the Swiss National Science Foundation (SNF) Postdoc Mobility Fellowship (P500PB_206833). 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). This research is partially funded by the EPSRC (The Engineering and Physical Sciences Research Council) Doctoral Training Partnership (DTP) Ph.D. studentship and supported by Cambridge University Press & Assessment (CUP&A), a department of The Chancellor, Masters, and Scholars of the University of Cambridge. This work was carried out in collaboration with The Observatoire Français de la Sclérose en Plaques (OFSEP), who is supported by a grant provided by the French State and handled by the "Agence Nationale de la Recherche," within the framework of the "Investments for the Future" program, under the reference ANR-10-COHO-002, by the Eugène Devic EDMUS Foundation against multiple sclerosis and by the ARSEP Foundation.

Appendix A. Definitions of quality metrics

Let $\#_{TP}, \#_{FP}, \#_{FN}$ be the number of true positive (TP), false positive (FP), and false negative (FN) voxels, respectively.

True positive rate:

$$TPR = \frac{\#_{TP}}{\#_{TP} + \#_{FN}}.$$

Positive predictive value:

$$PPV = \frac{\#_{TP}}{\#_{TP} + \#_{FP}}.$$

Dice similarity score or F$_1$-score:

$$DSC = F_1 = \frac{TPR \cdot PPV}{TPR + PPV} = \frac{2 \cdot \#_{TP}}{2 \cdot \#_{TP} + \#_{FP} + \#_{FN}}.$$

Normalized Dice similarity score [49]:

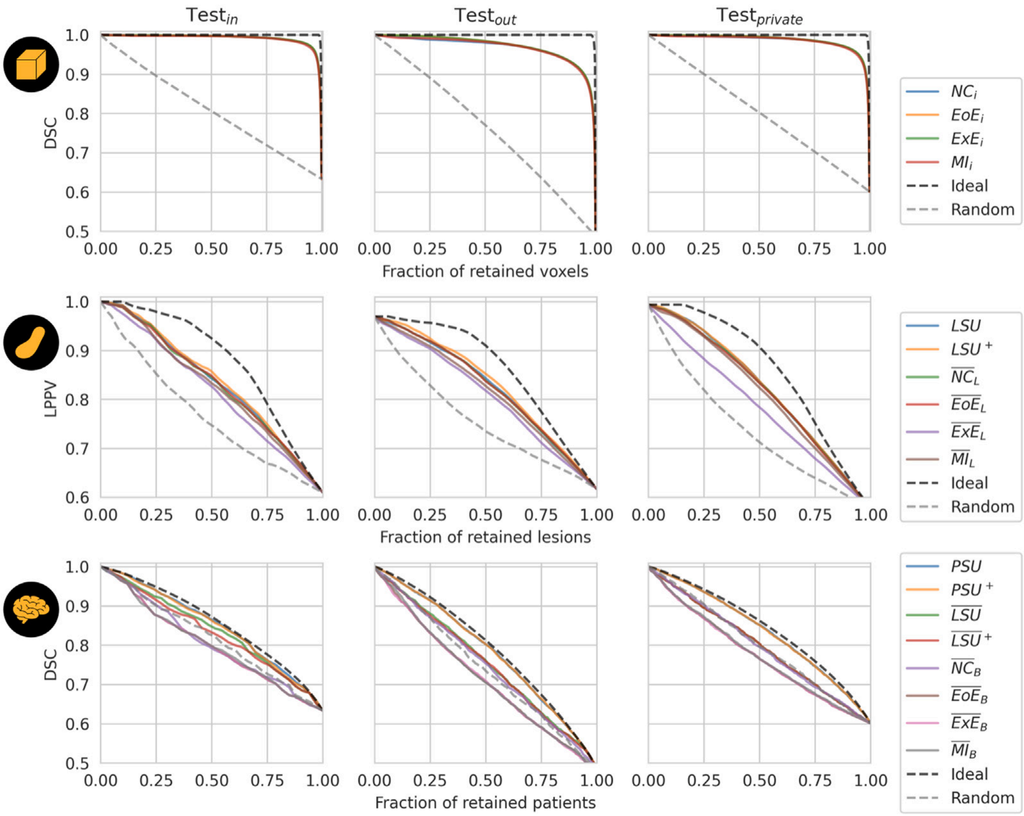


Fig. B.6. Error retention curves for the assessment of uncertainty measures at the voxel, lesion, and patient (rows one, two, and three, respectively) anatomical scales across the in-domain Test$_{in}$ (left column) and the out-of-domain Test$_{out}$ (center column) and Test$_{private}$ (left column) sets for the SB model. Different rows correspond to different anatomical scales indicated with icons on the left. The voxel-scale DSC-RCs and lesion-scale LPPV-RCs were obtained by averaging across the respective datasets. At each of the scales, the ideal (black dashed) line indicates the upper bound of an uncertainty measure performance in its ability to capture model errors; the random (gray dashed) indicates no relationship between an uncertainty measure and error; a worse-than-random performance indicates an inverse relationship.

$$nDSC = \frac{2 \cdot \#_{TP}}{2 \cdot \#_{TP} + \kappa \cdot \#_{FP} + \#_{FN}}, \kappa = h(r^{-1} - 1).$$

where h represents the ratio between the positive and the negative classes while $0 < r < 1$ denotes a *reference* value that is set to the mean fraction of the positive class, *i.e.* a lesion class in our case, across a large number of subjects.

Analogous, lesion-scale metrics can be defined by replacing $\#_{TP}$, $\#_{FP}$, $\#_{FN}$ with a number of TP, FP, and FN lesion (TPL, FPL, FNL). As mentioned before, the definition of lesion types can vary. This work uses 25% overlap to distinguish TPL and FPL among the predicted lesions. FNL is defined as the ground truth lesions that have no overlap with predictions.

Appendix B. Additional results

B.1. *Error retention curve analysis for the Shifts 2.0 baseline (SB) model*

Error retention curves for the SB model are shown in Fig. B.6.

B.2. *Patient-scale uncertainty as a proxy for segmentation quality*

Fig. B.7 extend the error retention curves analysis of the patient-scale uncertainty measures revealing more information about the relationship between the uncertainty measures and the segmentation quality measures by DSC.

B.3. *Generalizability analysis for white matter hyperintensity (WMH)*

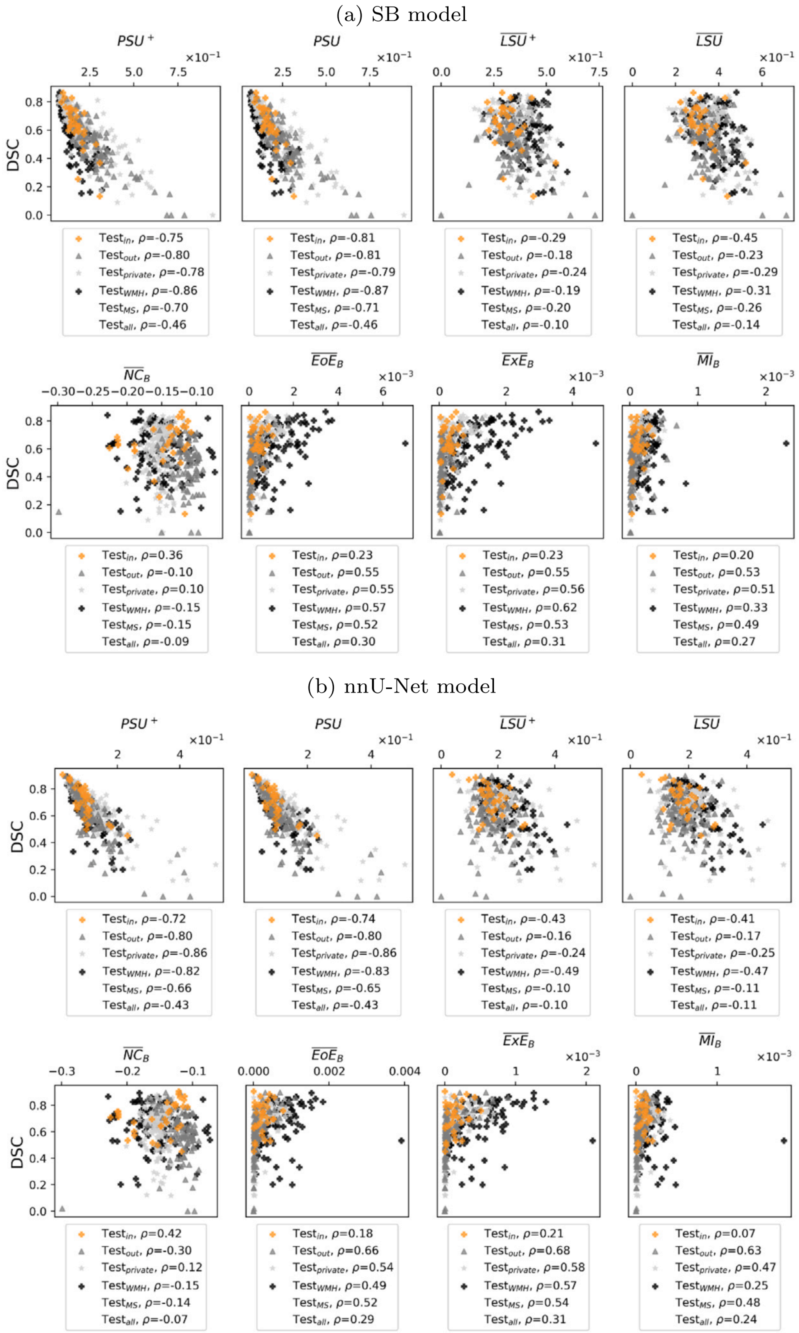
Areas under error retention curves for different anatomical scales are shown in Table B.6.

B.4. *Uncertainty relationship with lesion size and load*

Lesion-scale analysis of the relationship between the predicted lesion volumes and uncertainty are shown in violin plots in Fig. B.8(a) and (b) for SB and nnU-Net models, respectively. For all the lesion-scale uncertainty measures, lesions with smaller sizes tend to be more uncertain. For the nnU-Net model, the difference in medians of proposed $LSU^{(+)}$ uncertainty across different lesion volumes is less prominent compared to other measures.

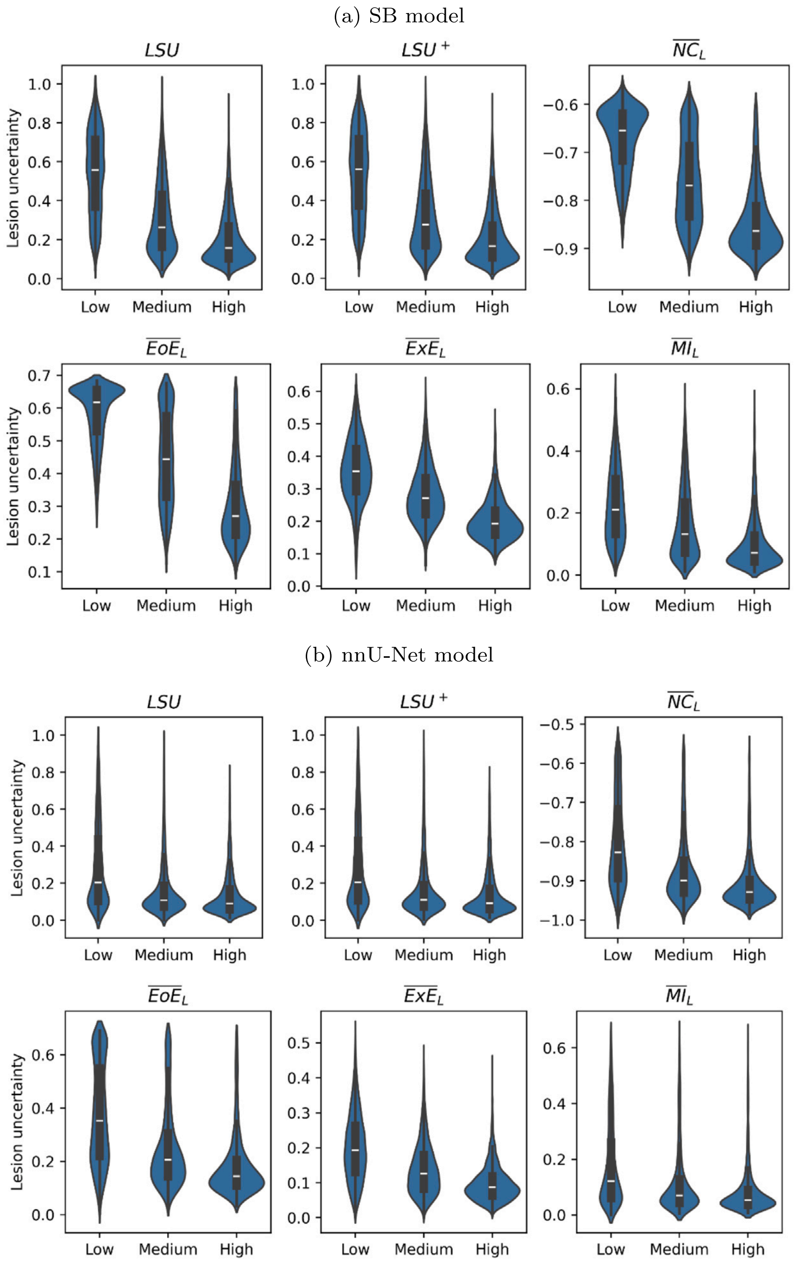
Patient-scale analysis of the relationship between the ground-truth total lesion volume and patient-scale uncertainty measure is given in Fig. B.9(a) and (b) for SB and nnU-Net models, respectively. Different measures have a different degree of associations with the ground-truth total lesion volume (TLV):

- $PSU^{(+)}$ values are negatively associated with the TLV: a patient with low uncertainty is more likely to have a high TLV;
- $\overline{LSU}^{(+)}$ and $\overline{NC}_B$ show a mild negative association with TLV;
- The rest of the aggregated voxel-scale measures ($\overline{EoE}_B$, $\overline{ExE}_B$ and $\overline{MI}_B$) have a strong positive association with the TLV: higher uncertainty for subjects with the higher TLV. This should explain a poor relationship with the overall segmentation quality.


(a) SB model

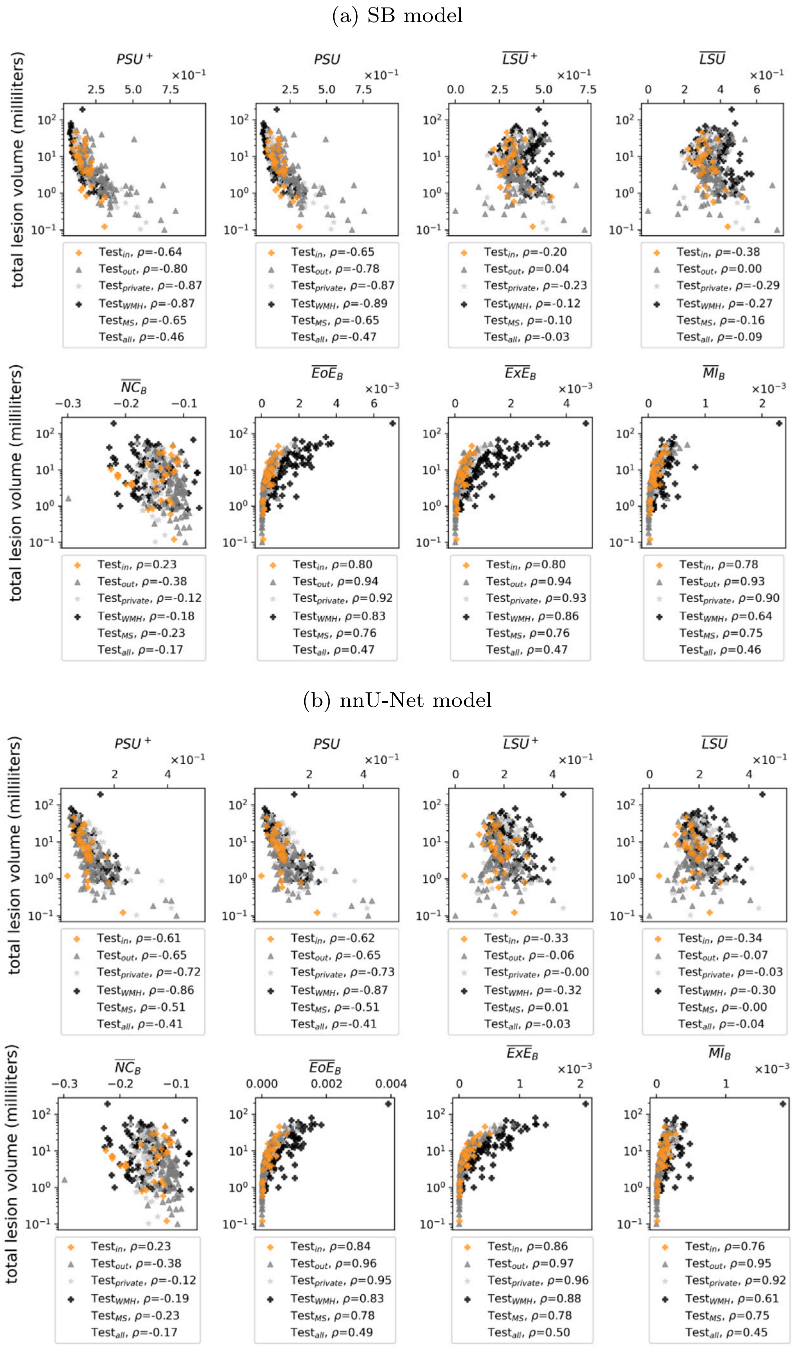
(b) nnU-Net model

Fig. B.7. The relationship between the total ground truth lesion volume in milliliters (logarithmic y-axis) and various patient uncertainty measures (x-axis). ρ (in the legend) is a Spearman's correlation coefficient.


(a) SB model

(b) nnU-Net model

Fig. B.8. The distributions of lesion uncertainty across 3 groups of predicted lesions in all the test sets jointly (Test$_{in}$, Test$_{out}$, Test$_{private}$, Test$_{WMH}$) defined through their volume percentiles: Low (0%–33%), Medium (33%–67%), High (67%–100%).


(a) SB model

(b) nnU-Net model

Fig. B.9. The relationship between the total ground truth lesion volume in milliliters (logarithmic y-axis) and various patient uncertainty measures (x-axis). ρ (in the legend) is a Spearman's correlation coefficient.

Table B.6
Mean average areas under error retention curves and 90% bootstrap confidence intervals for the assessment of the uncertainty measures at the voxel, lesion, and patient anatomical scales across the WMH Challenge dataset (Test$_{WMH}$). Results are presented for the Shifts Challenge Baseline (SB) and nnU-Net models. Highest AUC values for each dataset, model, and anatomical scale are highlighted in bold, lowest - in italic; ideal and random values are in gray color and indicate the upper and lower bounds of performance, respectively.

Measure	SB	nnU-Net
Voxel-scale DSC-AUC (↑)		
Ideal	99.81 [99.76, 99.83]	99.85 [99.80, 99.88]
NC_i	98.41 [98.11, 98.62]	99.40 [99.23, 99.50]
EoE_i	98.35 [98.05, 98.59]	99.40 [99.23, 99.51]
ExE_i	98.35 [98.05, 98.59]	99.39 [99.22, 99.50]
MI_i	98.16 [97.84, 98.41]	99.10 [98.97, 99.46]
Random	76.11 [74.06, 77.85]	80.15 [78.28, 81.70]
Lesion-scale LPPV-AUC (↑)		
Ideal	79.17 [75.94, 81.82]	91.98 [89.95, 93.48]
LSU	73.32 [70.13, 76.12]	86.64 [84.37, 88.43]
LSU^+	73.03 [69.87, 75.83]	86.64 [84.35, 88.44]
$\overline{NC}_L$	72.90 [69.76, 75.67]	86.81 [84.51, 88.61]
$\overline{EoE}_L$	72.94 [69.80, 75.71]	86.80 [84.52, 88.59]
$\overline{ExE}_L$	68.77 [65.72, 71.59]	85.27 [82.89, 87.17]
$\overline{MI}_L$	72.98 [69.79, 75.70]	86.70 [84.44, 88.50]
Random	66.38 [63.38, 69.08]	82.00 [79.72, 83.91]
Patient-scale DSC-AUC (↑)		
Ideal	84.17 [82.99, 85.32]	86.69 [85.62, 87.69]
PSU	83.51 [82.18, 84.75]	85.92 [84.62, 87.05]
PSU^+	83.47 [82.14, 84.72]	85.86 [84.56, 86.98]
$\overline{LSU}$	81.09 [79.82, 82.29]	84.60 [83.49, 85.70]
$\overline{LSU}^+$	80.59 [79.29, 81.80]	84.69 [83.59, 85.77]
$\overline{NC}_B$	80.06 [78.34, 81.53]	82.88 [81.19, 84.25]
$\overline{EoE}_B$	77.16 [75.83, 78.44]	80.66 [79.41, 81.84]
$\overline{ExE}_B$	76.93 [75.59, 78.22]	80.28 [79.00, 81.47]
$\overline{MI}_B$	78.27 [76.90, 79.56]	81.76 [80.49, 82.93]
Random	78.82 [77.10, 80.39]	81.76 [80.06, 83.20]

References

[1] C. Walton, R. King, L. Rechtman, W. Kaye, E. Leray, R.A. Marrie, N. Robertson, N.L. Rocca, B. Uitdehaag, I. van der Mei, M. Wallin, A. Helme, C.A. Napier, N. Rijke, P. Baneke, Rising prevalence of multiple sclerosis worldwide: Insights from the Atlas of MS, third edition, Multiple Scler. J. 26 (14) (2020) 1816–1821, http://dx.doi.org/10.1177/1352458520970841, arXiv:https://doi.org/10.1177/1352458520970841, PMID: 33174475.

[2] D. Reich, C. Lucchinetti, P. Calabresi, Multiple Sclerosis, in: D.L. Longo (Ed.), New Engl. J. Med. 378 (2) (2018) 169–180, http://dx.doi.org/10.1056/NEJMra1401483, URL http://www.nejm.org/doi/10.1056/NEJMra1401483.

[3] A.J. Thompson, B. Banwell, F. Barkhof, W.M. Carroll, T. Coetzee, G. Comi, J. Correale, F. Fazekas, M. Filippi, M.S. Freedman, K. Fujihara, S. Galetta, H.P. Hartung, L. Kappos, F. Lublin, R.A. Marrie, A. Miller, D.H. Miller, X. Montalbán, E.M. Mowry, P.S. Sørensen, M. Tintoré, A. Traboulsee, M. Trojano, B.M.J. Uitdehaag, S. Vukusic, E. Waubant, B.G. Weinshenker, S.C. Reingold, J.A. Cohen, Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria, Lancet Neurol. 17 (2) (2018) 162–173, http://dx.doi.org/10.1016/s1474-4422(17)30470-2.

[4] C. Hemond, R. Bakshi, Magnetic resonance imaging in multiple sclerosis, Cold Spring Harbor Perspect. Med. 8 (5) (2018) http://dx.doi.org/10.1101/cshperspect.a028969.

[5] M.P. Wattjes, O. Ciccarelli, D.S. Reich, B. Banwell, N. de Stefano, C. Enzinger, F. Fazekas, M. Filippi, J. Frederiksen, C. Gasperini, Y. Hacohen, L. Kappos, D.K.B. Li, K. Mankad, X. Montalban, S.D. Newsome, J. Oh, J. Palace, M.A. Rocca, J. Sastre-Garriga, M. Tintoré, A. Traboulsee, H. Vrenken, T. Yousry, F. Barkhof, À. Rovira, M.P. Wattjes, O. Ciccarelli, N. de Stefano, C. Enzinger, F. Fazekas, M. Filippi, J. Frederiksen, C. Gasperini, Y. Hacohen, L. Kappos, K. Mankad, X. Montalban, J. Palace, M.A. Rocca, J. Sastre-Garriga, M. Tintore, H. Vrenken, T. Yousry, F. Barkhof, A. Rovira, D.K.B. Li, A. Traboulsee, S.D. Newsome, B. Banwell, J. Oh, D.S. Reich, D.S. Reich, J. Oh, 2021 MAGNIMS–CMSC–NAIMS consensus recommendations on the use of MRI in patients with multiple sclerosis, Lancet Neurol. 20 (8) (2021) 653–670, http://dx.doi.org/10.1016/S1474-4422(21)00095-8, Publisher: Elsevier.

[6] C. Gramsch, F. Nensa, O. Kastrup, S. Maderwald, C. Deuschl, A. Ringelstein, J. Schelhorn, M. Forsting, M. Schlamann, Diagnostic value of 3D fluid attenuated inversion recovery sequence in multiple sclerosis, Acta Radiol. 56 (5) (2015) 622–627, http://dx.doi.org/10.1177/0284185114534413.

[7] A. Kaur, L. Kaur, A. Singh, State-of-the-art segmentation techniques and future directions for multiple sclerosis brain lesions, Arch. Comput. Methods Eng. 28 (2020) 1–27, http://dx.doi.org/10.1007/s11831-020-09403-7.

[8] X. Lladó, A. Oliver, M. Cabezas, J. Freixenet, J.C. Vilanova, A. Quiles, L. Valls, L. Ramió-Torrentà, À. Rovira, Segmentation of multiple sclerosis lesions in brain MRI: A review of automated approaches, Inform. Sci. 186 (1) (2012) 164–185, http://dx.doi.org/10.1016/j.ins.2011.10.011, URL https://www.sciencedirect.com/science/article/pii/S0020025511005548.

[9] C. Zeng, L. Gu, Z. Liu, S. Zhao, Review of deep learning approaches for the segmentation of multiple sclerosis lesions on brain MRI, Front. Neuroinform. 14 (2020) http://dx.doi.org/10.3389/fninf.2020.610967, URL https://www.frontiersin.org/articles/10.3389/fninf.2020.610967.

[10] F. Spagnolo, A. Depeursinge, S. Schädelin, A. Akbulut, H. Müller, M. Barakovic, L. Melie-Garcia, M. Bach Cuadra, C. Granziera, How far MS lesion detection and segmentation are integrated into the clinical workflow? A systematic review, NeuroImage: Clin. 39 (2023) 103491, http://dx.doi.org/10.1016/j.nicl.2023.103491, URL https://www.sciencedirect.com/science/article/pii/S2213158223001821.

[11] M. Reyes, R. Meier, S. Pereira, C.A. Silva, F.M. Dahlweid, H.v. Tengg-Kobligk, R.M. Summers, R. Wiest, On the interpretability of artificial intelligence in radiology: Challenges and opportunities, Radiology: Artif. Intell. 2 (3) (2020) e190043, http://dx.doi.org/10.1148/ryai.2020190043, arXiv:https://doi.org/10.1148/ryai.2020190043, PMID: 32510054.

[12] E. Begoli, T. Bhattacharya, D.F. Kusnezov, The need for uncertainty quantification in machine-assisted medical decision making, Nat. Mach. Intell. 1 (1) (2019) http://dx.doi.org/10.1038/s42256-018-0004-1, URL https://www.osti.gov/biblio/1561669.

[13] J. Gawlikowski, C. Tassi, M. Ali, J. Lee, M. Humt, J. Feng, A. Kruspe, R. Triebel, P. Jung, R. Roscher, M. Shahzad, W. Yang, R. Bamler, X. Zhu, A survey of uncertainty in deep neural networks, Artif. Intell. Rev. (2023) 1–77, http://dx.doi.org/10.1007/s10462-023-10562-9.

[14] M. Graziani, L. Dutkiewicz, D. Calvaresi, J.P. Amorim, K. Yordanova, M. Vered, R. Nair, P.H. Abreu, T. Blanke, V. Pulignano, J.O. Prior, L. Lauwaert, W. Reijers, A. Depeursinge, V. Andrearczyk, H. Müller, A global taxonomy of interpretable AI: unifying the terminology for the technical and social sciences, Artif. Intell. Rev. 56 (4) (2022) 3473–3504, http://dx.doi.org/10.1007/s10462-022-10256-8.

[15] S. Faghani, M. Moassefi, P. Rouzrokh, B. Khosravi, F.I. Baffour, M.D. Ringler, B.J. Erickson, Quantifying uncertainty in deep learning of radiologic images, Radiology 308 (2) (2023) e222217, http://dx.doi.org/10.1148/radiol.222217, arXiv:https://doi.org/10.1148/radiol.222217, PMID: 37526541.

[16] K. Zou, Z. Chen, X. Yuan, X. Shen, M. Wang, H. Fu, A review of uncertainty estimation and its application in medical imaging, Meta-Radiol. 1 (1) (2023) 100003, http://dx.doi.org/10.1016/j.metrad.2023.100003, URL https://www.sciencedirect.com/science/article/pii/S2950162823000036.

[17] B. Lambert, F. Forbes, A. Tucholka, S. Doyle, H. Dehaene, M. Dojat, Trustworthy clinical AI solutions: a unified review of uncertainty quantification in deep learning models for medical image analysis, 2022, http://dx.doi.org/10.48550/arXiv.2210.03736, Preprint.

[18] A.G. Roy, S. Conjeti, N. Navab, C. Wachinger, Bayesian QuickNAT: Model uncertainty in deep whole-brain segmentation for structure-wise quality control, NeuroImage 195 (2019) 11–22, http://dx.doi.org/10.1016/j.neuroimage.2019.03.042, URL https://www.sciencedirect.com/science/article/pii/S1053811919302319.

[19] G. Wang, W. Li, S. Ourselin, T. Vercauteren, Automatic brain tumor segmentation based on cascaded convolutional neural networks with uncertainty estimation, Front. Comput. Neurosci. 13 (2019) http://dx.doi.org/10.3389/fncom.2019.00056, URL https://www.frontiersin.org/articles/10.3389/fncom.2019.00056.

[20] M. Rottmann, P. Colling, T.P. Hack, R. Chan, F. Hüger, P. Schlicht, H. Gottschalk, Prediction error meta classification in semantic segmentation: Detection via aggregated dispersion measures of softmax probabilities, 2019, arXiv:1811.00648.

[21] T. Nair, D. Precup, D. Arnold, T. Arbel, Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation, Med. Image Anal. 59 (2020) 101557, http://dx.doi.org/10.1016/j.media.2019.101557, URL https://www.sciencedirect.com/science/article/pii/S1361841519300994.

[22] B. Lambert, F. Forbes, S. Doyle, A. Tucholka, M. Dojat, Fast Uncertainty Quantification for Deep Learning-based MR Brain Segmentation, in: EGC 2022 - Conference francophone pour l'Extraction et la Gestion des Connaissances, Blois, France, 2022, pp. 1–12, URL https://hal.archives-ouvertes.fr/hal-03498120.

[23] A. Jungo, F. Balsiger, M. Reyes, Analyzing the quality and challenges of uncertainty estimations for brain tumor segmentation, Front. Neurosci. 14 (2020) http://dx.doi.org/10.3389/fnins.2020.00282, URL https://www.frontiersin.org/articles/10.3389/fnins.2020.00282.

[24] L. Whitbread, M. Jenkinson, Uncertainty categories in medical image segmentation: A study of source-related diversity, in: Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, Springer Nature, Switzerland, 2022, pp. 26–35, http://dx.doi.org/10.1007/978-3-031-16749-2_3.

[25] R. McKinley, R. Wepfer, L. Grunder, F. Aschwanden, T. Fischer, C. Friedli, R. Muri, C. Rummel, R. Verma, C. Weisstanner, B. Wiestler, C. Berger, P. Eichinger, M. Mühlau, M. Reyes, A. Salmen, A.T. Chan, R. Wiest, F. Wagner, Automatic detection of lesion load change in Multiple Sclerosis using convolutional neural networks with segmentation confidence, NeuroImage: Clin. 25 (2020) 102104, http://dx.doi.org/10.1016/j.nicl.2019.102104.

[26] A. Malinin, A. Athanasopoulos, M. Barakovic, M. Bach Cuadra, M.J.F. Gales, C. Granziera, M. Graziani, N. Kartashev, K. Kyriakopoulos, P.J. Lu, N. Molchanova, A. Nikitakis, V. Raina, F.L. Rosa, E. Sivena, V. Tsarsitalidis, E. Tsompopoulou, E. Volf, [Dataset] Shifts 2.0: Extending the dataset of real distributional shifts, 2022, arXiv:2206.15407.

[27] B. Lakshminarayanan, A. Pritzel, C. Blundell, Simple and scalable predictive uncertainty estimation using deep ensembles, in: I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), in: Advances in Neural Information Processing Systems, vol. 30, Curran Associates, Inc., 2017, pp. 1–12, URL https://proceedings.neurips.cc/paper_files/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38-Paper.pdf.

[28] Y. Gal, Z. Ghahramani, Dropout as a Bayesian approximation: Representing model uncertainty in deep learning, in: Proceedings of the 33rd International Conference on Machine Learning, 2015.

[29] B. Lambert, F. Forbes, S. Doyle, A. Tucholka, M. Dojat, Beyond voxel prediction uncertainty: Identifying brain lesions you can trust, in: International Workshop on Interpretability of Machine Intelligence in Medical Image Computing, Springer, 2022, pp. 61–70.

[30] N. Molchanova, V. Raina, A. Malinin, F. La Rosa, H. Müller, M. Gales, C. Granziera, M. Graziani, M. Bach Cuadra, Novel structural-scale uncertainty measures and error retention curves: Application to multiple sclerosis, in: 2023 IEEE 20th International Symposium on Biomedical Imaging, ISBI, 2023, pp. 1–5, http://dx.doi.org/10.1109/ISBI53787.2023.10230563.

[31] M.J. Fartaria, A. Todea, T. Kober, K. O'brien, G. Krueger, R. Meuli, C. Granziera, A. Roche, M. Bach Cuadra, Partial volume-aware assessment of multiple sclerosis lesions, NeuroImage: Clinical 18 (2018) 245–253, http://dx.doi.org/10.1016/j.nicl.2018.01.011, URL https://www.sciencedirect.com/science/article/pii/S2213158218300111.

[32] M.J. Fartaria, T. Kober, C. Granziera, M. Bach Cuadra, Longitudinal analysis of white matter and cortical lesions in multiple sclerosis, NeuroImage: Clinical 23 (2019) 101938, http://dx.doi.org/10.1016/j.nicl.2019.101938, URL https://www.sciencedirect.com/science/article/pii/S2213158219302888.

[33] A. Malinin, M. Gales, Uncertainty estimation in autoregressive structured prediction, in: International Conference on Learning Representations, 2021, URL https://openreview.net/forum?id=jN5y-zb5Q7m.

[34] A. Malinin, Uncertainty Estimation in Deep Learning with Application to Spoken Language Assessment (Ph.D. thesis), University of Cambridge, United Kingdom, 2019.

[35] R. Mehta, A. Filos, U. Baid, C. Sako, R. McKinley, M. Rebsamen, K. Dätwyler, R. Meier, P. Radojewski, G.K. Murugesan, S.S. Nalawade, C. Ganesh, B. Wagner, F. Yu, B. Fei, A.J. Madhuranthakam, J.A. Maldjian, L. Daza, C. Gómez, P. Arbeláez, C. Dai, S. Wang, H. Reynaud, Y. Mo, E.D. Angelini, Y. Guo, W. Bai, S. Banerjee, L. Pei, M. Ak, S. Rosas-González, I. Zemmoura, C. Tauber, M.H. Vu, T. Nyholm, T. Löfstedt, L.M. Ballestar, V. Vilaplana, H. McHugh, G.D.M. Talou, A. Wang, J. Patel, K. Chang, K. Hoebel, M. Gidwani, N. Arun, S. Gupta, M. Aggarwal, P. Singh, E.R. Gerstner, J. Kalpathy-Cramer, N. Boutry, A. Huard, L. Vidyaratne, M.M. Rahman, K.M. Iftekharuddin, J. Chazalon, É. Puybareau, G. Tochon, J. Ma, M. Cabezas, X. Lladó, A. Oliver, L. Valencia, S. Valverde, M. Amian, M. Soltaninejad, A. Myronenko, A. Hatamizadeh, X. Feng, D. Quan, N.J. Tustison, C.H. Meyer, N. Shah, S.N. Talbar, M.A. Weber, A. Mahajan, A. Jakab, R. Wiest, H.M. Fathallah-Shaykh, A. Nazeri, M. Milchenko, D.S. Marcus, A. Kotrotsou, R.R. Colen, J. Freymann, J. Kirby, C. Davatzikos, B.H. Menze, S. Bakas, Y. Gal, T. Arbel, QU-BRATS: MICCAI BRATS 2020 Challenge on Quantifying Uncertainty in Brain Tumor segmentation – Analysis of ranking scores and benchmarking results, J. Mach. Learn. Biomed. Imaging 1 (August 2022) (2022) 1–54, http://dx.doi.org/10.59275/j.melba.2022-354b.

[36] O. Ronneberger, P. Fischer, T. Brox, U-NET: Convolutional Networks for Biomedical Image Segmentation, in: Lecture Notes in Computer Science, 2015, pp. 234–241, http://dx.doi.org/10.1007/978-3-319-24574-4_28.

[37] O. Çiçek, A. Abdulkadir, S.S. Lienkamp, T. Brox, O. Ronneberger, 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation, Lecture Notes in Computer Science, Springer, 2016, pp. 424–432, http://dx.doi.org/10.1007/978-3-319-46723-8_49.

[38] O. Commowick, A. Istace, M. Kain, B. Laurent, F. Leray, M. Simon, S. Camarasu-Pop, P. Girard, R. Améli, J.c. Ferré, A. Kerbrat, T. Tourdias, F. Cervenansky, T. Glatard, J. Beaumont, S. Doyle, F. Forbes, J. Knight, A. Khademi, A. Mahbod, C. Wang, R. McKinley, F. Wagner, J. Muschelli, E. Sweeney, E. Roura, X. Lladó, M.M. Santos, W.P.D. Santos, A.G. Silva-Filho, X. Tomas-Fernandez, H. Urien, I. Bloch, S. Valverde, M. Cabezas, F.J. Vera-Olmos, N. Malpica, C.R. Guttmann, S. Vukusic, G. Edan, M. Dojat, M. Styner, S.K. Warfield, F. Cotton, C. Barillot, Objective Evaluation of Multiple Sclerosis Lesion Segmentation using a Data Management and Processing Infrastructure, Sci. Rep. 8 (1) 2018. http://dx.doi.org/10.1038/s41598-018-31911-7.

[39] F. La Rosa, A. Abdulkadir, M.J. Fartaria, R. Rahmanzadeh, P.J. Lu, R. Galbusera, M. Baraković, J.P. Thiran, C. Granziera, M.B. Cuadra, Multiple sclerosis cortical and WM lesion segmentation at 3T MRI: a deep learning method based on FLAIR and MP2RAGE, NeuroImage: Clin. 27 (2020) 102335, http://dx.doi.org/10.1016/j.nicl.2020.102335.

[40] C. Granziera, Imaging the interplay between axonal damage and repair in multiple sclerosis (INsIDER), 2018. https://classic.clinicaltrials.gov/ct2/show/NCT05177523.

[41] H. Kuijf, J.M. Biesbroek, J. de Bresser, R. Heinen, C. Chen, M.W. van der Flier, F. Barkhof, A.M. Viergever, G.J. Biessels, Data of the white matter hyperintensity (WMH) segmentation challenge, DataverseNL, 2022. http://dx.doi.org/10.34894/AECRSD.

[42] D. Erten-Lyons, R. Woltjer, J. Kaye, N. Mattek, H.H. Dodge, S. Green, H. Tran, D.B. Howieson, K. Wild, L.C. Silbert, Neuropathologic basis of white matter hyperintensity accumulation with advanced age, Neurology 81 (11) (2013) 977–983, http://dx.doi.org/10.1212/wnl.0b013e3182a43e45.

[43] F. Isensee, M. Schell, I. Pflueger, G. Brugnara, D. Bonekamp, U. Neuberger, A. Wick, H.P. Schlemmer, S. Heiland, W. Wick, M. Bendszus, K. Maier-Hein, P. Kickingereder, Automated brain extraction of multisequence MRI using artificial neural networks, Hum. Brain Mapp. 40 (2019) http://dx.doi.org/10.1002/hbm.24750.

[44] N.J. Tustison, B.B. Avants, P.A. Cook, Y. Zheng, A. Egan, P.A. Yushkevich, J.C. Gee, N4ITK: Improved N3 bias correction, IEEE Trans. Med. Imaging 29 (6) (2010) 1310–1320, http://dx.doi.org/10.1109/tmi.2010.2046908.

[45] A. Carass, S. Roy, A. Jog, J.L. Cuzzocreo, E. Magrath, A. Gherman, J. Button, J. Nguyen, F. Prados, C.H. Sudre, M. Jorge Cardoso, N. Cawley, O. Ciccarelli, C.A. Wheeler-Kingshott, S. Ourselin, L. Catanese, H. Deshpande, P. Maurel, O. Commowick, C. Barillot, X. Tomas-Fernandez, S.K. Warfield, S. Vaidya, A. Chunduru, R. Muthuganapathy, G. Krishnamurthi, A. Jesson, T. Arbel, O. Maier, H. Handels, L.O. Iheme, D. Unay, S. Jain, D.M. Sima, D. Smeets, M. Ghafoorian, B. Platel, A. Birenbaum, H. Greenspan, P.L. Bazin, P.A. Calabresi, C.M. Crainiceanu, L.M. Ellingsen, D.S. Reich, J.L. Prince, D.L. Pham, Longitudinal multiple sclerosis lesion segmentation: Resource and challenge, NeuroImage 148 2017. 77–102, https://www.sciencedirect.com/science/article/pii/S1053811916307819,http://dx.doi.org/10.1016/j.neuroimage.2016.12.064.

[46] O. Commowick, A. Istace, M. Kain, B. Laurent, F. Leray, M. Simon, S. Pop, P. Girard, R. Ameli, J.C. Ferré, A. Kerbrat, T. Tourdias, F. Cervenansky, T. Glatard, J. Beaumont, S. Doyle, F. Forbes, J. Knight, A. Khademi, C. Barillot, Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure, Sci. Rep. 8 2018. 13650–13666, http://dx.doi.org/10.1038/s41598-018-31911-7.

[47] Z. Lesjak, A. Galimzianova, A. Koren, M. Lukin, F. Pernus, B. Likar, Žiga piclin, A novel public MR image dataset of multiple sclerosis patients with lesion segmentations based on multi-rater consensus, Neuroinformatics 16 (2017) 51–63.

[48] G. Bonnier, A. Roche, D. Romascano, S. Simioni, D. Meskaldji, D. Rotzinger, Y.C. Lin, G. Menegaz, M. Schluep, R. Du Pasquier, T.J. Sumpf, J. Frahm, J.P. Thiran, G. Krueger, C. Granziera, Advanced MRI unravels the nature of tissue alterations

in early multiple sclerosis, Ann. Clin. Transl. Neurol. 1 (6) 2014. 423–432, https://onlinelibrary.wiley.com/doi/abs/10.1002/acn3.68, http://dx.doi.org/10.1002/acn3.68, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/acn3.68.

[49] V. Raina, N. Molchanova, M. Graziani, A. Malinin, H. Muller, M. Bach Cuadra, M. Gales, Novel structural-scale uncertainty measures and error retention curves: Application to multiple sclerosis, in: 2023 IEEE 20th International Symposium on Biomedical Imaging, ISBI, 2023, pp. 1–5, http://dx.doi.org/10.1109/ISBI53787.2023.10230755.

[50] N. Kumar, R. Verma, D. Anand, Y. Zhou, O.F. Onder, E. Tsougenis, H. Chen, P.A. Heng, J. Li, Z. Hu, Y. Wang, N.A. Koohbanani, M. Jahanifar, N.Z. Tajeddin, A. Gooya, N. Rajpoot, X. Ren, S. Zhou, Q. Wang, D. Shen, C.K. Yang, C.H. Weng, W.H. Yu, C.Y. Yeh, S. Yang, S. Xu, P.H. Yeung, P. Sun, A. Mahbod, G. Schaefer, I. Ellinger, R. Ecker, O. Smedby, C. Wang, B. Chidester, T.V. Ton, M.T. Tran, J. Ma, M. N, S. Graham, Q.D. Vu, J.T. Kwak, A. Gunda, R. Chunduri, C. Hu, X. Zhou, D. Lotfi, R. Safdari, A. Kascenas, A. O'Neil, D. Eschweiler, J. Stegmaier, Y. Cui, B. Yin, K. Chen, X. Tian, P. Gruening, E. Barth, E. Arbel, I. Remer, A. Ben-Dor, E. Sirazitdinova, M. Kohl, S. Braunewell, Y. Li, X. Xie, L. Shen, J. Ma, K.D. Baksi, M.A. Khan, J. Choo, A. Colomer, V. Naranjo, L. Pei, K.M. Iftekharuddin, K. Roy, D. Bhattacharjee, A. Pedraza, M.G. Bueno, S. Devanathan, S. Radhakrishnan, P. Koduganty, Z. Wu, G. Cai, X. Liu, Y. Wang, A. Sethi, A Multi-Organ nucleus segmentation challenge, IEEE Trans. Med. Imaging 39 (5) (2019) 1380–1391, http://dx.doi.org/10.1109/tmi.2019.2947628.

[51] A. Colombo, G. Saia, A.A. Azzena, A. Rossi, F. Zugni, P. Pricolo, P.E. Summers, G. Marvaso, R. Grimm, M. Bellomi, B.A. Jereczek-Fossa, A.R. Padhani, G. Petralia, Semi-Automated Segmentation of Bone Metastases from Whole-Body MRI: Reproducibility of Apparent Diffusion Coefficient Measurements, Diagnostics 11 (3) (2021) 499, http://dx.doi.org/10.3390/diagnostics11030499.

[52] M. Afnouch, O. Gaddour, Y. Hentati, F. Bougourzi, M. Abid, I. Alouani, A.T. Ahmed, BM-Seg: A new bone metastases segmentation dataset and ensemble of CNN-based segmentation approach, Expert Syst. Appl. 228 (2023) 120376, http://dx.doi.org/10.1016/j.eswa.2023.120376.

[53] C. Hassani, K. Tran, S.L. Palmer, K.M. Patel, Vascularized lymph node transfer: A primer for the radiologist, Radiographics 40 (4) (2020) 1073–1089, http://dx.doi.org/10.1148/rg.2020190118.

[54] M. Malova, E. Morelli, V. Cardiello, D. Tortora, M. Severino, M.G. Calevo, A. Parodi, L.C. De Angelis, D. Minghetti, A. Rossi, L.A. Ramenghi, Nosological differences in the nature of punctate white matter lesions in preterm infants, Front. Neurol. 12 (2021) http://dx.doi.org/10.3389/fneur.2021.657461.