Evaluation of Medical Image Segmentation Models for Uncertain, Small or Empty Reference Annotations

Sophie Ostmeier¹, Brian Axelrod, Jeroen Bertels², Fabian Isensee³, Maarten G.Lansberg⁴, Soren Christensen², Gregory W. Albers⁴, Li-Jia Li², and Jeremy J. Heit¹

¹ Department of Radiology, Stanford School of Medicine, Palo Alto, USA
² KU Leuven, Medical Imaging Research Center, Leuven, Belgium
³ Division of Medical Image Computing, German Cancer Research Center (DKFZ), Heidelberg, Germany
⁴ Department of Neurology, Stanford School of Medicine, Palo Alto, USA
⁵ gray number analytics, Lomma, Sweden
⁶ Department of OHNS / Otolaryngology / Head & Neck, Stanford School of Medicine, Palo Alto, USA

Abstract. Performance metrics for medical image segmentation models are used to measure agreement between the reference annotation and the prediction. A common set of metrics is used in the development of such models to make results more comparable. However, there is a mismatch between the distributions in public data sets and cases encountered in clinical practice. Many common metrics fail to measure the impact of this mismatch, especially for clinical data sets containing uncertain, small or empty reference annotation. Thus, models may not be validated for clinically meaningful agreement by such metrics.

Dimensions of evaluating clinical value include independence from reference annotation volume size, consideration of uncertainty of reference annotations, reward of volumetric and/or location agreement and reward of correct classification of empty reference annotations.

Unlike common public data sets, our in-house data set is more representative. It contains uncertain, small or empty reference annotations. We examine publicly available metrics on the predictions of a deep learning framework in order to identify for which settings common metrics provide clinical meaningful results. We compare to a public benchmark data set without uncertain, small or empty reference annotations. https://github.com/SophieOstmeier/UncertainSmallEmpty.git

1 Introduction

Generally, machine learning practitioners define a metric as a function that describe the performance of a model. The optimal choice of metrics depends on the data set and the machine learning task to guarantee that the predictions accurately describe the intended phenomenon [1]. Metrics can be used in two different ways. First, as the criteria that the model tries to optimize as a loss functions. Second, as a way validating and evaluating the performance of the model. This work focuses on the latter, referred to as performance metrics.

Performance metrics differ in their characteristics. The correlations between them determines the additional information revealed, and therefore, the suitability at measuring the performance of models on various clinical data set. Selecting suitable performance metrics for a specific machine learning task ensures consistency in model performance between the development and deployment. For example, physicians that potentially use model predictions for treatment decisions of patients rely on an optimization and evaluation process of the model towards reliable and meaningful clinical information.

For data sets with uncertain, small (< 1% of organ, 1-2ml) or empty reference annotations established performance metrics penalize or misinterpret clinically meaningful information. Several works describe the potential dependency of metric values to the segmentation size and degree of class imbalance [1][2], the equal weighing of all regions of misplaced delineation independently of their distance from the surface [4] or missing definition for empty reference annotations [3][5]. The failure to describe uncertain, small, or empty segmentations lead to irrelevant and misleading optimization and evaluation procedures of segmentation models.

We propose clinical meaningful metrics and their implementation. Our recommendations are backed by analysis of established metrics on two medical image segmentation data sets with opposing properties; the
public data set BraTS 19 with certain, large and no empty reference annotations and the in-house NCCT data set with uncertain, small, or empty reference annotations.

1.1 Uncertain Reference Annotations

We define uncertainty as underlying inter-expert variability of reference annotations. Although volume and location of the segmented object between experts differ, every expert is equally correct [6].

For volume agreement, the reference annotation’s classification of a voxel can be true, and the segmentation of another expert or the prediction of the model can be false or vice versa. In practice, the spectrum ranges from a worst-case to a best case-scenario. In the best-case scenario all \( FP \) are true positives (\( TP \)). In the worst-case scenario all false positives (\( FP \)) are \( TP \). For example, in Fig. 1(a), the segmentation of an acute stroke from an expert A (green, 2. row) is larger than the segmentation of expert B (pink, 3. row). Some green voxels at the border of the segmentation might falsely or truly be part of the lesion (\( FP \) or \( TP \)). However, the underlying low signal-to-noise ratio of NCCT and the continuous transition from healthy to ischemic brain tissue inevitable prevent a precise segmentation of acute ischemic brain tissue. Both, expert A and B are correct. Another example is shown in Fig. 1(c). The expert A’s segmentation is empty (2. row), however the prediction (yellow, 3. row) is not empty. Visual investigation shows an ambiguous lesion that was not segmented by the expert A making all voxels \( FP \) but truly \( TP \).

For location agreement, the distance between voxels from the reference annotation to another expert or prediction might be longer or shorter. For example, in Fig. 1(a), the surface voxels of expert A (green, 2. row) and expert B (pink, 3. row) will have a different distance to the surface voxels of a predicted segmentation.

The BraTS 19 data set uses a consensus vote on the manual segmentations of four experts to approximate a more certain binary reference annotation. In contrast, for the in-house NCCT data set only expert A is chosen as reference annotation. A fusion of all three available expert’s segmentation would result in a binary segmentation not reflective of pathophysiological and modality related image characteristics of NCCT.

We compare how metrics behave in the setting of fused reference annotations of the BraTS 19 data set and uncertain reference annotations of expert A in the NCCT data set.

1.2 Small Reference Annotations

Small volumes can be defined as less than 1% of the organ volume. For the intracranial space, 1% is about 13 ml [7]. The distribution and median of reference annotation volumes vary across medical image date sets (Fig. 1(b)). The BraTS 19 data set has a median (IQR) volume of 48 (24-81) ml compared to the in-house NCCT data set with 15 (6-39) ml [8][10]. We hypothesize that the distribution of reference annotation volumes influence the value of metrics independently from the model’s performance (Fig. 1(a)) [5].

In the clinical context, small segmentations present a challenge for established metrics. For example, an acute ischemic stroke patient with a suspected large vessel occlusion undergoes emergent imaging to identify
(i) the presence and volume of ischemic brain tissue, (ii) the presence and volume of salvageable brain tissue if blood flow to the brain is restored (penumbra), and (iii) the presence of a large vessel occlusion. The volume of the ischemic core is currently the most important factor in guiding treatment decisions [11]. The volume of acute ischemic brain tissue on NCCT of stroke patients with a large vessel occlusion patient is often quite small (1-5 ml in volume [11]). A model may segment a 1-2 ml lesion volume that has poor overlap with the segmentation by a neuroradiologist and have a low performance metric despite properly identifying the volume. A slight difference in volume location within the brain is highly unlikely to influence a physician’s decision to treat the patient with endovascular therapy or not.

We describe how the distribution of reference annotation volumes produce different metrics values regardless of their degree of location and volume agreement.

1.3 Empty Reference Annotations

Empty reference annotations are described as segmentations where the object of interest could not be identified by the annotator. The object might have been invisible at the time of the segmentation (Fig. 1(c)).

Segmentation of an object within an image is a different task than classification of an image. A classification task confirms the presence or absence of an object in the image (image-level), while a segmentation task assigns each voxel of the image to an object class (voxel-level) [5]. An image-level classification task can also be formulated as a segmentation task by checking if the segmentation is empty. Therefore, when using a segmentation model in this way, it is important for the performance metrics to capture behavior on empty segmentations. However, many metrics for image segmentation are not defined for empty segmentations and return a no or a meaningless value.

For clinical deployment, empty images are possible. The predictions of segmentation models need to be optimized and evaluated for correct image-level classification [3]. For example, it is possible that a stroke lesion in an early time window (0-4h after symptom onset) has a very low contrast and cannot be segmented. In this case, the reference annotation and the predicted segmentation should both be empty and metrics should return the optimal value. No visible and no predicted lesion would result in a treatment decision in favor of endovascular therapy [11]. Another example is a population of patients with suspected brain tumors. Similarly, empty segmentations are possible and influence clinical implications. The absence of a visible brain tumor combined with clinical typical symptoms like headaches, vision problems, nausea, and dizziness would guide clinician towards other brain conditions such as pseudotumor cerebri.

We introduce a method to modify established performance metrics by setting a volumetric threshold tailored to each clinical context. Below the threshold the agreement measure of metrics is expected to go beyond clinical relevance.

1.4 Clinical Meaningfulness

To apply challenge winning segmentation models in clinical practice, the focus on clinically meaningful optimization and performance metrics for each clinical context is desirable. Meaningful encompasses:

- Independence from reference annotation volume
- Consideration of uncertainty in the reference annotation
- Reward of volumetric and location agreement between the reference annotation and prediction
- Reward of correct classification of empty reference annotations

2 Metrics

2.1 Fundamentals

A 3D image consists of a \( (w_i \times h_i \times d_i) \) voxel grid with width \( w_i \), height \( h_i \), and depth \( d_i \). We refer to the set of voxels as \( X \) with \( |X| = w_i \times h_i \times d_i = n \). A segmentation mask \( M \) is defined as division of the image grid into meaningful parts like organs, pathologies, or any object. Each voxel \( x \in X \) is assigned a membership to an object class \( k \) by a function \( f^k(x) \) [1].
The segmentation mask $M$ can be evaluated by volume and location agreement of the segmented object. On a voxel-level, the agreement between the reference annotation mask $M_{ra}$ and the predicted mask of a model $M_p$ can be measured in distance or similarity for each voxel.

For the similarity agreement, all voxels of $M_{ra}$ have been assigned to their true classes $M_{ra} = \{ M_{ra}^0, M_{ra}^1, \ldots M_{ra}^k \}$. The model’s classification for each voxel results in a prediction of all voxels in $M_p = \{ M_p^0, M_p^1, \ldots M_p^k \}$. $k = 0$ is referred to as background class. Each classified voxel of $M_p$ is further compared to its reference annotation class in $M_{ra}$. For a binary classification task ($k = \{ 0, 1 \}$) a confusion matrix of four cardinalities namely $TP$, $FP$, false negatives ($FN$), and true negatives ($TN$) can be defined $\square$ (Table 1).

### Table 1: Four Cardinalities

| $M_{ra}^1$ | $M_{ra}^0$ |
|------------|------------|
| $TP_{ra}$  | $FP_{ra}$  |
| $FN_{ra}$  | $TN_{ra}$  |

Therefore, $TP + FP + TN + FN = |X|$. The volume in ml of the $M_{ra}$ and $M_p$ is defined as

$$V_{ra} = M_{ra}^1 \times V_x = (TP + FN) \times V_x$$
$$V_p = M_p^1 \times V_x = (TP + FP) \times V_x$$

where $V_x = w_x \times h_x \times d_x$, $w_x$, $h_x$ and $d_x$ refer to the width, height, and depth of a voxel $x$. The inherent class imbalance of 3D medical image segmentation ($M_{ra}^1 << M_{ra}^0$) prevent meaningful performance evaluation by any metric that includes $TN$ (Specificity, ROC, Accuracy, Kappa etc.). Therefore, metrics that include $TN$ in their function should be avoided. Volume metrics without $TN$ are VS and AVD. Overlap metrics measure $TP$ relative to a combination of $TP$, $FP$, $TN$ and $FN$. Overlap metrics that exclude $TN$ are Dice, Recall and Precision (Table 2).

For distance agreement, the set of surface voxels $S^k$ is defined $\int_S ds^k$. Where $s^k$ is a voxel on the border of the voxels assigned to object class $k$ within $M$. The distance for a voxel $x$ to $S^k$ is defined as

$$d(x,S^k) = \min_{s \in S} d(x,s^k).$$

Where $d(x, S) = ||x - s||$ is the Euclidean distance. The unit of $d$ is mm, where a value of 0 means no distance and a high value a greater distance between $x$ and $S^k$.

For a binary segmentation task with $k = \{ 0, 1 \}$, the set of voxels $S_{ra}^1$ are defined as surface of $M_{ra}^1$ and $S_p^1$ as surface of $M_p^1$. Metric definitions for common volume, overlap, distance metrics, that were used for the experiments can be found with their implementations in github link and in Table 2.

#### 2.2 Surface Dice at Tolerance

A new evaluation metric first introduced by Nikolov et al. $\square$ is the Surface Dice. It describes which portion of voxels of $S_p^k$ that have the same spatial location as the $S_{ra}^k$ relative to the sum of $S_p^k$ and $S_{ra}^k$. Therefore, it measures first the non-symmetric distances between the $S_p^k$ and $S_{ra}^k$ and then normalizes by the overlapping surface area. Thus, the Surface Dice rewards true positive surface parts between $S_p^k$ and $S_{ra}^k$.

Adding a tolerance with depth of $t$ establishes a border volume $B^t$. A voxel $x$ of $S_p^k$ within the tolerated distance $t$ to $S_{ra}^k$ would be categorized as true positive voxel.

$$SDT = \frac{|S_{ra}^k \cap B^t_{ra}| + |S_{ra}^k \cap B^t_p|}{|S_p^k| + |S_{ra}^k|}.$$
Table 2: Definitions of Performance Metrics for Medical Image Segmentation

| Category          | Metric                        | Abbr | Usage | Definition                                                                 |
|-------------------|-------------------------------|------|-------|---------------------------------------------------------------------------|
| **Volume**        | Volumetric Similarity         | VS   | 12    | $1 - \frac{||V_p^1 - V_{ra}^1||}{|V_p^1| + |V_{ra}^1| + \epsilon}$       |
|                   | Absolute Volume Difference    | AVD  | 17    | $\frac{1}{M} \sum_{i=1}^{M} |V_i^1 - V_{ra}^1| + \epsilon$          |
| **Overlap**       | Dice Similarity Coefficient   | Dice | 16    | $\frac{2 \times TP}{2 \times TP + FN + FP}$                            |
|                   | Recall = Sensitivity          | Recall| 16    | $\frac{TP}{TP + FN}$                                                    |
|                   | Precision                     | Precision | 16 | $\frac{TP}{TP + FP}$                                                  |
| **Distance**      | Hausdorff Distance, $q =$    | HD 95| 16    | $\text{max} (h(A, B), h(B, A))$ with $h(A, B) = \text{max_{a \in A, b \in B}} ||b - a||$ |
|                   | 95th percentile               | ASD  | 16    | $\frac{1}{|S_{ra} | + | S_p |} \sum_{x \in S_p} d(x, S_{ra}) + \sum_{y \in S_{ra}} d(y, S_p)$ |
|                   | Average Surface Distance      | ASD  | 16    | $\frac{|S_{ra} \cap B_p^1| + |S_{ra} \cap B_p^1|}{|S_{ra} | + | S_p |}$       |
|                   | Surface Dice at Tolerance     | SDT  | 4     | $\frac{|S_{ra} \cap B_p^1| + |S_{ra} \cap B_p^1|}{|S_{ra} | + | S_p |}$       |
| **Image-level**   | Lesion Detection Rate         | LDR  | 2     | $\frac{\text{number of subjects detected with lesion}}{\text{number of all subjects with lesion}}$ |
| classification    | Correct Classification Rate   | CCR  | -     | $\frac{\text{number of correctly detected subjects}}{\text{number of all subjects with lesion}}$ |

### 3 Methods

#### 3.1 Data Set

This study was approved by the IRB and the requirement for written informed consent was waived.

A deidentified dataset of 200 NCCT images of acute ischemic stroke patients presenting within 6-16 hours of symptom onset from DEFUSE3 trial [22] was provided to three neuroradiologists with 8, 8 and 9 years of experience. For the study design please refer to [https://clinicaltrials.gov/ct2/show/NCT02586415](https://clinicaltrials.gov/ct2/show/NCT02586415).

The experts were instructed to segment abnormal hypodensity on the NCCT that is consistent with acute ischemic stroke within 6 to 16-hours of symptom onset $M_{ra}^1$. Detailed instructions and videos, as well as oral explanation of the task where given. Any missed lesions or missed slices were not corrected.

Expert A (= reference annotation) has 21 empty segmentations and additional 4 segmentations < 1ml. The median (IQR) of $V_{ra}$ is 15 (6-39) ml. Expert B has 0 empty segmentations and 6 segmentations < 1ml, expert C has 38 empty segmentation, 63 segmentations ¡1ml, the prediction of the model include 11 empty segmentations and additional 8 segmentations < 1ml.

The BraTS 19 public data set was used as reference data set to compare results obtained from the NCCT data set [8][10]. It includes MRI of high and low grade glioma patients. The $M_{ra}$ were segmentations by four experts. The $M_{ra}$ had been fused with a consensus vote. The reference annotation consisted of five object classes $M_{ra} = \{M_{ra}^0, M_{ra}^1, M_{ra}^2, M_{ra}^3, M_{ra}^4\}$. We only used the object class $k = 1$, $M_{ra}^1$, corresponding to whole tumors.

#### 3.2 Model

The nnUNet was selected as the baseline model [23]. The input was the NCCT image and the output a predicted segmentation $M_p$. The manual segmentation of expert A was used as reference annotation $M_{ra}$ to train the model. The number of epochs was set to 500. The automatically configured patch size was
$1 \times 28 \times 512 \times 512$ and spacing of $0.00, 0.45, 0.45$ stages, two 3D convolutions per stage, leaky ReLU as activation function. 5 splits were randomly determined for 5-fold cross validation with 160 for the training set and 40 for the validation set. The evaluation is made on the aggregated 5-fold cross validation segmentations ($n = 200$). For the extraction the intracranial space form the skull and soft tissue on NCCT the BET_CT was used according to Akkus et al. [24].

A second baseline nnUNet was trained as described in the previous paragraph. The input image encompasses the MRI sequences T1, T1 CE, T2 and FLAIR as four channels. The output of the model is the segmentation of the whole tumor $M_{ra} = \{M_{ra}^0, M_{ra}^1\}$. For the extraction the intracranial space form the skull and soft tissue on MRI, the HD_BET was applied [24].

### 3.3 Inter-expert Agreement

The chosen metrics are computed with an extended and modified framework based on [4,23] (Table 2). The python code can be applied to folders with reference annotations and predictions in .nii.gz format and produces an .xlsx file with all results as well as with the mean and median. The inter-expert agreements were evaluated with the same framework and displayed as average of the three experts. The code will be published.

### 3.4 Evaluation of Models

The model’s performance (agreement between $M_{ra}$ and $M_p$) was measured with the proposed evaluation framework [https://github.com/SophieOstmeier/UncertainSmallEmpty.git]. In addition, the agreement between $M_p$ to $M_{ra}$, $M_{expert2}$ and $M_{expert3}$ was averaged. All cases where $M_{ra}$ and corresponding $M_p$ disagreed in the image-level classification of a stroke were visually compared to the segmentations of the two experts B and C, the laterality and Computed Tomography Perfusion (CTP). The BraTS 19 data set was evaluated with the same modified framework.

### 3.5 Evaluation of Metrics

We evaluate the metrics on both the NCCT and BraTS 19 data sets. We quantitatively and qualitatively analyzed their correlations among each other and to $V_{ra}$ and $V_p$. The original Dice is not defined if $V_{ra}$ and $V_p$ are empty and therefore not included in the following graphs. Based on the above, we propose a guideline about selecting a clinical meaningful metrics for segmentations with uncertainty, small volumes and possible empty reference annotations.

|          | Inter-expert Agreement Average | NCCT Agreement $M_{ra}$ to $M_p$ | NCCT Average Agreement all experts to $M_p$ | BraTS19 Agreement $M_{ra}$ to $M_p$ |
|----------|-------------------------------|----------------------------------|---------------------------------------------|-----------------------------------|
| VS       | 0.66                          | 0.75                             | 0.61                                        | 0.95                              |
| AVD      | 0.75                          | 5.61                             | 7.03                                        | 4.8                               |
| Dice     | 0.37                          | 0.52                             | 0.37                                        | 0.85                              |
| Recall   | 0.66                          | 0.65                             | 0.4                                         | 0.86                              |
| Precision| 0.35                          | 0.53                             | 0.58                                        | 0.86                              |
| HD 95    | 18.2                          | 14.6                             | 17.84                                       | 3.16                              |
| ASD      | 2.92                          | 2.96                             | 5.59                                        | 0.74                              |
| SDT 5mm  | 0.67                          | 0.72                             | 0.62                                        | 0.98                              |

Table 3: Inter-Expert Agreement vs. Model
4 Results and Discussion

4.1 Inter-expert Agreement and Evaluation of Models

Upon negative tests for normal distribution, results of the aggregated validation examples for each metric are shown as medians, unless specified as average over three experts. For the NCCT data set, volume and overlap metrics are highest for the model performance (VS 0.75, Dice 0.52, Table 3).

Fig. 2(a) shows the distribution of data points for the Dice with more than 15% allocated to the first bin (0.0 - 0.02). A subgroup analysis for 31 cases with Dice = 0 and 31 cases of the Dice > 0.73 on the x-axis and with \( V_{ra} \) in ml on the y-axis is shown in Fig. 2(b). The cases with Dice = 0 have a median (IQR) \( V_{ra} \) of 0.9 (0-3.2) ml whereas the cases with the highest Dice have a median \( V_{ra} \) of 32 (16-78) ml.

The average Dice between reference annotations of expert A, B and C is equal to the Dice of experts to the predicted masks \( M_p \) (0.37 and 0.37, respectively). The probability for voxels to be classified as positive within a stroke lesion is similar between experts (Recall 0.65 vs. 0.66), whereas the probability within the prediction to be truly positive is higher for \( M_p \) (Precision 0.53 vs. 0.35).

Distance metrics are comparable between the model performance and inter-expert agreement with HD 95 (14.6 vs. 18.2mm), ASD (2.96 vs. 2.92 mm) and SDT 5mm (0.72 vs. 0.67). The SDT show increasing agreement with increasing tolerance.

Overall, model’s segmentation performance is equal or higher than the average agreement between experts. Even, the average agreement of three experts to \( M_p \) is similar to the inter-expert agreement (Table 3 column 2 vs. 4). Therefore, the experts agree at least as good with the prediction of a model that was trained only on expert A than between each other.

For the image-level classification task of the NCCT data set, LDR and CCR were calculated (0.98 and 0.91, respectively). 7 of the 21 empty reference annotations were correctly classified as empty. The correct classification of subjects was higher compared to the inter-expert LDR and CCR (0.89 and 0.85, respectively). Visual inspection of \( M_p \), \( M_{ra} \), CTP and laterality for the 18 cases where CCR indicates disagreement reveals at least 8 cases are correctly classified by the model.

For the BraTS 19, no empty reference annotations or predictions were observed. Therefore, the LDR and CCR returned optimal values. An image-level classification task cannot be evaluated and how the algorithm performs on a clinical data set with empty reference annotations remains unknown.
4.2 Evaluation of Metrics

4.2.1 Quantitative analyses  Fig. 3 displays the Spearman correlation matrix between the metric values derived from NCCT data set. Dice, Recall, SDT, VS, \( V_p \) positively correlate with each other and ASD and HD 95 correlate negatively.

As defined in Section 4.1, metrics should ideally not be correlated to the value of \( V_{ra} \) to effectively judge model performance independent of lesion size \( V_{ra} \). Fig. 4 displays the Spearman correlation of each metric to each quartile of \( V_{ra} \) with 0 being the optimal value. Metrics are ordered from left to right by the value of correlation for the 1st quartile of \( V_{ra} \) (NCCT data set: 0.5 – 8.8ml, \( n = 45 \), BraTS 19 data set: 0.0 – 24.5ml, \( n = 84 \), red dots). For the NCCT data set the 1st quartile of \( V_{ra} \) has a significant correlation for all metrics, while the 2nd to 4th quartile of \( V_{ra} \) does not (Fig. 4(a), \( > 8.8 \)ml, orange and grey dots). For the BraTS 19 data set only Dice (0.23) and Precision (0.39) as well as ASD show significant correlation for the 1st quartile of \( V_{ra} \) (Fig. 4(b)).

Overlap Metrics
All overlap metric show strong correlation to \( V_{ra} \) which is indicated by large and dark circles in Figure 3. For the BraTS 19 data set Dice and Precision have a strong correlation to \( V_{ra} \) (> 0.5).

Dice: In Fig. 4 Dice has the highest correlation (0.65) for the 1st quartile of \( V_{ra} \) in the NCCT. These findings are consistent with Liu et al. [2], BraTS 19 data set shows a lower correlation (0.23). The following explanations refer to \( M_{1ra} \) instead of \( V_{ra} \) to emphasize the voxel-level (see equation 1).

In theory, an image segmentation task can be perfectly balanced binary classification problem with \( \alpha = \frac{M_{1ra}}{M_{0ra}} = 0.5 \) and \( M_{1ra} = M_{0ra} \). The smaller \( M_{1ra} \) the higher is the class imbalance between voxels of \( M_{1ra} \) and \( M_{0ra} \). In practice, \( M_{1ra} << M_{0ra} \) and \( \alpha << 0.5 \). \( M_{0ra} \) include mostly TN with a low probability of FP or FN. Therefore, we define a reasonable region of interest (ROI) with \( M_{ROI} \subseteq M_{ra} \), here the intracranial space. The degree of balance is equal to the fraction of positive voxels in ROI and can be defined as \( \alpha = \frac{M_{1ra}}{M_{0ra}} \). For example, the 1st quartile of \( V_{ra} \) has a lower \( \alpha \) than the 4th.

We hypothesize that the value of \( \alpha \) influences the probability of a voxel to be classified as TP. Because the Dice primarily rewards the assignment of voxels to TP, it does not evaluate small and large \( M_{1ra} \) equally.

Now, consider a binary segmentation model that decides each voxels’ membership in \( M_{1p} \) and \( M_{0p} \) with a coin toss. We plot the expected Dice \( E_D \) versus the underlying image class imbalance \( \alpha \) with
Fig. 4: For (a) and (b) the positions and heights of the points show how metrics correlate depending on the $V_{ra}$ and how the correlation depends on the magnitude of $V_{ra}$. X indicate a non-significant correlation (p-value < 0.05). The metrics are ordered by value of the correlation Q1 to $V_{ra}$. $V_{ra}$ is equally split in 4 groups. Empty reference annotations were excluded.

\[ E_D = \int \frac{2^{-n} \left( TP + FN \right)}{TP + FN} \left( TN + FP \right) \times \alpha^{TP+FP} (1-\alpha)^{FN+TP} \ D \]  

where D is the Dice function \(\left(\begin{array}{cc} 2xTP \\ 2xTP+TN+FP \end{array}\right)\) (Fig. 5). For example, if $\alpha$ is very low at 0.01 (1% of the organ), then the expected Dice is 0.02. If $\alpha$ is 0.5 (50% of the organ), the expected Dice is much higher at 0.5 (dashed line). The regression lines for the Dice over $\alpha$ for the NCCT (blue line) and BraTS 19 data set (green line) confirm the monotonic tendency of the expected Dice in practice. The Spearman correlations are 0.63 and 0.50 for the NCCT and BraTS data set, respectively. Therefore, the Dice value depend on $\alpha$ for both data sets, however this behaviour is more present in the NCCT data set. Since class imbalance produces lower Dice values (Fig. 5), location error for small $M_{ra}$ is more penalized than for larger $M_{ra}$. The Dice does not evaluate small and large $M_{ra}$ equally.

The Dice does not consider uncertainty. The denominator contains the sum of $M_{ra}^1 = TP + FN$ and $M_p^1 = TP + FP$ (Table 2). If in the best case-scenario expert B or a model is superior to expert A (reference annotation) in classifying voxels correctly, all FP are TP and all FN are TN. However, the denominator would disproportionately increase and lead to a lower Dice value. As a result the performance of the model is underestimated. In the worst case-scenario expert B or a model is inferior to expert A in classifying voxels correctly; all FP are FP and all FN are FN. The Dice value does not change. Therefore, the Dice always considers the worst case-scenario and does not consider uncertainty in $M_{ra}^1$.

The Dice does not reward correct image-level classification. It returns zero or "NaN" if $M_{ra}^1$ or overlap is zero ($TP = 0$) without considering the distance between $M_{ra}^1$ and $M_p^1$. Therefore, the Dice is zero, if both $M_{ra}^1$ and $M_p^1$ are zero, $M_{ra}^1$ and $M_p^1$ are right next to each other and is also zero if $M_{ra}$ and $M_p$ are fare from each other. Simply, eliminating very small $V_{ra}$ from evaluation of overlap metrics, for example < 1ml,
Fig. 5: The red line represents the expected Dice $E_D$ over class imbalance $\alpha$ under the random model defined here. Blue and green dots and lines represent each data point and a regression line for the NCCT and BraTS 19 data set, respectively. The dashed line indicates the expected Dice $E_D$ for a balanced data set.

and setting the metric value to 1, if $V_{ra}$ and $V_p < 1$ml, enables measure of image-level classification. For example, a patient with $V_{ra} < 1$ml is reliably evaluated by the modified Dice if $V_p < 1$ml has no relevant ischemic brain tissue. Model optimization procedures and complex model architectures that aim to increase performance for stroke segmentation on NCCT $V_{ra} < 1$ml will not have a clinical impact.

**Precision and Recall:** Precision is stronger correlated to $V_{ra}$ (0.62) compared to Recall (0.25). The relative number of $FP$ increases for smaller $V_{ra}$ while the relative number of $FN$ do less (Spearman correlation -0.36 vs. -0.24). This correlation pattern is less reproducible with the BraTS 19 data set (0.53 and 0.27).

Mathematically, Precision and Recall are similar except for their consideration of $FP$ for Precision and $FN$ for Recall in the denominator. Precision rewards $TP$ relative to $TP + FP = M^1_p$ and Recall $TP$ relative to $TP + FN = M_{ra}$ (Table 2). Given the above mentioned positive correlations to $V_{ra}$, the model classifies voxels to $M^1_p$ more than an expert to $M^1_{ra}$ (Fig. 3). This behaviour can especially seen for small $V_{ra}$ (Fig. 4(a)). One can argue that the model learns to classify voxels less frequently to $M^1_p$, because the chance of being correct is higher. This would imply for all overlap metrics, that in the presence of small $V_{ra}$ and higher class imbalance within the image, small $V_{ra}$ are underestimated in their size. However, the expert’s assessment is even more guided by class imbalance with small volumes being less obvious to the expert’s eye. Therefore, $M^1_{ra}$ would be equal to $TP + FN + FP$, which explains a lower value of Recall compared to Precision.

**Volume Metrics**

**VS:** VS show insignificant correlation to $V_{ra}$ overall. For the 1st quartile of $V_{ra}$ VS has lower correlation (0.46) than Precision and Dice (0.65, 0.64). For the BraTS 19 data set VS has insignificant correlation overall and for the 1st quartile of $V_{ra}$. In theory, VS penalizes the difference between $FN$ and $FP$ relative to the sum of $V_{ra}$ and $V_p$. Analysis shows that the relative $FP$ and $FN$ to $M^1_{ra}$ correlate negatively with $V_{ra}$ (see section 4.2.1). Assuming that for small $V_{ra}$ one source of $FP$ and $FN$ is uncertainty (see section 1.1); VS does not penalize uncertainty as long as their difference has a linear relationship to $V_{ra}$. However, VS ignores location error. Each $FN$ and $FP$ can be anywhere in the image without an influence on the value of VS. VS therefore might overestimate the performance of the model if the location agreement is clinically relevant.

**AVD:** AVD has one of the highest correlations to $V_{ra}$ (0.62). The AVD correlation to $V_{ra}$ for the BraTS 19 data set is lower (0.34). For the 1st and 2nd quartile of $V_{ra}$ AVD has no significant correlation to $V_{ra}$ in the NCCT and BraTS 19 2019 data set. AVD measures the absolute difference between $FN$ and $FP$. In practice, it remains unclear whether the reduction in AVD slightly improved the performance for large $V_{ra}$ or substantially for small $V_{ra}$, because AVD does not normalize to the sum of $V_{ra}$ and $V_p$. Overall, VS is less
correlated to $V_{ra}$ as AVD. Volume metrics allow uncertainty in location and only reward volume agreement. VS returns the optimal value if $V_{ra} = 0$ and $V_p = 0$.

**Distance Metrics**

**HD 95 and ASD:** HD 95 and ASD, have negative and low correlation to $V_{ra}$ (insignificant and -0.29). For $V_{ra}$’s 1st quartile HD 95 show less correlation than ASD (-0.36 vs. -0.52). For the BraTS 19 data set, ASD and HD 95 also correlate negatively to $V_{ra}$ (-0.19 and -0.19). Low to insignificant correlation is shown for $V_{ra}$’s 1st quartile. The negative correlation can be explained by the possible larger distance between smaller $V_{ra}$ and increased uncertainty for smaller $V_{ra}$. HD 95 does not penalize volume error. Uncertainty in volume is considered.

**SDT:** SDT 0mm and 5mm have insignificant correlation to $V_{ra}$. However SDT 10mm is correlated to $V_{ra}$ for the NCCT data set. For the BraTS 19 data set, SDT at all tolerance levels has the lowest correlation to $V_{ra}$ overall as well as for the $V_{ra}$’s 1st quartile. SDT measures both the location by the distance and the volume by surface agreement. It weights errors in the outer region of $M_{ra}$ less by introducing a tolerance. This is especially useful, if the pathophysiological and modality related lower signal cause more uncertainty in the outer compared to the inner of $M_{ra}$. Higher correlation of SDT 10mm can be explained by the volume’s growth rate of the tolerance border being proportional to the thickness. This results in larger thicknesses such from $t = 10mm$ having a stronger correlation to $V_{ra}$. Contrary to overlap measures, if $V_{ra}$ and $V_p$ are right next to each other and within the border region $B_{t}^p$, SDT still measures a meaningful agreement.

For image-level classification, distance metrics have an infinite distance if $V_{ra} = 0$, therefore a threshold for irrelevant small volumes is also justified (see 4.2.1).

**Image-level classification metrics**

A LDR is a slight modification of the SDR proposed [2]. Liu et al.’s study only included patients with $V_{ra} > 0$. However, the agreement between the classification is not assessed in case $V_{ra} = 0$. 97% for the NCCT data set and 100% for the BraTS 19 data set correctly classified as having a lesion. The proposed CCR evaluates the agreement of classification to $V_{ra} = 0$ or $V_{ra} > 0$. Therefore, agreement is rewarded with a value of 1 if both the expert and the prediction agree on stroke or no lesion. However, both LDR and CCR suffer from case class imbalance with 179 cases with a lesion and 21 without a lesion. For example, 33% of the empty reference annotations are classified correctly, but LDR and CCR indicate very good performance for image-level classification.

In conclusion, separate metrics for image-level classification to evaluate segmentation models carry misleading information, because data sets include at most only a few patients with $V_{ra} = 0$. This class imbalance prevents an meaningful evaluation on image-level classification task. A solution present segmentation metrics with a volumetric threshold, below which agreement is rewarded by setting the metric value to the optimum (e.g. < 1ml, see 4.2.1).

**4.2.2 Qualitative analyses**

Inspection of the NCCT images of the $V_{ra}$’s 1st quartile revealed cases with moderate overlap having sufficient visual agreement for clinical application (e.g. Dice = 0.53, Fig. 6(a)). For this case, a small displacement is not reflected in a lower SDT 5mm value (0.99). For images of 4th quartile (Fig. 6(b)), the Dice indicated higher overlap compared to Fig. 6(a) (0.78 vs. 0.53). But $V_{ra}$ and $V_p$ do not overlap in the region of the basal ganglia. Whether a lesion extents to the basal ganglia has an influence on the patient’s functional outcome and should be penalized with a lower metric score. SDT 5mm returns a lower value (0.77 vs. 0.99) [25].

**5 Guidelines to Choose Clinically Meaningful Metrics for Data Sets with Uncertain, Small and Empty Reference Annotation**

**5.1 Volume agreement**

VS measures volume agreement less dependent on $V_{ra}$. VS allows uncertainty about the location of $V_{ra}$ (Table 3). VS is suitable, if volume agreement is the most important clinical concern, like for stroke imaging [11]. VS is less suitable for clinical data set and segmentations tasks, where location is crucial. In addition, for
segmentation with $V_{ra} = 0$, VS returns the optimal value of 1. Therefore, VS is also suitable for data set with expected empty reference annotations.

AVD does not offer advantages over VS for data set with a small median of $V_{ra}$ and uncertainty. AVD makes comparison between cases of the same data set impossible.

5.2 Overlap Agreement

Overlap metrics (Dice, Recall, Precision) reveal the highest correlation to $V_{ra}$ compared to volume and distance metrics. Underlying uncertainty of the reference annotation is not considered. Overlap metrics simultaneously measure volumetric and location agreement. Overlap metrics do not measure image-level classification. A modified overlap metrics with a clinical driven threshold are more suitable. Cases with empty or very small $V_{ra}$ and $V_p$ should be set to 1, because overlap metrics are expected to penalize location and volume error beyond the clinical relevance and agreement of empty or very small $V_{ra}$ and $V_p$ has clinical impact (Table 4). Given this analysis, the Dice is a sub-optimal choice of metric for data sets with uncertain, small or empty reference annotation.

The choice for Precision or Recall over Dice as performance metric depends on the clinical context. Recall measures the percentage of $V_{ra}$ that overlaps with $V_p$ and Precision the percentage of $V_p$ that overlaps with $V_{ra}$. Recall is a more suitable metrics if a high sensitivity for each voxel to be classified to $k = 1$. For example, intracranial hemorrhage on a NCCT of an acute stroke patient should be reliably segmented, because a positive case changes the decision of intravenous thrombolysis [11].

5.3 Distance agreement

The HD 95 measures the maximum distance between $S_{ra}$ and $S_p$ and does not correlate $V_{ra}$. Uncertainty in volume size is not penalized. HD 95 is valuable metric for clinical questions, where the exact location is crucial (Table 4). For example Multiple Sclerosis patient are assessed with the McDonald criteria [26]. For the dissemination in space the exact location (periventricular, cortical, juxtacortical, infratentorial, or spinal
Table 4: Guideline to Choose Clinically Meaningful Metrics for Data Sets with Uncertain, Small and Empty Reference Annotation

| Metric | Independence from Volume of Reference Annotation | Consideration of Uncertainty in Reference Annotation | Reward of Volume and Location agreement | Reward of Absence Agreement |
|--------|-----------------------------------------------|-----------------------------------------------|----------------------------------------|---------------------------|
| VS     | ✓                                             | ✓                                            | -                                      | ✓                         |
| AVD    | -                                             | ✓                                            | -                                      | -                         |
| Dice   | -                                             | -                                            | ✓                                      | -                         |
| Recall | -                                             | -                                            | ✓                                      | -                         |
| Precision | -                                           | -                                            | ✓                                      | -                         |
| HD 95  | ✓                                             | ✓                                            | -                                      | set threshold             |
| ASD    | -                                             | -                                            | ✓                                      | set threshold             |
| SDT 0mm | ✓                                          | -                                            | ✓                                      | set threshold             |
| SDT 5mm | ✓                                          | ✓                                            | ✓                                      | -                         |
| SDT 10mm | -                                         | ✓                                            | ✓                                      | -                         |

The value of ASD moderately correlated to $V_{ra}$ and does not offer superior characteristics over HD 95. The SDT show an insignificant correlation to $V_{ra}$ for tolerance level of 0 and 5mm. A higher tolerance comes with the cost of stronger correlation to $V_{ra}$. The degree of considered uncertainty is well represented by the tolerance level and can be adjusted according clinical needs of precise object location. SDT with 5mm tolerance seems to be the most clinical meaningful optimizing metric for stroke imaging. The SDT 5mm highly correlates with volume metrics such as VS (0.73) and at the same time with HD 95 (0.86). Therefore, reward of volume and location agreement is interfered. With a threshold for empty and very small reference annotations to enable image-level classification, SDT shows a superior clinical meaningfulness compared to all other metrics.

6 Limitations

The first limitation is that the evaluation is based on one expert. However, using a single expert allows us to analyze how metrics behave when applied to uncertain reference annotations. The second limitation is that the choice of baseline model might influence the correlations between metrics. In order to mitigate this, we choose nnUNet, a model that is generalizable to many clinical problems \cite{23}, and BraTS 19, one of the most benchmarked data sets. Furthermore, correlation does not prove causation. A correlation between $V_{ra}$ and the value of Dice does not imply that higher $V_{ra}$ value cause higher Dice value. The correlation of Dice and $V_{ra}$ has been found in previous works \cite{1,2,5} however in-depth explanations were missing to date.
7 Conclusion

There is a mismatch between dataset properties in medical image segmentation challenges and cases encountered in clinical practice.

For machine learning practitioners, many common metrics do not capture whether models generalize well to distribution of images encountered in clinical practice. In particular, (i) uncertainty is not considered and cause misleading the values, (ii) small target regions cause unreasonable low metric values, (iii) empty target regions cause return of "NaN", "inf" or meaningless values. We demonstrate that a model evaluated by a Surface Dice at Tolerance generalizes better to clinical practice for a data set with uncertain, small and empty reference annotations. Furthermore, each result provided by a segmentation model needs to be evaluated with a set of metrics before clinical deployment that capture the relevant properties of the specific clinical question. Since these properties vary depending on the area of clinical practice, we recommend using table II to select metrics that will ensure good model performance in a clinical deployment.

For clinicians, it highlights the difficulty of comparing models trained to address different clinical problems. Since the intrinsic demands of these problems vary so much, one cannot make such comparisons based a singular metric or standardized set of metrics.

Ultimately, patient outcome measures of clinical trials should be used to judge significant improvement of such artificial intelligence models for enhanced decision making in clinical practice.

References

1. A. A. Taha and A. Hanbury, “Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool,” BMC medical imaging, vol. 15, pp. 29–29, 2015. 10.1186/s12880-015-0068-x[pii]. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4533825/pdf/12880_2015_Article_68.pdf

2. C.-F. Liu, J. Hsu, X. Xu, S. Ramachandran, V. Wang, M. I. Miller, A. E. Hillis, A. V. Faria, M. Wintermark, S. J. Warach, G. W. Albers, S. M. Davis, J. C. Grotta, W. Hacke, D.-W. Kang, C. Kidwell, W. J. Koroshetz, K. R. Lees, M. H. Lev, D. S. Liebeskind, A. G. Sorensen, V. N. Thijs, G. Thomalla, J. M. Wardlaw, M. Luby, S. The, and V. I. investigators, “Deep learning-based detection and segmentation of diffusion abnormalities in acute ischemic stroke,” Communications Medicine, vol. 1, no. 1, p. 61, 2021. [Online]. Available: https://doi.org/10.1038/s43856-021-00062-8https://www.nature.com/articles/s43856-021-00062-8.pdf

3. O. Commowick, A. Istace, M. Kain, B. Laurent, F. Leray, M. Simon, S. C. Pop, P. Girard, R. Améli, C.-C. Ferré, A. Kerbrat, T. Toudias, F. Cervenansky, T. Glotard, J. Beaumont, S. Doyle, F. Forbes, J. Knight, A. Khademi, A. Mahbod, C. Wang, R. McKinley, F. Wagner, J. Muschelli, E. Sweeney, E. Soura, X. Lladó, M. M. Santos, W. P. Santos, A. G. Silva-Filho, X. Tomas-Fernandez, H. Uren, I. Bloch, S. Valverde, M. Cabezas, F. J. Vera-Olmos, N. Madjic, C. Guttmann, S. Vukusic, G. Eden, M. Dojat, M. Styner, S. K. Warfield, F. Cotton, and C. Barillot, “Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure,” Scientific Reports, vol. 8, no. 1, p. 13650, 2018. [Online]. Available: https://doi.org/10.1038/s41598-018-31911-7https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6135867/pdf/41598_2018_Article_31911.pdf

4. S. Nikolov, S. Blackwell, A. Zverovitch, R. Mendes, M. Livne, J. De Fauw, Y. Patel, C. Meyer, H. Askham, and B. Romera-Paredes, “Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy,” arXiv preprint arXiv:1809.04430, 2018.

5. L. Maier-Hein, A. Reinke, E. Christodoulou, B. Glocker, P. Godau, F. Isensee, J. Klessiek, M. Kozubek, M. Reyes, and M. A. Riegler, “Metrics reloaded: Pitfalls and recommendations for image analysis validation,” arXiv preprint arXiv:2006.01653, 2022.

6. A. Jungo, R. Meier, E. Ernis, M. Blatti-Moreno, E. Herrmann, R. Wiest, and M. Reyes, “On the effect of inter-observer variability for a reliable estimation of uncertainty of medical image segmentation,” in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2018, Conference Proceedings, pp. 682–689.

7. K. Akeret, C. H. Bas van Niftrik, M. Schöck, G. Muscas, T. Visser, V. E. Staartjes, F. Marioni, C. Serra, L. Regli, N. Krayenbühl, M. Piccirelli, and J. Fierstra, “Topographic volume-standardization atlas of the human brain,” medRxiv, p. 2021.02.26.21251901, 2021. [Online]. Available: https://www.medrxiv.org/content/medrxiv/early/2021/03/01/2021.02.26.21251901.full.pdf

8. S. Bakas, M. Reyes, A. Jakab, S. Bauer, M. Rempfler, A. Crimi, R. T. Shinohara, C. Berger, S. M. Ha, and M. Rozyczki, “Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge,” arXiv preprint arXiv:1811.02629, 2018.
at 6 to 16 hours with selection by perfusion imaging,” *New England Journal of Medicine*, vol. 378, no. 8, pp. 708–718, 2018. [Online]. Available: https://www.nejm.org/doi/full/10.1056/NEJMoa1713973

23. F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, “nnu-net: a self-configuring method for deep learning-based biomedical image segmentation,” *Nat Methods*, vol. 18, no. 2, pp. 203–211, 2021. isensee, Fabian Jaeger, Paul F Kohl, Simon A A Petersen, Jens Maier-Hein, Klaus H eng Research Support, Non-U.S. Gov't 2020/12/09 Nat Methods. 2021 Feb;18(2):203-211. doi: 10.1038/s41592-020-01008-z. Epub 2020 Dec 7. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/33288961

24. M. Schell, I. Tursunova, I. Fabian, D. Bonekamp, U. Neuberger, W. Wick, M. Bendszus, K. Maier-Hein, and P. Kickingereder, “Automated brain extraction of multi-sequence mri using artificial neural networks.” European Congress of Radiology-ECR 2019, 2019, Conference Proceedings.

25. C. Laredo, Y. Zhao, S. Rudilosso, A. Renu, J. C. Pariente, A. Chamorro, and X. Urra, “Prognostic significance of infarct size and location: The case of insular stroke,” *Scientific Reports*, vol. 8, no. 1, p. 9498, 2018. [Online]. Available: https://doi.org/10.1038/s41598-018-27883-3

26. A. J. Thompson, B. L. Banwell, F. Barkhof, W. M. Carroll, T. Coetzee, G. Comi, J. Correale, F. Fazekas, M. Filippi, M. S. Freedman, K. Fujihara, S. L. Galetta, H. P. Hartung, L. Kappos, F. D. Lublin, R. A. Marrie, A. E. Miller, D. H. Miller, X. Montalban, E. M. Mowry, P. S. Sorensen, M. Tintoré, A. L. Traboulsee, M. Trojano, B. M. J. Uitdehaag, S. Vukusic, E. Waubant, B. G. Weinshenker, S. C. Reingold, and J. A. Cohen, “Diagnosis of multiple sclerosis: 2017 revisions of the mcdonald criteria,” *The Lancet Neurology*, vol. 17, no. 2, pp. 162–173, 2018, doi: 10.1016/S1474-4422(17)30470-2. [Online]. Available: https://doi.org/10.1016/S1474-4422(17)30470-2