

Domain Shifts Between Clinical Annotators

Task 1 at the MICCAI Hackathon 2022

Amith Kamath

September 2022

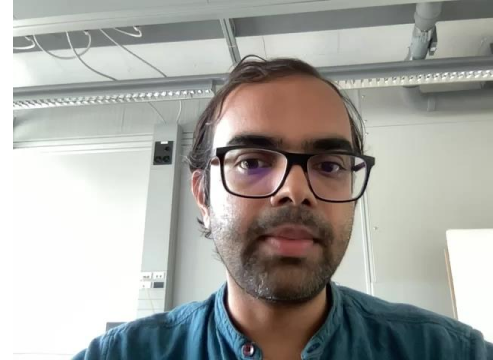
Universität Bern

https://github.com/amithjkamath/miccaihackathon_shifts



Overview

- Introduction
 - Experiments with Annotation Histograms and Disagreement scores
 - Visualizing disagreements versus Uncertainty metrics
 - Is there a relationship between Lesion size and disagreement?
 - Possible next steps beyond the Hackathon
 - Acknowledgements
- 0.5 minute
 - 1 minute
 - 1 minute
 - 1 minute
 - 1 minute



Motivations for experiments



Why are we doing this?

- Quantify the impact of **annotator variance** within the multiple sclerosis dataset of WM lesions in the Shifts 2.0 publicly available [dataset](#) .
- How does model uncertainty compare against annotator variance – how is annotator variance even quantified?
- Are there correlations between the model uncertainty metrics, model prediction accuracy, and lesion sizes?
- Personally: interested in knowing more & networking with the community of researchers here.
- Also personally: Reasonably close to my PhD research topic :-

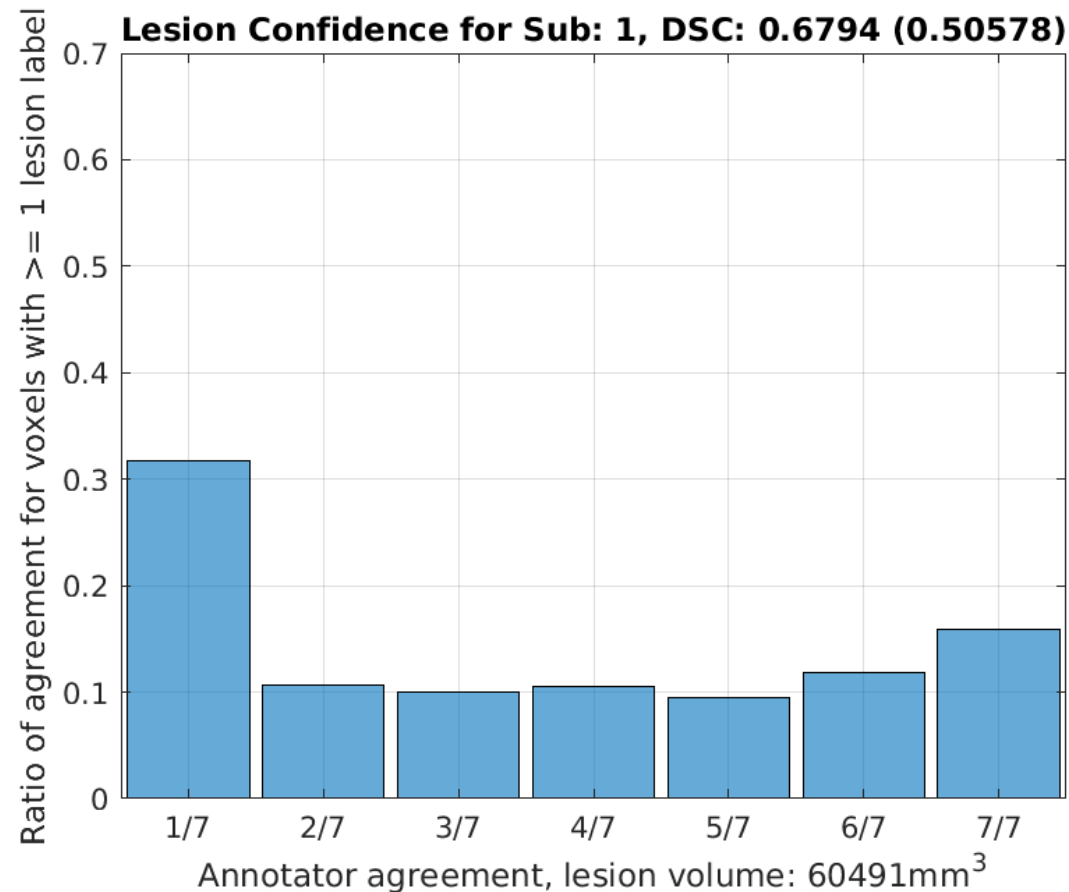


Understanding Annotator variance:

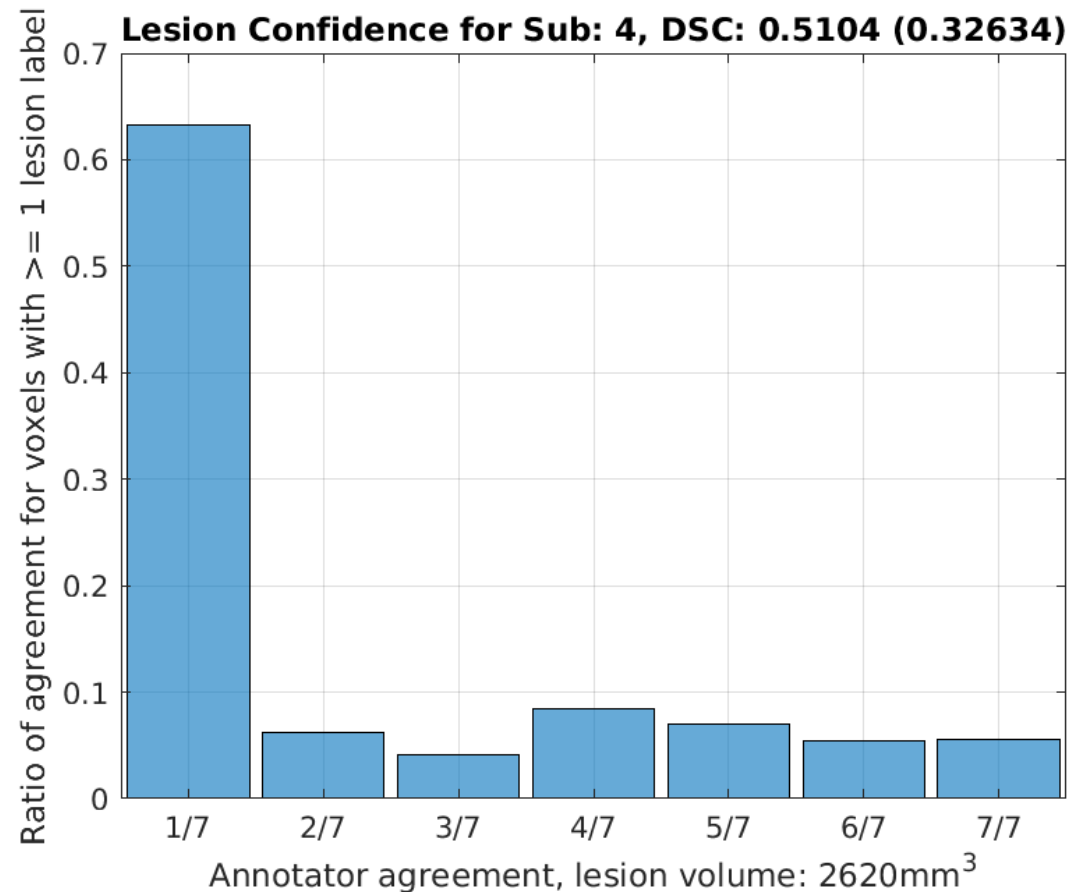
agreements and
disagreements



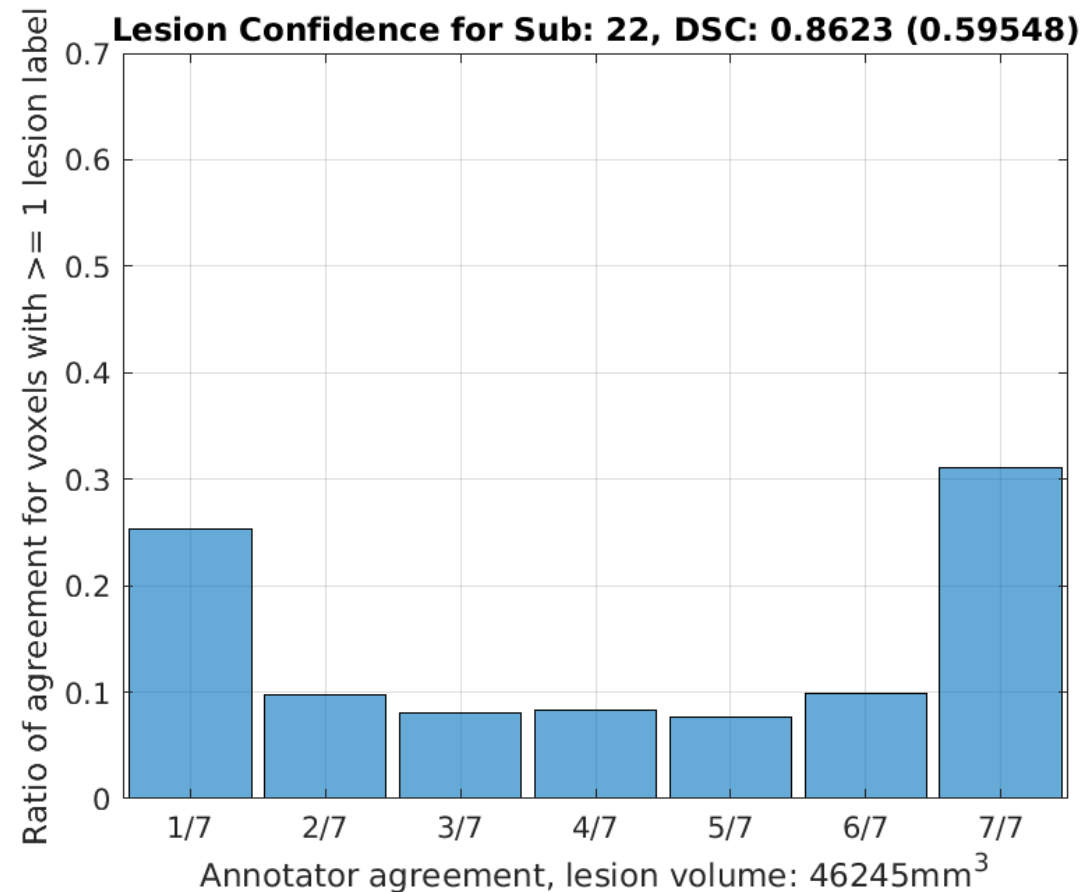
What is the level of agreement among annotators for marking a voxel as lesion?



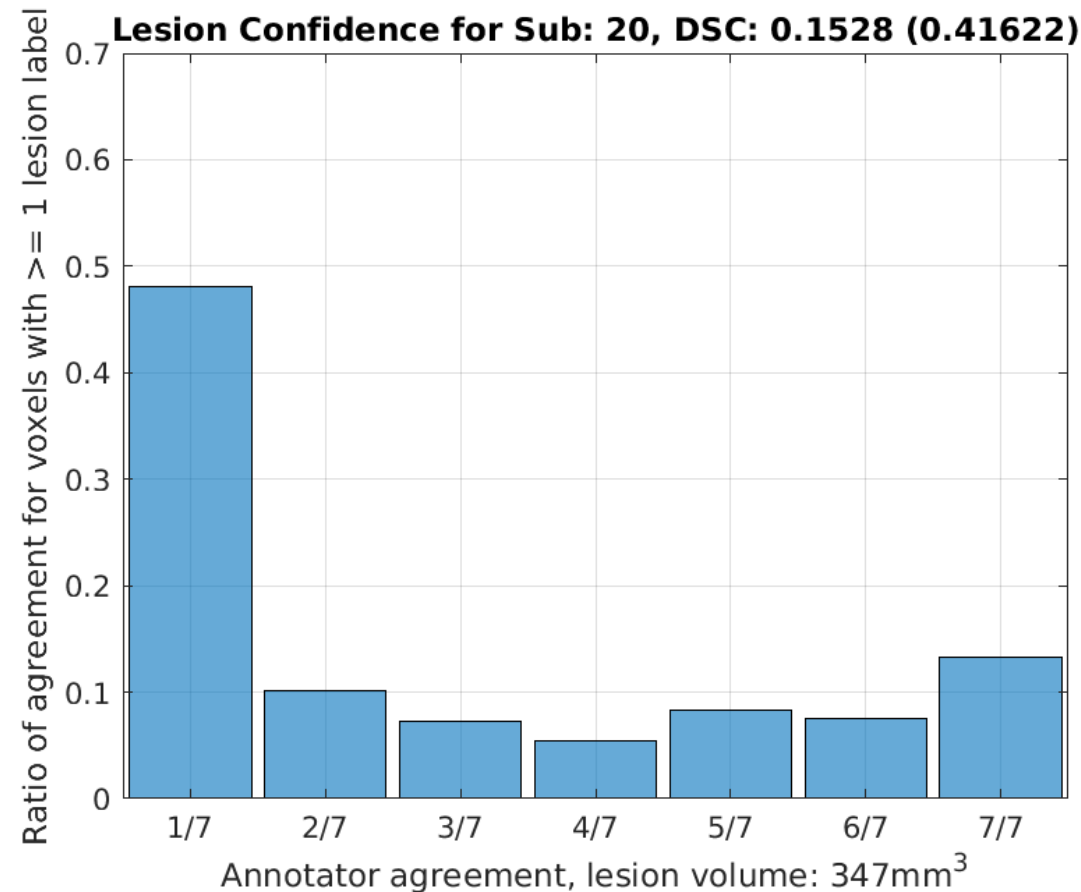
Example of low agreement: smaller lesion volume as well!



Higher agreement: better model DSC as well!



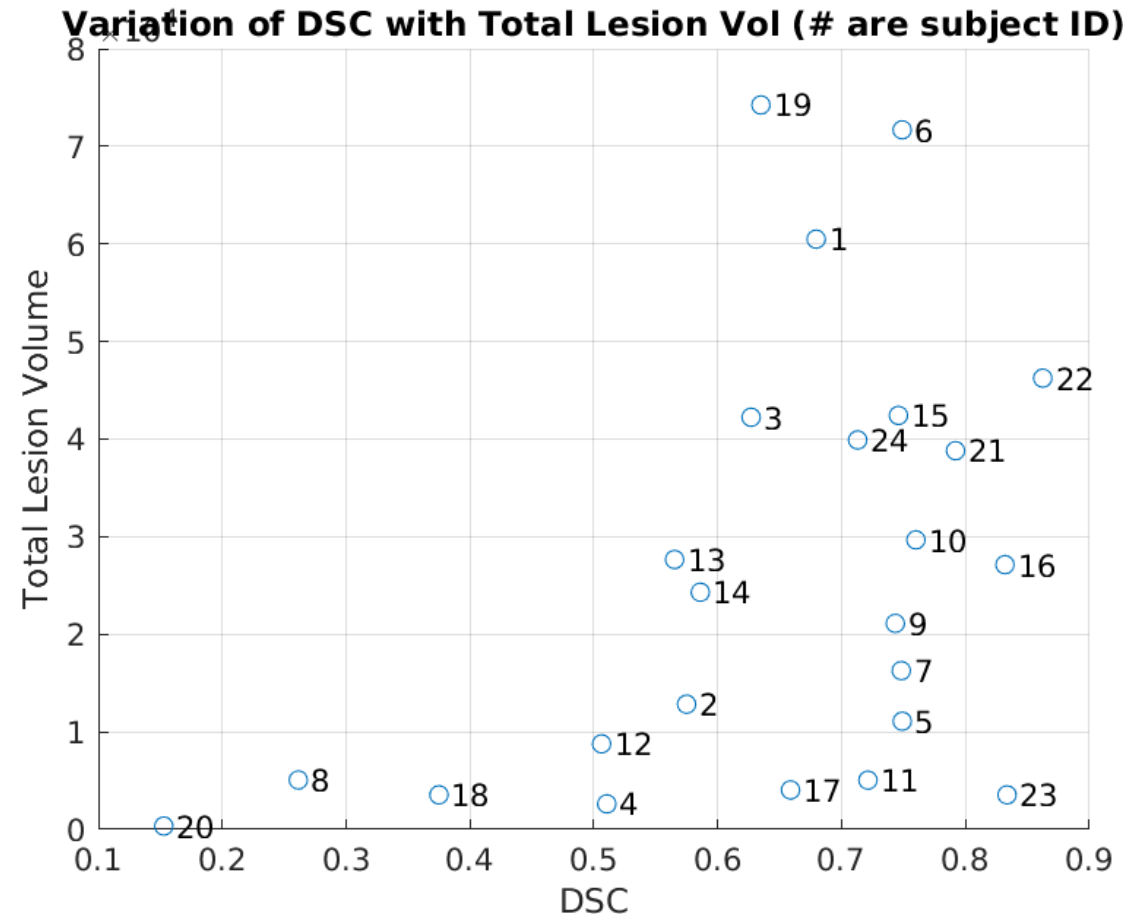
Extremely small total lesion volume – more on this later ... (also worst DSC in the eval set)



Even though this volume is smaller than Subject #4 earlier, there appears to be less agreement (e.g., for example)



How does model DSC relate to total lesion volume?



Subjects on the top right are better performing than bottom left.

Interesting that even for small lesion volumes (say subject 23), the DSC is still high.

Total lesion volume = includes any voxel that any of the 7 annotators have indicated as lesion (not majority vote, not STAPLE; a superset of these).

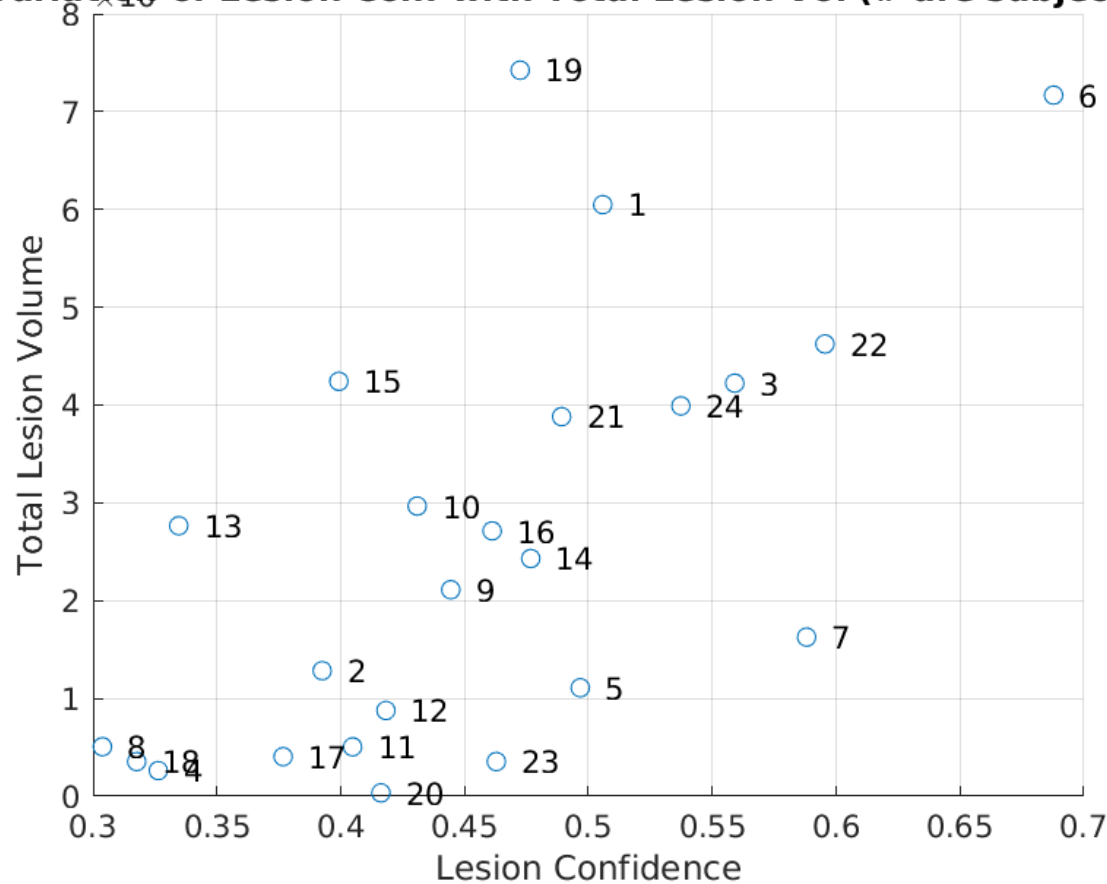
Can we build another metric to capture the annotator variance?

$\text{corrcoef}(\text{DSC}, \text{vol}) = 0.42$



Introducing "Lesion Confidence"

Variation⁴ of Lesion Conf with Total Lesion Vol (# are subject ID)

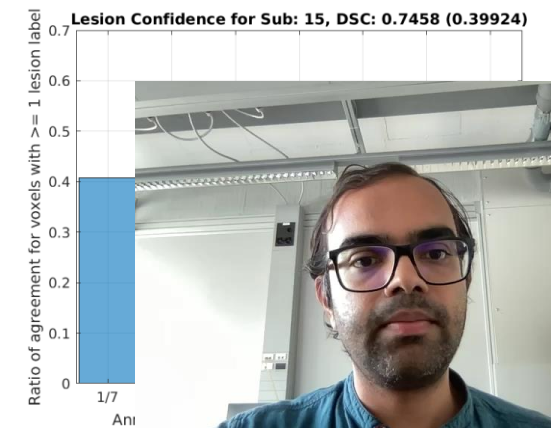
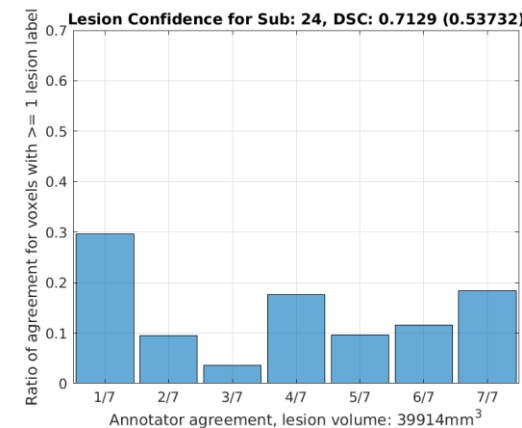


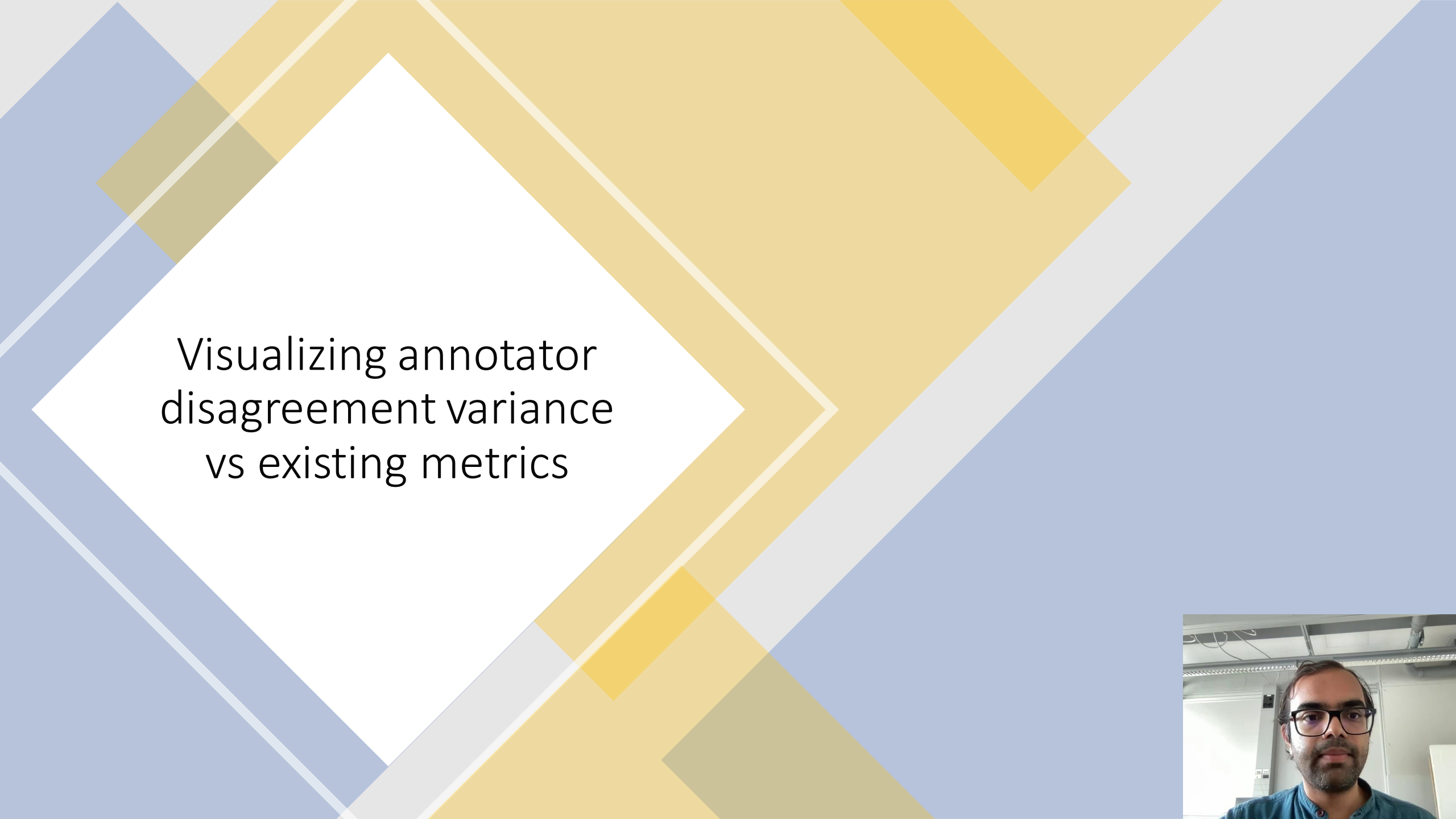
Lesion confidence = (proportion of annotators agreeing that a voxel is a lesion) * (normalized histogram ratio of all the voxels at that level)

Appears to be more linearly related to the total lesion volume (see no bottom right entries)

Separates #24 and #15 better (#15 is has less agreement than #24)

$\text{corrcoef}(\text{lesion_conf}, \text{vol}) = 0.63$





Visualizing annotator disagreement variance vs existing metrics



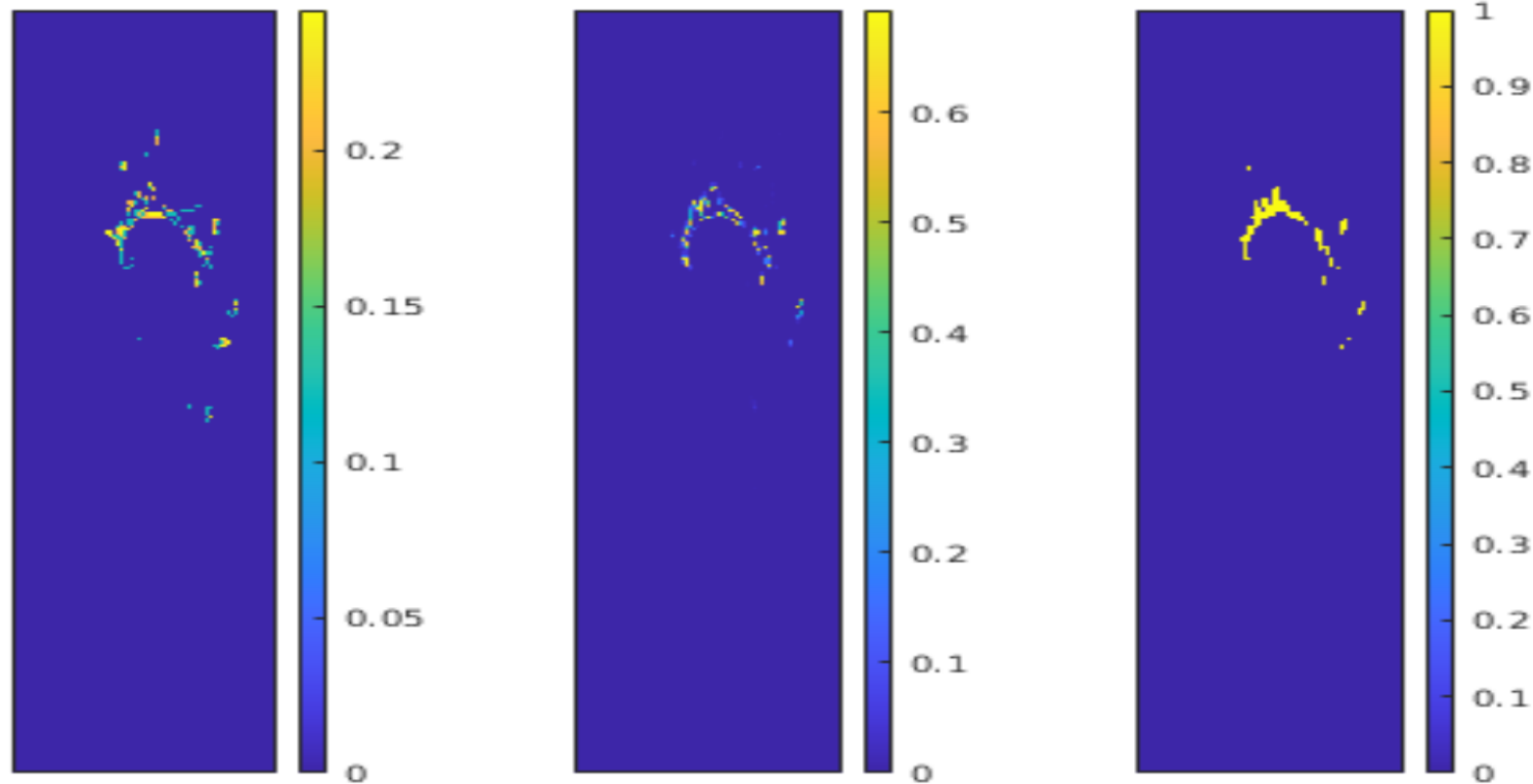
Computing annotator disagreement variance

- From the eval_in set of the msseg data set: average voxel-wise all the annotator masks, and then do a voxel-wise $p*(1-p)$ where p is the value of the voxel.
- This is an unbiased measure of the variance: 0 is good (certainly FG or BG), 0.25 is the worst (most uncertain).
- The provided code generates uncertainty estimates using "confidence", "entropy_of_expected", "epkl", "expected_entropy", "mutual_information", "pred_bin", "pred_prob", "reverse_mutual_information": we compare these with anno variance.



Visual verification of location correspondences between disagreement, EoE metric, and maj. Vote GT used for training

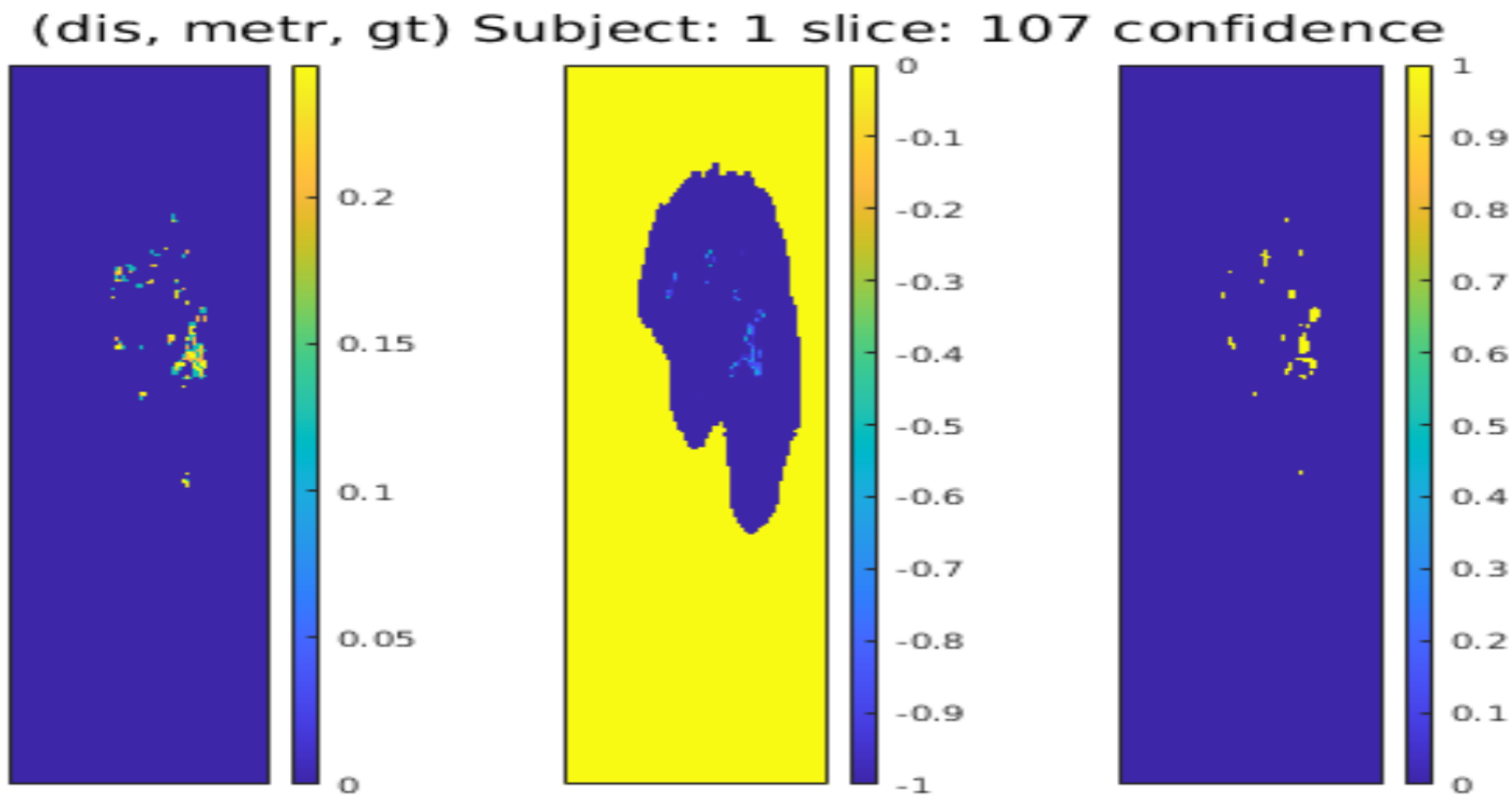
(dis, metr, gt) Subject: 1 slice: 64 entropy_of_expected



Full video here: https://unibe365-my.sharepoint.com/:v:/r/personal/amith_kamath_unibe_ch/Documents/MICCAI/5_entropy_of_expected_disagreement_annotation.avi?csf=1&web=1&e=dVCxpn



However, scales for metrics vary!



Challenges with Quantitative measurements

- Scales of each uncertainty metric are different (confidence: -1 to 0; EoE: 0 to 1; ...): direct numeric correlation may not be accurate.
- Disagreement also between 0 and 0.25: need to normalize to compare with metrics too.
- Hence, rescale these metrics and disagreement volumes to $[0, 1]$ and then do a simple MAE computation to measure which metric is closest to the disagreement, and, if it is consistent across subjects.



Quantitative comparisons between rescaled disagreements and metrics

Computed as MSE between volumes

| Confidence | EoE | EPKL | Exp. Entropy | Mutual Info | Reverse Mutual Info |
|------------|-----------|----------|--------------|-------------|---------------------|
| 0.8697 | 0.0009268 | 0.001037 | 0.0008931 | 0.0009995 | 0.001048 |

Computed as MAE between volumes

| Confidence | EoE | EPKL | Exp. Entropy | Mutual Info | Reverse Mutual Info |
|------------|----------|----------|--------------|-------------|---------------------|
| 0.8701 | 0.001321 | 0.001386 | 0.001303 | 0.001363 | 0.001392 |

Admittedly without caring much for what each of these mean, it appears the Expected Entropy is closest on average using both MSE and MAE as compared to our estimate of the annotator disagreement. Need to investigate further (not over a weekend ;-)) if this really makes sense!



Does lesion size matter for annotator disagreement?

Quick and dirty
analysis

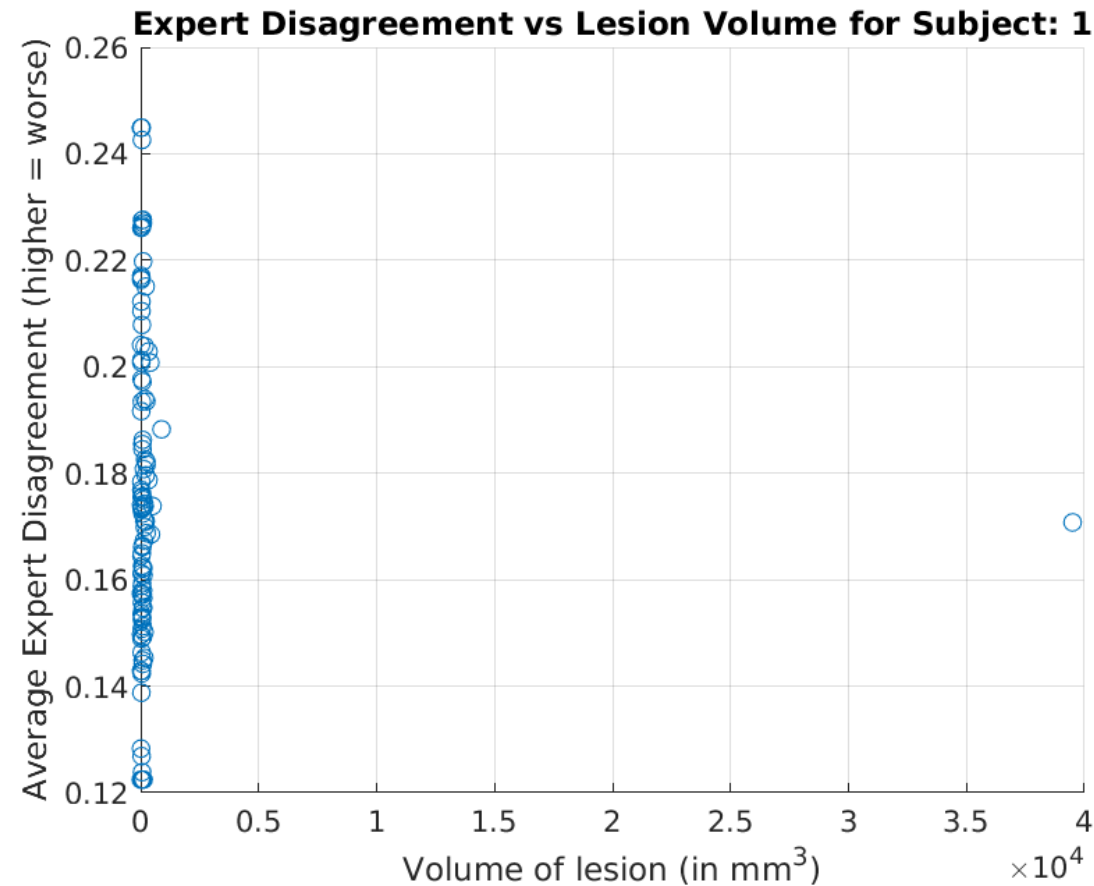


How is this computed?

- Using the disagreement volume earlier: we do a connected component separation, thresholded for volumes > 9 voxels.
- For each of these volumes, we compute average disagreement values, and plot this average versus the size of these volumes (in mm^3 , since isotropic).
- The expectation is that smaller volumes lead to more disagreement, larger volumes are easier to segment, hence less disagreement.



Examples of results: subject #1

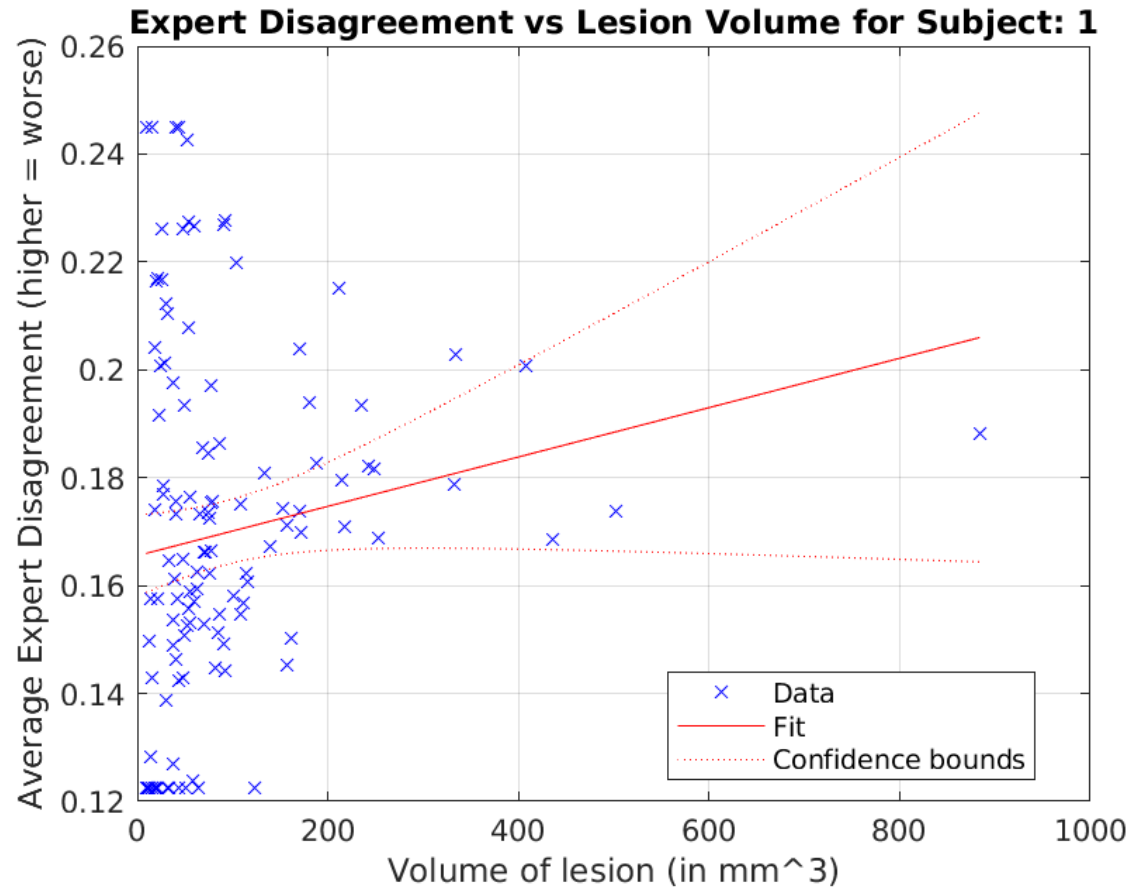


There's one dominant lesion that is $> 3.5 \times 10^4$ mm³ in size.

Hence, we filter out to only include the smallest 90% of the lesions so as to see the trends.



Examples of results: subject #1



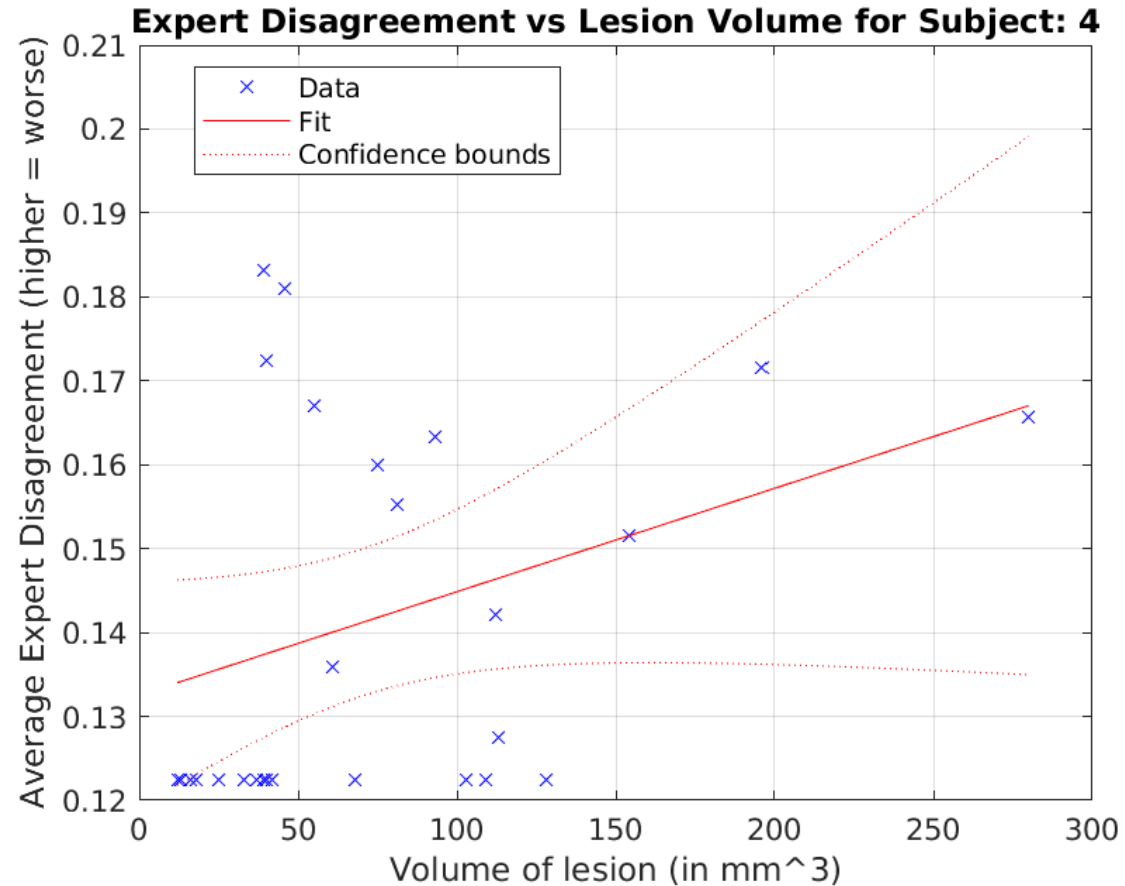
Our expectation of a negative trend is not true, at least for this subject.

There doesn't appear really to be any relation between lesion size and the disagreement within that region.

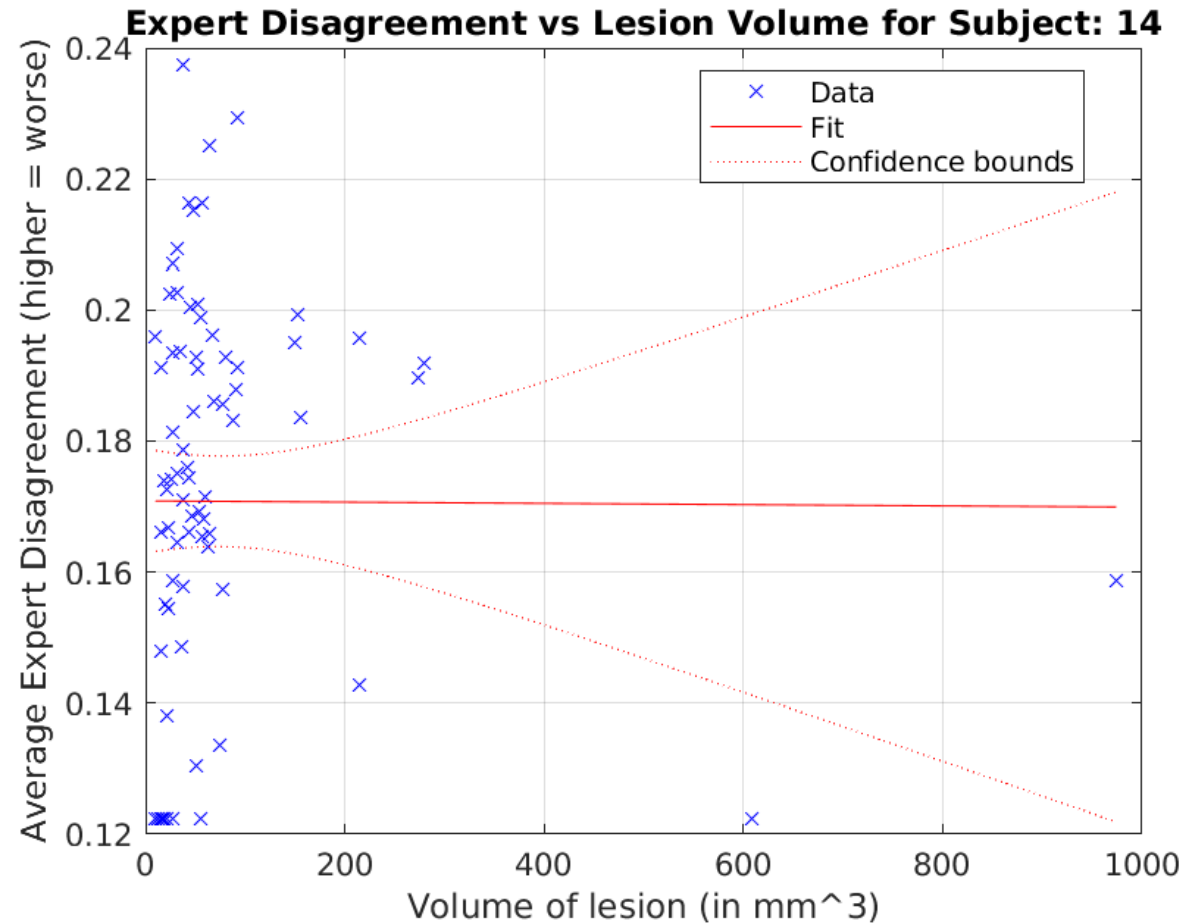
How do other subjects look?



Examples of results: subject #4



Examples of results: subject #14



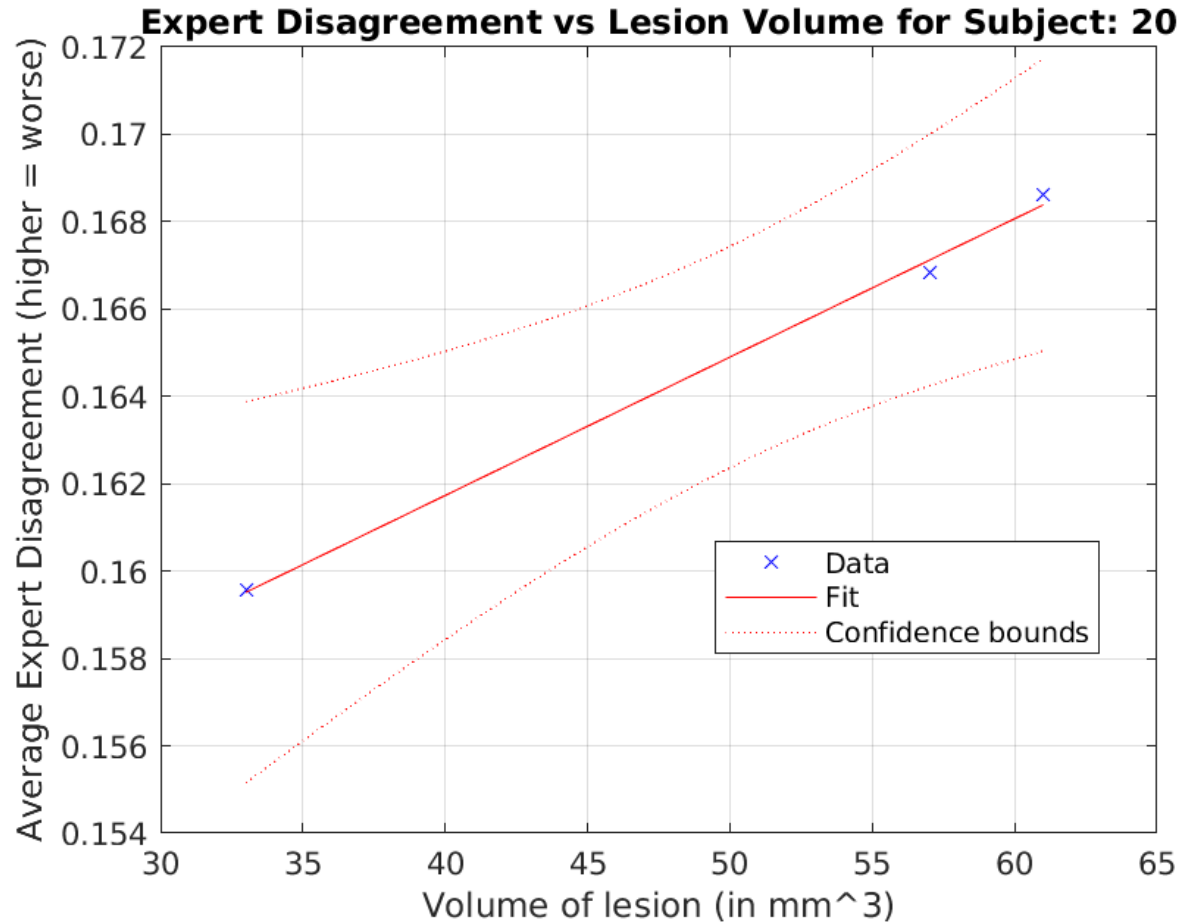
Here's a counter example: where the slope of the regression line is nearly 0.

This could be due to multiple factors:

1. Our estimate of disagreement may not be the best; there could be other ways to evaluate this, and could then really correlate with volume.
2. The large lesions could possibly bias this analysis and we could filter even more aggressively to see trends at smaller volumes.



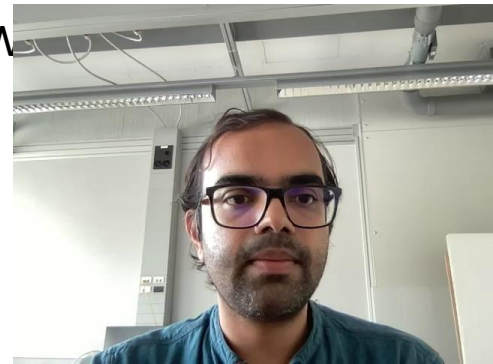
Examples of results: subject #20 (extremely small region)



Here's a tricky case: there are only three lesion regions (independent connected components), and the entire volume is also super small (as seen earlier).

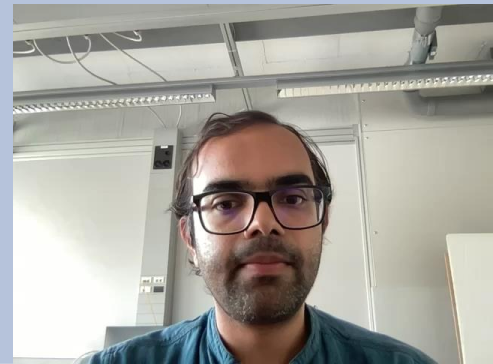
This points to potential errors in how the ground truth is computed for this subject, possibly it being a super Out-of-Distribution situation for the model, hence also the DSC of prediction being so low (0.15).

Exploratory data analysis FTW



Potential next steps

After the Hackathon



More analysis

- Better measures of disagreements between annotators: is there a way to be even more accurate than simply average? Include annotator-level scoring/weighting?
- Consider the effect of Leave-one-out annotator and finding out how it could change the estimates?
- Are there other ways to measure model uncertainty that more closely matches data uncertainty?
- Extrapolate to datasets from other centers: how wide is the gap between intra-center variations and inter-center variations?



Acknowledgements



Many thanks to:

- The MICCAI Hackathon organizers: for the platform and fun discord messages.
- Task 1 mentors, Mara, Andrey, Nataliia and Vatsal: for the starter code, introductory presentation and timely and async thoughts through the discord channel.
- The shifts project: for sharing the data to run this analysis on.

