The MS-Multi-Spine Challenge

Evaluation Results, Discussion, and Conclusions

https://portal.fli-iam.irisa.fr/MS-Multi-Spine/

OFSEP
Observatoire Français de la Sclérose en Plaques

EMPENN

FLI
France Life Imaging

CREATIS

Outline

1. Evaluation process, training data and testing data

1.1 The annotation process
1.2 Overall characteristics of the ground-truth
1.3 Detailed characteristics of the ground-truth
1.4 Experts performances

2. Results

2.1 Results on the different sequences combinations
2.2 Results for overall test set
2.3 Inter-methods variability and Results example

3. More Results

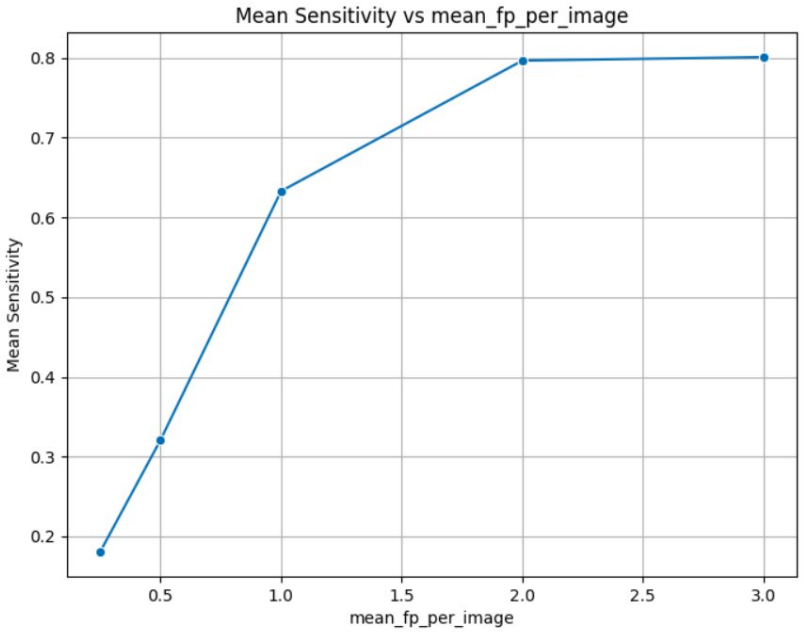
3.1 Dealing with the three-sequences combination
3.2 Added value of multi-sequence over T2 alone: some insights
3.3 Performances and dependency to IoU thresholds

4. Discussion and conclusions

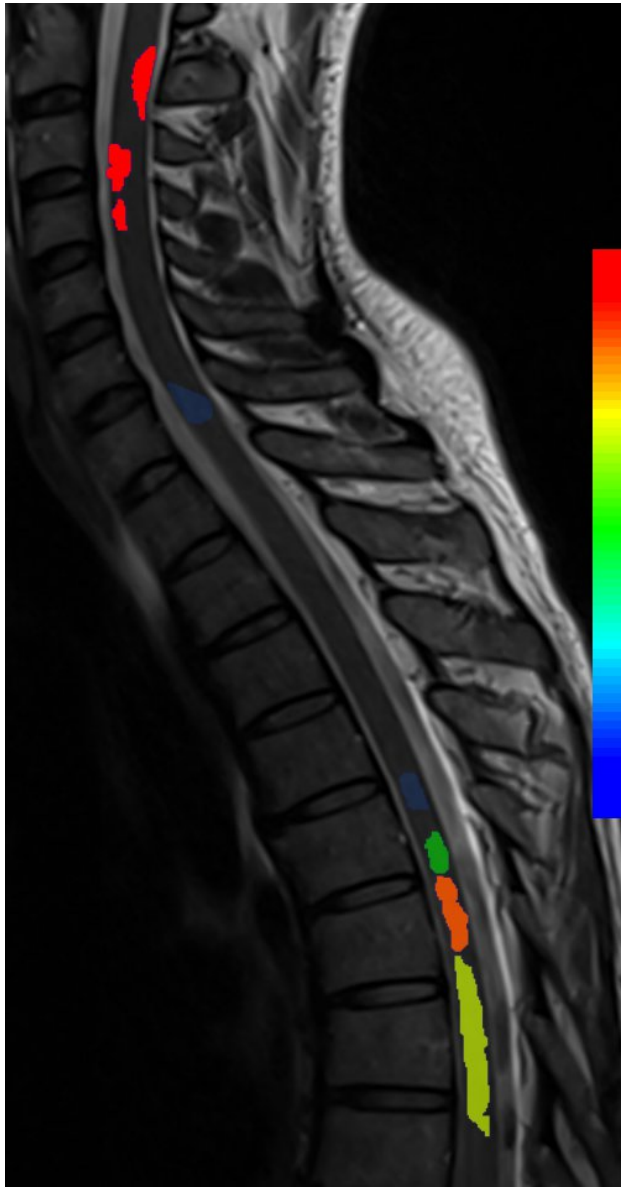
1. Evaluation process, training data and testing data

A calibrated detection-based evaluation

Participating methods must
detect lesions
— AND —
associate a probability to each detected lesion

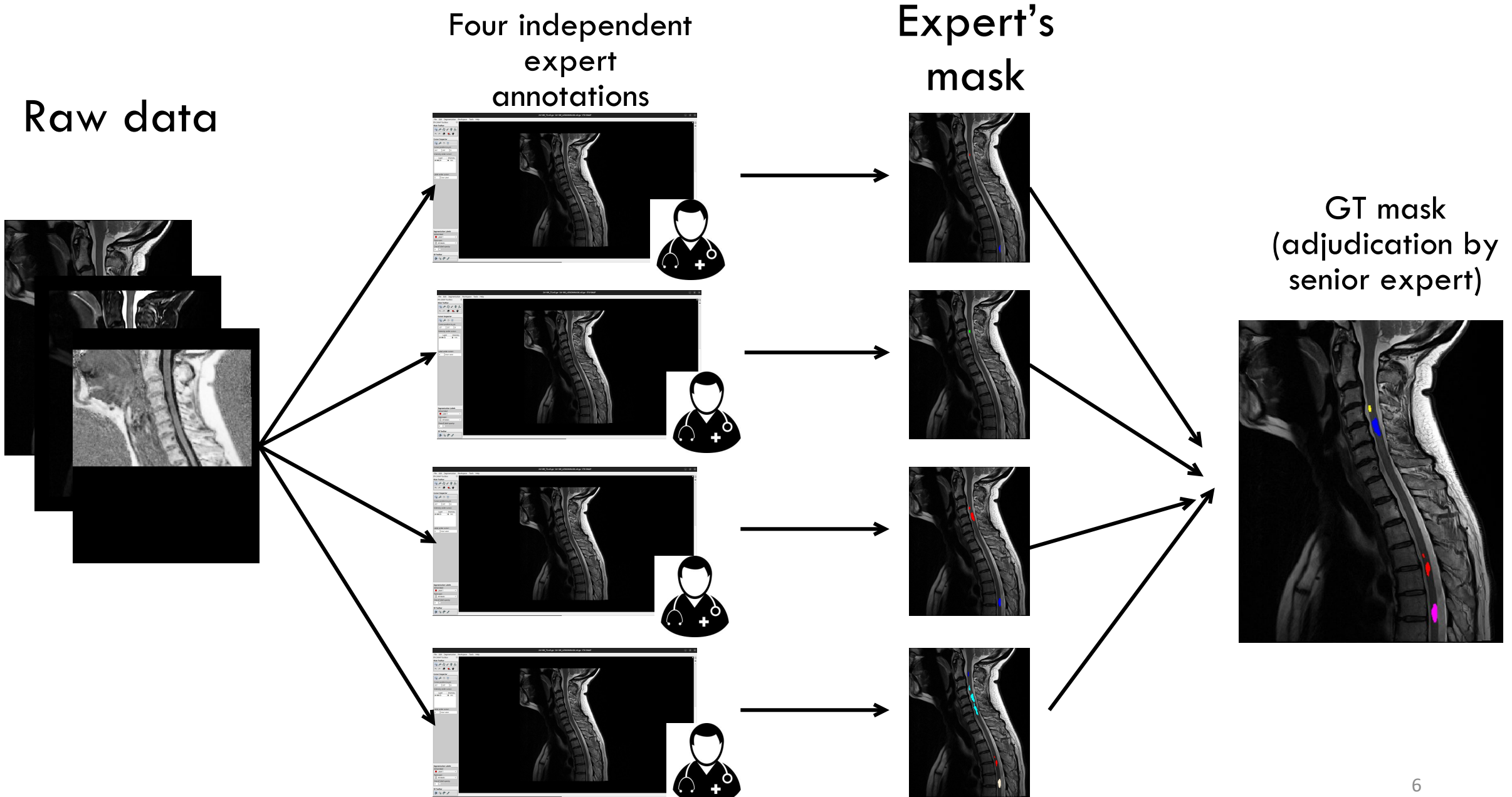


The challenge metric is the
mean sensitivity averaged among the five
false positive rates 0.25, 0.5, 1, 2 and 3.



1.1 The Annotation Process

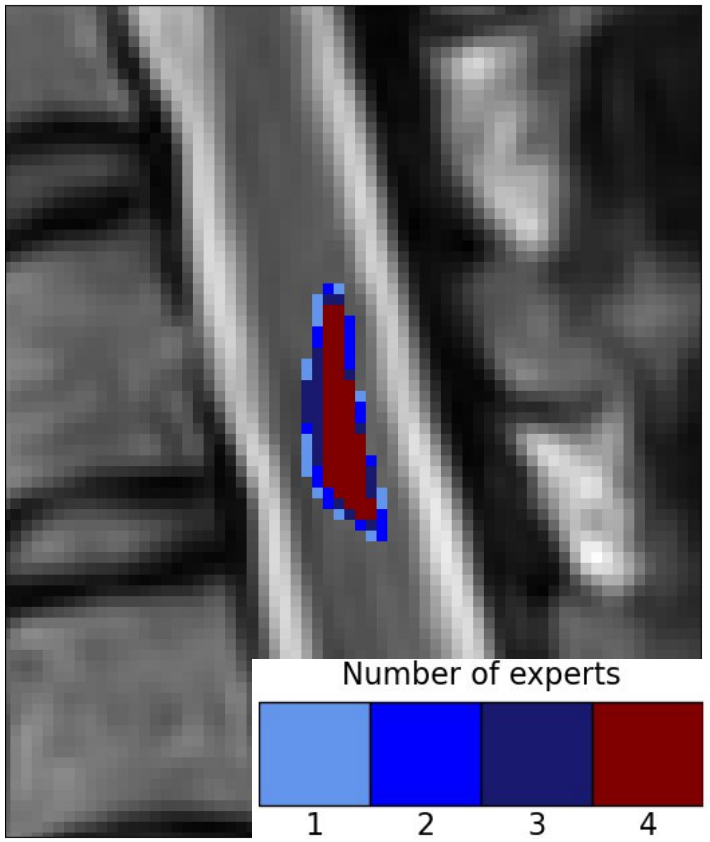
The annotation process



Adjudication: Lesion kept

T2-w

Voxel-wise agreement between experts



Majority vote and adjudication

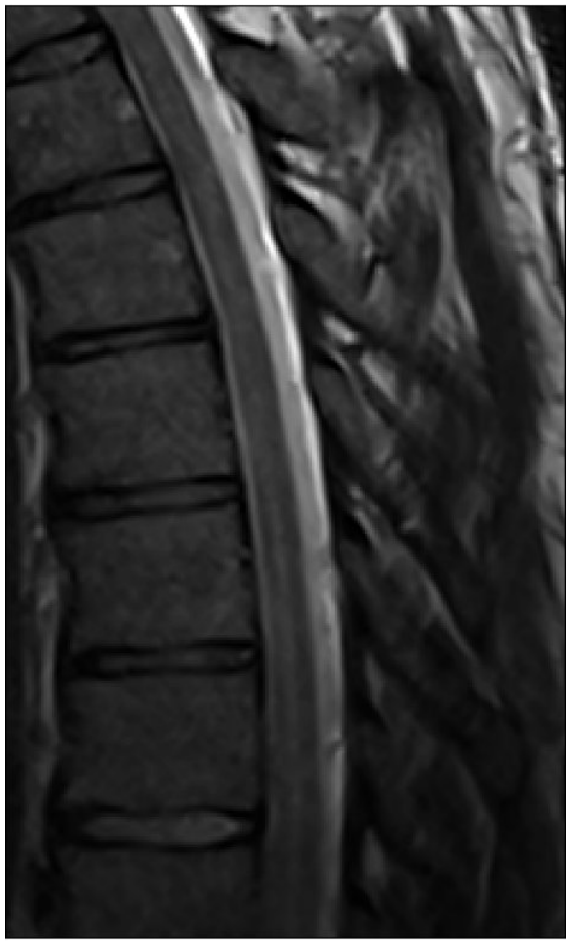
Adjudication: Lesions removed

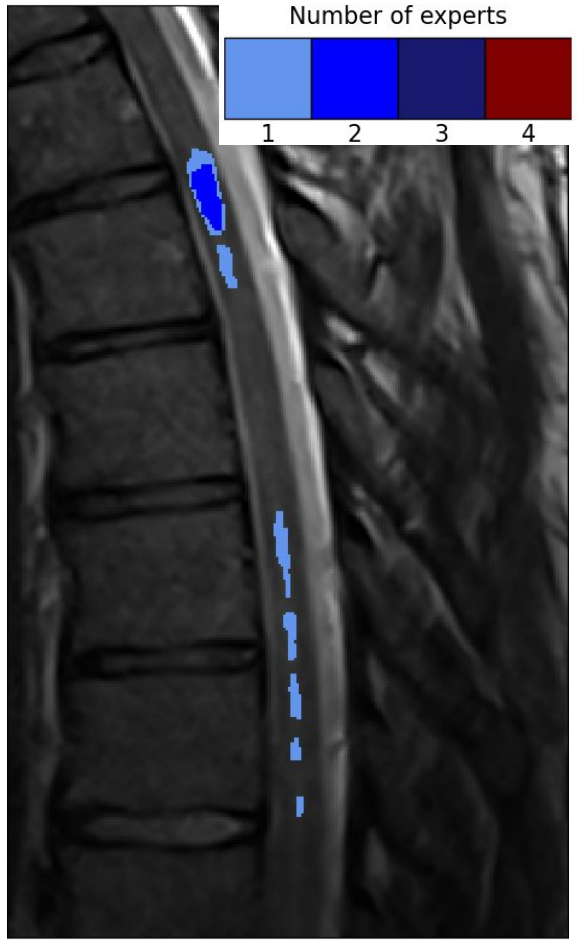
T2-w

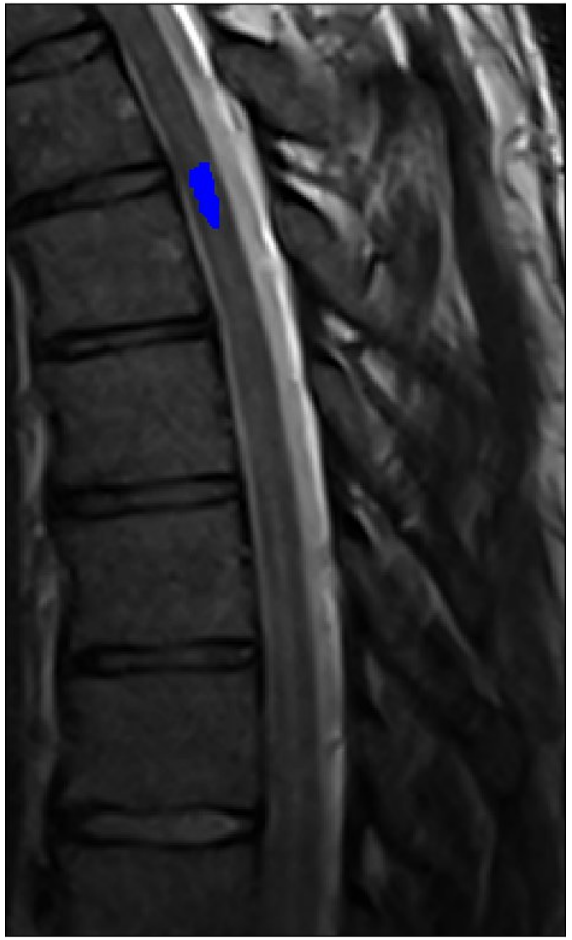
Voxel-wise agreement between experts

Majority vote

Adjudication

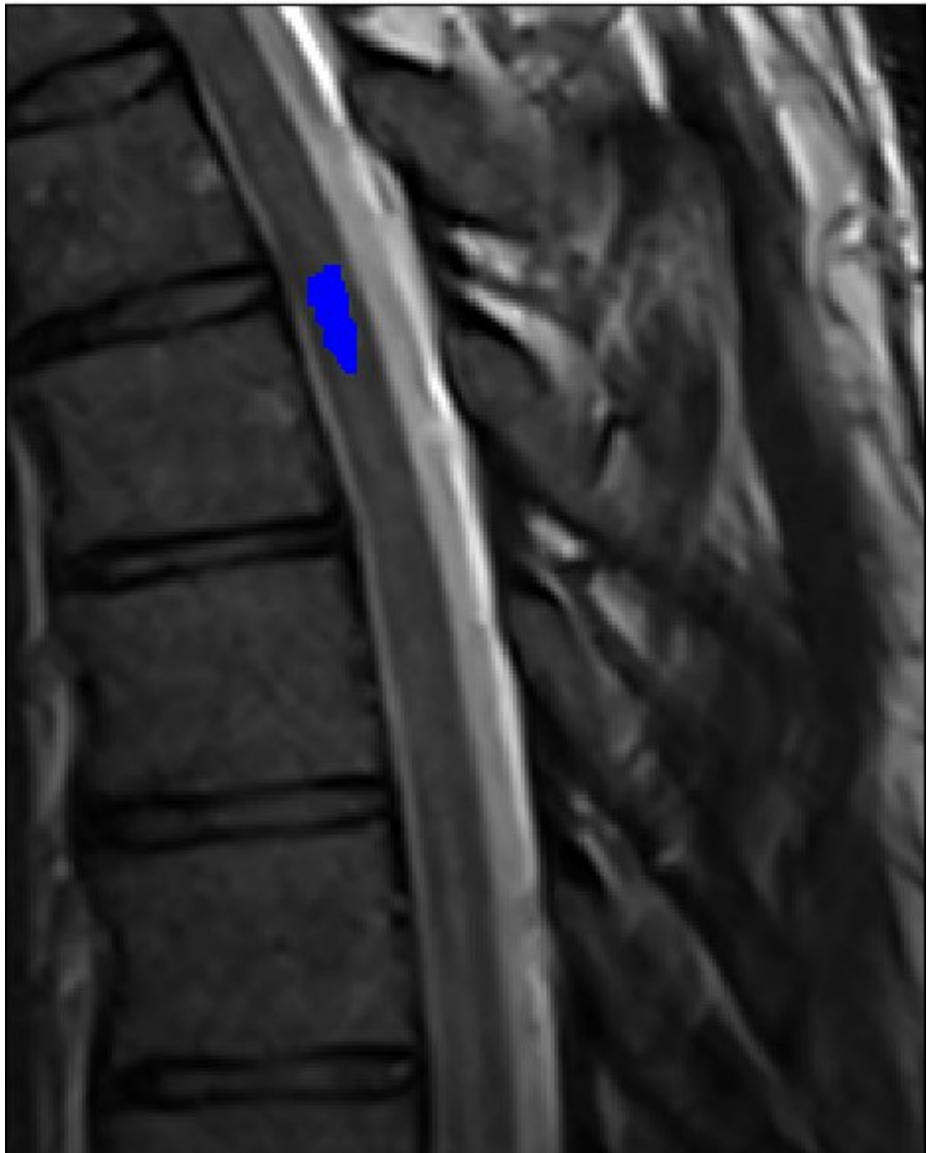




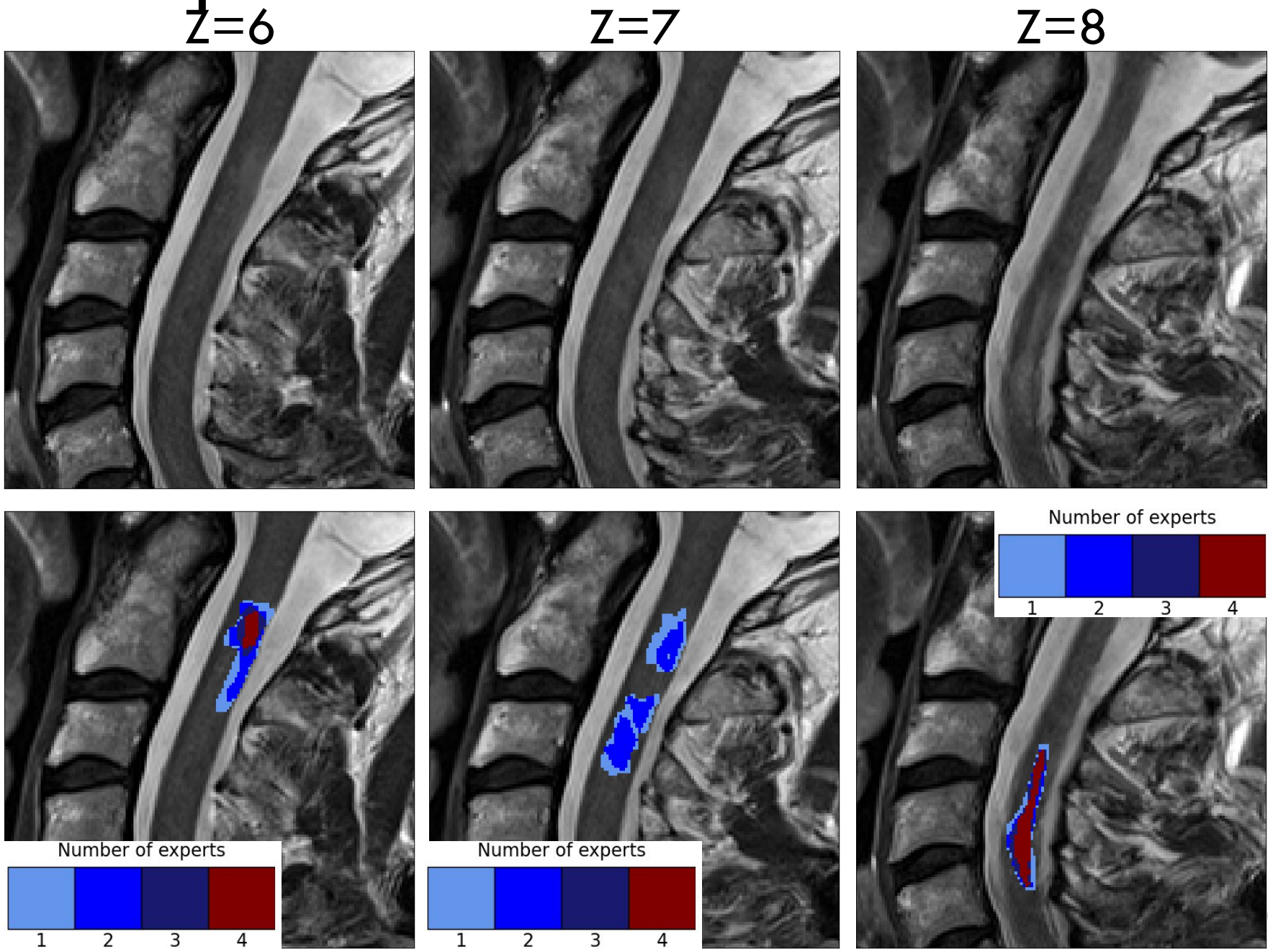


Adjudication: Lesions removed

Adjudication



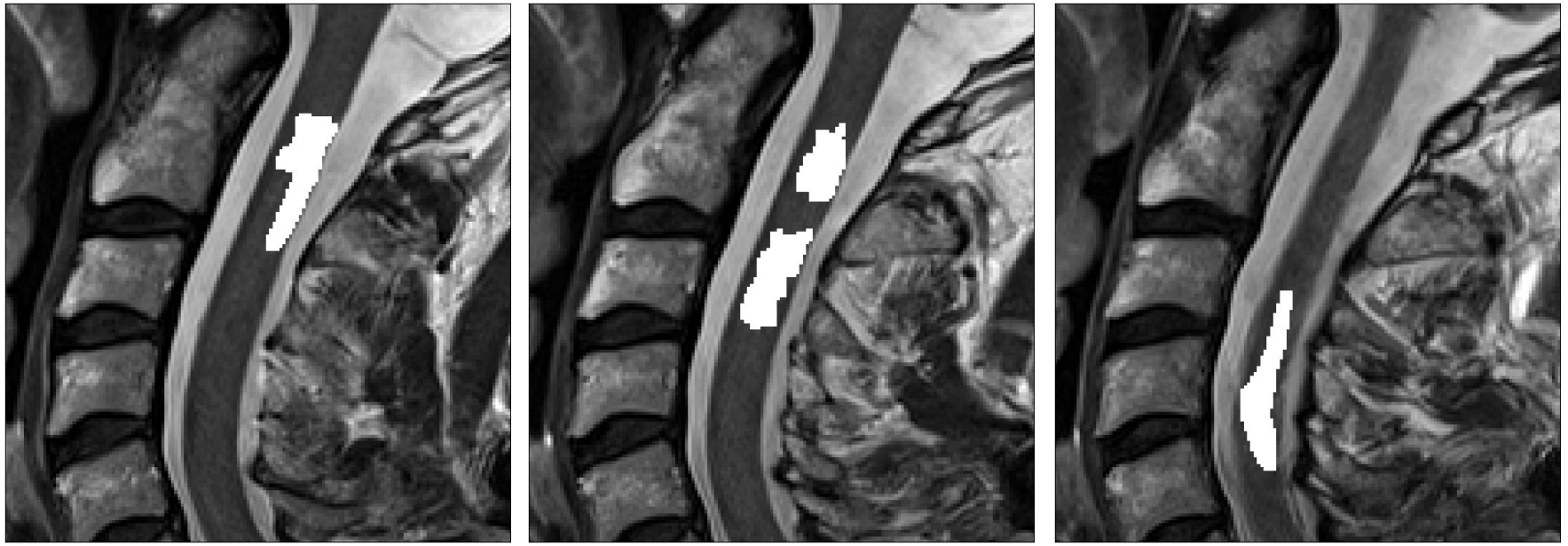
Adjudication: Complex cases



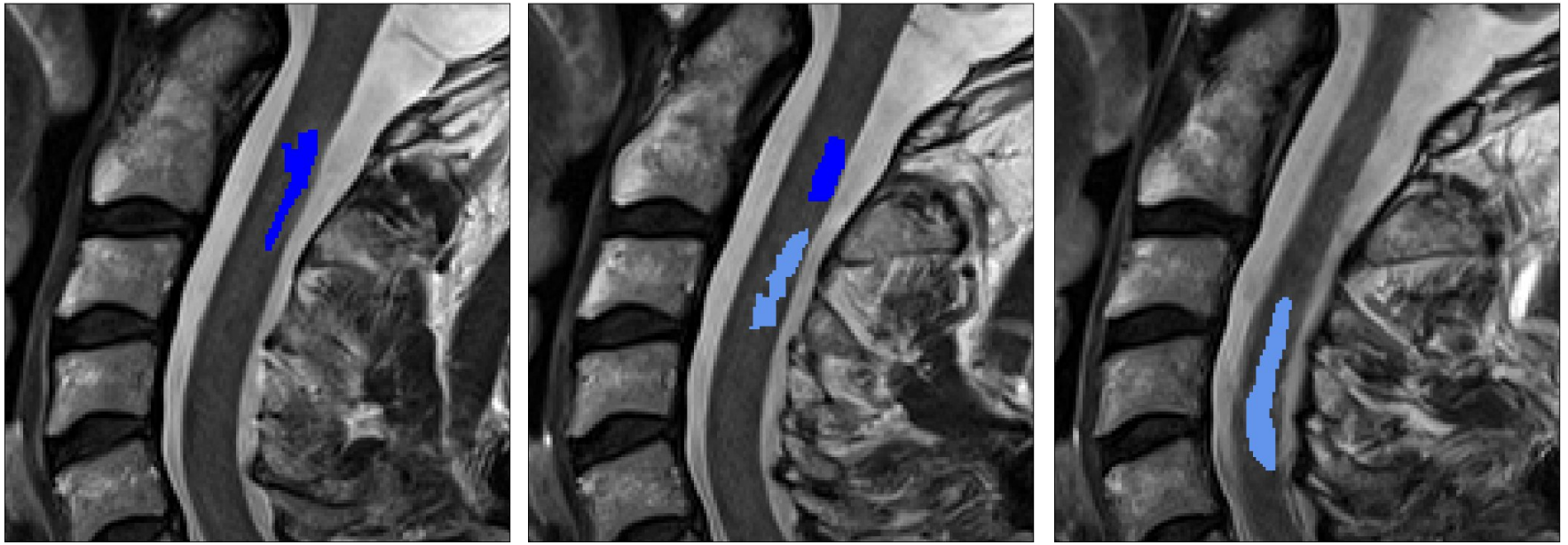
Adjudication: Complex cases

Z=6 Z=7 Z=8

Majority vote (Lesions marked as problematic)



Adjudication

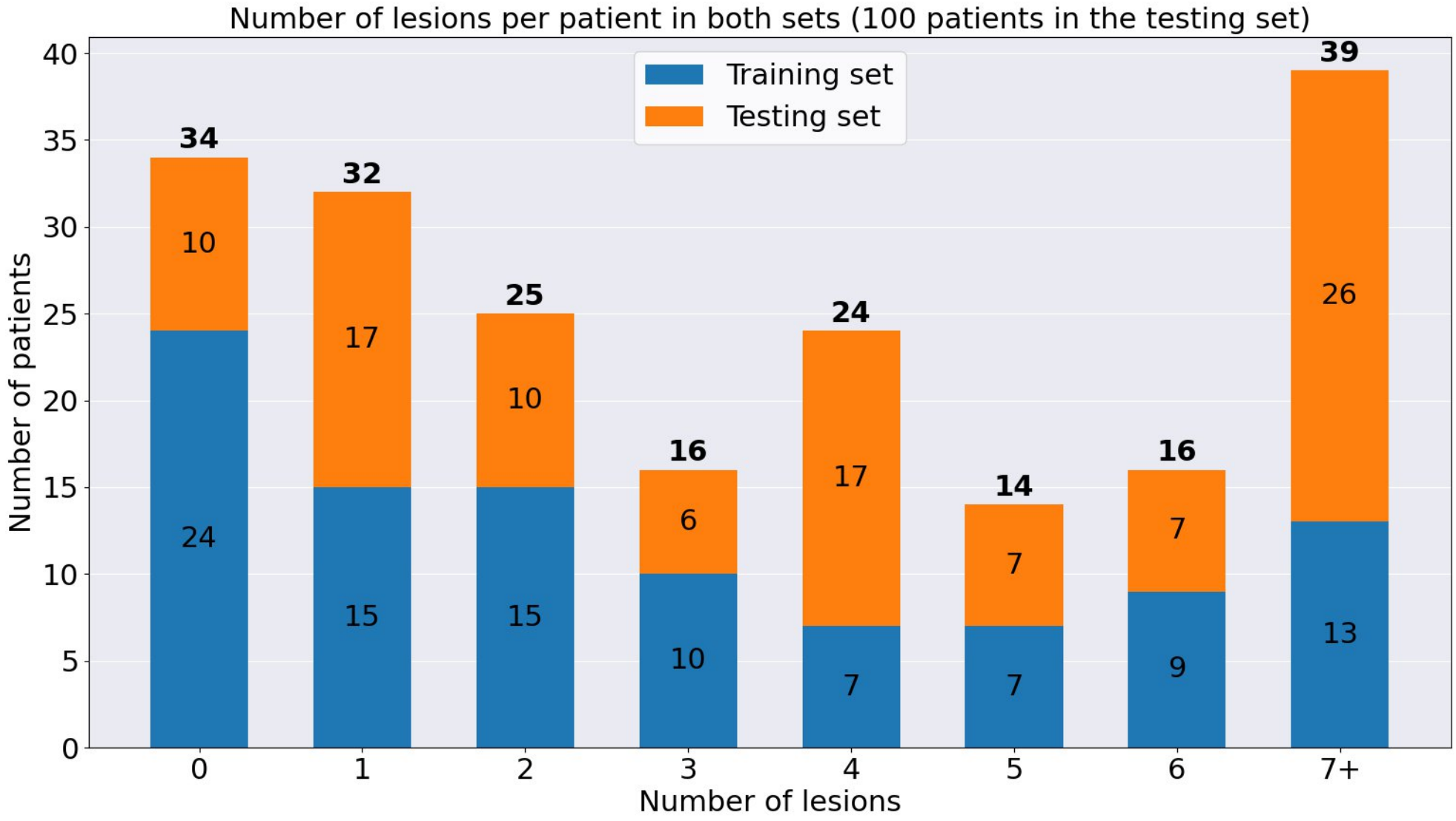


1.2 Characteristics of the ground-truth

Number of lesions in both sets

Training set:
Mean lesion Nb=
3.01±2.89

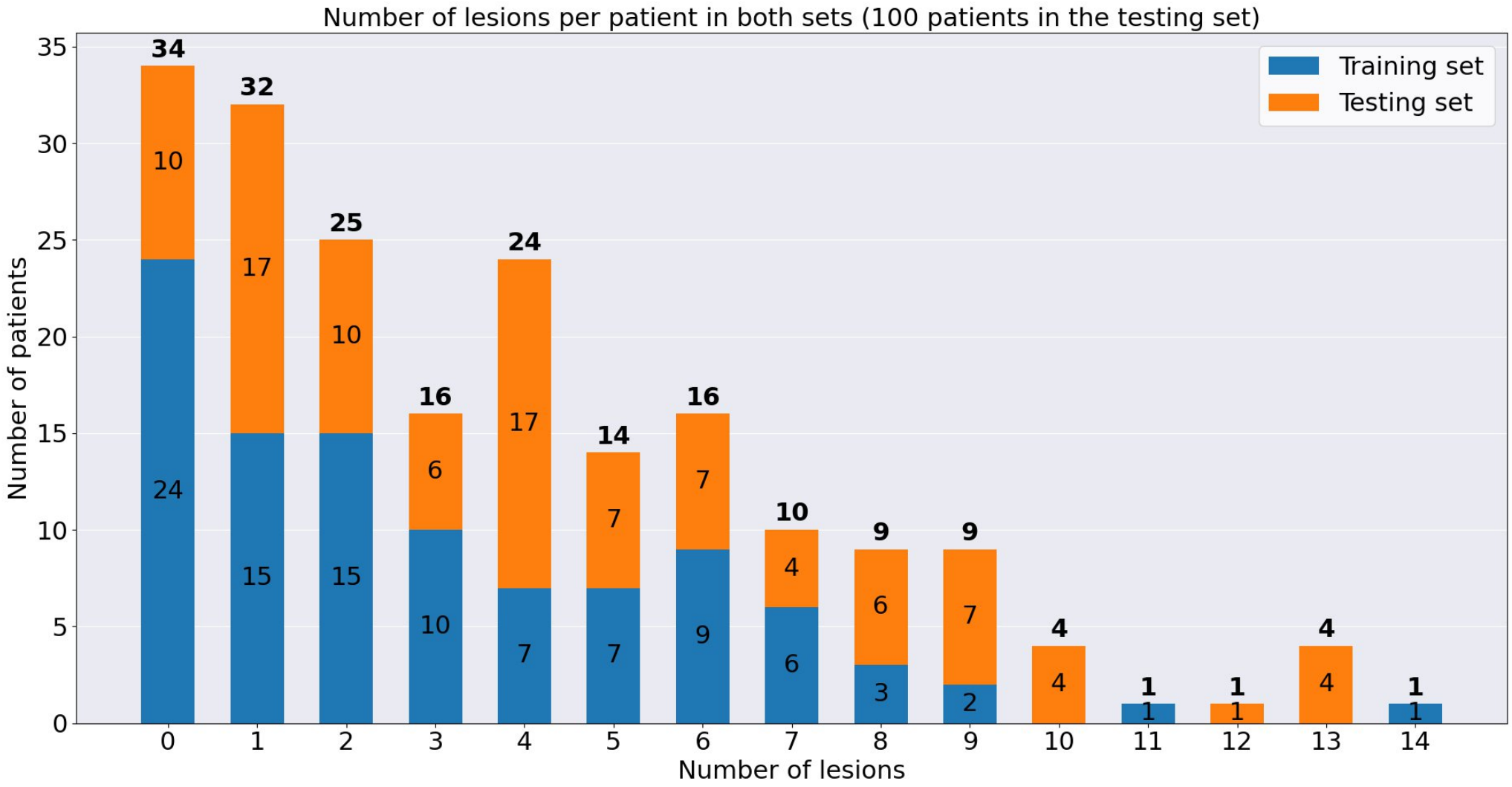
Testing set:
Mean lesion Nb=
4.43±3.48



Number of lesions in both sets

Training set:
Mean lesion Nb=
3.01±2.89

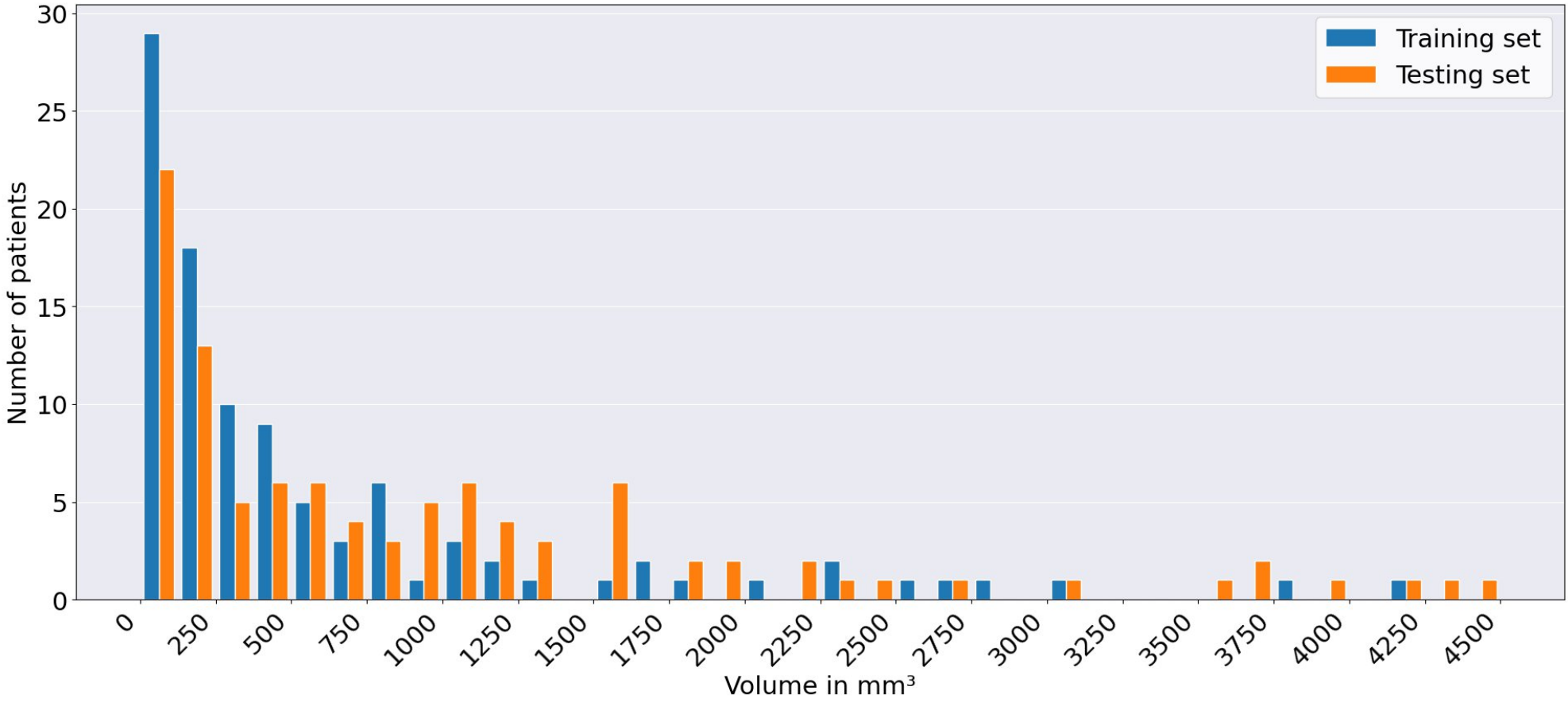
Testing set:
Mean lesion Nb=
4.43±3.48



Volume of lesions per patient (in mm^3) in both sets

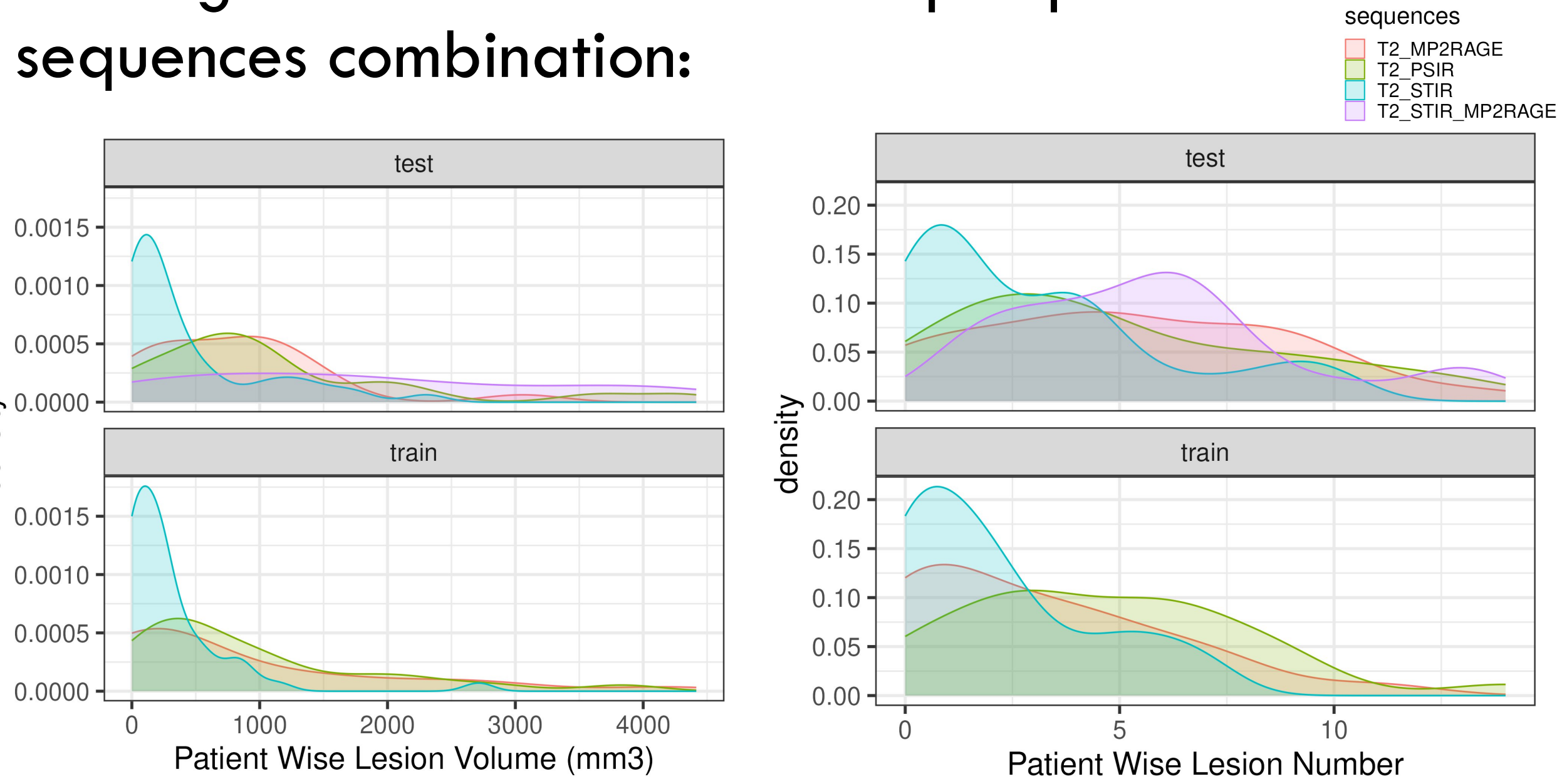
Training set:
Mean lesion Volume = $602 \pm 847 mm^3$

Testing set:
Mean lesion Volume = $951 \pm 1076 mm^3$



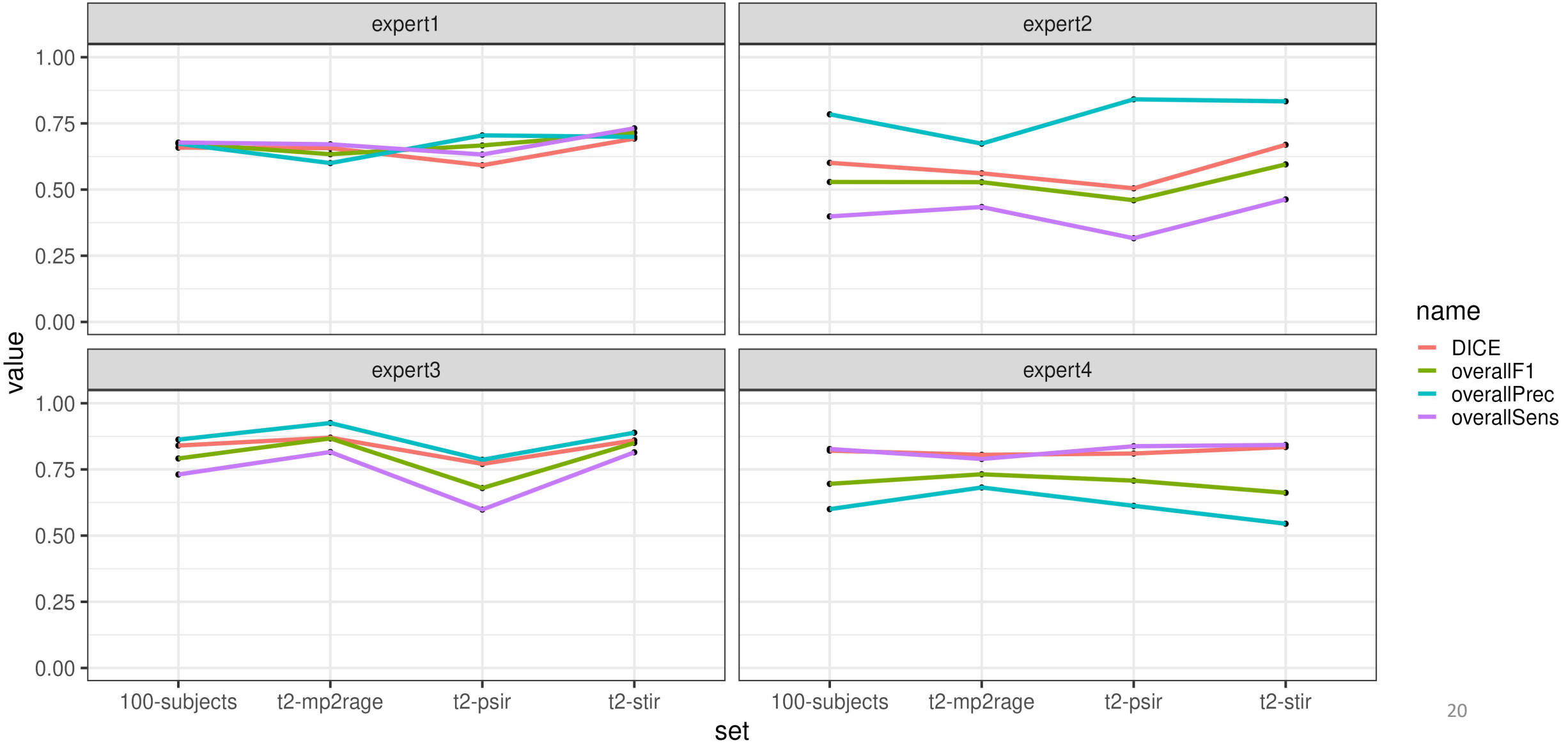
1.3 Characteristics of the ground-truth for each sequences combination

Average lesions characteristics per patient for each sequences combination:

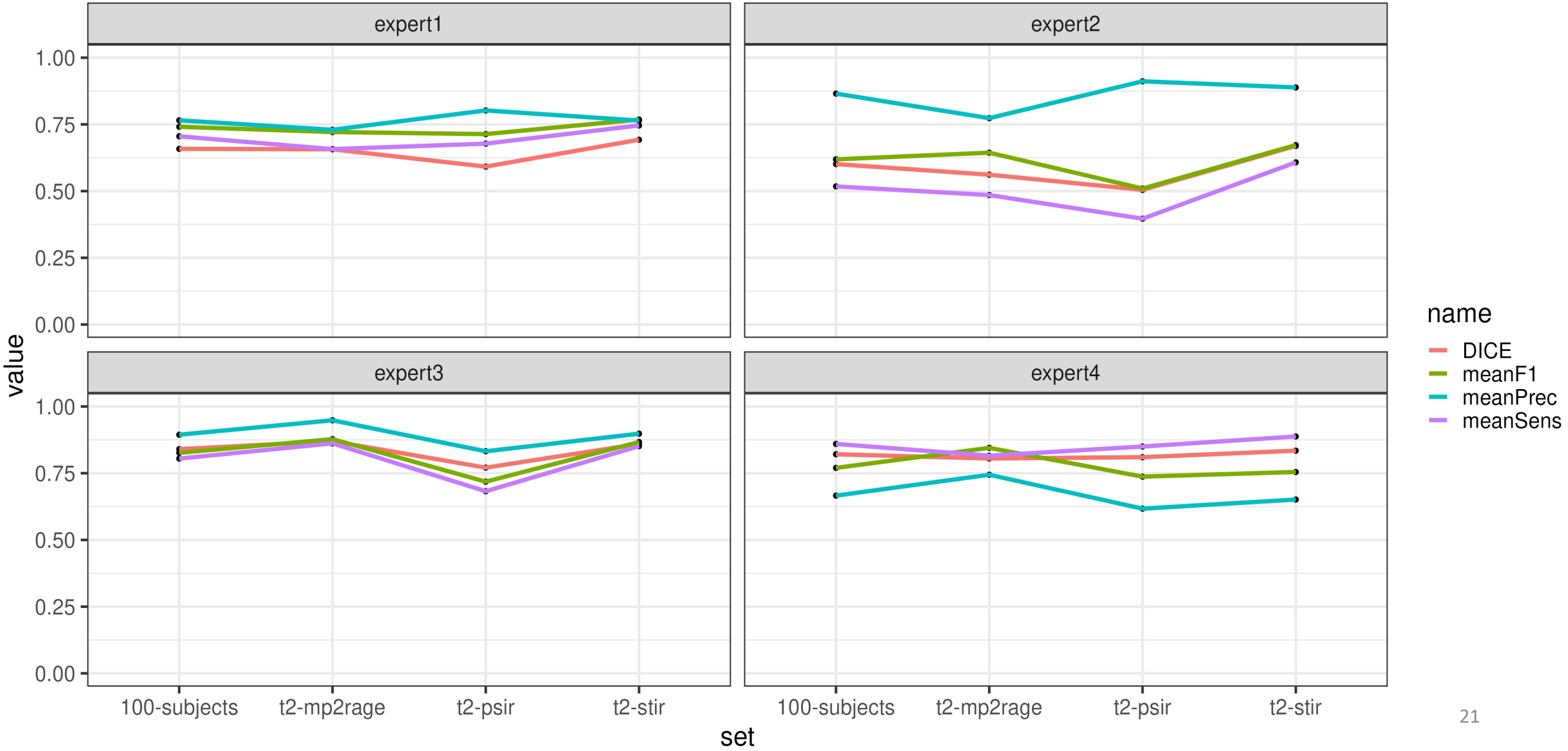


1.4 Performances of experts

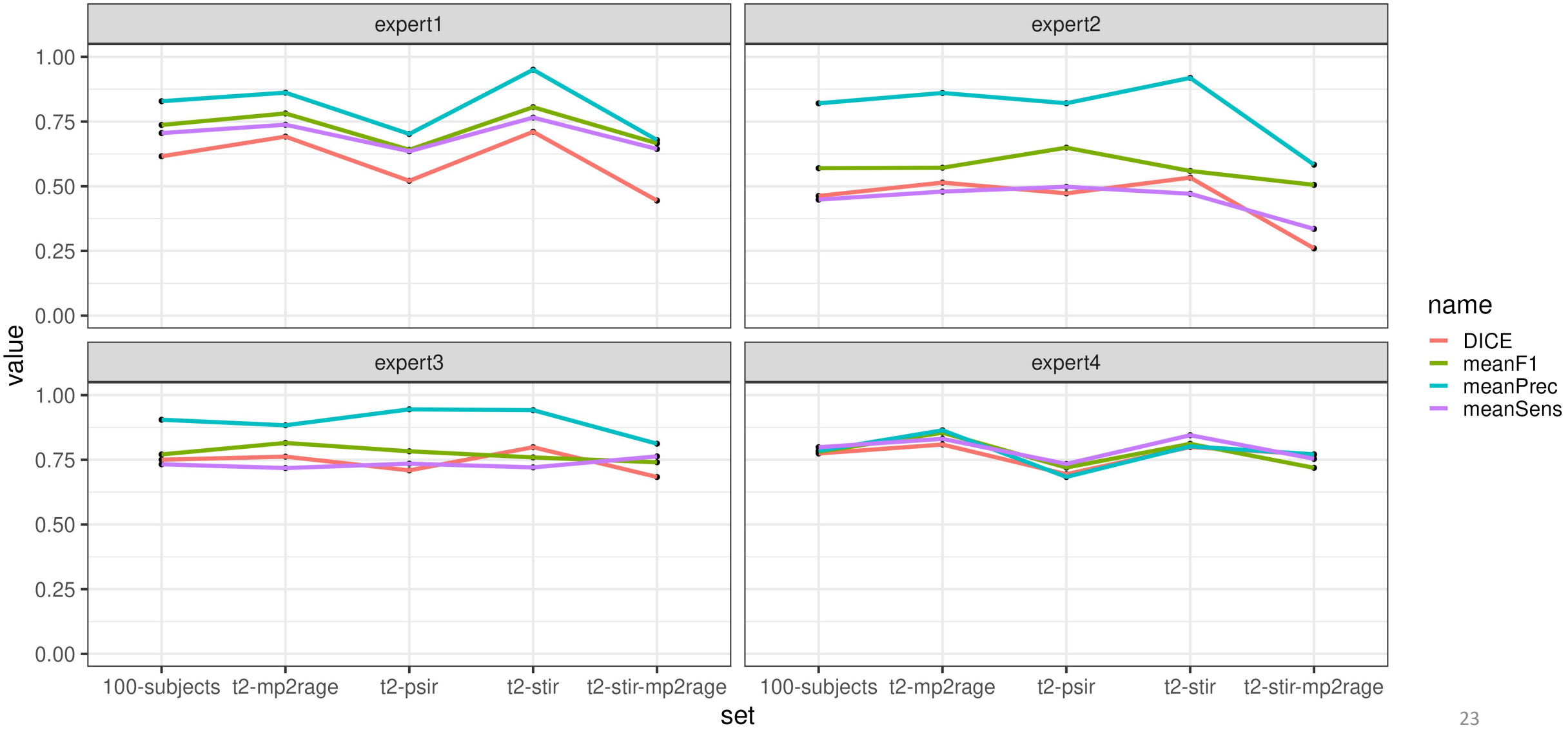
Performances of experts against final Ground Truth (with IoU threshold 0.2) on the train set



Performances of experts against final Ground Truth (with IoU threshold 0.2) on the train set



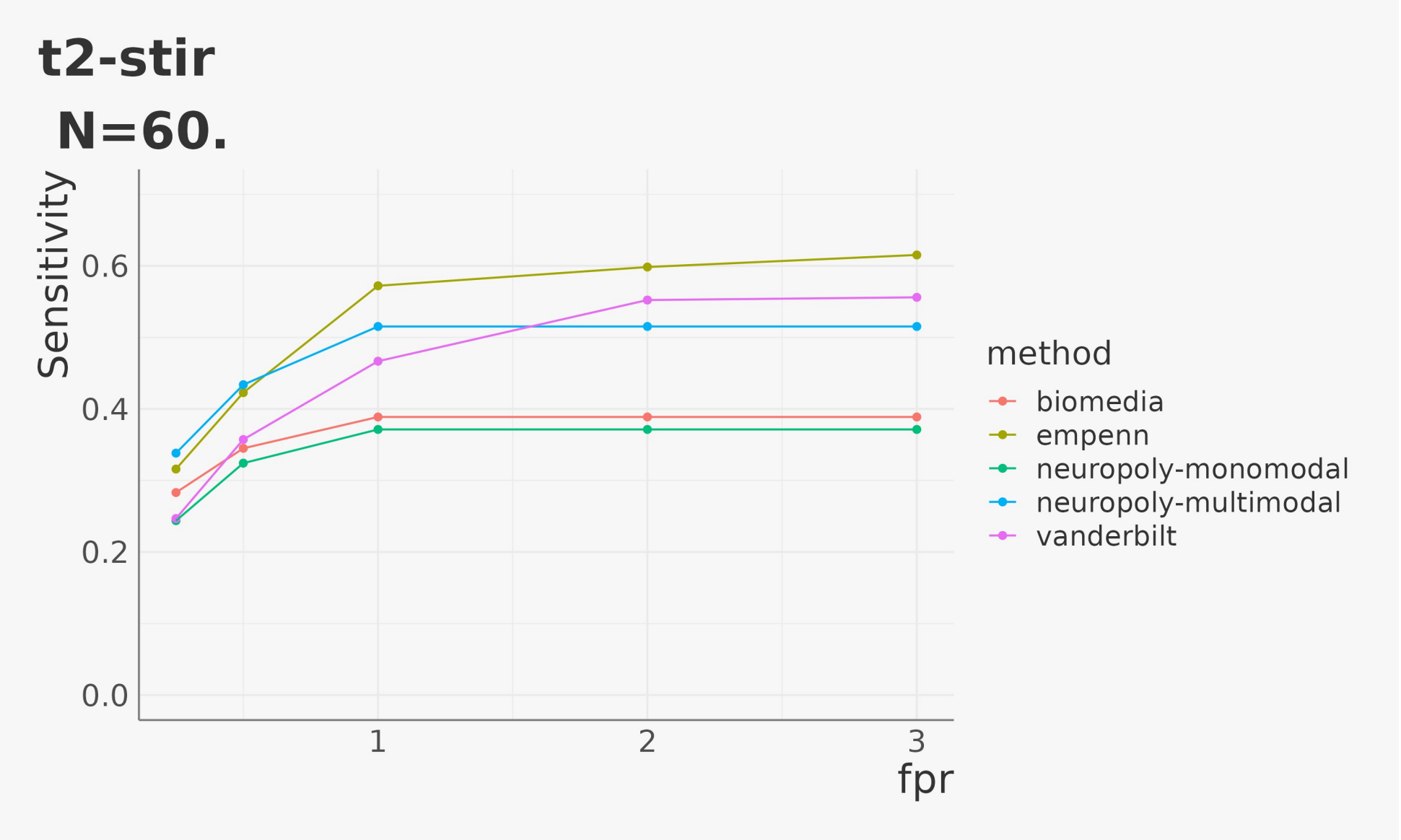
Performances of experts against final Ground Truth (with IoU threshold 0.2) on the test set



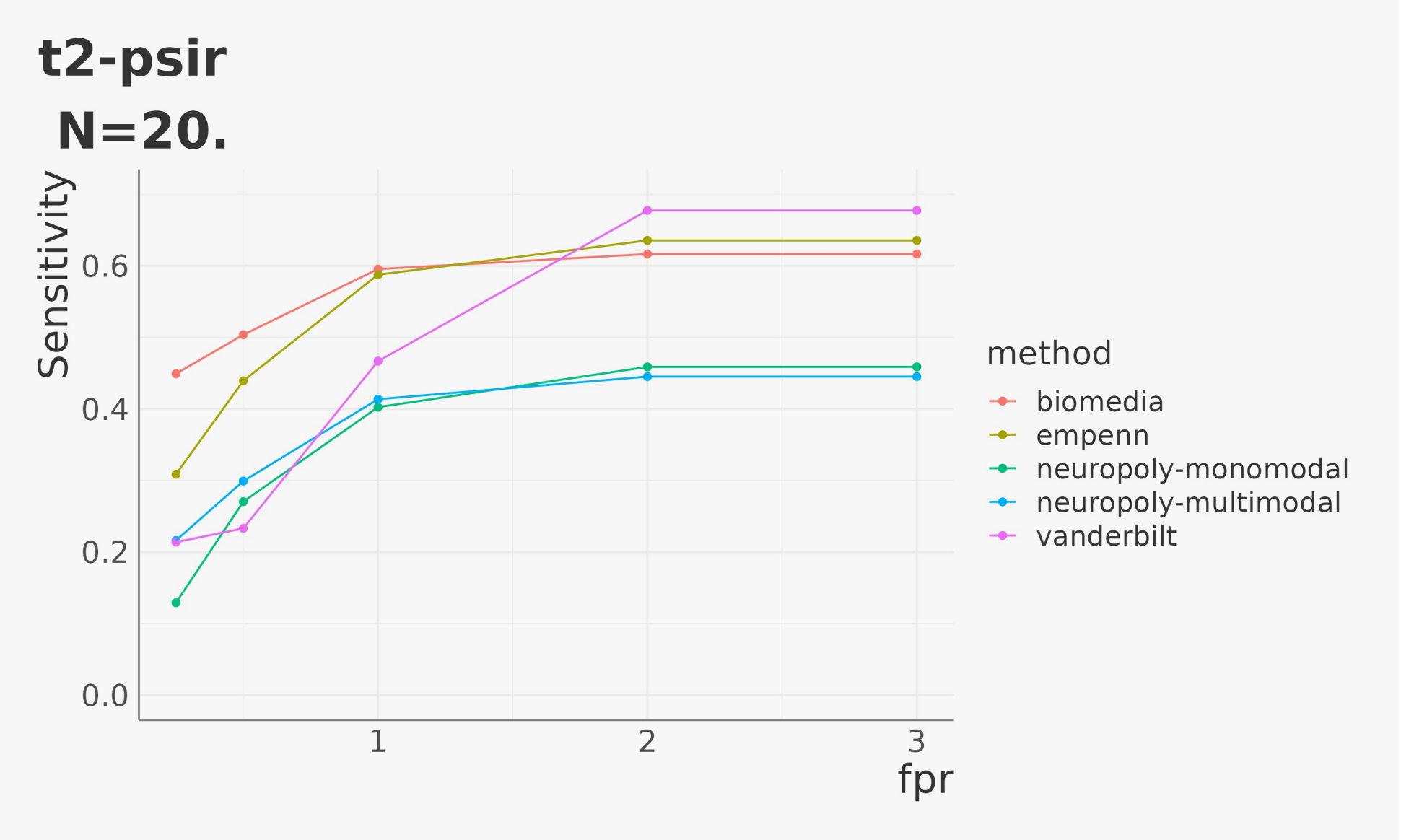
2. Results

2.1 Results on the different sequences combinations

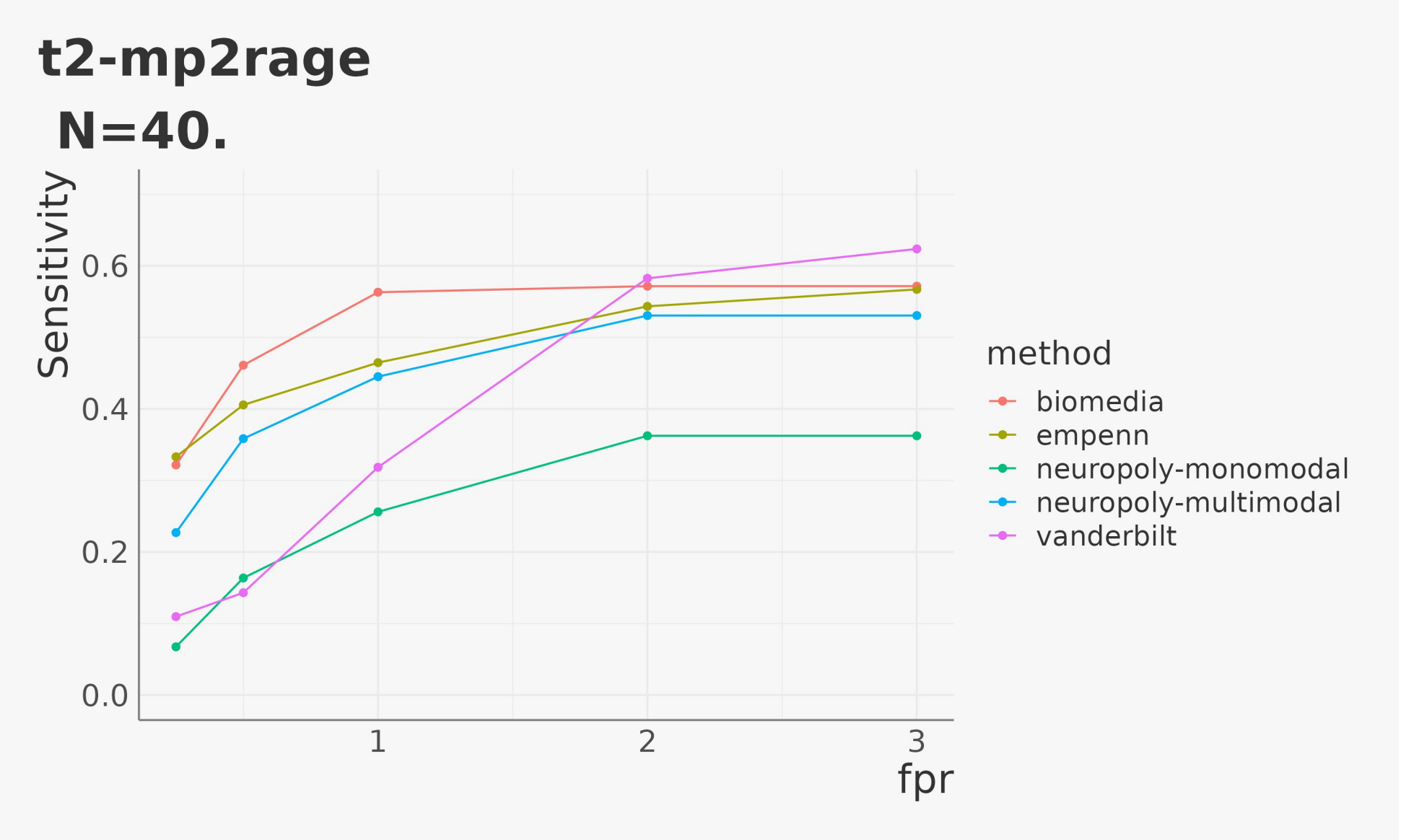
Results for the combination t2+stir (N=60)



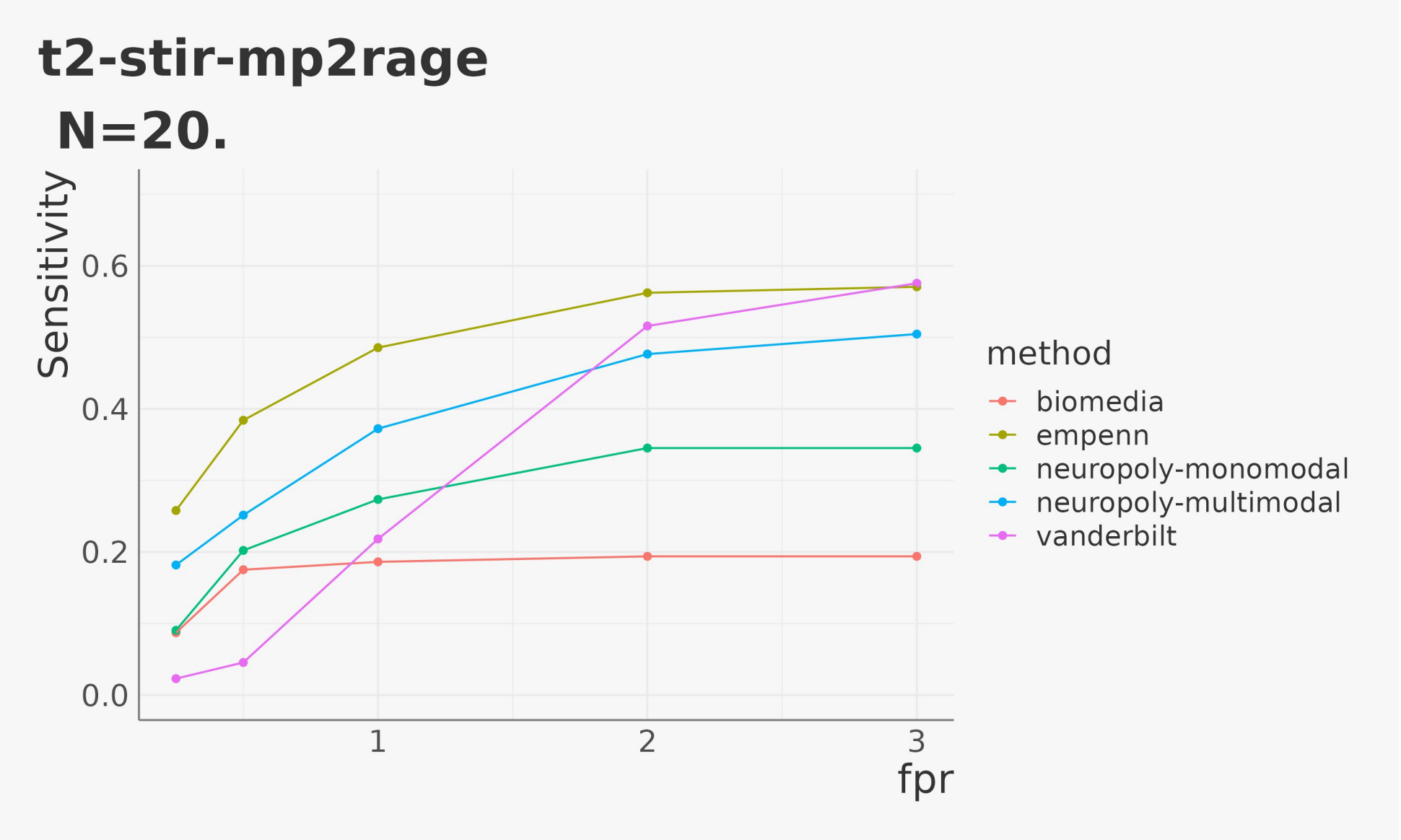
Results for the combination t2+psir (N=20)



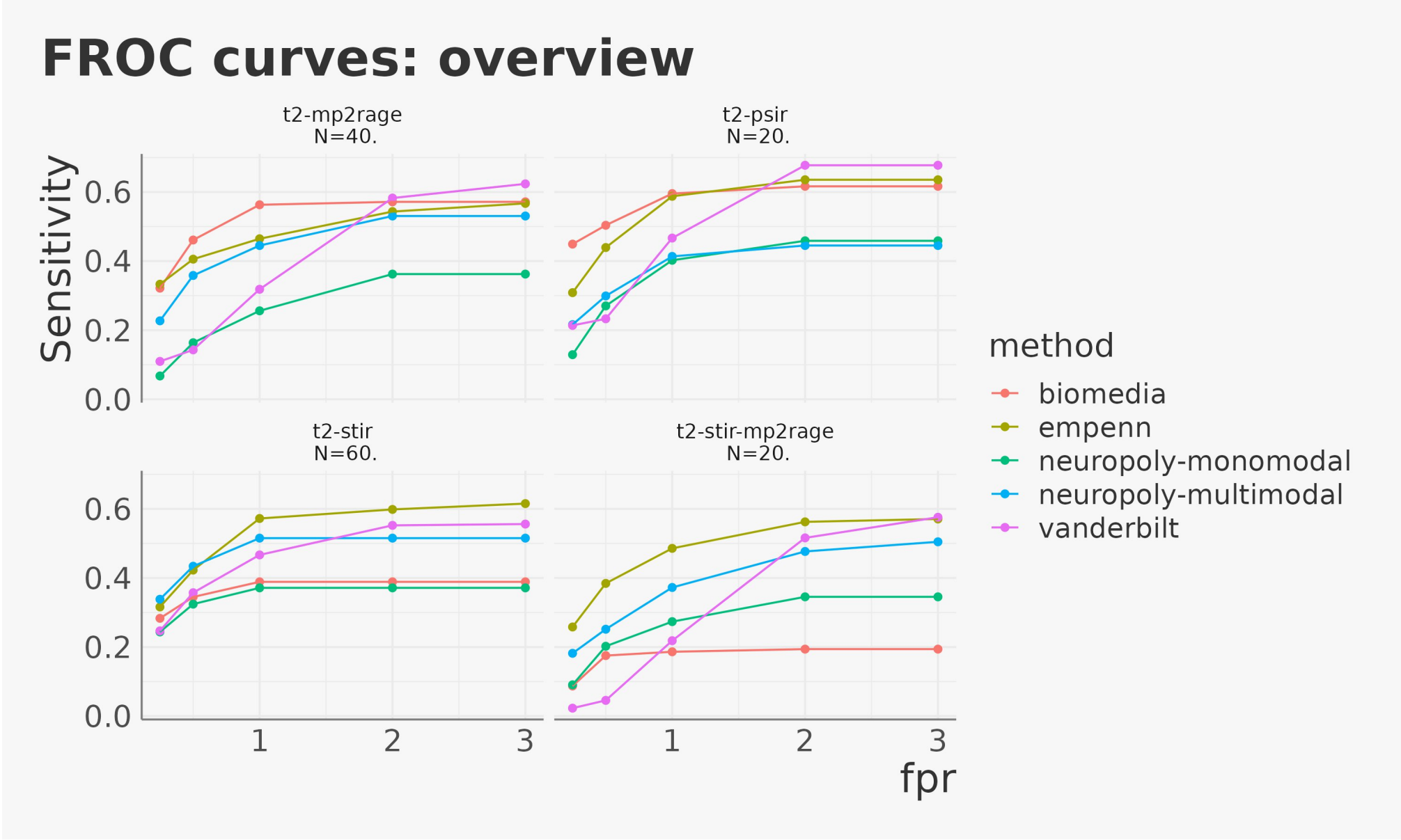
Results for the combination t2+mp2rage (N=40)



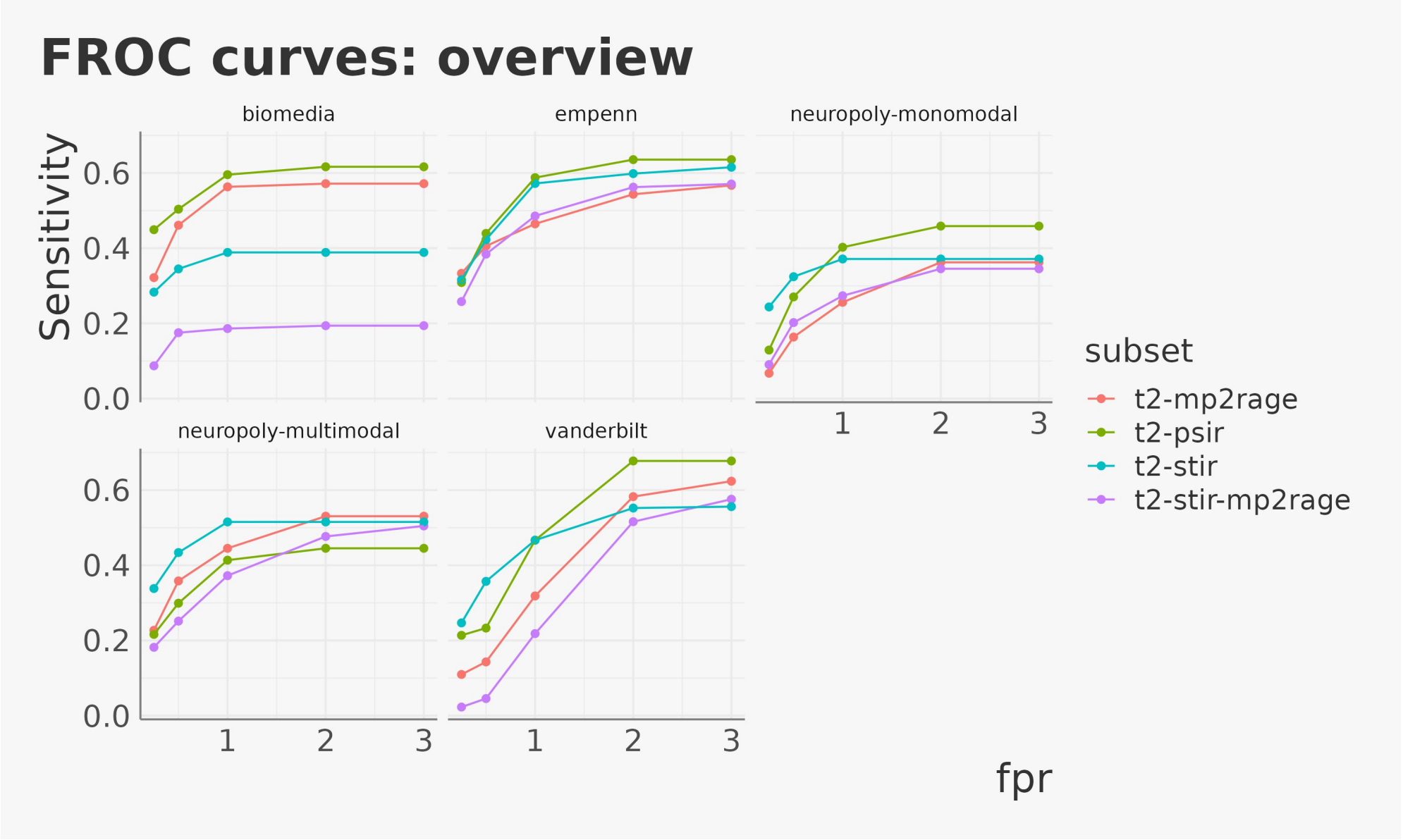
Results for the combination t2+mp2rage+stir (N=20)



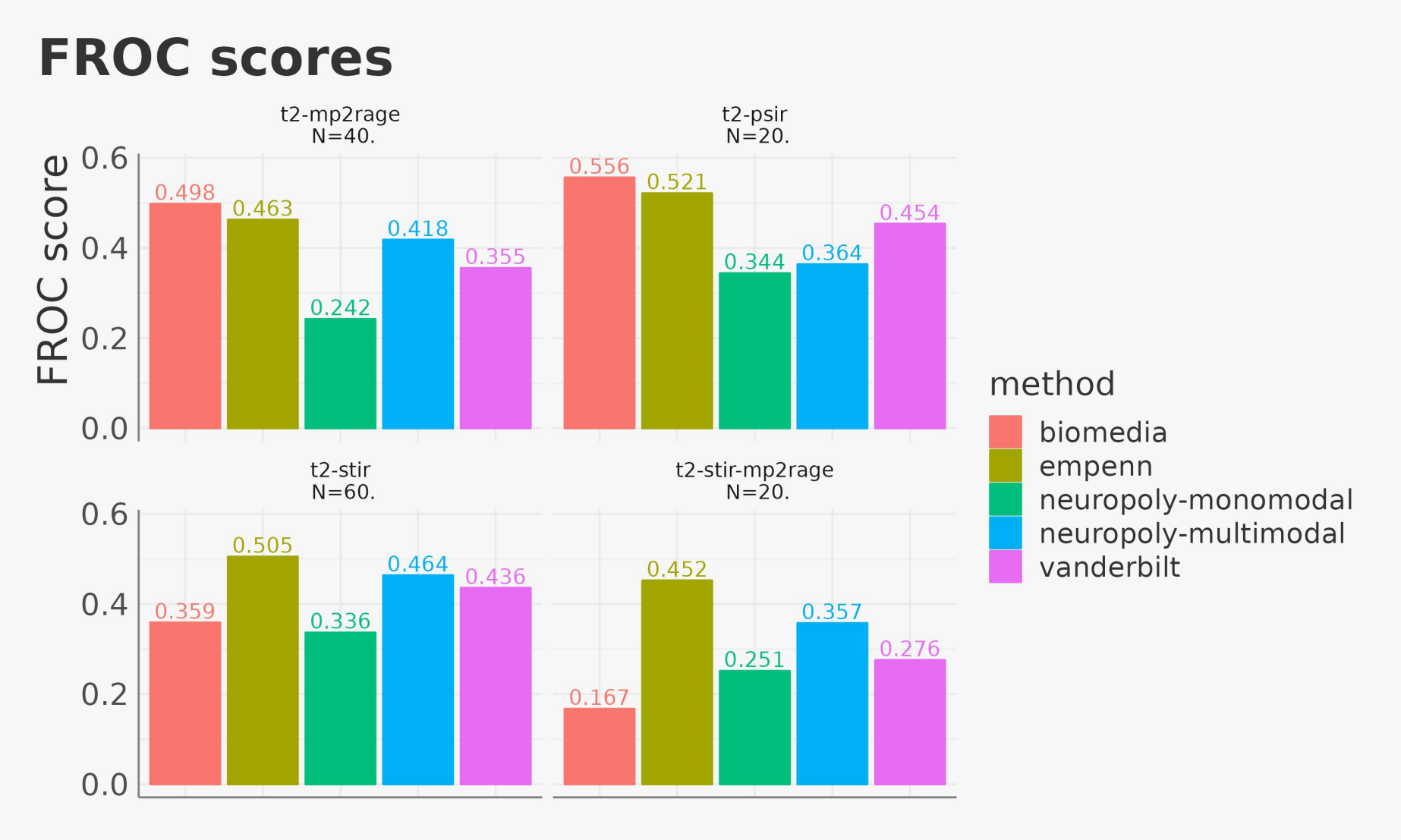
Summary for the different combinations



Summary for the different combinations

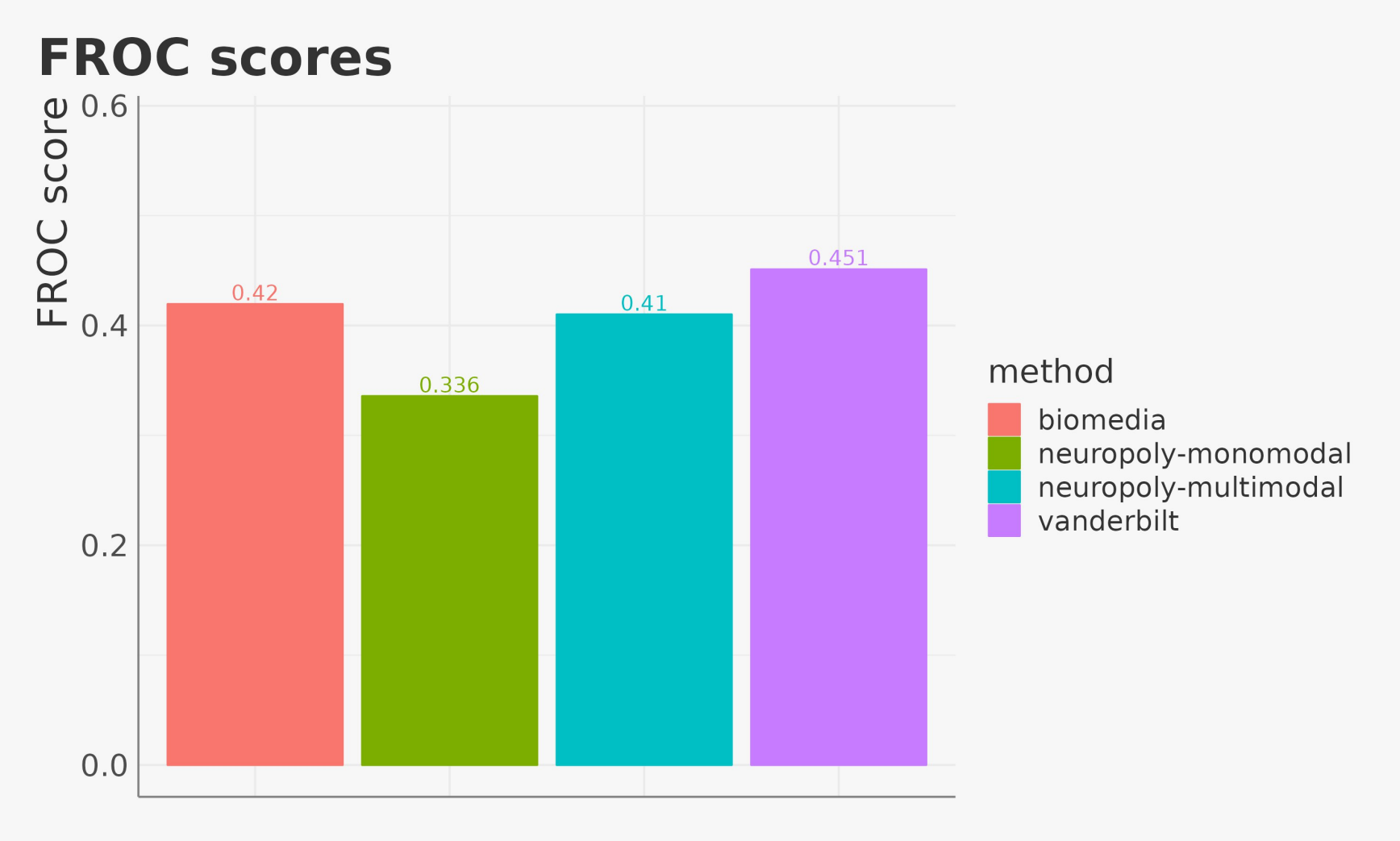


Summary for the different combinations

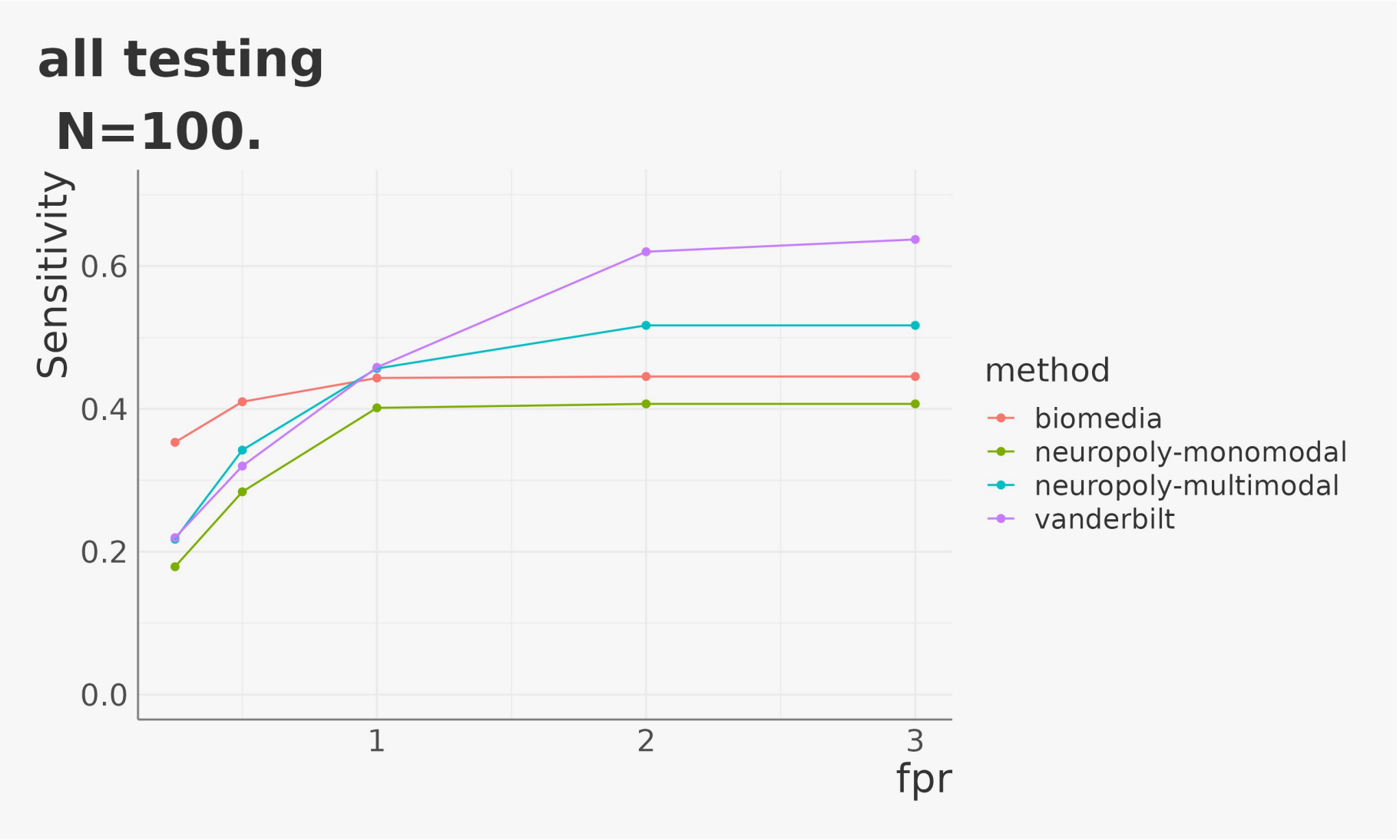


2.2 Results for overall test set (N=100)

Results for overall test set (N=100)



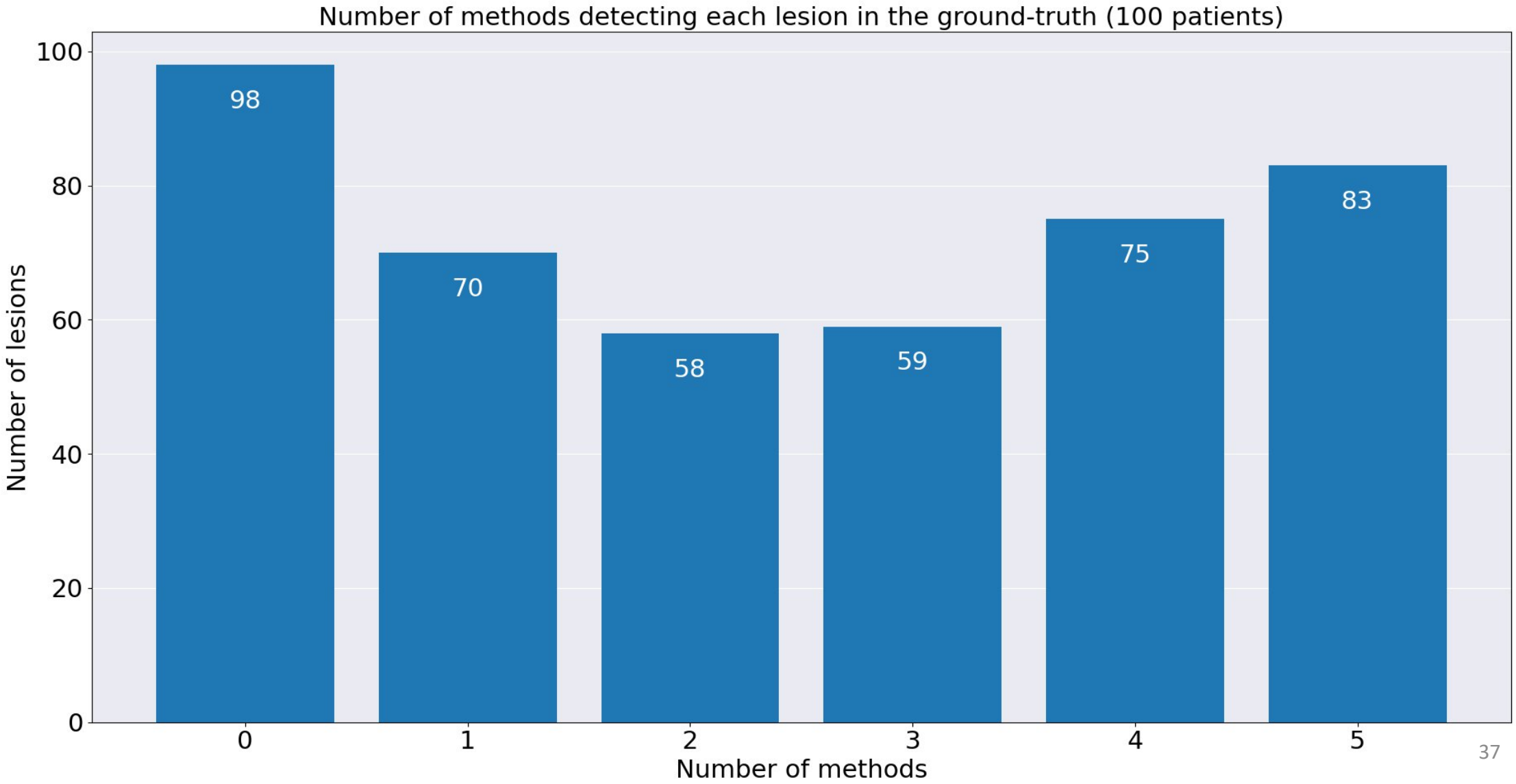
Results for overall test set (N=100)



2.3 Inter-method variability with examples

Results for overall test set (N=100)

IoU threshold 0.2, decision threshold=0

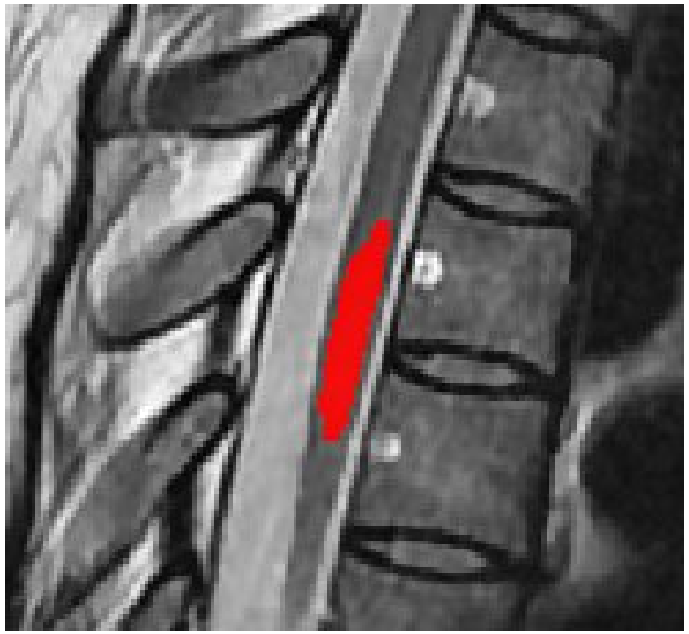


Lesions detected by all methods

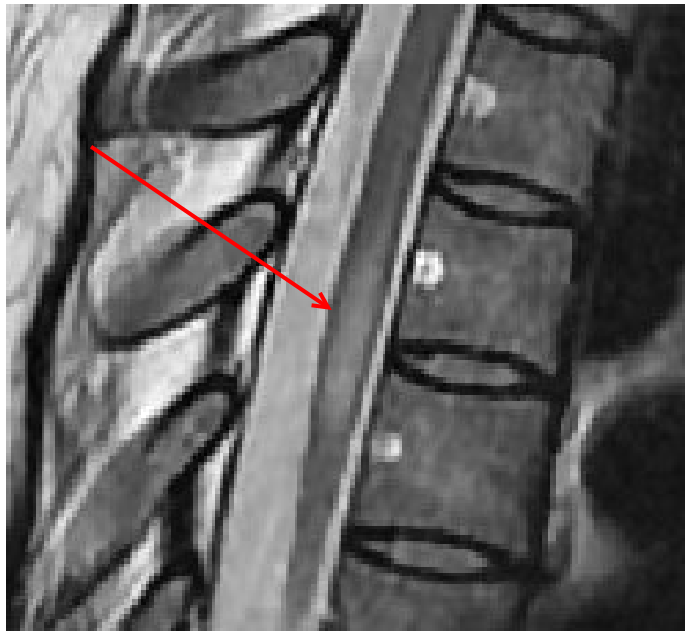
Lesions detected by all methods (IoU threshold 0.2):

Example: T2-w + STIR (4/4 experts)

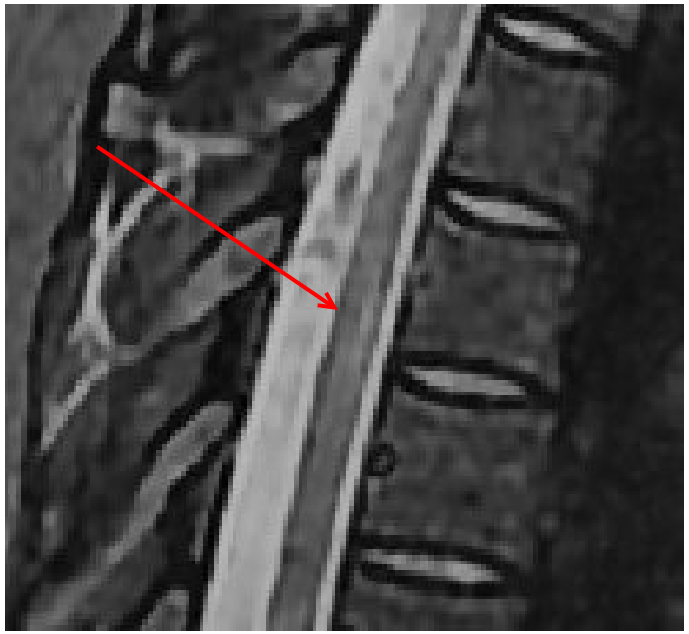
T2-w rawdata + GT segmentation:



T2-w rawdata:



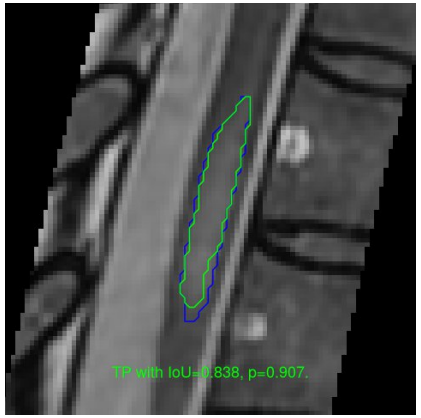
STIR rawdata:



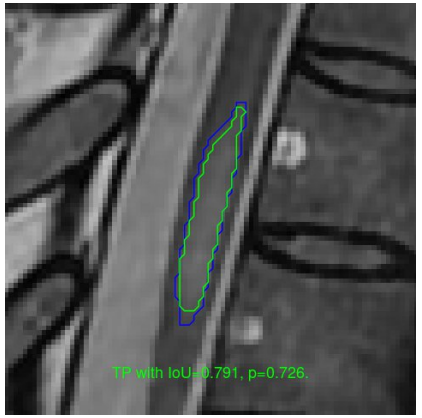
Lesions detected by all methods (IoU threshold 0.2):

Example: T2-w + STIR (4/4 experts)

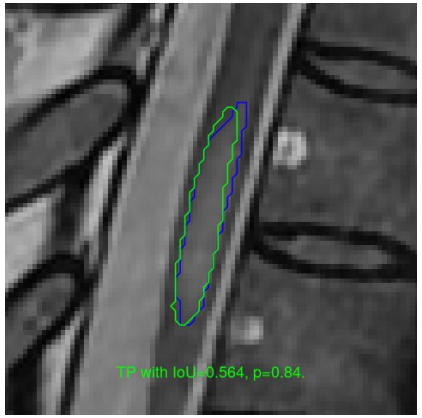
BioMedIA



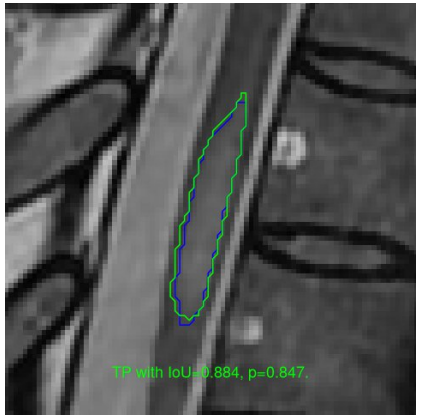
Empenn



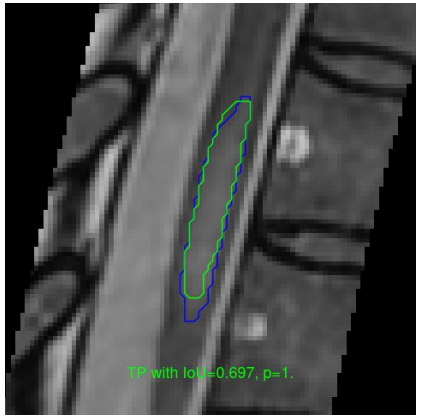
Neuropoly (monomodal)



Neuropoly (multimodal)



Vanderbilt



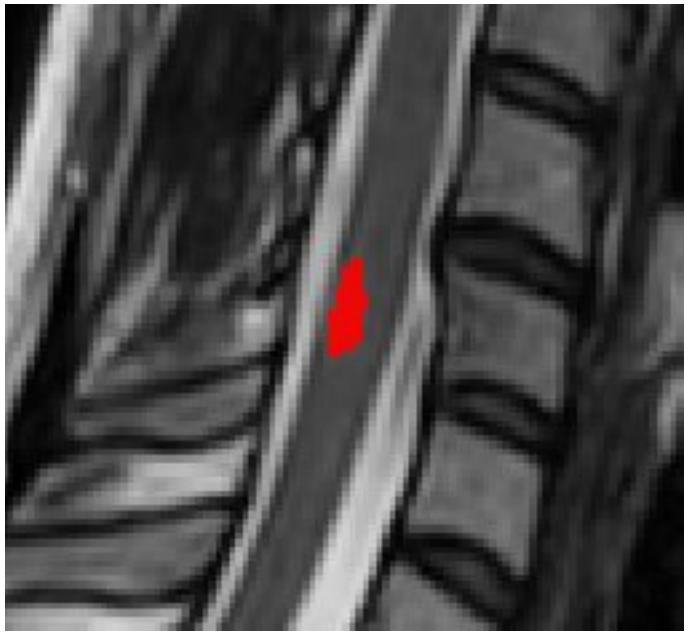
predicted lesion outline

GT lesion outline

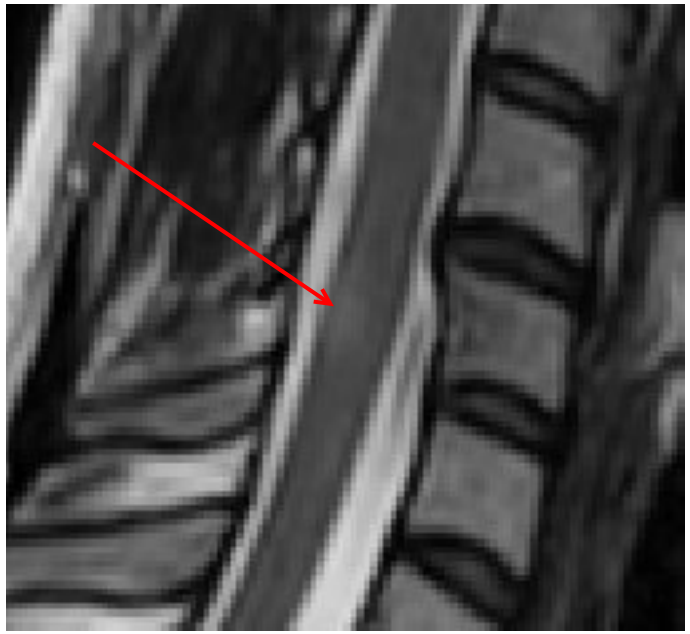
Lesions detected by all methods (IoU threshold 0.2):

Example: T2-w + PSIR (4/4 experts)

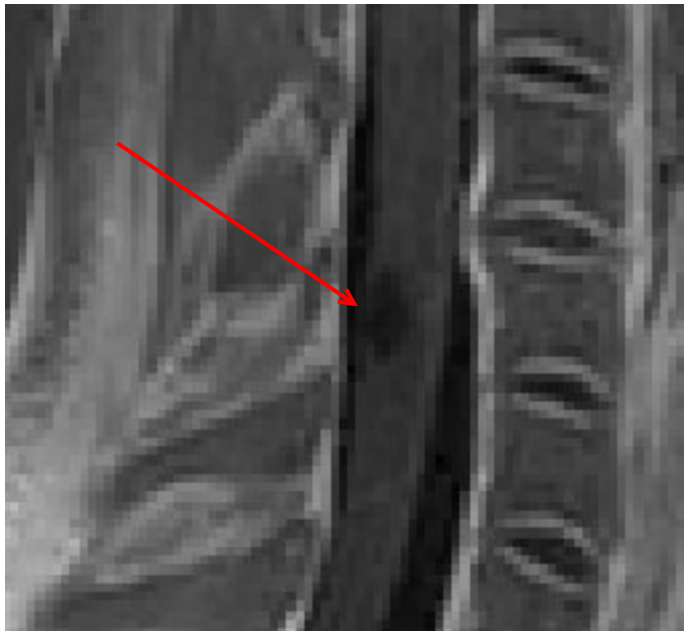
T2-w rawdata + GT segmentation:



T2-w rawdata:



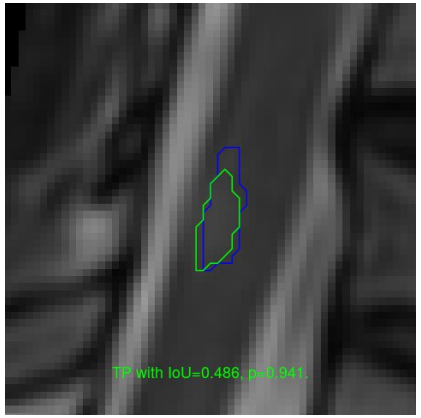
PSIR rawdata:



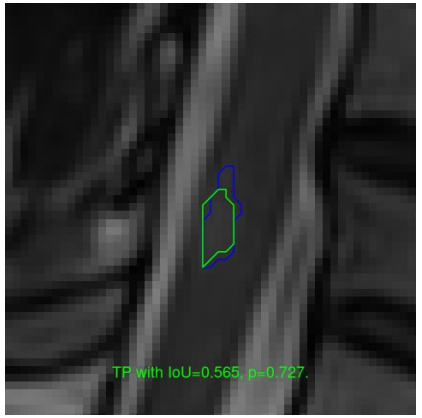
Lesions detected by all methods (IoU threshold 0.2):

Example: T2-w + PSIR (4/4 experts)

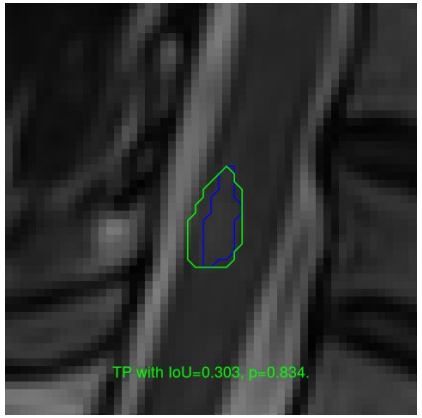
BioMedIA



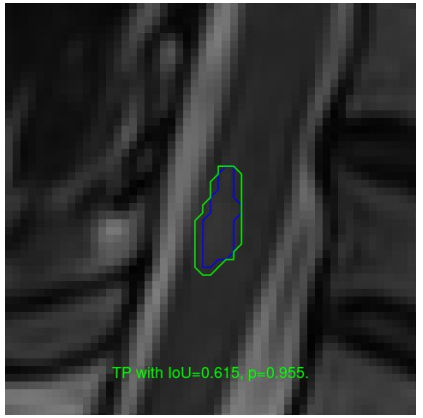
Empenn



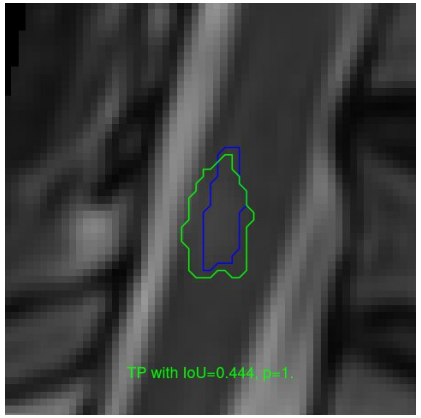
Neuropoly (monomodal)



Neuropoly (multimodal)



Vanderbilt



predicted lesion outline

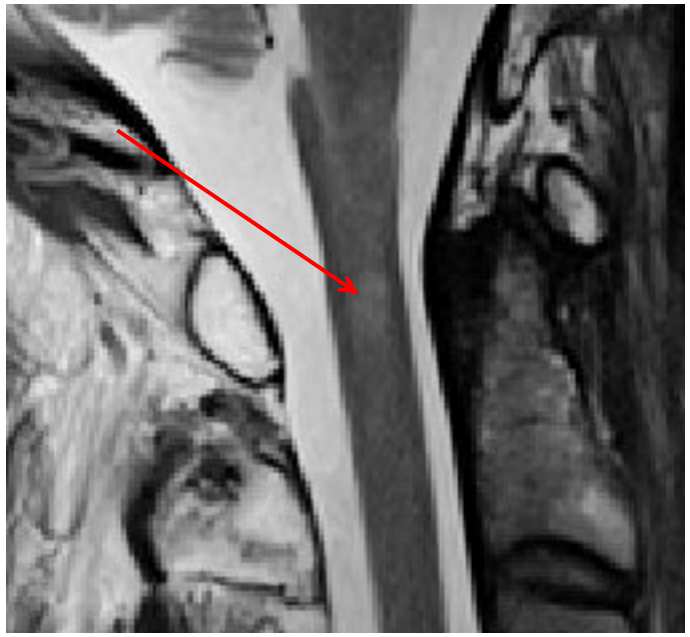
GT lesion outline

Lesions detected by all methods (IoU threshold 0.2):

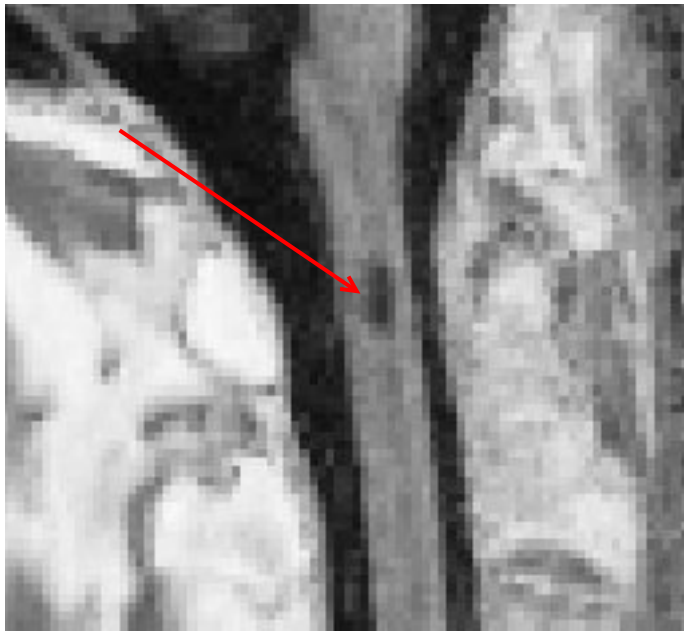
Example: T2-w + MP2RAGE (4/4 experts)

T2-w rawdata + GT segmentation:

T2-w rawdata:



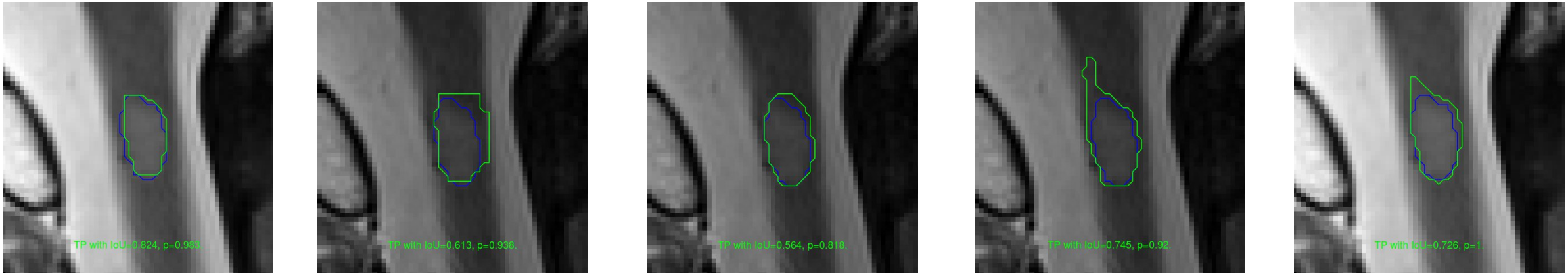
MP2RAGE rawdata:



Lesions detected by all methods (IoU threshold 0.2):

Example: T2-w + MP2RAGE (4/4 experts)

BioMedIA | Empenn | Neuropoly (monomodal) | Neuropoly (multimodal) | Vanderbilt



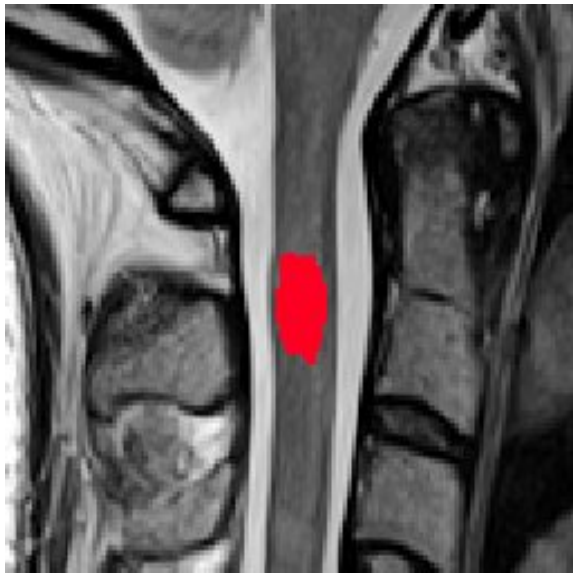
predicted lesion outline

GT lesion outline

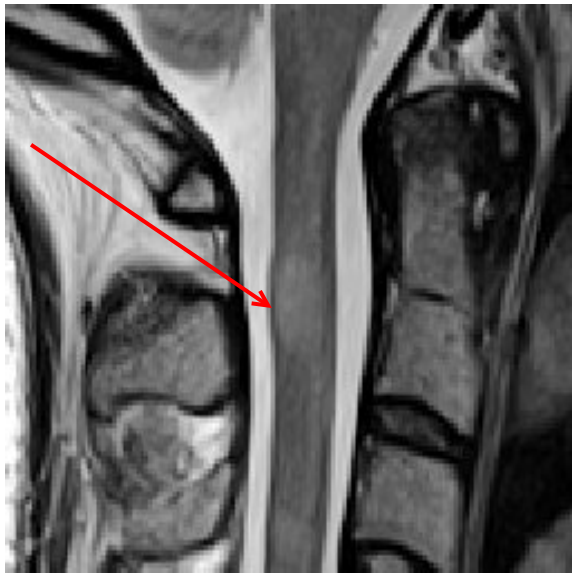
Lesions detected by all methods (IoU threshold 0.2):

Example: T2-w + STIR + MP2RAGE (4/4 experts)

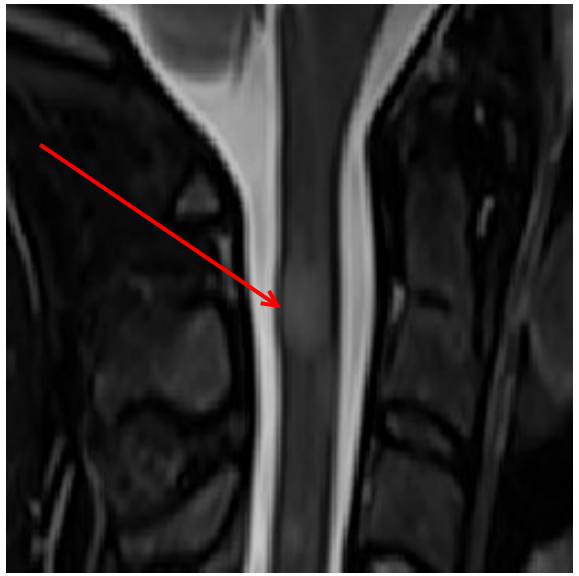
T2-w rawdata + GT segmentation:



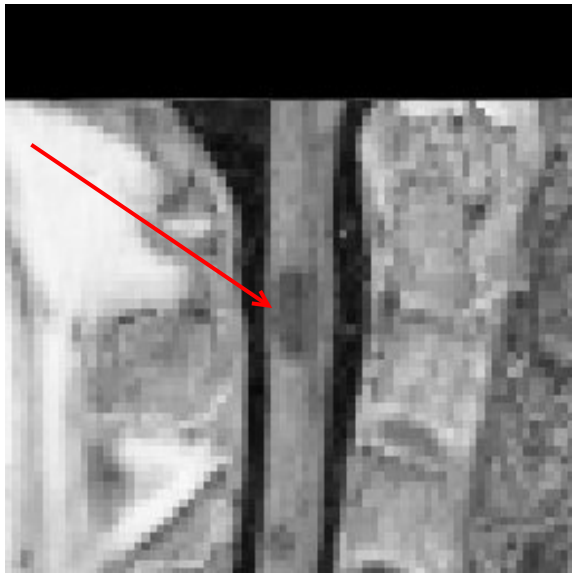
T2-w rawdata:



STIR rawdata:



MP2RAGE rawdata:

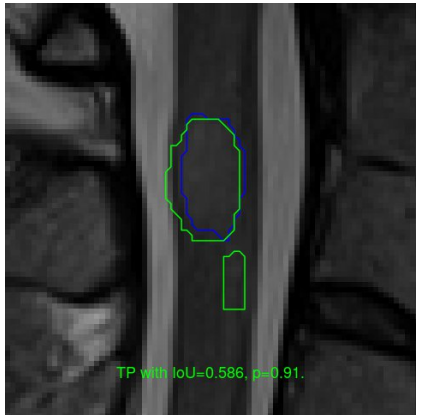


Lesions detected by all methods (IoU threshold 0.2):

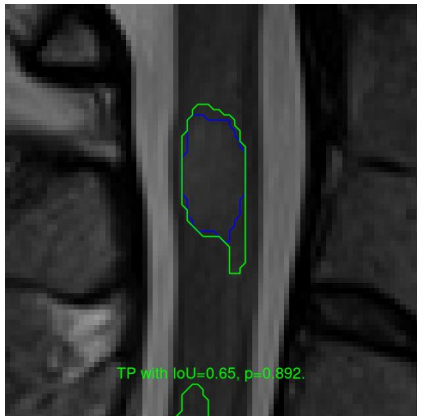
Example: T2-w + STIR + MP2RAGE (4/4 experts)

BioMedIA

Empenn

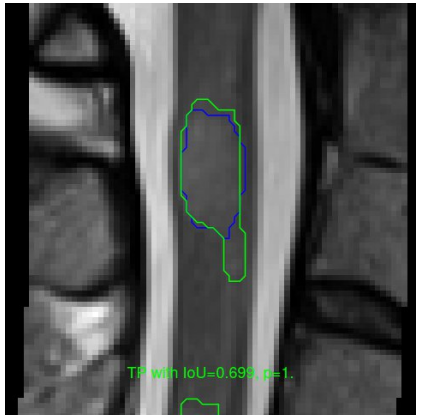


Neuropoly (monomodal)



Neuropoly (multimodal)

Vanderbilt



predicted lesion outline

GT lesion outline

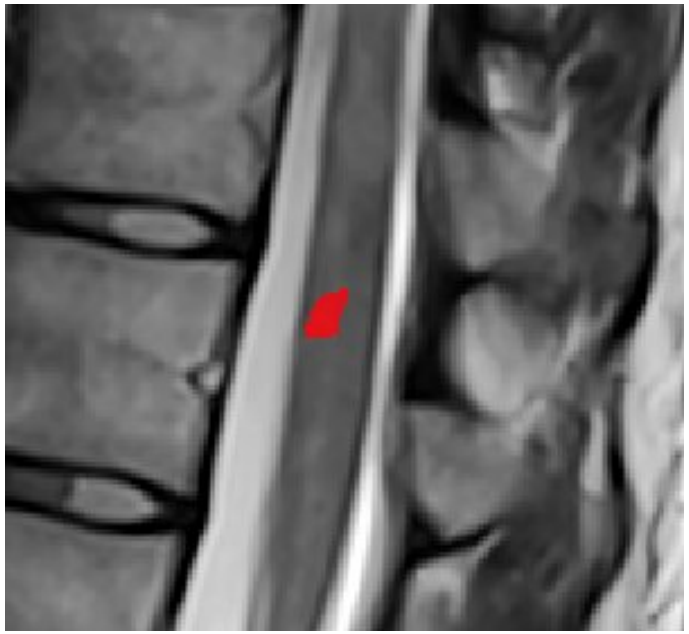
Lesions detected by no method

Lesions detected by no method

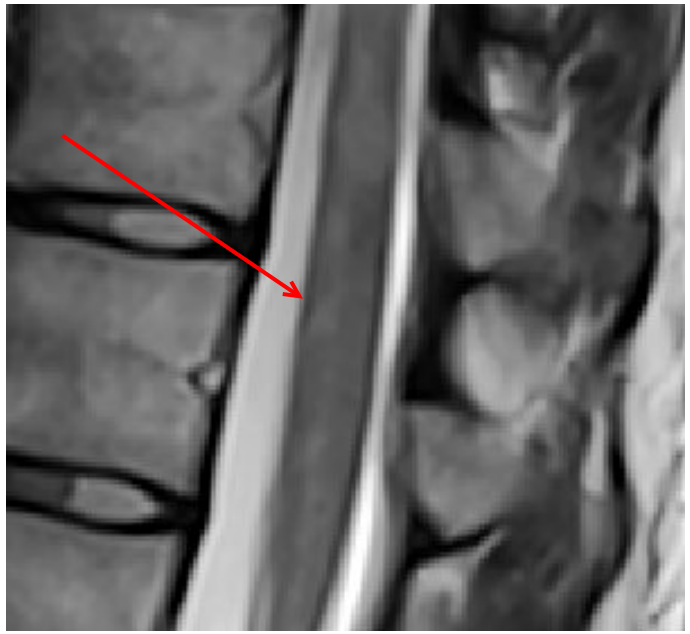
(IoU threshold=0, decision proba thresold=0)

Example: T2-w + STIR (4/4 experts)

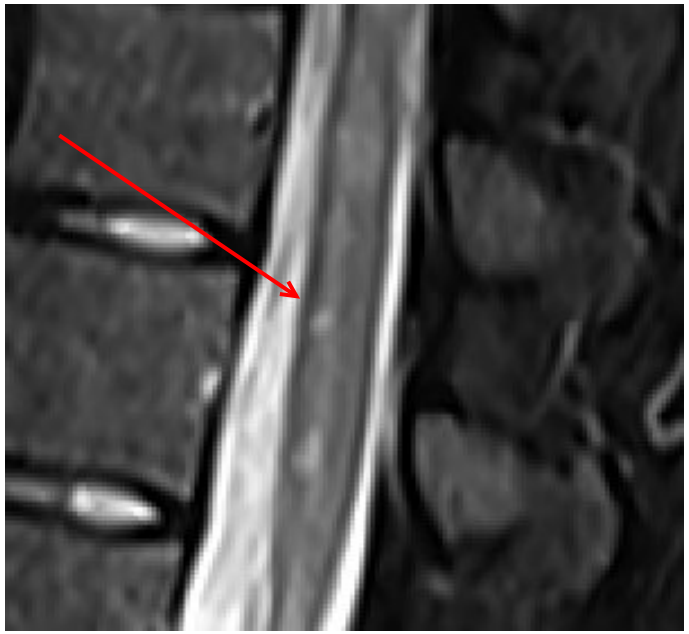
T2-w rawdata + GT segmentation:



T2-w rawdata:



STIR rawdata:

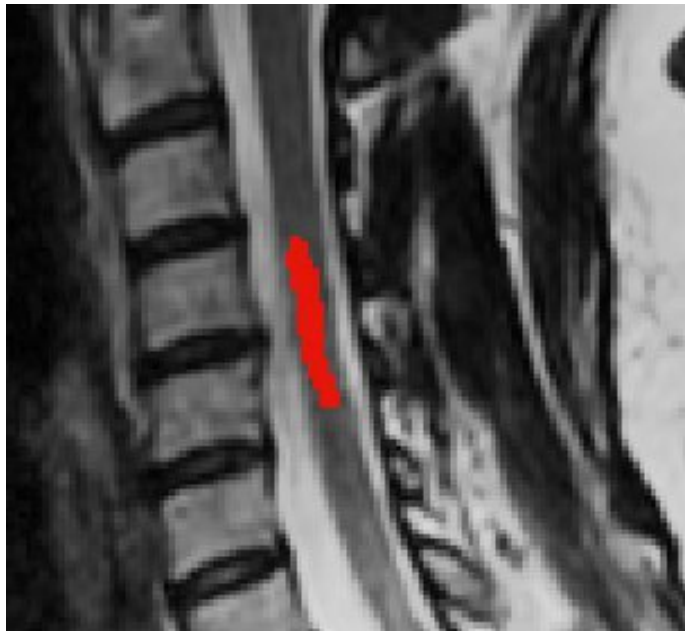


Lesions detected by no method

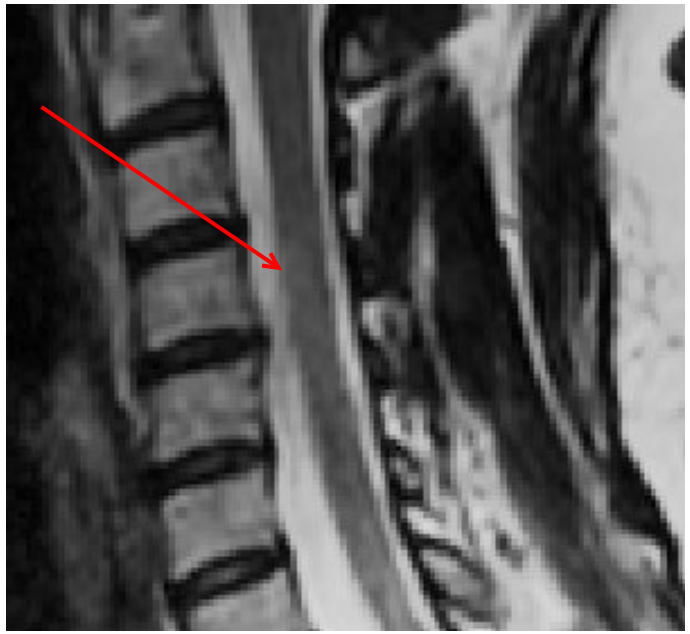
(IoU threshold=0, decision proba thresold=0)

Example: T2-w + PSIR (3/4 experts)

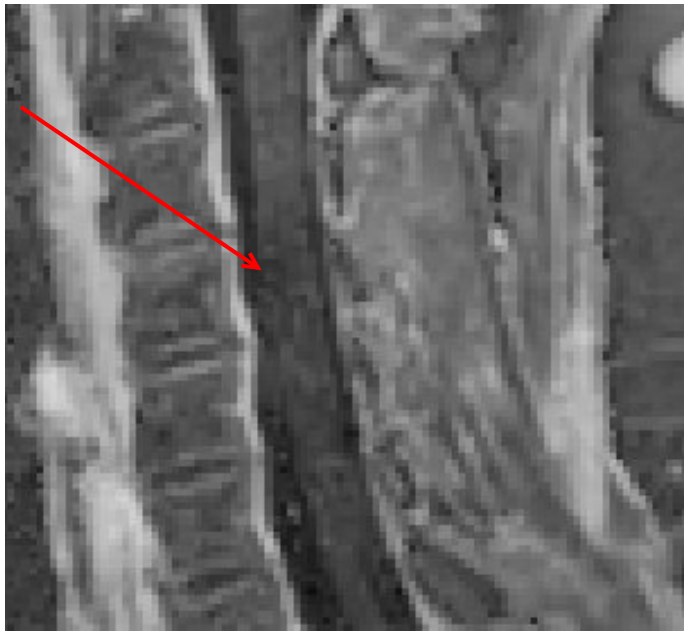
T2-w rawdata + GT segmentation:



T2-w rawdata:



PSIR rawdata:

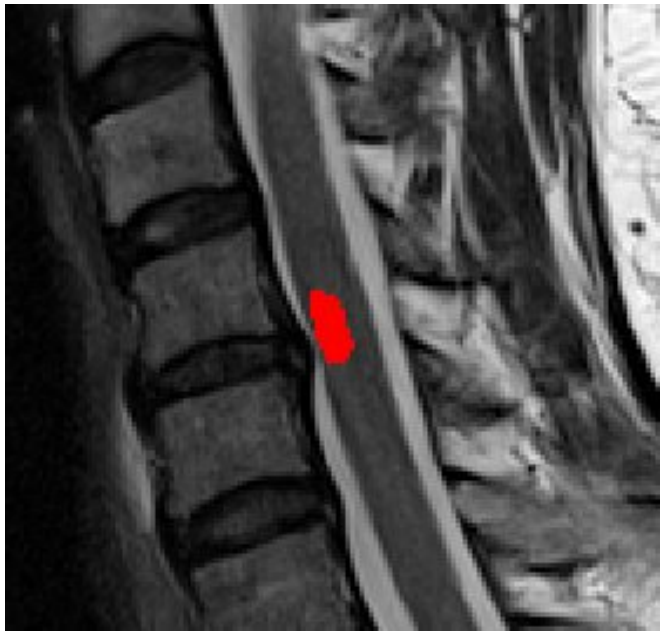


Lesions detected by only one method

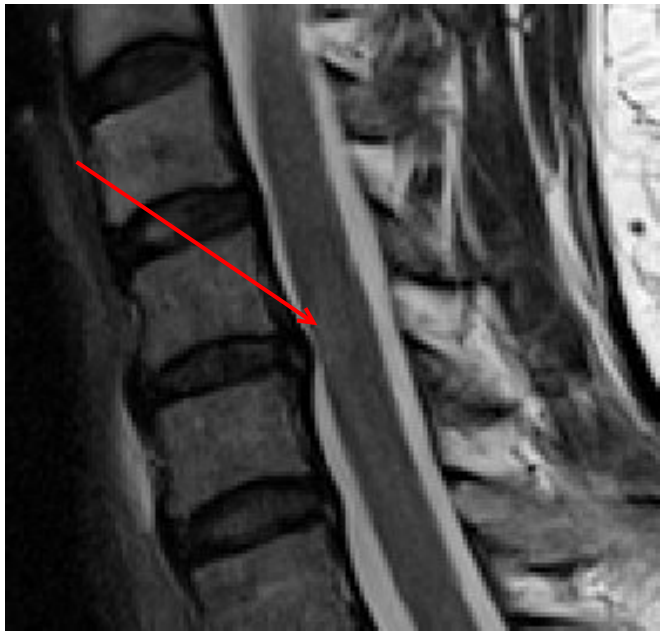
(IoU threshold=0, decision proba thresold=0)

Example: T2-w + MP2RAGE (3/4 experts)

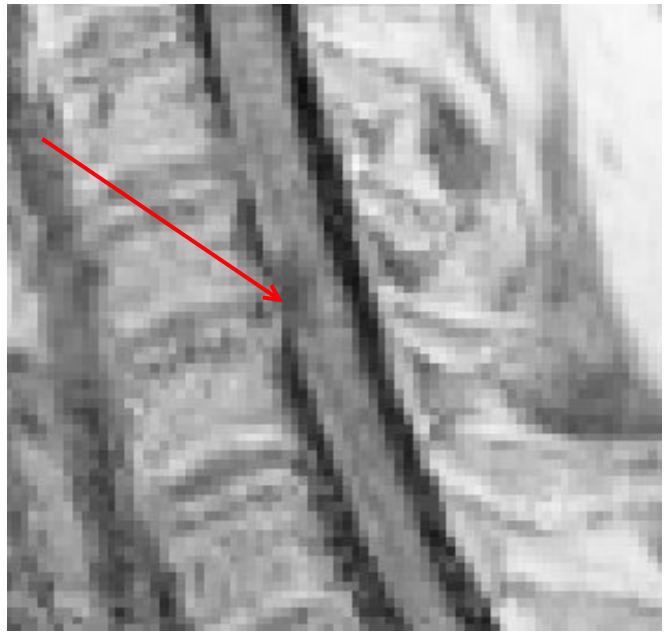
T2-w rawdata + GT segmentation:



T2-w rawdata:



MP2RAGE rawdata:

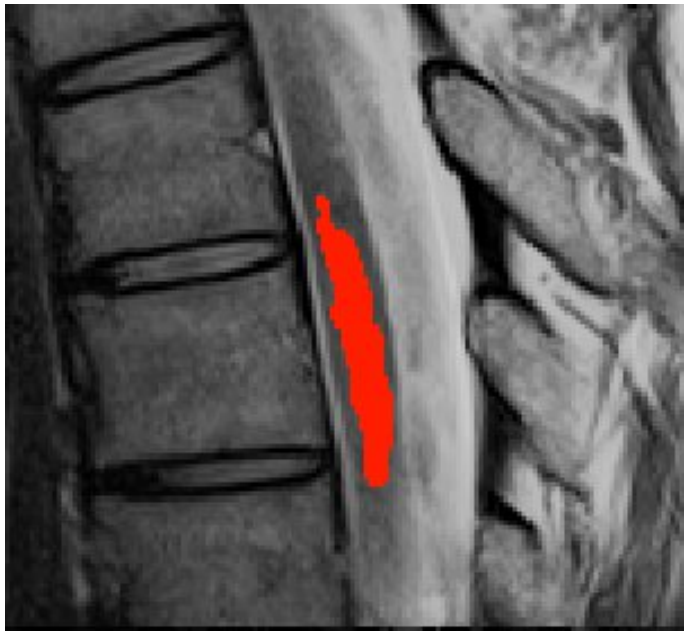


Lesions detected by no method

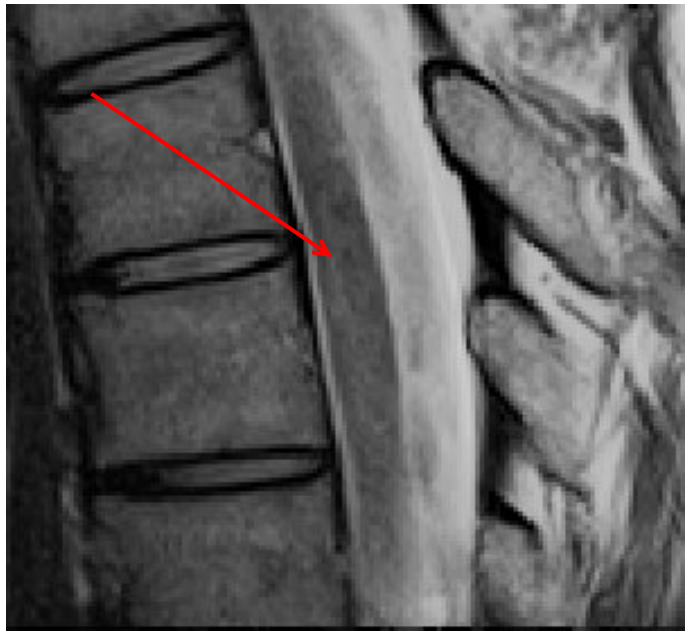
(IoU threshold=0, decision proba thresold=0)

Example: T2-w + STIR + MP2RAGE (1/4 experts)

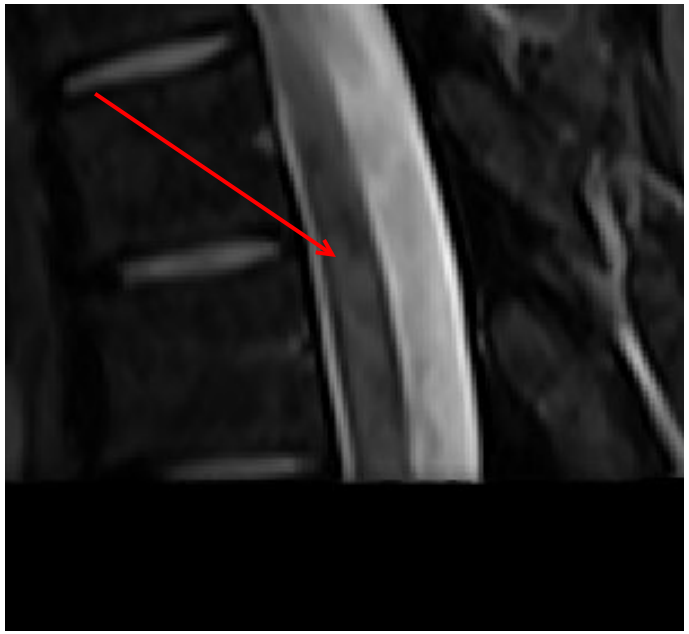
T2-w rawdata + GT segmentation:



T2-w rawdata:



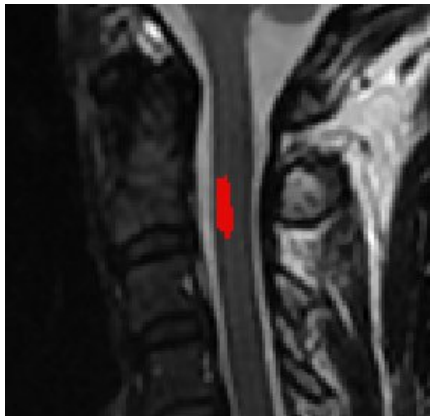
STIR rawdata:



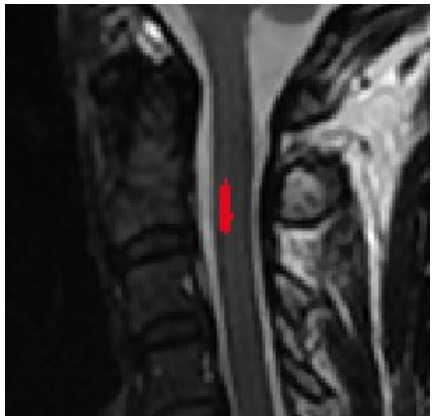
Lesions detected by all methods that are not in the ground truth

Subject without any lesion in ground truth:

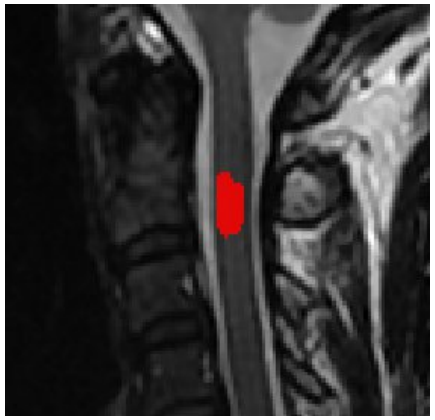
BioMedIA



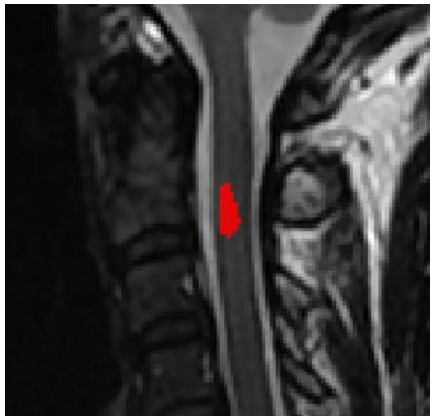
Empenn



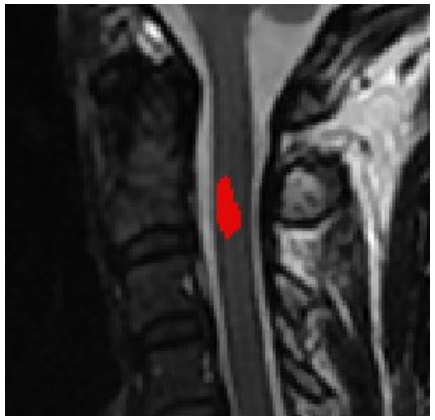
Neuropoly (monomodal)



Neuropoly (multimodal)

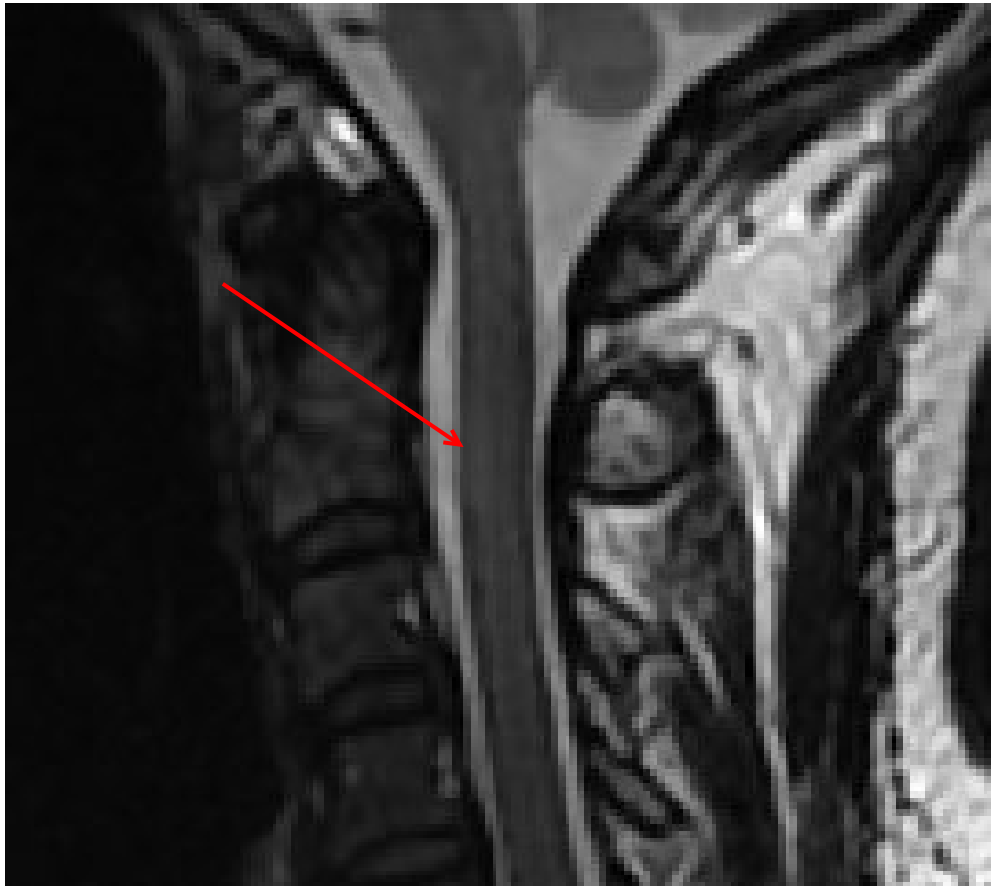


Vanderbilt

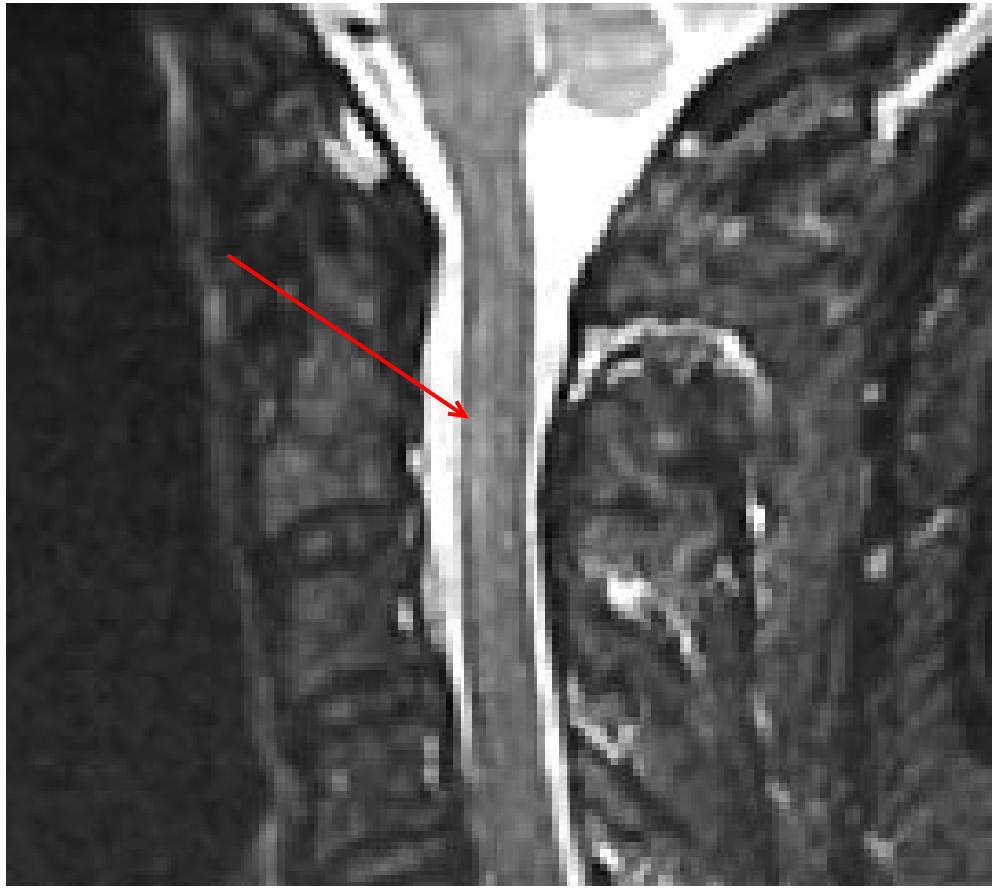


Subject without any lesion in ground truth:

T2-w rawdata:



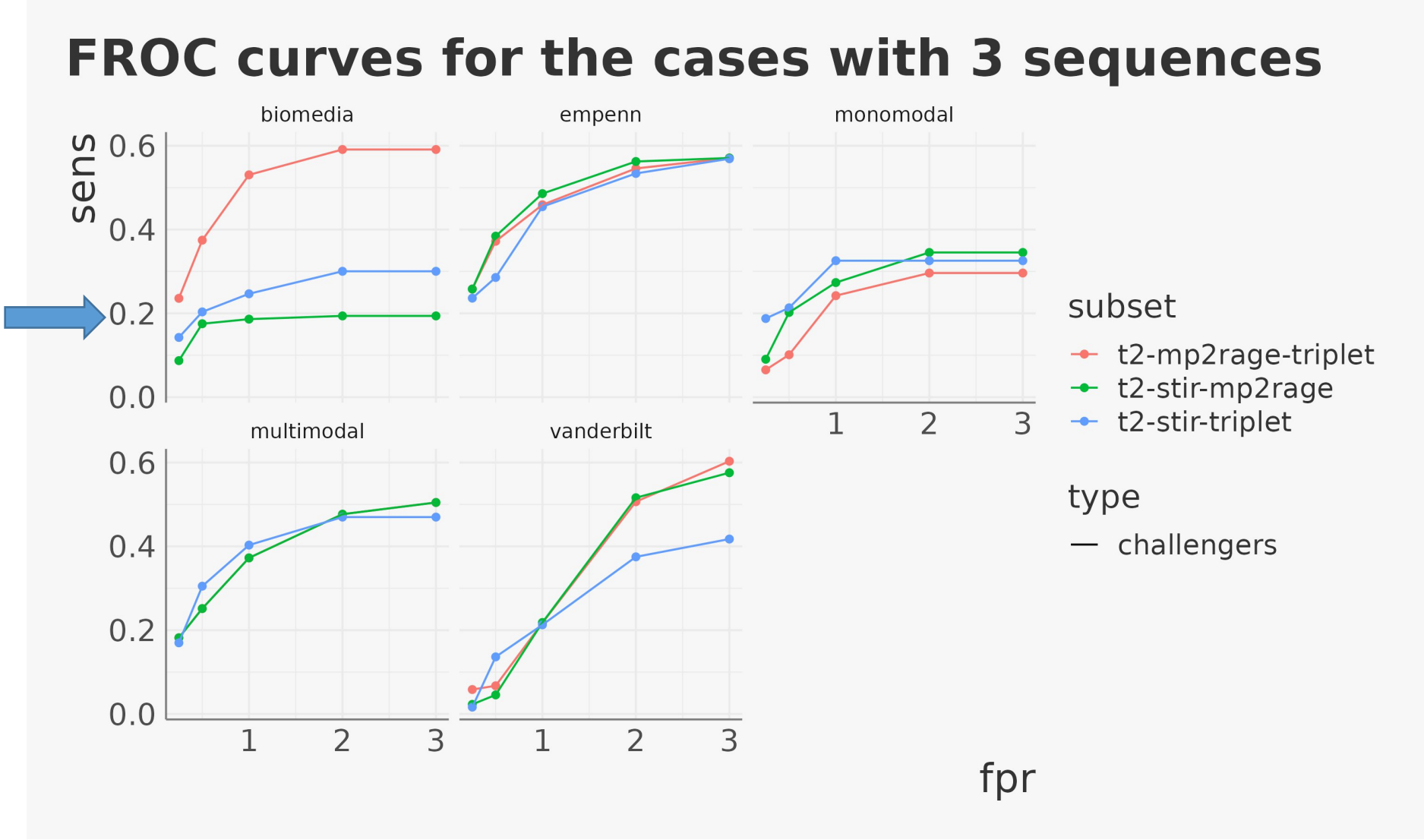
STIR rawdata:



3. More Results

3.1 Dealing with the three-sequences combination

The case with 3 sequences (only at testing time)



3.2 Added value of multi-sequence over T2 alone: some insights

Do we improve performance wrt to t2Sag only?

We compare to predictions using only t2Sag inputs with two models:

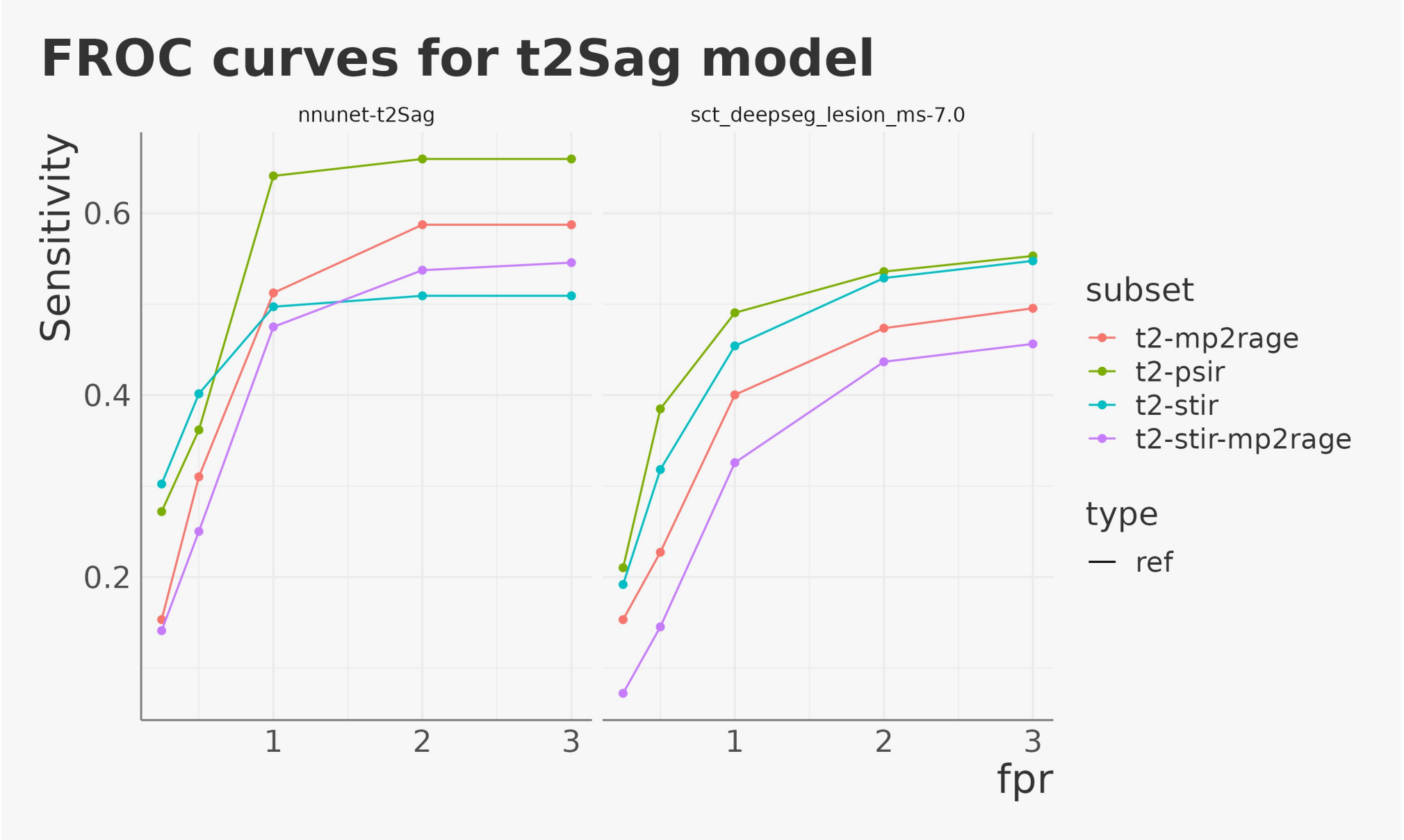
nnUnet-t2Sag:

- nnU-Net (3D U-Net) with batch size: 2, patch size: [256, 64, 128], features per stage: 32, 64, 125, 256, 320, 320, kernel size: [3,3,3] for all stages, strides: [1,1,1] puis [2,2,2] for the other 5 stages, 1000 epochs
- Trained on the 100 preprocessed t2Sag in the training set.
- 0.5 binarisation threshold on softmax outputs to create instances.
- Instance probabilities are assigned as the maximum softmax score within the region.

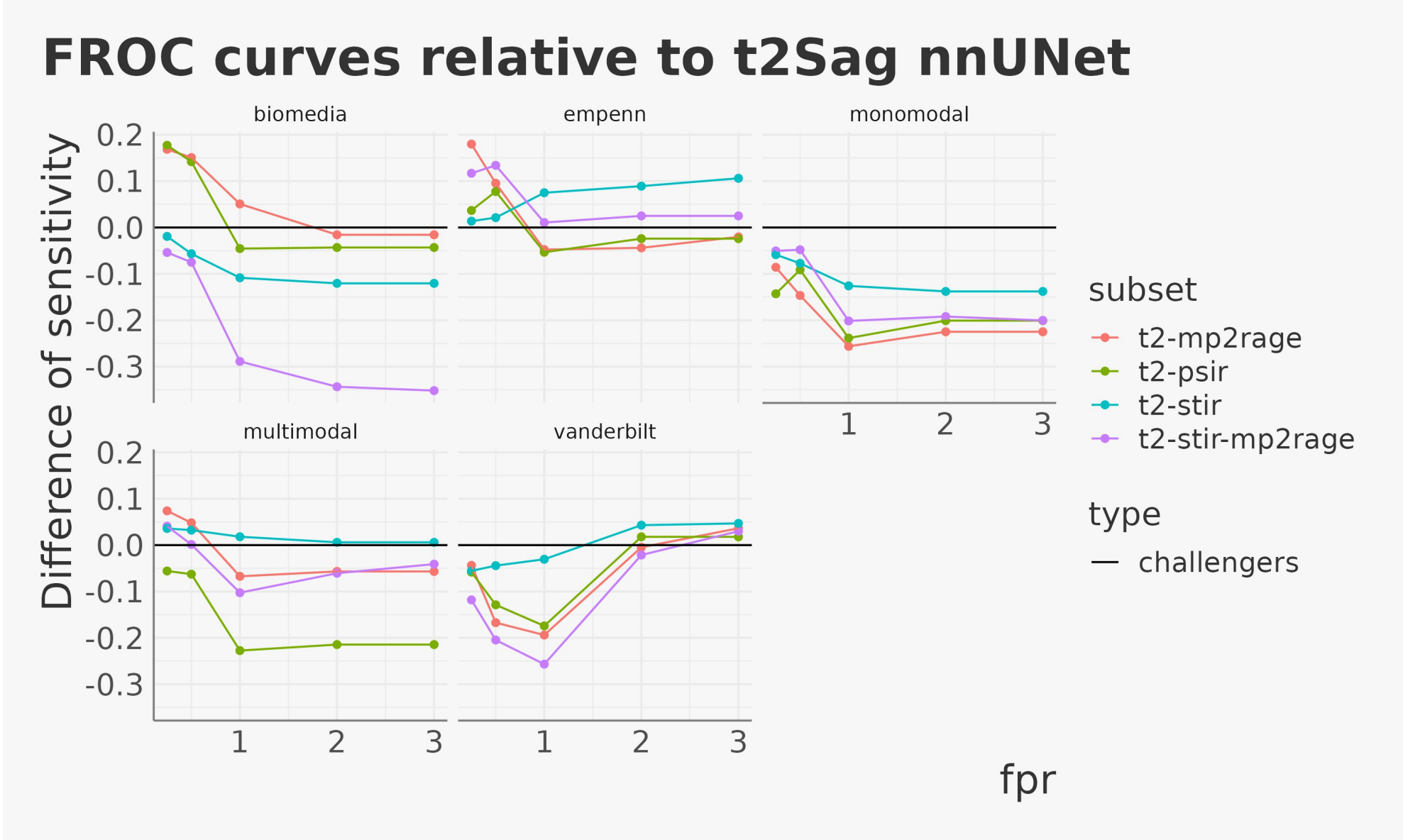
sct_deepseg_lesion_ms_7.0:

- Softmax outputs are produced using the freely available `sct_deepseg lesion_ms` (v7.0).
- 0.5 binarisation threshold on softmax outputs to create instances.
- Instance probabilities are assigned as the maximum softmax score within the region.

Do we improve performance wrt to t2Sag only?



Do we improve performance wrt to t2Sag only?

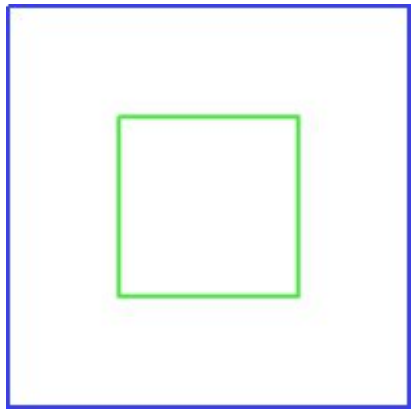


3.3 Performances and dependency to IoU thresholds

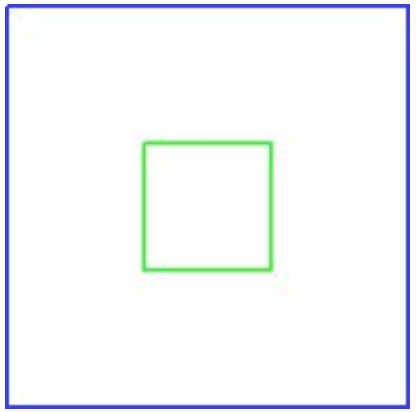
Dependency to IoU thresholds

prediction
ground-truth

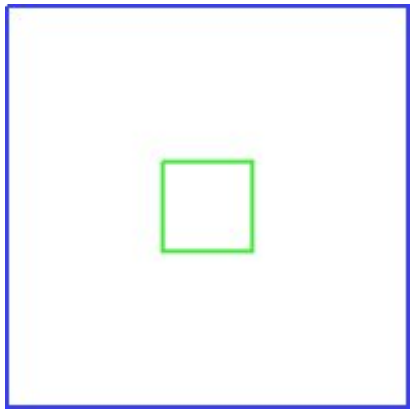
IoU score:
0.20



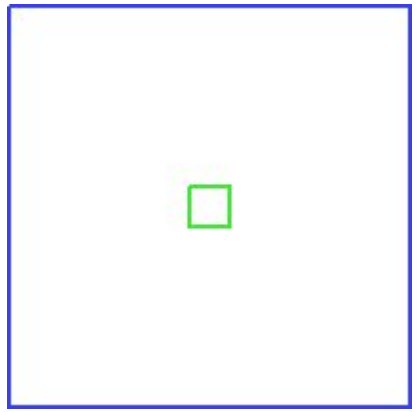
IoU score:
0.10



IoU score:
0.05



IoU score:
0.01



Dependency to IoU thresholds

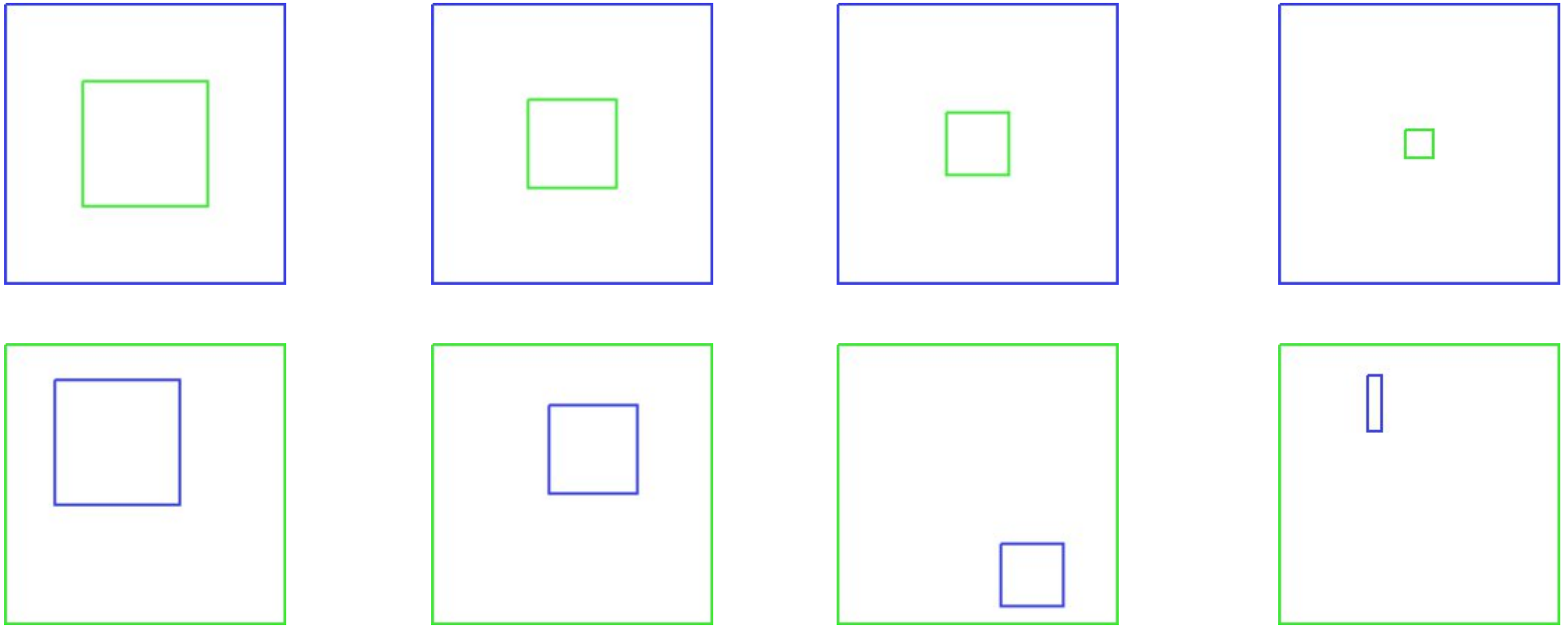
prediction
ground-truth

IoU score:
0.20

IoU score:
0.10

IoU score:
0.05

IoU score:
0.01

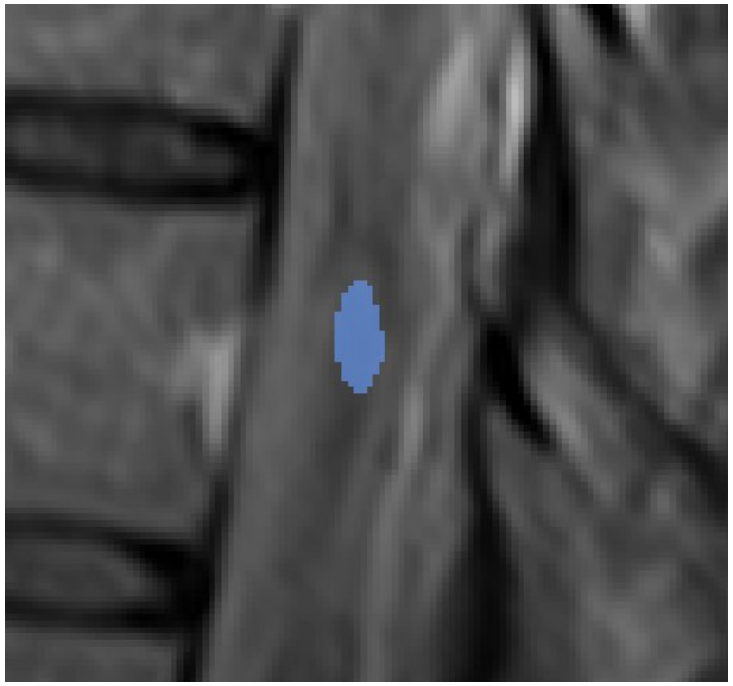


Dependency to IoU thresholds (IoU = 0.191)

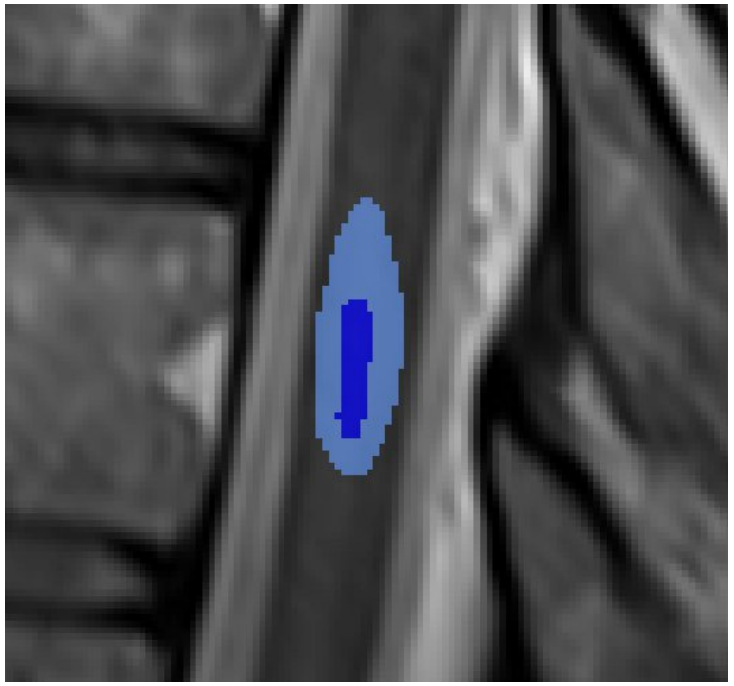
Prediction:

Ground-truth:

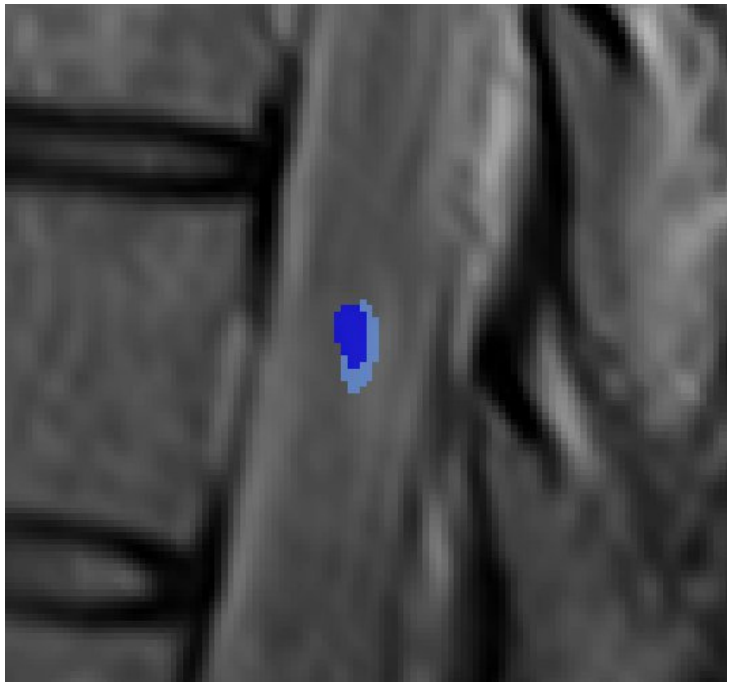
Z=7



Z=8

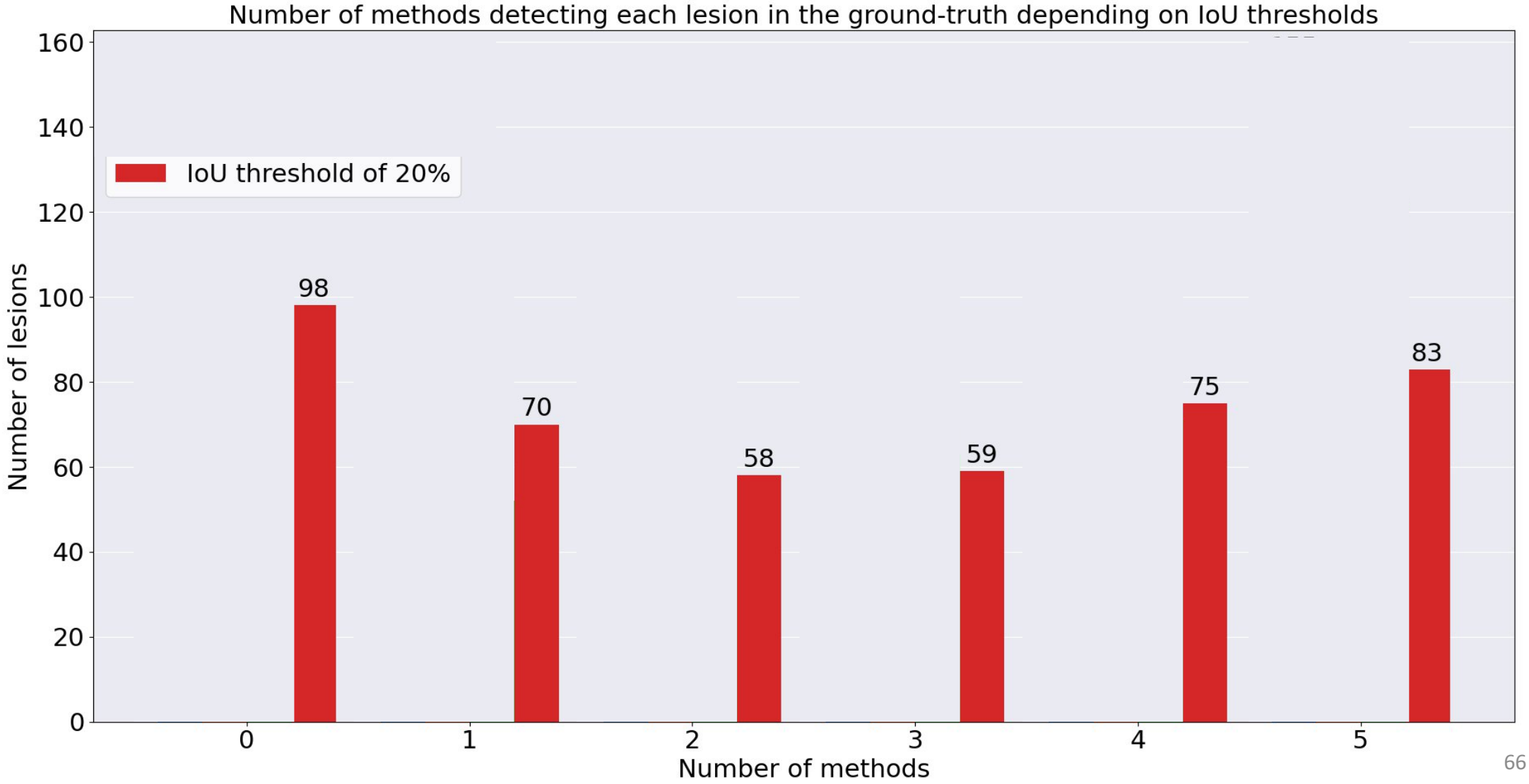


Z=9



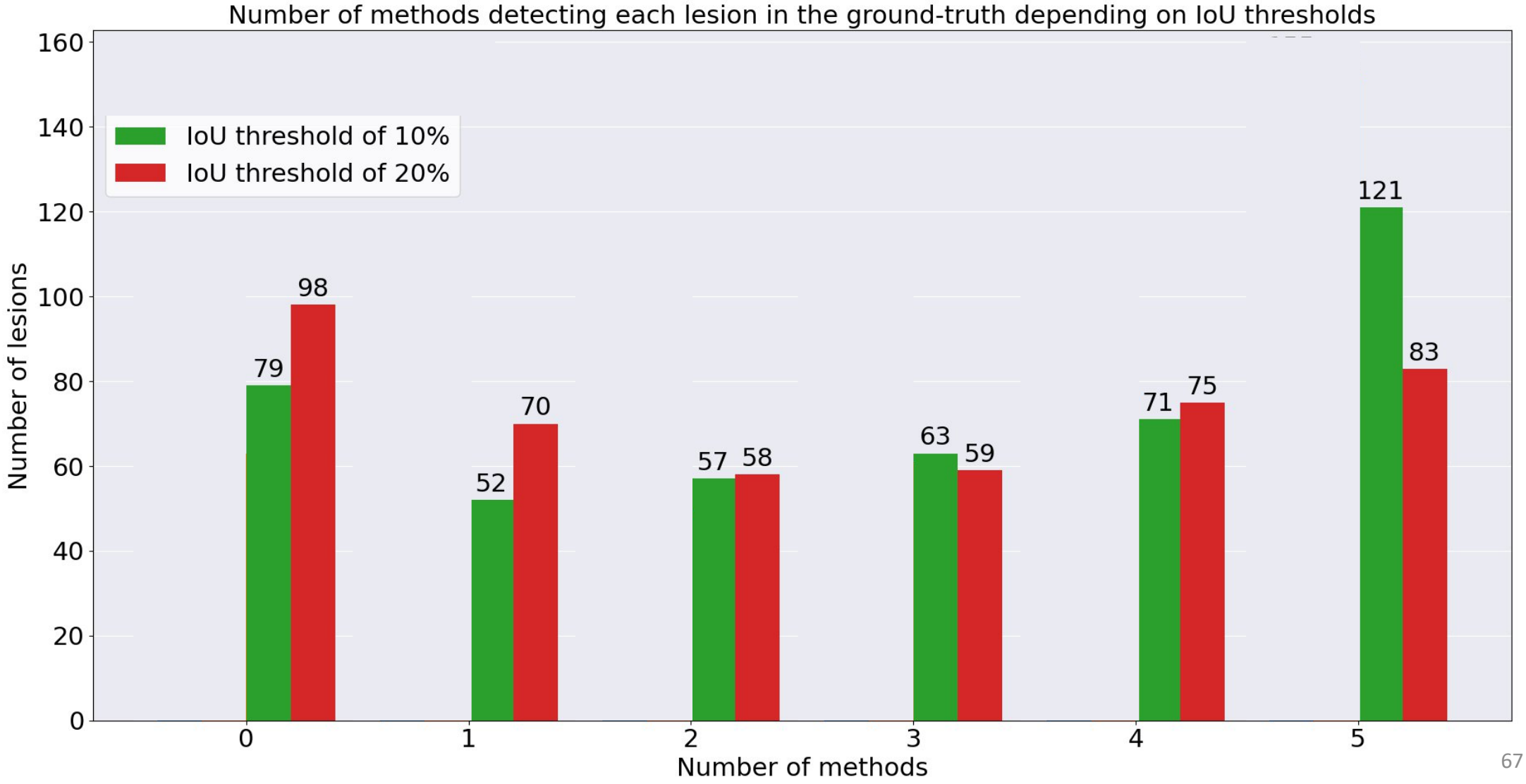
Results for overall test set (N=100)

Multiple IoU thresholds: 0%, 1%, 10% and 20% without decision threshold



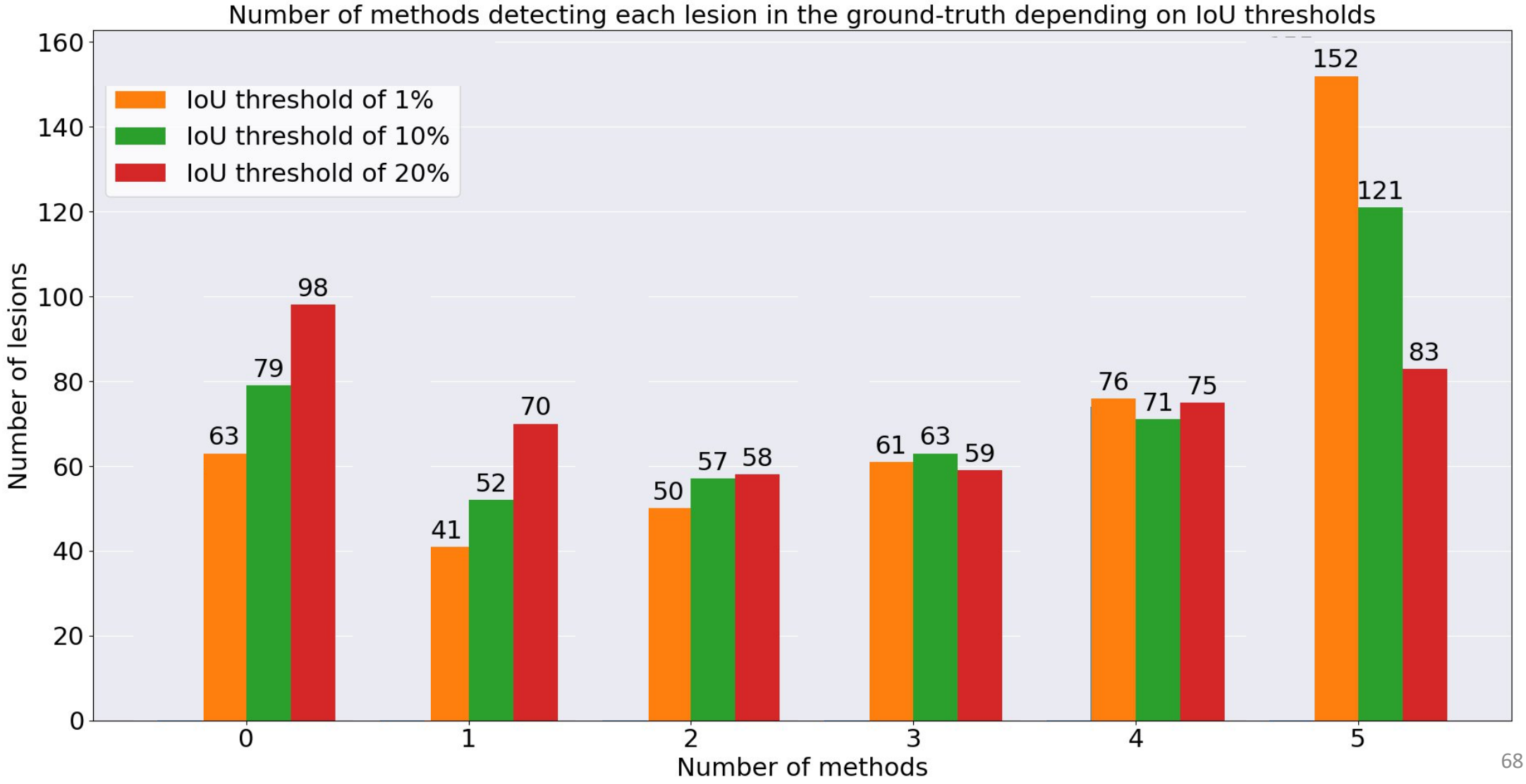
Results for overall test set (N=100)

Multiple IoU thresholds: 0%, 1%, 10% and 20% without decision threshold



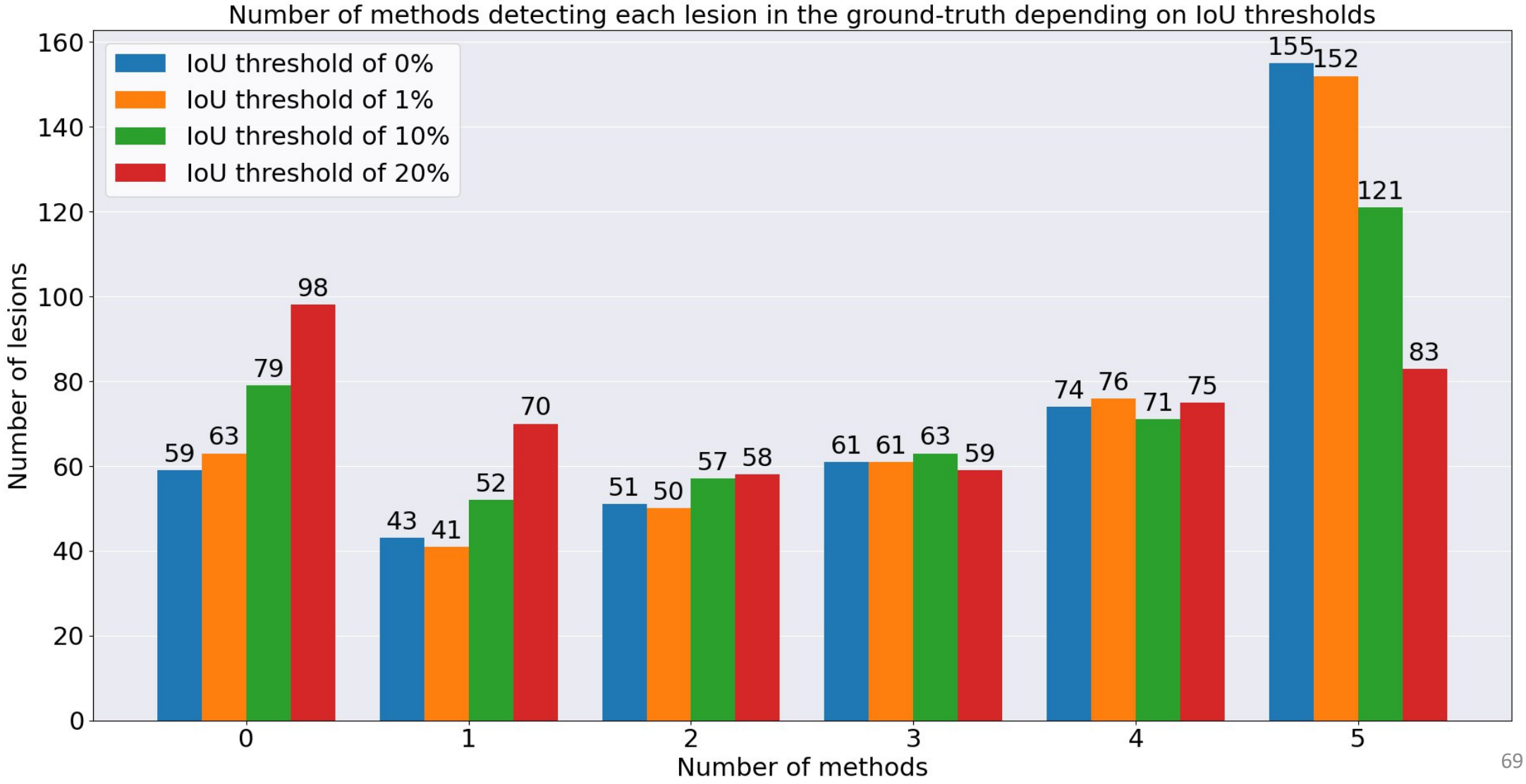
Results for overall test set (N=100)

Multiple IoU thresholds: 0%, 1%, 10% and 20% without decision threshold

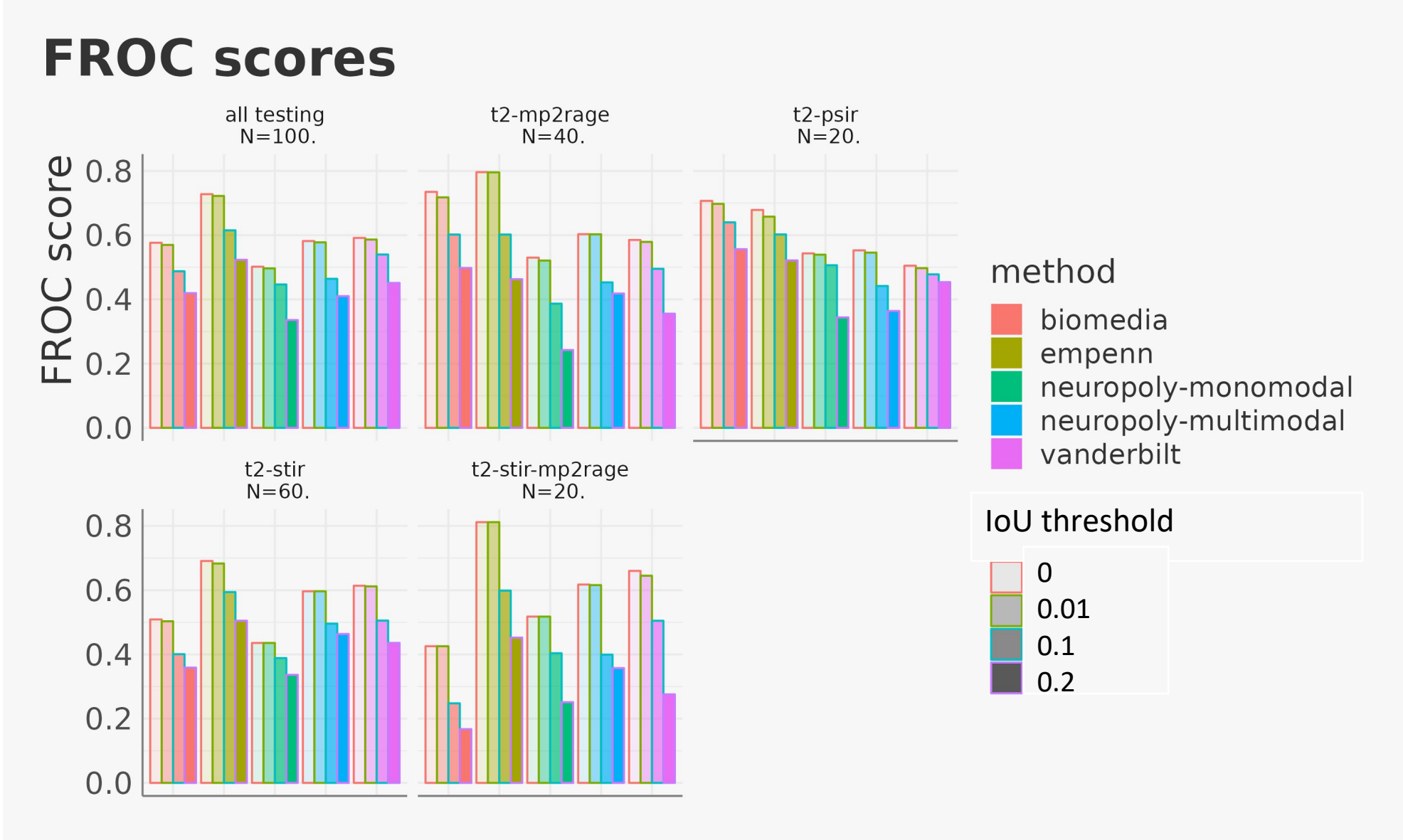


Results for overall test set (N=100)

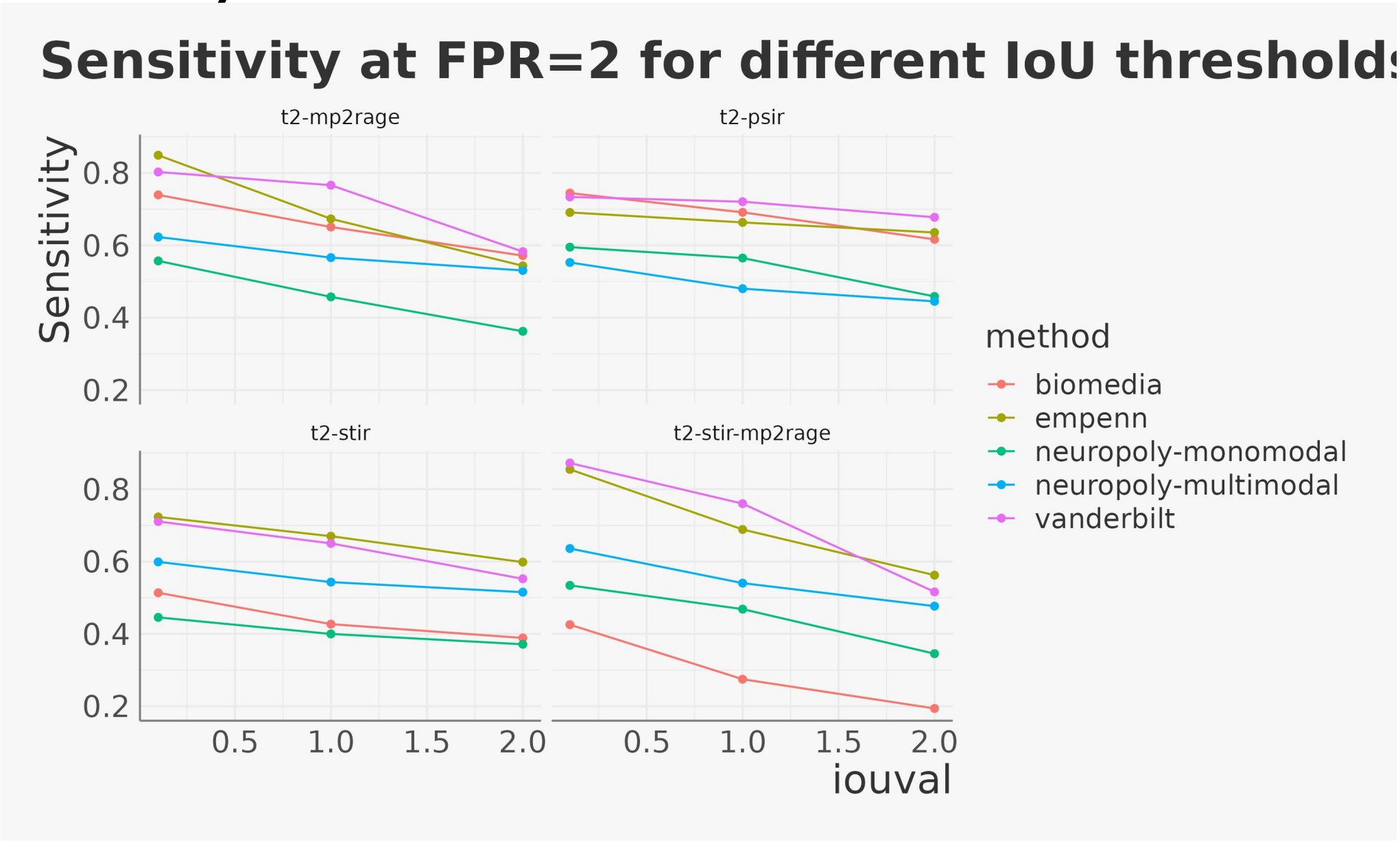
Multiple IoU thresholds: 0%, 1%, 10% and 20% without decision threshold



Dependency to IoU thresholds



Dependency to IoU thresholds



3.5 Characteristics of methods at FPR=1 and FPR=2

Performances of methods at FPR=1 (with IoU threshold 0.10)

Method	Sensitivity	Precision	F1	DICE	Median Lesion load error (mm^3)
BioMedIA	0.504	0.737	0.550	0.476	253.361
Empenn	0.638	0.711	0.612	0.584	202.535
Neuropoly (monomodal)	0.484	0.753	0.534	0.443	242.714
Neuropoly (multimodal)	0.549	0.768	0.588	0.524	270.197
Vanderbilt	0.614	0.722	0.578	0.532	211.224
nnUnet	0.604	0.719	0.601	0.543	191.146
sct	0.540	0.671	0.520	0.511	320.651

Performances of methods at FPR=2 (with IoU threshold 0.10)

Method	Sensitivity	Precision	F1	DICE	Median Lesion load error (mm^3)
BioMedIA	0.507	0.617	0.497	0.476	253.361
Empenn	0.682	0.602	0.581	0.588	213.746
Neuropoly (monomodal)	0.484	0.753	0.534	0.443	242.714
Neuropoly (multimodal)	0.549	0.768	0.588	0.524	270.197
Vanderbilt	0.739	0.597	0.588	0.590	144.328
nnUnet	0.617	0.591	0.549	0.547	184.615
sct	0.619	0.572	0.519	0.535	301.652

5. Discussion & Conclusion

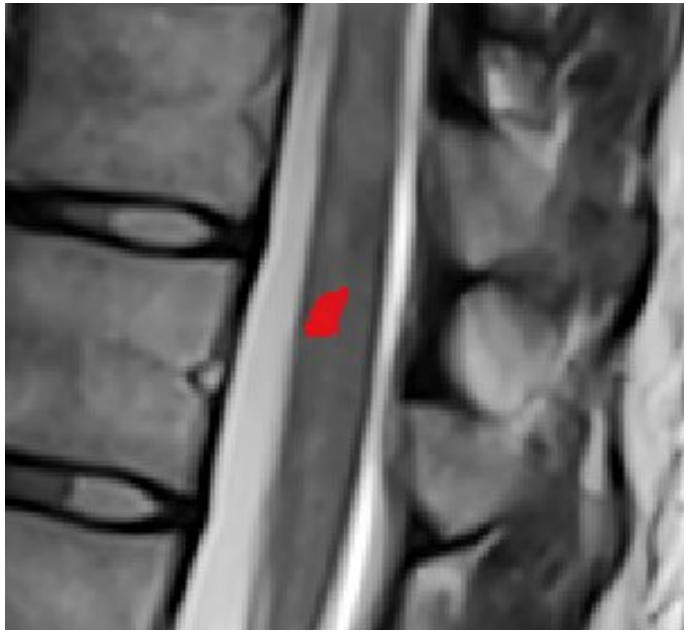
Overall

- First challenge on spinal cord lesion segmentation.
- Expert annotations are consistent with model inferences (as for now).
- We hope that the dataset will allow more research in this topic.

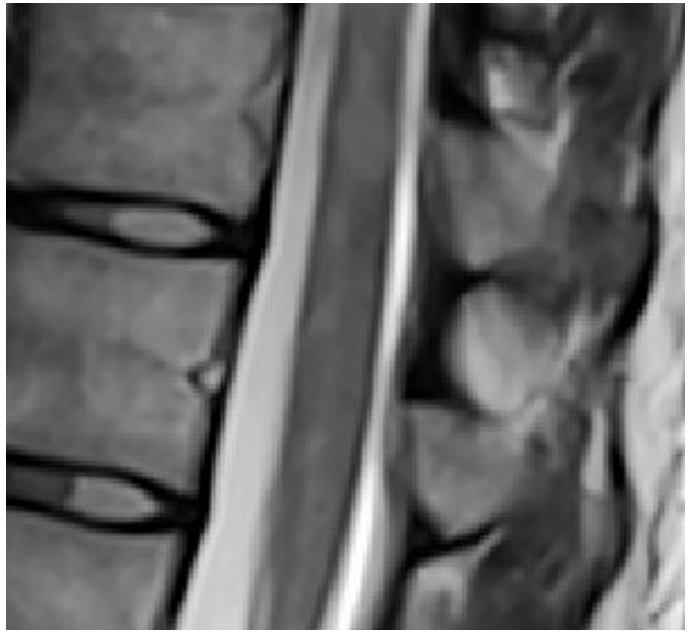
Results

- Overall, still a room for improving sensitivity (*see the examples of lesions detected by none of the methods even for the highest FPR*).

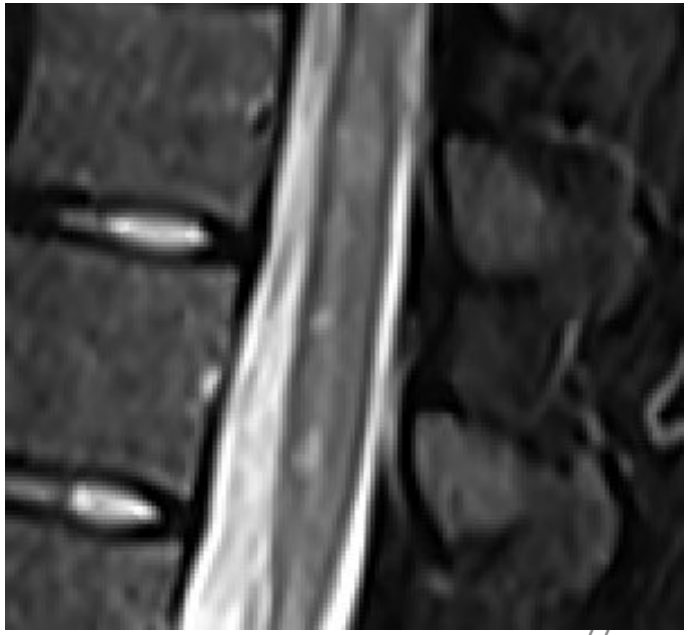
T2-w rawdata + GT segmentation:



T2-w rawdata:



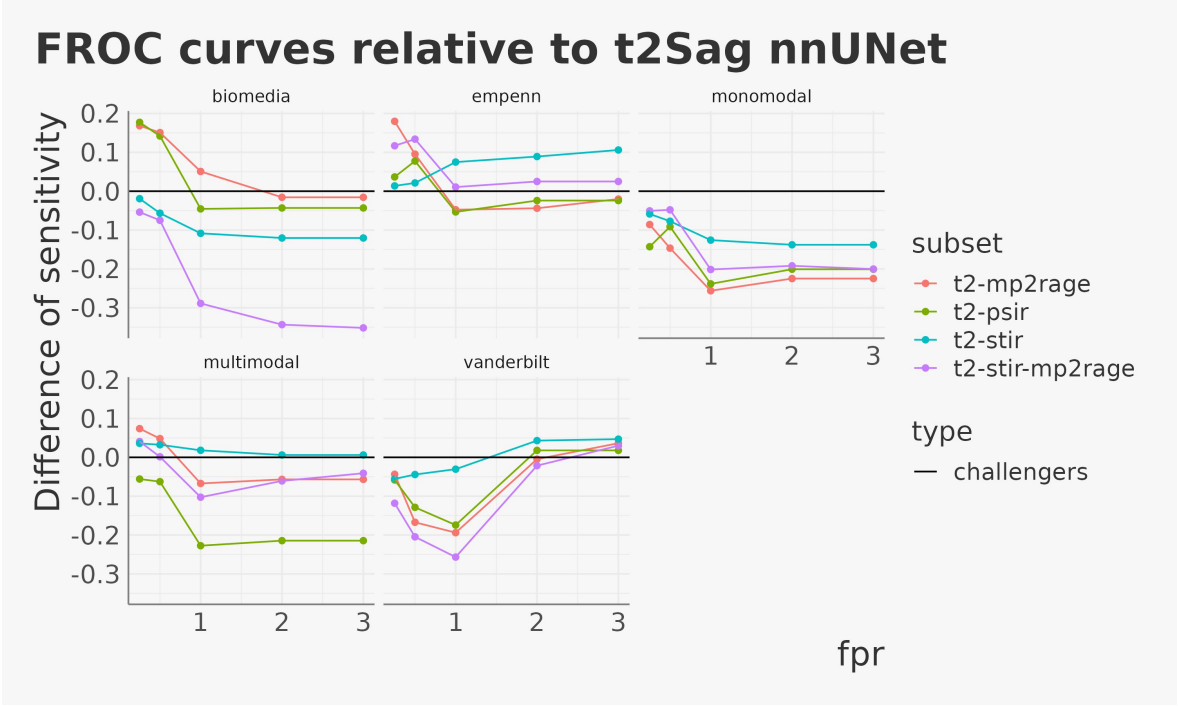
STIR rawdata:



Results

- Overall, still a room for improving sensitivity (*see the examples of lesions detected by none of the methods even for the highest FPR).*

- Still difficult to evidence that automated methods take the best of multisequences (but more experiments are needed in particular for different FPR).

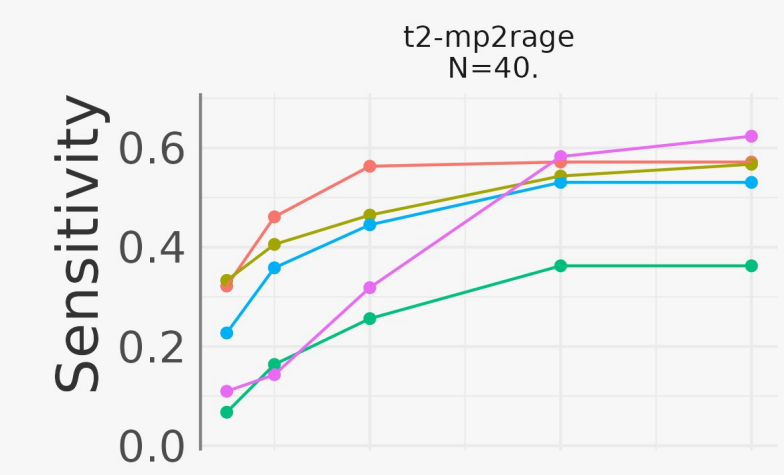


Results

- Overall, still a room for improving sensitivity (*see the examples of lesions detected by none of the methods even for the highest FPR).*

- Still difficult to evidence that automated methods take the best of multisequences (but more experiments are needed in particular for different FPR).

- Calibration is different from one method to another: some method performs better to other at lower FPR while other at higher FPR. The results suggest that this not just about having the best model, “probability calibration” may be an important factor.



Metrics

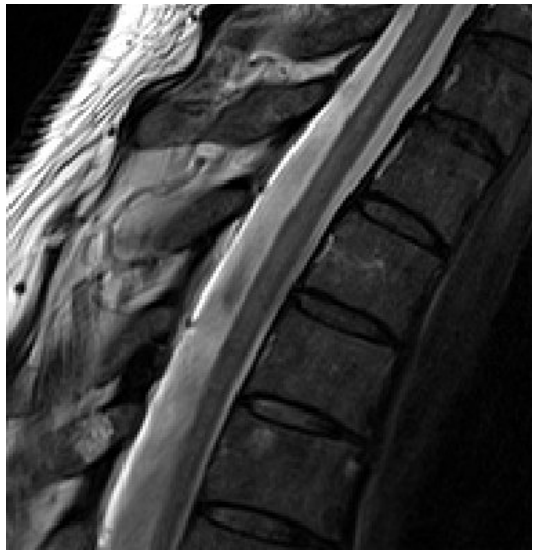
- The multi-threshold metric allowed to compare methods for some desired FPRs, which is very convenient for a principled comparison.

Metrics

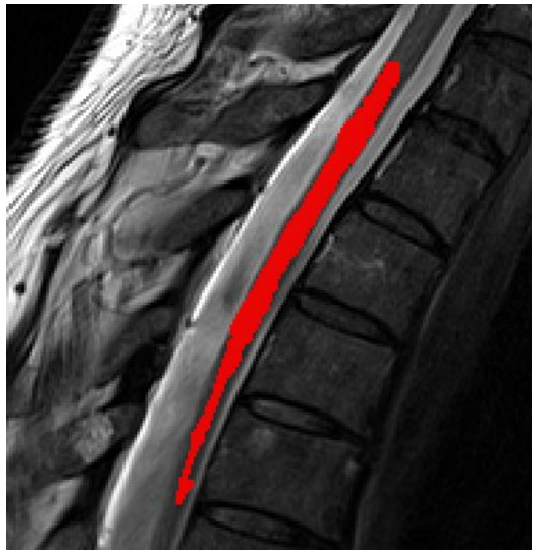
- The multi-threshold metric allowed to compare methods for some desired FPRs, which is very convenient for a principled comparison.

- Difficult to be fully satisfied with what the IoU based detection metric reflect. IoU threshold was a bit too high but whatever its values, cases with big “semi-contiguous” lesions are problematic.

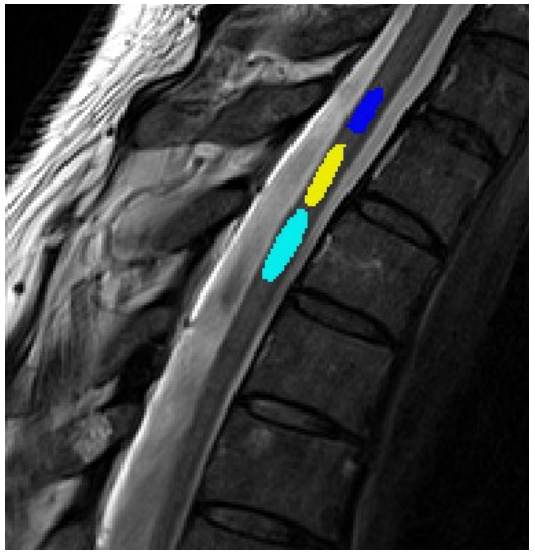
T2-w rawdata:



T2-w rawdata + GT segmentation:



Prediction:



Ongoing work

- Investigate deeper on the added-value of multisequence (which of the submitted pipelines can run with T2 alone?).

- Investigate effect of data characteristics on model performance (scanner brand, lesions characteristics, sagittal coverage, sequence combination).

- More principled statistical comparisons.

Thanks all for this challenge !

Organization team:

R. Casey, B. Combès, F. Cotton, M. Dojat, M. Kain, A. Kerbrat, S.Pop.

Actively working with:

L. Padé, G. Ambrosino

A. Bonnet, C. Meurée, G. Soisnard