Evaluating the lesion detection and segmentation performance of **LION v0.14.0** **in early breast cancer patients**

What is LION ?

- Central platform for segmenting tumors from whole-body **PET/CT** datasets.
- Developed by a team in the **Medical University of Vienna** (Manel Pires, Lalith Kumar Shiyam Sundar, Thomas Beyer).

MEDICAL UNIVERSITY OF VIENNA

What is LION ?

- Central platform for segmenting tumors from whole-body **PET/CT** datasets.
- Developed by a team in the **Medical University of Vienna** (Manel Pires, Lalith Kumar Shiyam Sundar, Thomas Beyer).
- Optimized for **FDG and PSMA PET/CT** images.
- Trained on multiple pathologies, lymphoma, melanoma, lung and breast cancers.

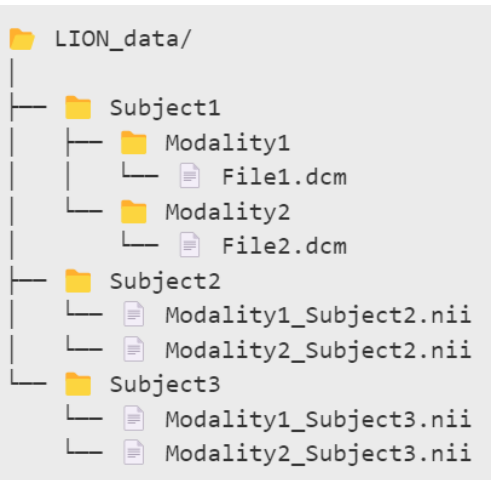
AutoPET Challenge	Medical University of Leipzig	Azienda Ospedaliero Universitaria Careggi	Institut Curie	Azienda Ospedaliera Hospital S. Croce e Carle	Universitätsklinikum Essen
N = 914	N = 370	N = 193	N = 108	N = 2763	N = 327
Lymphoma Melanoma Lung Cancer	Lung Cancer	Lung Cancer	Breast Cancer	Lymphoma	Mixed pathology

What is LION ?

- Central platform for segmenting tumors from whole-body **PET/CT** datasets.
- Developed by a team in the **Medical University of Vienna** (Manel Pires, Lalith Kumar Shiyam Sundar, Thomas Beyer).
- Optimized for **FDG and PSMA PET/CT** images.
- Trained on mixed pathologies, lymphoma, melanoma, lung and breast cancers.
- **Aim:** to streamline medical imaging tasks and enhance diagnostic capabilities.

How to use LION ?

- Installation: `pip install lionz`
- Python library: `import lionz`
- Command in a terminal: `lionz -d <path_to_image_dir> -m <model_name>`
- Thresholding option of SUV 4 for FDG and SUV 1 for PSMA: `lionz -d <path_to_image_dir> -m <model_name> -t`
- Compatible with both DICOM and NIFTI formats
- Need for specific data structure and naming conventions



Data

Population used to test LION:

- **Early (n=188) breast cancer (BC) patients** (age: 50±12 years, height: 163±14 cm, weight: 67±14 kg) **enrolled at Institut Curie (Paris, France)**, and for whom an FDG PET/CT scan (voxel size: 13.38±7.82 mm^3) before treatment, clinical and follow-up information were available. Molecular subtype was:

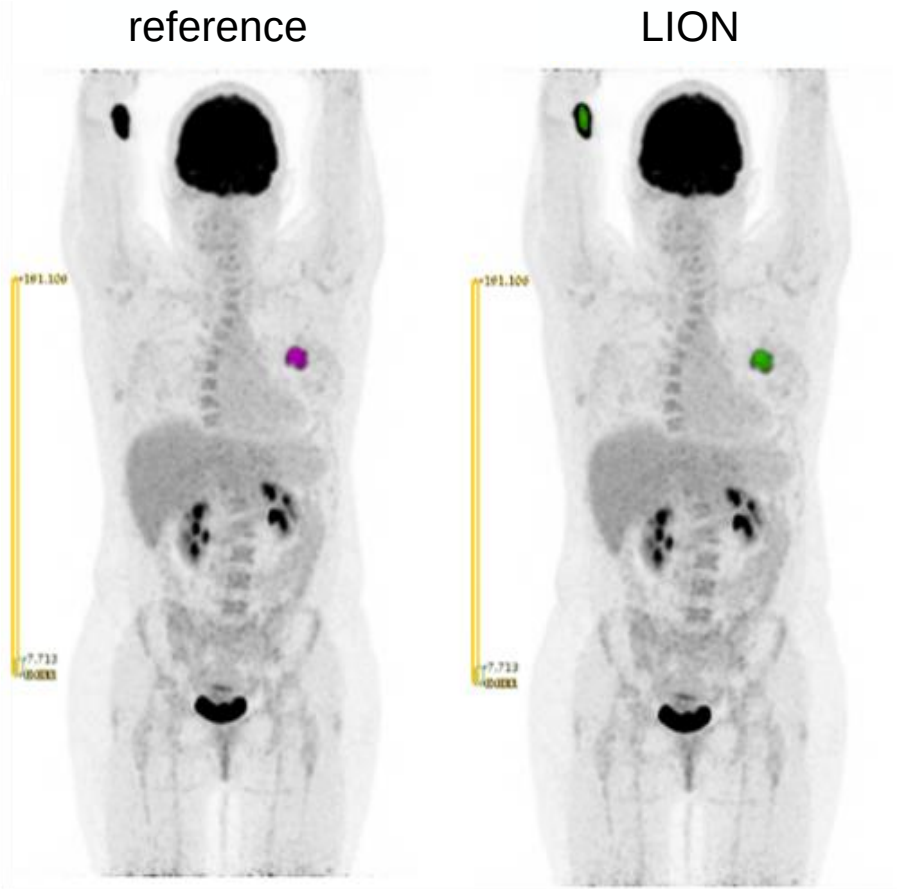
Database	Number of patients
Triple Negative (TNBC)	97
HR+HER2- (HR+)	52
HR+/-HER2+ (HER2)	39

Methods

Methods: Segmentation

Tumor segmentation:

- **Reference tumor segmentation:**
  - LION v0.8.1 segmentation
  - SUV>4 threshold with LIFEx*
  - Manual corrections performed using LIFEx*, based on nuclear medicine reports, by adding or removing regions. Each lesion was labelled as primitive tumor, lymph nodes or metastases.
- **LION tumor segmentation:**
  - LION v0.14.0 segmentation
  - SUV>4 threshold with LIFEx*



Methods: Segmentation

Segmentation of regions of interest: To assess segmentation performance at a **compartment-specific level**, we segmented the following **organs and tissues** with **TotalSegmentator***:

- total : 117 compartments
- tissues : 3 compartments
  - Subcutaneous fat (FatSC)
  - Torso/Visceral fat (FatV)
  - Skeletal muscles (Muscles)
- breast : 1 compartment

⇨ **121 regions of interest**

* Wasserthal, J., Breit, H.-C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D., Cyriac, J., Yang, S., Bach, M., Segeroth, M., 2023. TotalSegmentator: Robust Segmentation of 104 Anatomic Structures in CT Images. Radiology: Artificial Intelligence. https://doi.org/10.1148/ryai.230024

Methods: Evaluation methodology

1. Detection Performance

We checked whether each tumor lesion was detected (true positive), missed (false negative), added (false positive)

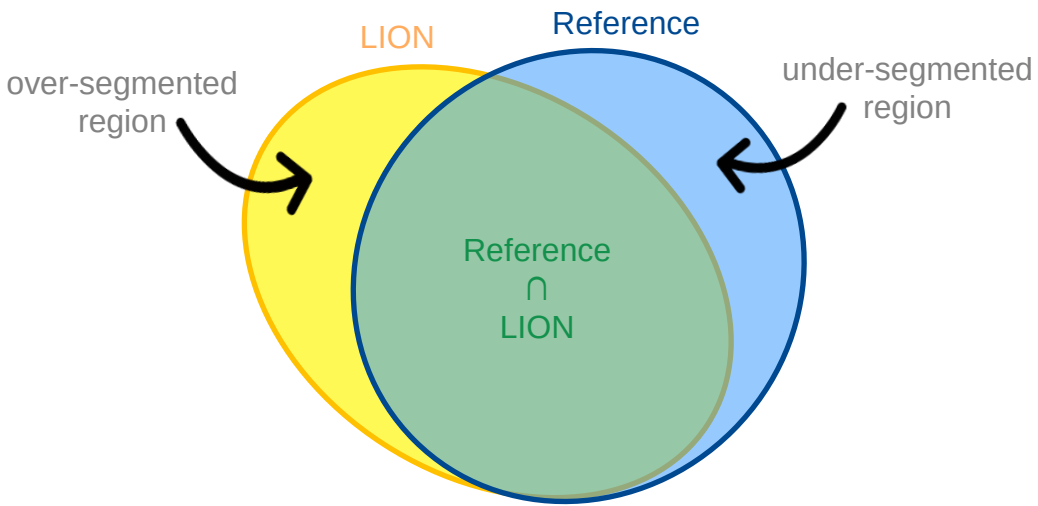
- at patient/body scale and at compartment scale.
- for all patients and by separating them into single-site and multi-site patients.

Methods: Evaluation methodology

2. Segmentation Accuracy, allowing a 10% volume tolerance or a 10 mL tolerance for tumors larger than 100 mL

We determined whether segmentations were accurate (within 10% or 10 mL), too large (over-segmented) or too small (under-segmented) by calculating the volume common to the reference and LION segmentations,

- at patient/body scale and at compartment scale.
- for all patients and by separating them into single-site and multi-site patients.



Methods: Evaluation methodology

3. PET Features analysis

We compared for both reference and LION segmentations (at lesion scale and WB scale):

- Metabolic Tumor Volume (MTV)
- Total Metabolic Tumor Volume (TMTV),
- Maximum Distance Between Tumor Lesions (Dmax),
- SUVmax,
- SUVmean.

Methods: Evaluation methodology

3. PET Features analysis

We compared for both reference and LION segmentations (at lesion scale and WB scale):

- Metabolic Tumor Volume (MTV)
- Total Metabolic Tumor Volume (TMTV),
- Maximum Distance Between Tumor Lesions (Dmax),
- SUVmax,
- SUVmean.

4. Dice Similarity Coefficient Calculation

We compared the reference and LION segmentations with DICE coefficient.

Results

Results: Data

Lesion distribution:

- 8 patients without any lesions with SUV ≥ 4 (in reference segmentation).

- Among patients with a lesion with a SUV ≥ 4:
  - 70 patients had only a primary tumor = single-site group.
  - 110 patients had several lesions (lymph nodes or metastases) = multi-site group.

- 24 patients with an activation of brown fat according to their medical report.

- 32 patients with multifocal cancer in the breast (presence of several lesions in the breast area).

Results: Detection performance

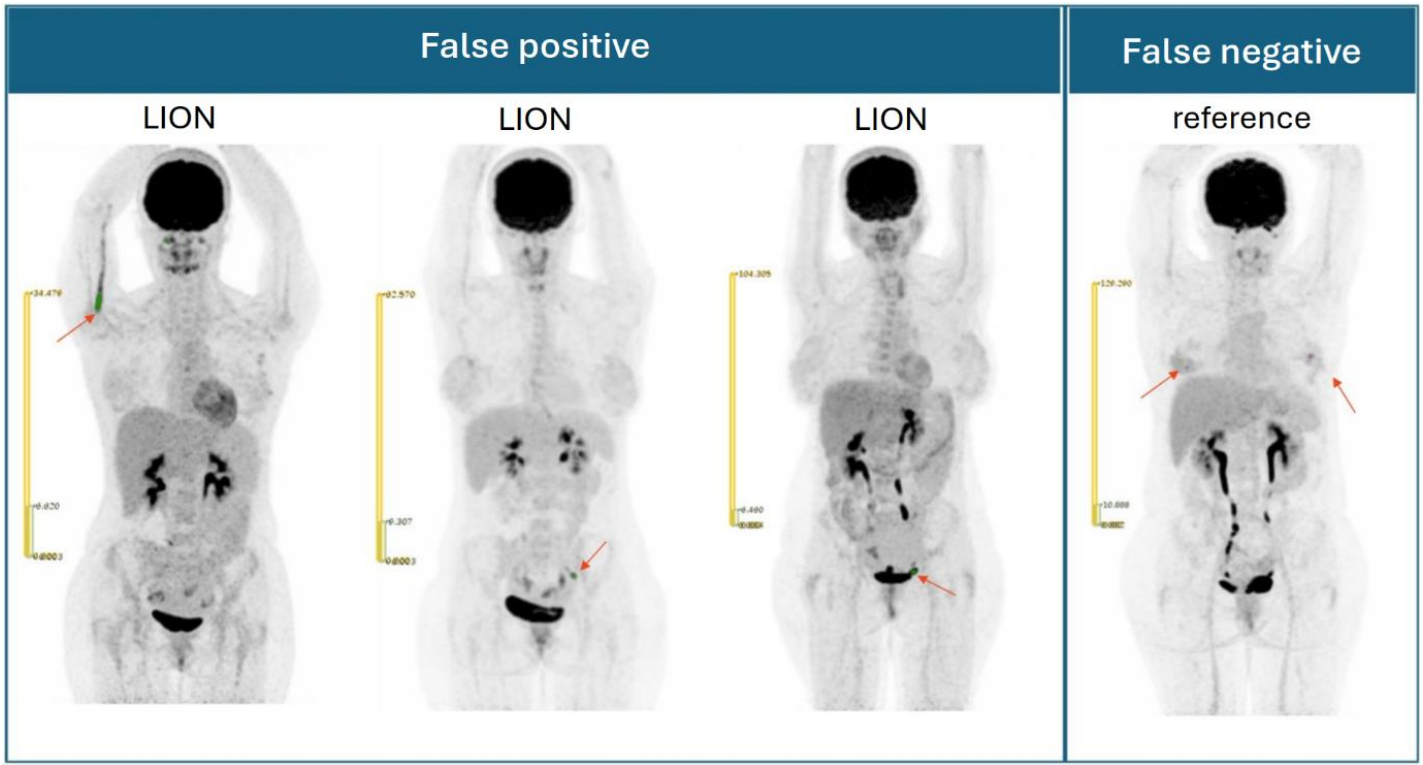
- **LION detected most lesions**, particularly in patients with a single lesion.
- **Twice as many false positives as false negatives per patient.**

FOR ALL PATIENTS	ALL	SINGLE-SITE	MULTI-SITE
Average % lesions detected (sensitivity)	88%	97%	82%
Average number of false positive	1.9	1.5	2.1
Average number of false negative	0.6	0.1	1.0
Number of patients	183	70	110

Results: errors at the patient level

Among the 188 patients:

- **3 patients were false positives** (no lesion at all while LION found some)
- **1 patient was a false negative** = small lesion missed by LION

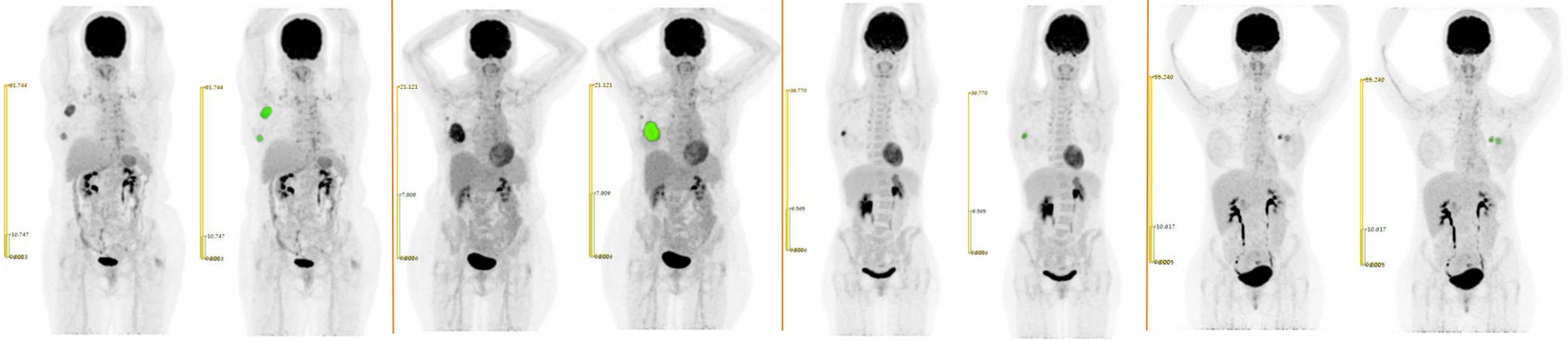


Results: Brown fat issue

24 cases with brown fat activation:

- **Correct segmentation** for **7 patients** (SUVlesion > SUVbrownfat).

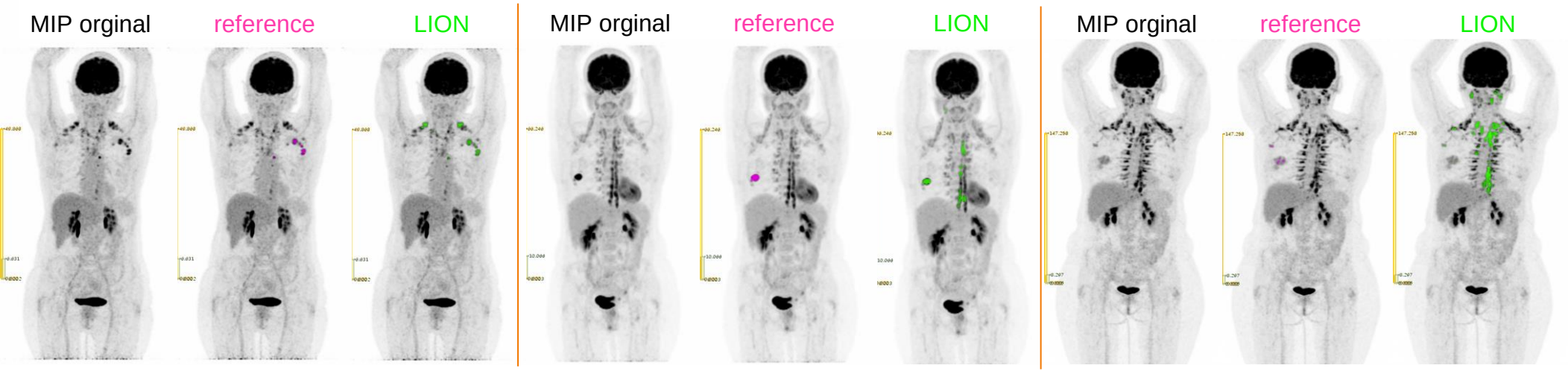
MIP orginal LION MIP orginal LION MIP orginal LION MIP orginal LION



Results: Brown fat issue

24 cases with brown fat activation:

- In **15 other patients**, **LION segmented lesions <u>and</u> brown fat**
  - **overestimation of TMTV (8±14 mL)**



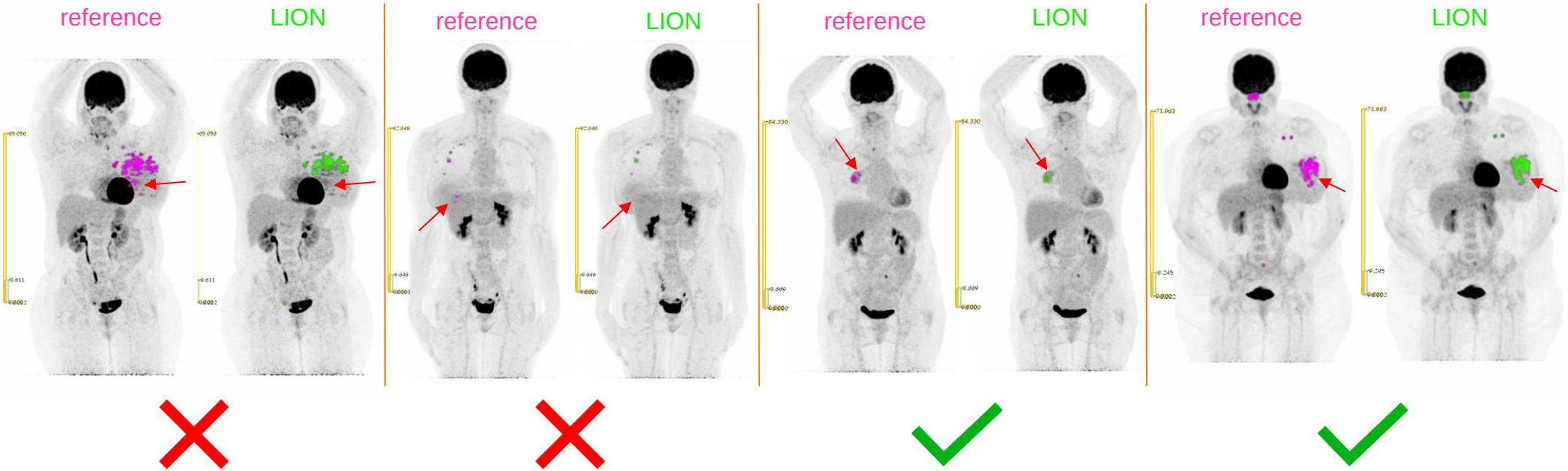
Results: Brown fat issue

24 cases with brown fat activation:

- **LION missed the lesions** and segmented brown fat **for 2 patients** with **SUVlesion < SUVbrownfat**.

Results: Fragmented lesions

Some lesions have a fragmented pattern and sometimes LION is unable to segment these lesions correctly:



Results: Detection performance

- **LION better detected primary tumors compared to lymph nodes**.

	Primary tumor	Lymph nodes
% detected at lesion scale	95%	82%
% detected at patient scale	94%	62%
Number of patients	173	86

Results: Detection performance

At the compartment scale:

	Muscles	FatSC	FatV	Breast
Average percentage of detected lesions *	83%	68%	45%	87%
Average number of false positive	2	1	1	0
Number of patients with lesions in this tissue compartment	160	159	9	177

- **Lesions in the lung and bone compartments were rare** in these early-stage patients; however, **bone lesions were generally well detected, while lung lesions were often missed**.

** Calculated only on patients with lesions in the specific tissue compartment (last row).*

Results: False positive

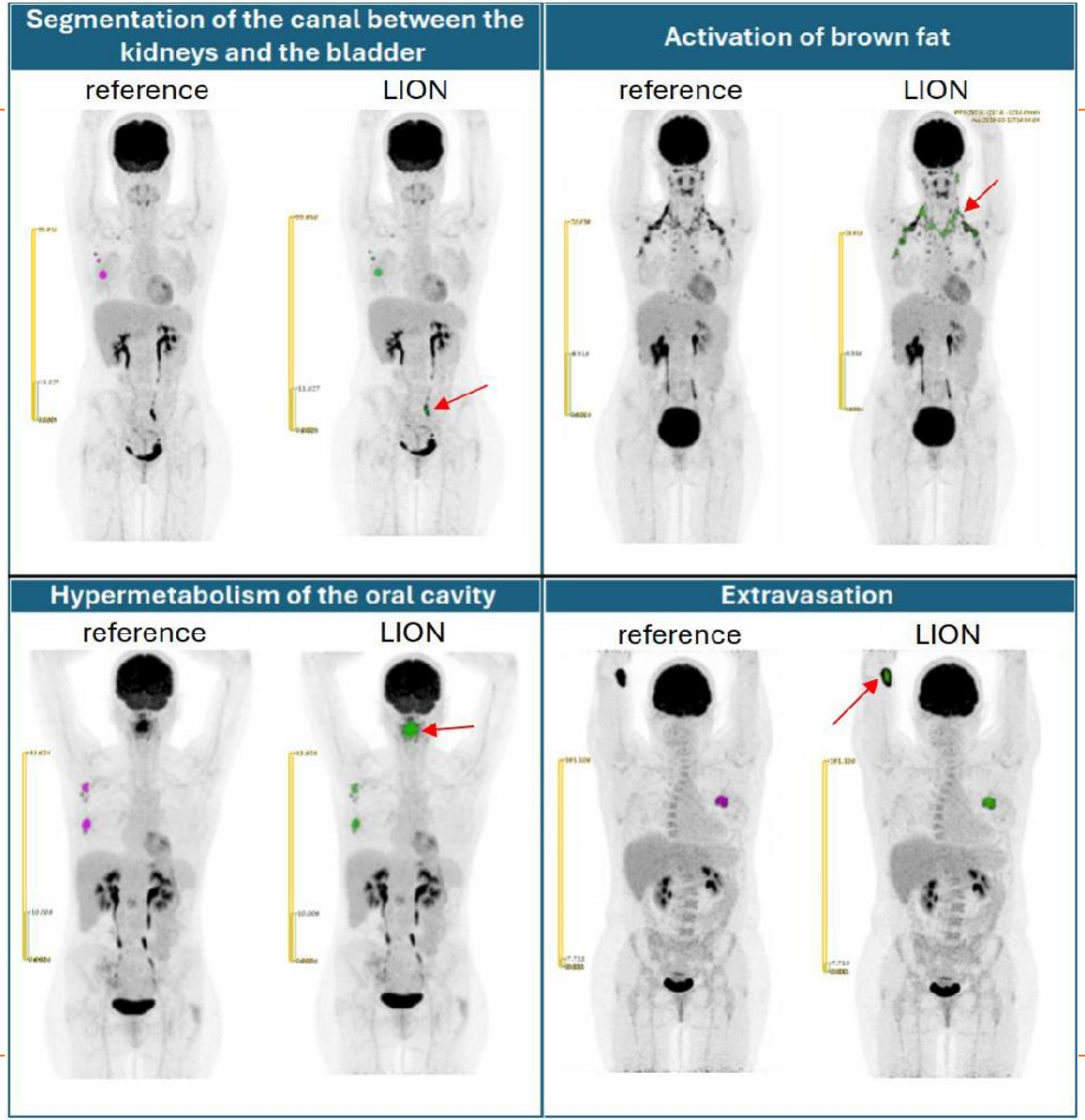
- False positives **mainly in the fat and muscle compartments outside the breast region.**

	Muscles	FatSC	FatV	Breast
Average number of false positive per patient	2	1	1	0
Number of patients with false positive lesions	65	46	22	3
Number of patients with lesions in this tissue compartment	160	159	9	177

- **False positive characteristics:**
  - found in 95 patients
  - MTV: 1.01±2.56 mL (84±222 voxels)
  - SUVmax: 9.38±10.95

Results: False positive

- False positive:
  - **Extravasation**
  - **Physiological uptake**
  - **Brown fat activation**
  - **Non pathological uptake**



Results: False negative

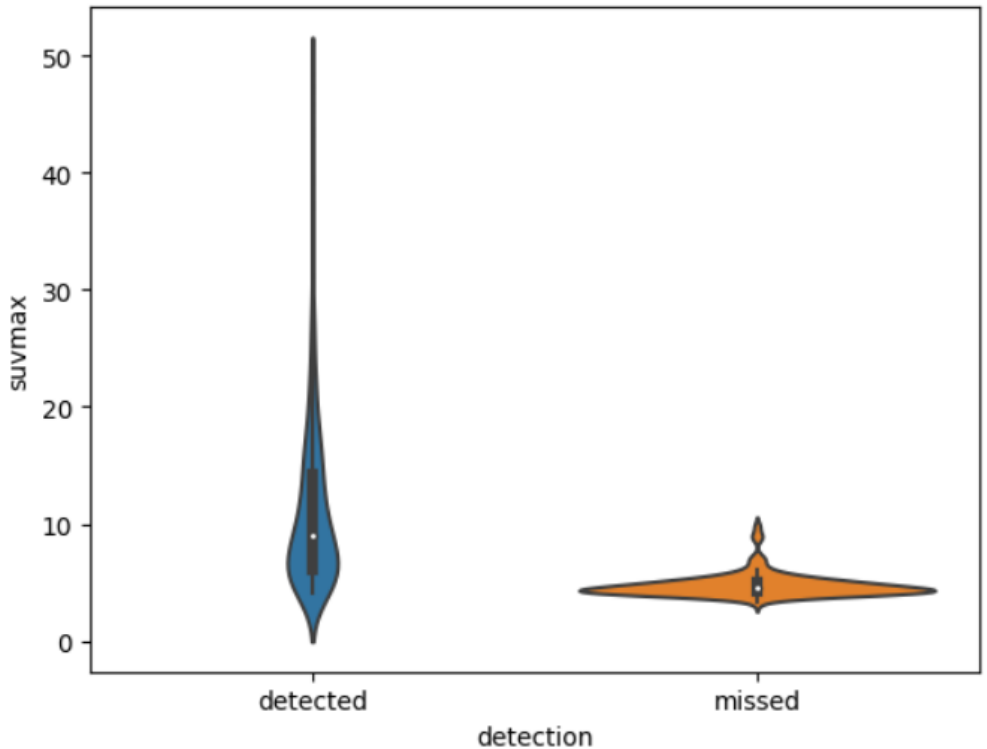
- False negatives **occurred mainly in the fat** compartments (fatSC and fatV).

	Muscles	FatSC	FatV	Breast
Number of patients with false negative lesions	66	128	6	48
Number of patients with lesions in this tissue compartment	160	159	9	177

- False negatives mainly occurs for **fragmented/diffuse lesions, small lesions or lesions with low intensity:**
  - found in 54 patients
  - MTV: 0.38±1.27 mL (31±138 voxels)
  - SUVmax: 4.77±0.98

Results: local PET Features

- **SUVmax was correctly estimated** in 97% of detected lesions.
- Lesions are **less well detected when SUVmax is low**.

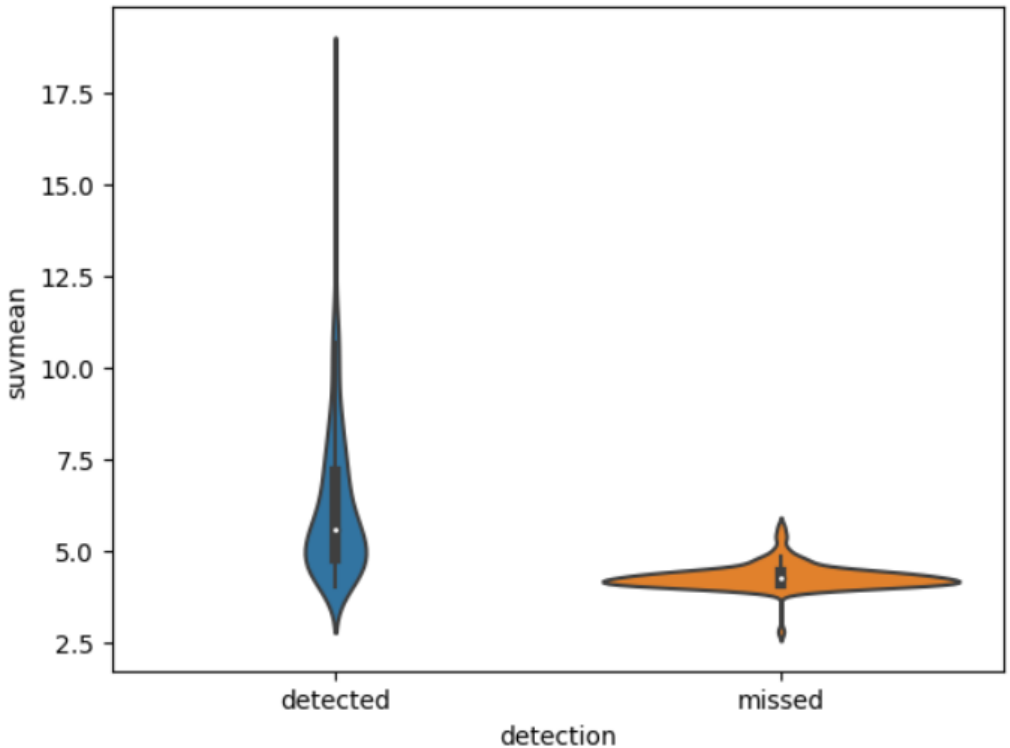


Quartiles (SUVmax)	% of detected lesions
Q1 (<=4.98)	39%
Q2 (4.98-7.38)	81%
Q3 (7.38-12.48)	98%
Q4 (>12.48)	100%

Note: 142 lesions per quartile

Results: local PET Features

- **SUVmean was well estimated** within 10% in 86% of the detected lesions.
- Lesions are **less well detected when SUVmean is low.**

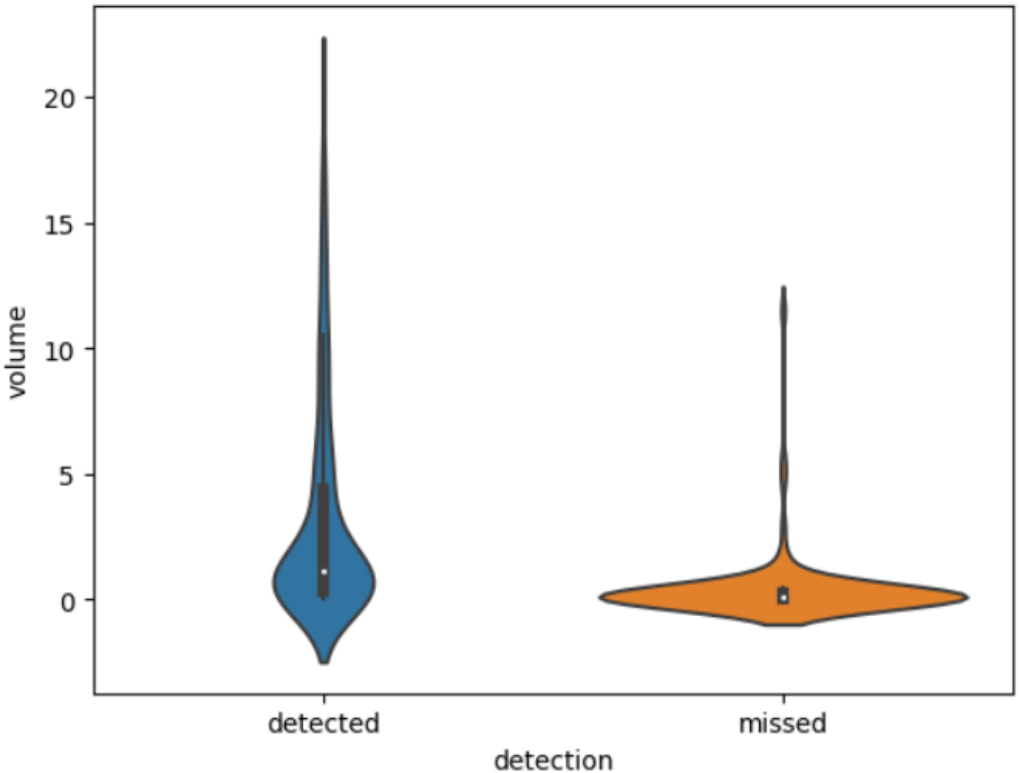


Quartiles (SUVmean)	% of detected lesions
Q1 (<=4.41)	37%
Q2 (4.41-5.14)	82%
Q3 (5.14-6.66)	98%
Q4 (>6.66)	100%

Note: 142 lesions per quartile

Results: local PET Features

- **MTV was well estimated** within 10% in 41% of the detected lesions.
- **Lesions** are **less well detected** when the reference **MTV is low**.



Quartiles (MTV in mL)	Quartiles (MTV in vx)	% of detected lesions
Q1 (<=0.18)	Q1 (<=14)	40%
Q2 (0.18-0.74)	Q2 (14-62)	84%
Q3 (0.74-4.19)	Q3 (62-303)	96%
Q4 (>4.19)	Q4 (>303)	98%

Note: 142 lesions per quartile

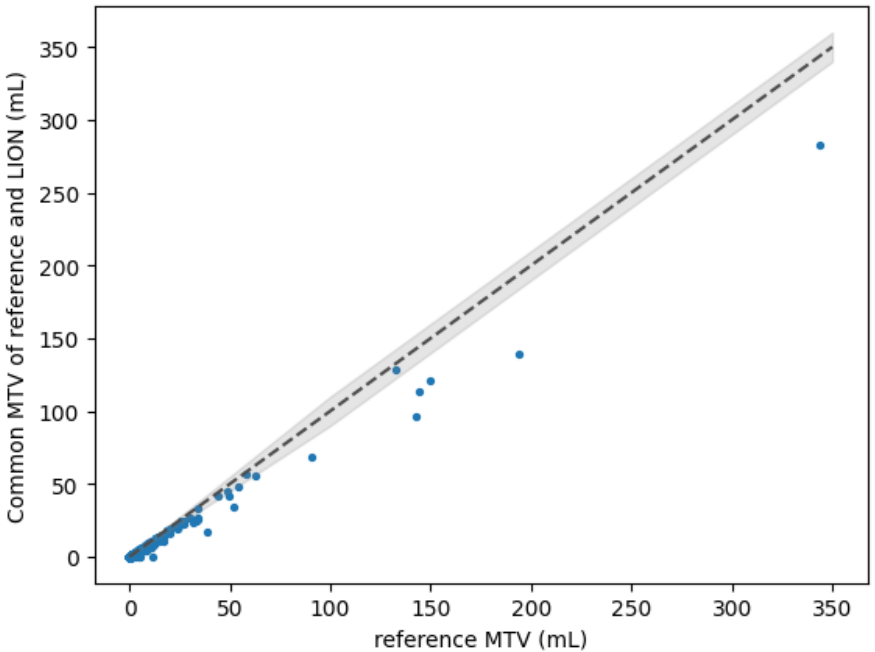
Results: Segmentation

- **LION consistently under-segmented** the lesions.

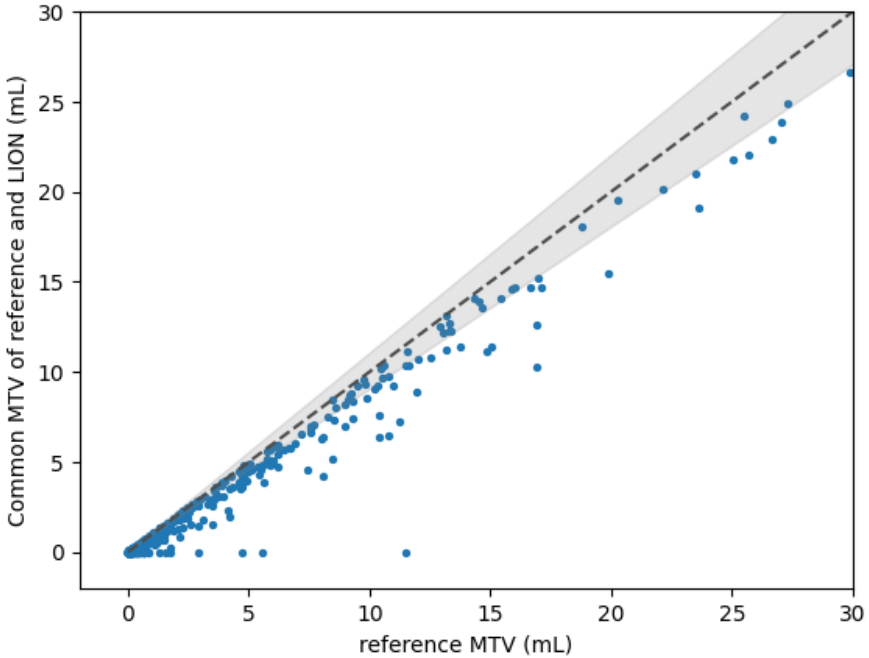
	ALL
Lesions well segmented	38%
Lesions under-segmented	50%
Lesions over-segmented	0%
Number of lesions (reference-LION)	570-817
Number of patients	179

Results: Segmentation and local PET features

- **LION consistently under-segmented** the lesions.

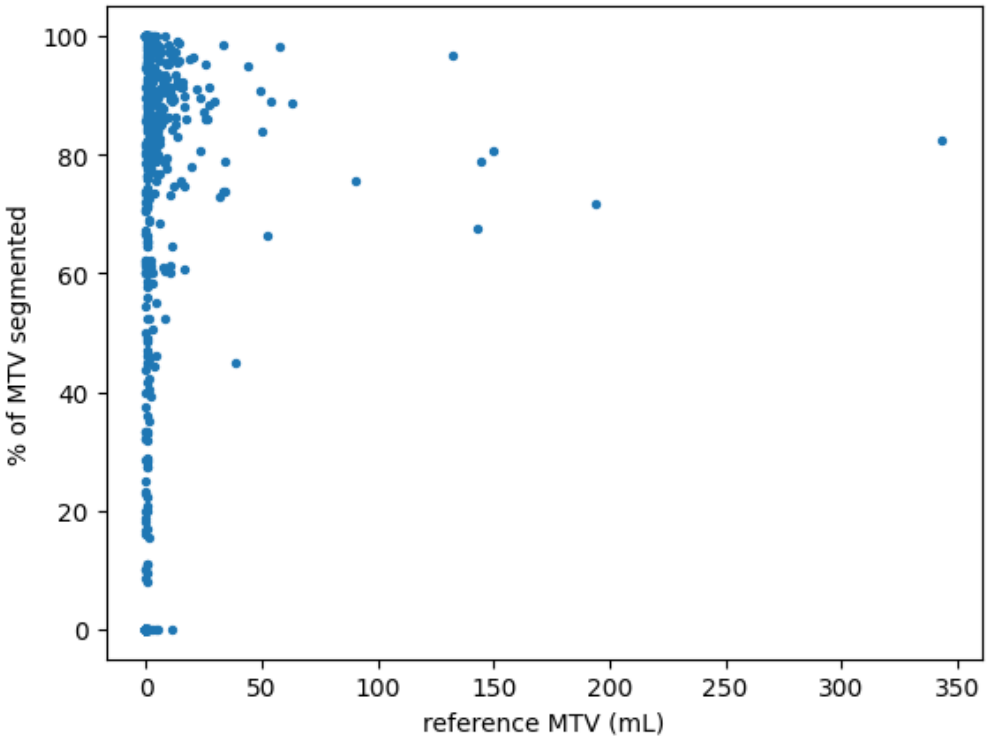


ZOOM



Results: Segmentation and local PET features

- **MTV** is **less well segmented** when the reference **MTV is low**.



Quartiles (MTV in mL)	Quartiles (MTV in vx)	% of MTV segmented for detected lesions
Q1 (<=0.18)	Q1 (<=14)	29%
Q2 (0.18-0.74)	Q2 (14-62)	60%
Q3 (0.74-4.19)	Q3 (62-303)	81%
Q4 (>4.19)	Q4 (>303)	84%

Note: ~142 lesions per quartiles

Results: Segmentation

- **Better segmentation performance for primary tumors than for ADPs**, with a success rate twice as high for primary tumors.

	PRIMITIVE	ADP
Lesion well segmented	45%	26%
Lesion under-segmented	50%	56%
Lesion over-segmented	0%	0%
Number of lesions	220*-799	270-534
Number of patients	173	86

* 12 patients had more than one lesion in their reference PRIMITIVE segmentation because, during correction, the lesions were grouped together as a single lesion because they were very close and appeared to come from the same lesion.

Results: global PET Features and Dice Score

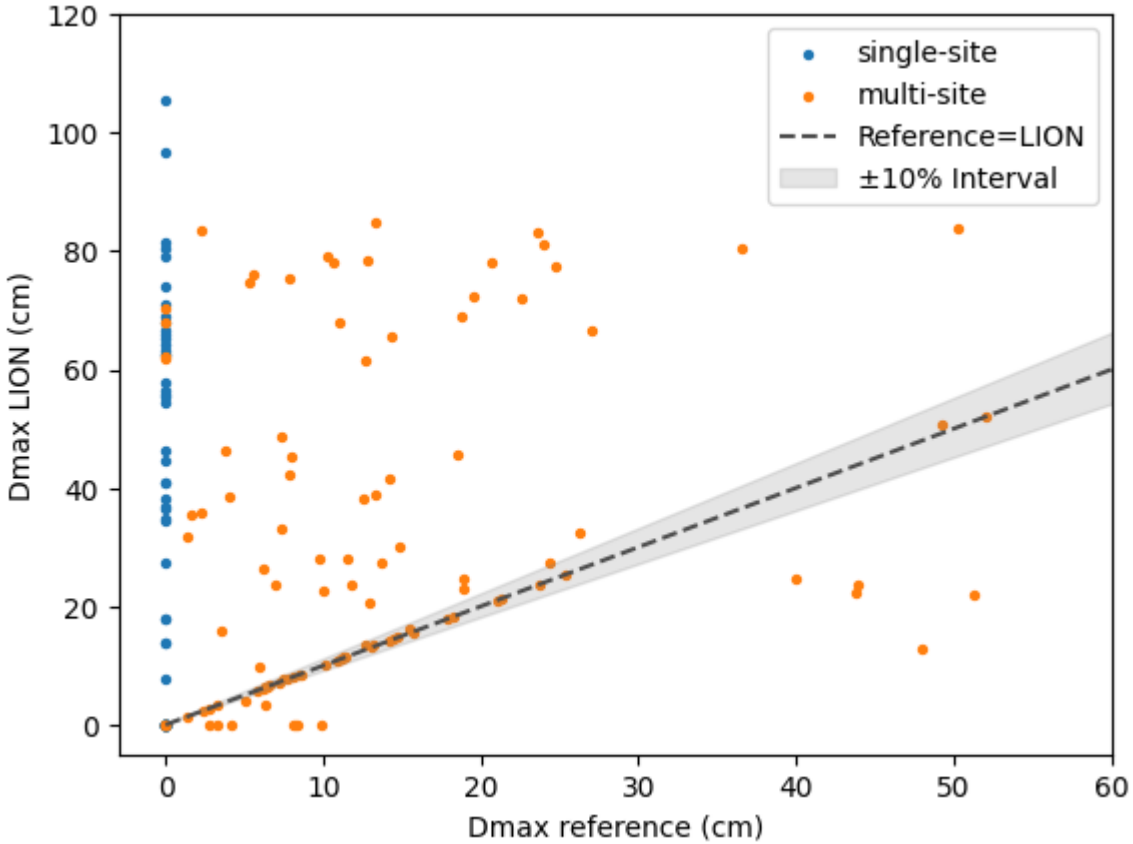
- Mean **Dice score** **of 0.84 for all groups**.
- **TMTV** **was well estimated** within 10% in 42% of the patients.
- **Dmax** **and** **TMTV** **had relatively low concordance rates (~40%).**

	ALL
% of patients with correct **Dmax** within 10%	42%
Mean±std difference in the other patients	40±25 cm
Number of patients by group	180

	ALL
% of patients with correct **TMTV** within 10%	42%
Mean±std difference in the other patients	8±13 mL
Number of patients by group	180

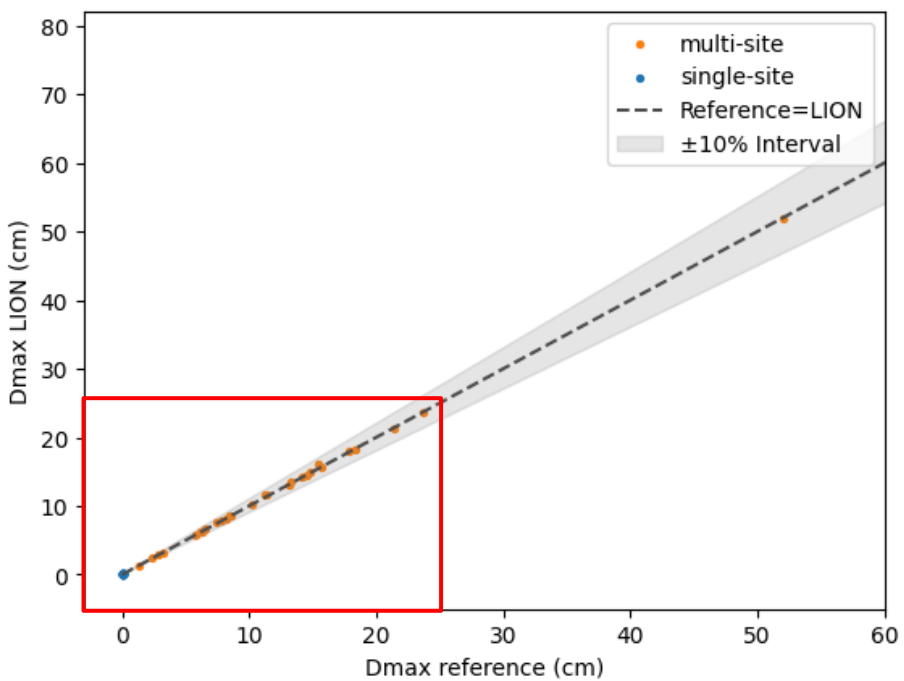
Results: global PET Features

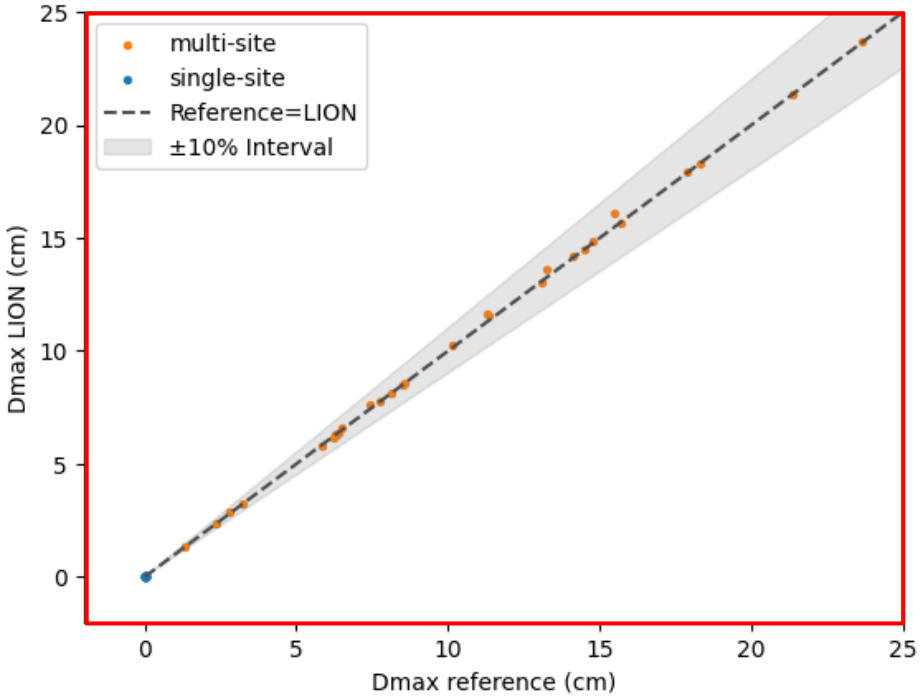
- **Errors in Dmax** due to false positive and false negative lesions



Results: global PET Features

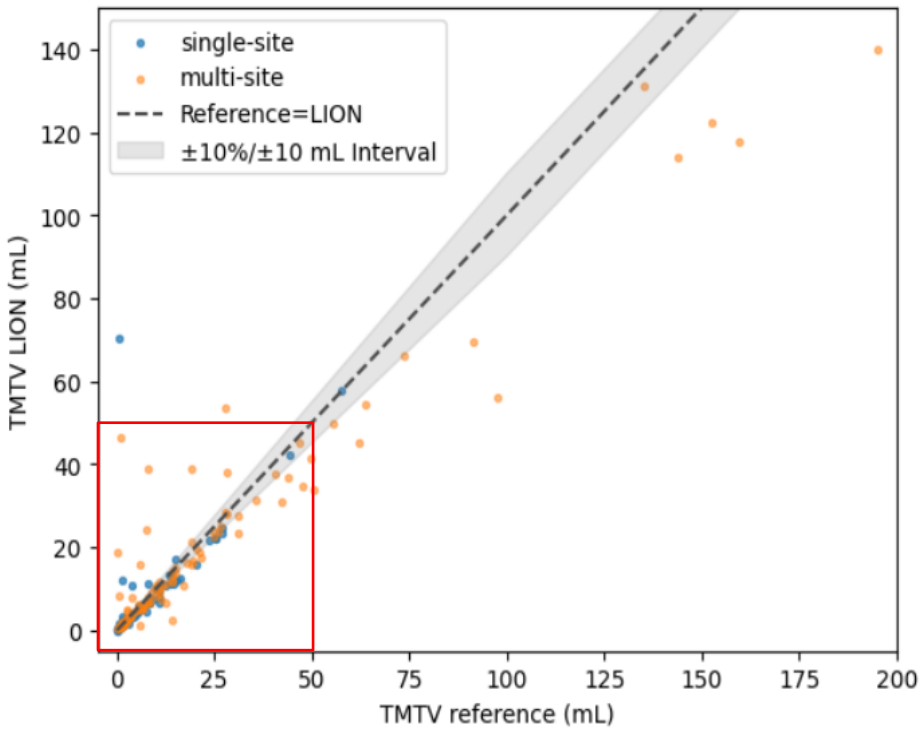
- **Dmax accuracy of 100%** within 10%, for patients **without false positive and false negative lesions** (n = 64 = 32 single-site + 32 multi-site)

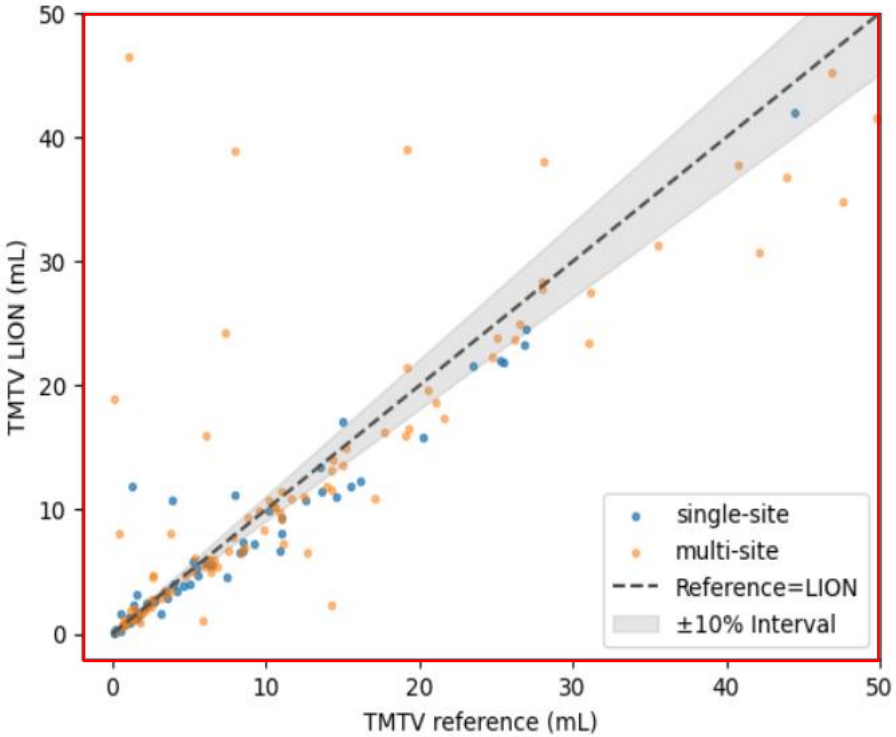




Results: global PET features

- **Yet TMTV was not always underestimated because of false positive lesions**





Results comparison

Similar results to the preprint:

Maria C. Ferrández, Sandeep S. V. Golla, Sara C. A. De Visser et al. **Evaluation of an artificial intelligence method for lesion segmentation of baseline FDG PET studies of DLBCL patients**, 26 March 2025, PREPRINT (Version 1) available at Research Square [https://doi.org/10.21203/rs.3.rs-6294601/v1]

In their method, they compare manual correction with a threshold of 4 SUV (**SUV4** method) with 2 other methods:

- **LIONZ**: LION without post-processing
- **LIONZ^{SUV4}**: LION + shrinking/growing region based on threshold of 4 SUV **+ manual correction of false positive lesions** (but not false negative lesions).

Results comparison

Similar results to the preprint:

- Fragmented lesions and under segmentation: *"Moreover, in cases with **largely disseminated tumors and smooth tumor borders**, LIONZ tended to **under-segment lesions** leaving out regions of interest that should be included in the segmentation (i.e. under-segmentation). [...] In largely **diffused tumors**, LIONZ **failed to identify the tumor borders** and leads to a large underestimation of the tumor region."*

- DICE (mean=0.84): *"The DSC was calculated for all segmentations from LIONZ and LIONZ^{SUV4} with SUV4.0 as the reference segmentation. The median DSC and interquartile range (IQR) resulted in 0.77 (0.64 - 0.84) for LIONZ and **0.87 (0.80 - 0.93) for LIONZ^{SUV4}**."*

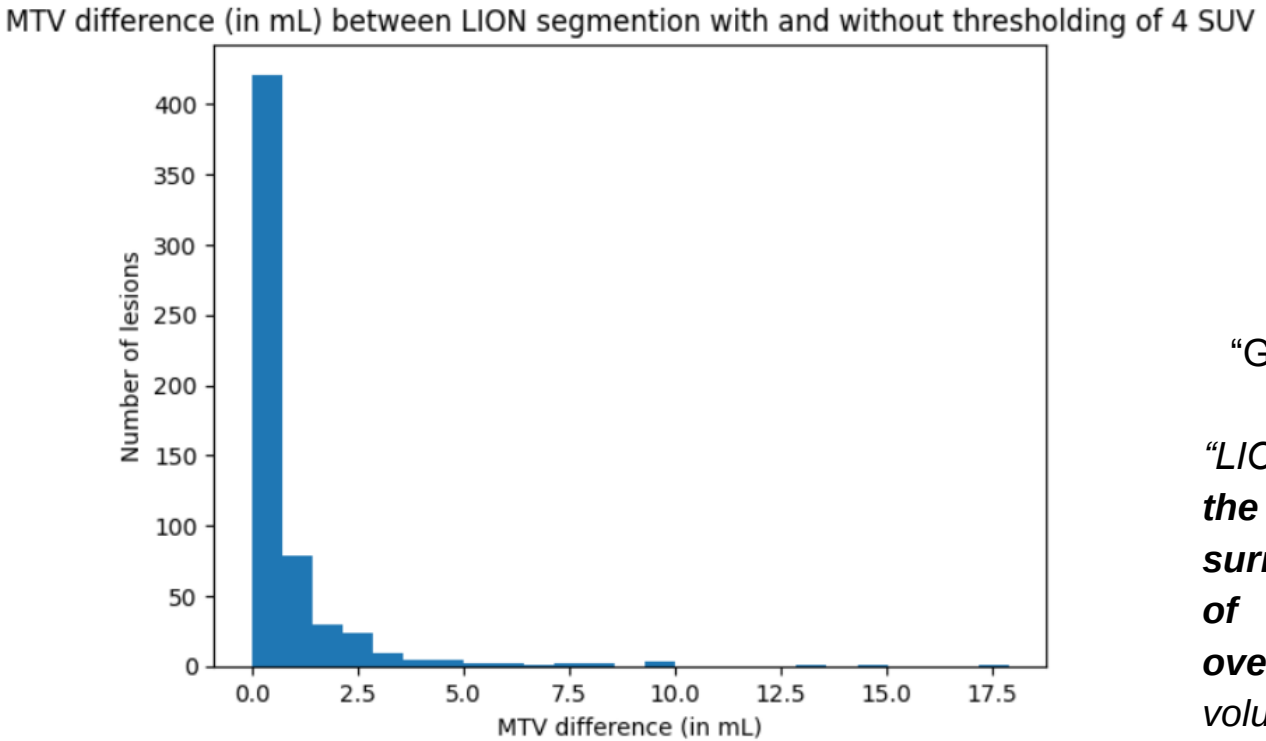
- Detection of small lesions: "*There were 6 segmentations for which the DSC was equal to 0. These 6 cases corresponded to very **small lesions (<3ml)** in the SUV4.0 segmentations. For 4 of these cases, **LIONZ failed in detecting any lesions**."*

Limitations

- **Use of SUV≥4 for defining the reference.**
  - How were the training data segmented? Segmented by experts with no common guidelines

Limitations

- **Use of SUV≥4 for defining the reference.**
  - How were the training data segmented? Segmented by experts with no common guidelines



Note: **Only** 19% of lesions (142/729) had the **same MTV with and without thresholding**, otherwise MTV_LION > MTV_LION_SUV4.

Same observation in the preprint:

"Generally, **LIONZ overestimates tumor volume** compared to SUV4."

*"LIONZ consistently identified **areas beyond the actual lesion borders and labeling surrounding healthy tissue areas as part of the lesion**. This led to an **overestimation** of the tumor size and volume."*

Conclusion

Conclusions

- High detection sensitivity of primary breast tumors (95%)
- Lower sensitivity in small lesions (40% for lesion less than 0.18 mL or 14 voxels)
- Poorer sensitivity for low SUVmax (<5.0) or low SUVmean (<4.4)
- Still some errors when brown fat is activated (false positive)
- Good detection of skeletal metastases
- False positives occurred mainly outside the breast.
- False negatives occurred mainly in the fat compartments (fatSC and fatV).
- SUVmax and Dmax were correctly estimated for the detected lesions.
- LION tended to under-segment with respect to SUV>4, probably because of the reference segmentation used for training

Perspectives

How to improve LION?

To **reduce false positive (FP) and false negative (FN)** cases by **adding cases in the training set** with :

- brown fat activation
- small lesion volumes
- low uptake values
- lesions in specific regions (FP and FN)

Perspectives of this work:

- Rerun the analysis with the new version of LION to **publish this analysis** as an evaluation of the algorithm for early breast cancer (in the LION publication article and/or in a separate article).

Annexes Methodology

Annexe: Evaluation methodology

3. Detection Performance

We checked whether each tumor lesion was detected (true positive), missed (false negative), added (false positive) by the software,

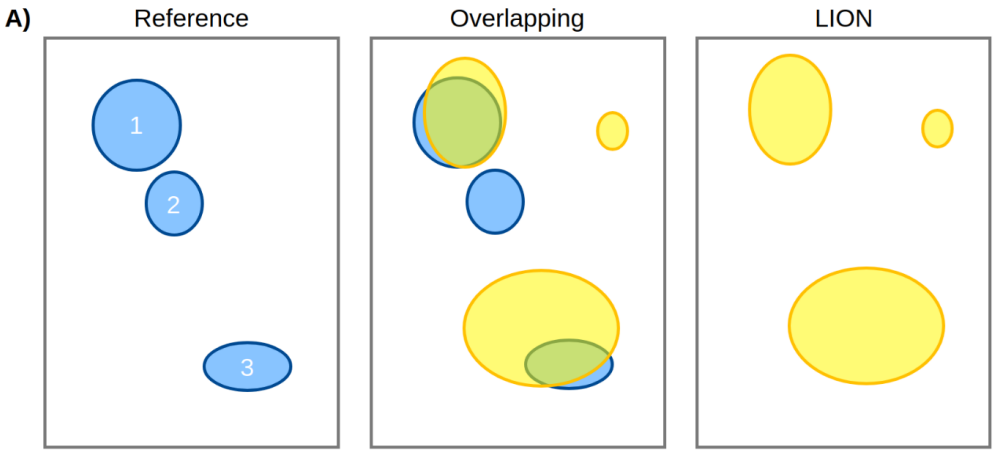
- at patient/body scale and at compartment scale.
- for all patients and by separating them into single-site and multi-site patients.

```
Code:
For all ref_lesion
        If at least one LION_lesion has at least one voxel in common with ref_lesion
                Then ref_lesion is true positive
        Else (=If none of LION_lesions has one voxel in common with ref_lesion)
                Then ref_lesion is false negative

For all LION_lesion
        If none of ref_lesions has one voxel in common with LION_lesion
                Then LION_lesion is a false positive
```

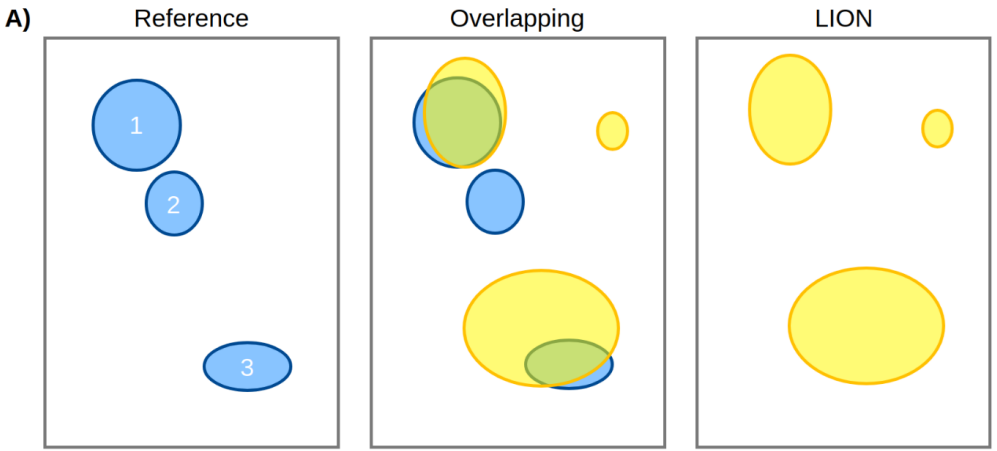
Annexe: Evaluation methodology

3. Detection Performance



Annexe: Evaluation methodology

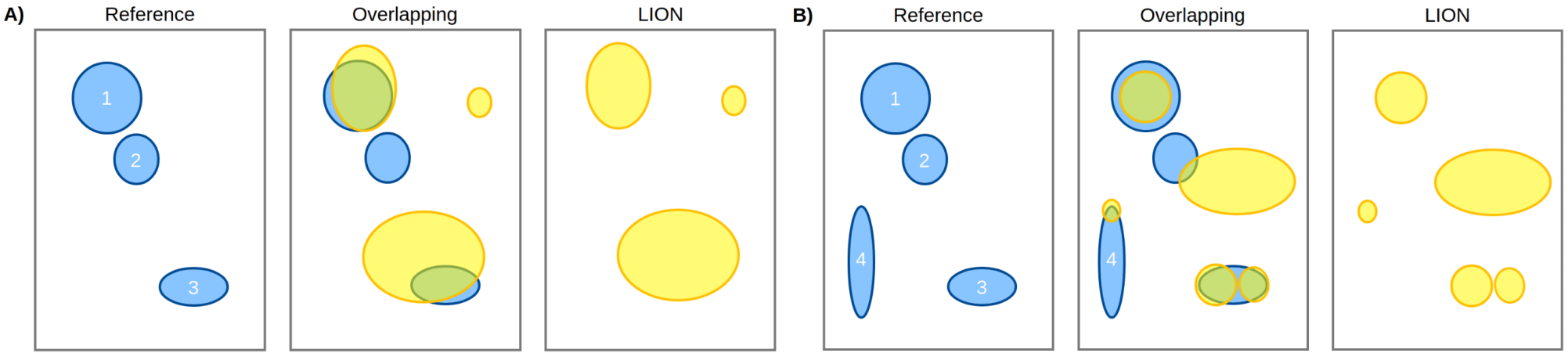
3. Detection Performance



Lesion 1 is detected
Lesion 2 is missed
Lesion 3 is detected
There is an extra lesion

Annexe: Evaluation methodology

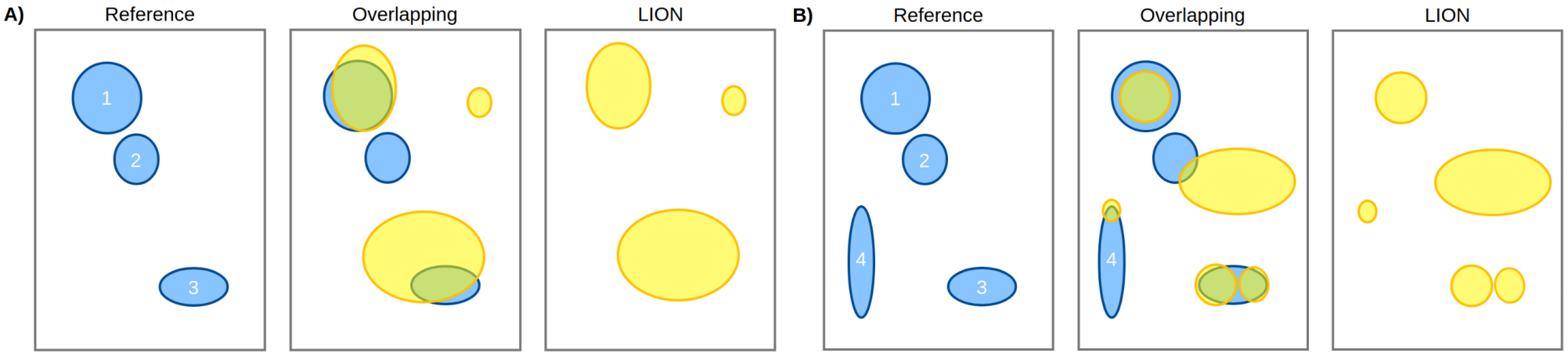
3. Detection Performance



Lesion 1 is detected
Lesion 2 is missed
Lesion 3 is detected
There is an extra lesion

Annexe: Evaluation methodology

3. Detection Performance



Lesion 1 is detected
Lesion 2 is missed
Lesion 3 is detected
There is an extra lesion

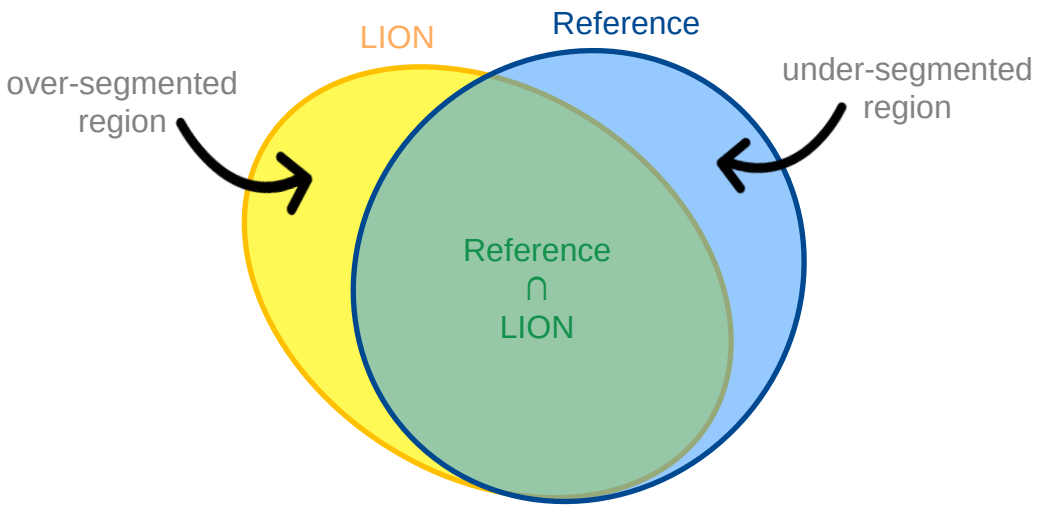
Lesion 1 is detected
Lesion 2 is detected
Lesion 3 is detected
Lesion 4 is detected
There are no missed or extra lesions

Annexe: Evaluation methodology

4. Segmentation Accuracy, allowing a 10% volume tolerance or a 10 mL tolerance for tumors larger than 100 mL

We determined whether segmentations were accurate (within 10% or 10 mL), too large (over-segmented) or too small (under-segmented) by calculating the volume common to the reference and LION segmentations,

- at patient/body scale and at compartment scale.
- for all patients and by separating them into single-site and multi-site patients.



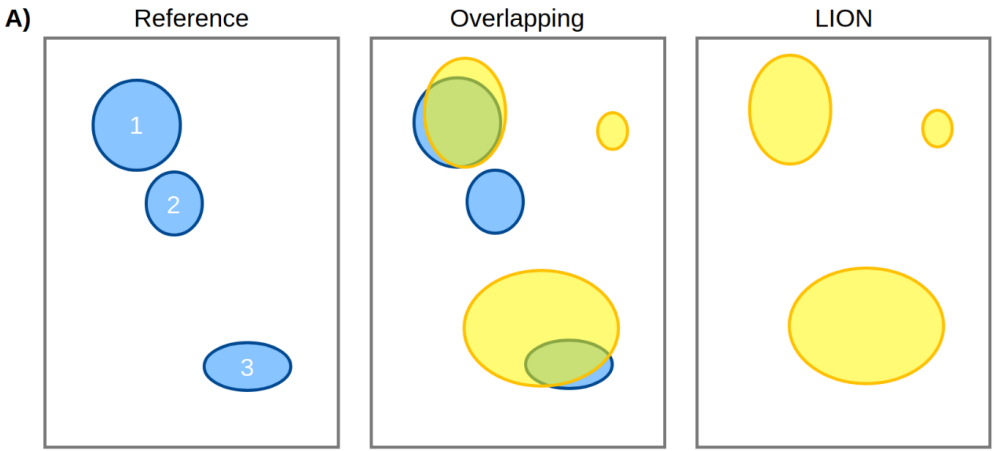
Annexe: Evaluation methodology

4. Segmentation Accuracy

Code:
FOR all ref_lesion
    Threshold = 10 if volume_ref_lesion >= 100 else 0.1*volume_ref_lesion
    IF at least one LION_lesion has at least one voxel in common with ref_lesion:
        THEN: specific_volume_LION_lesion = sum of volume_LION_lesions that have at least one voxel in common with volume_ref_lesion
            intersection = common volume between specific_volume_LION_lesion and volume_ref_lesion
            IF intersection < volume_ ref_lesion - threshold:
                THEN: ref_lesion is under-segmented
            ELSE:
                IF specific_volume_LION_lesion > volume_ ref_lesion + threshold:
                    THEN: ref_lesion is over-segmented
                ELSE:
                    ref_lesion is well segmented

Annexe: Evaluation methodology

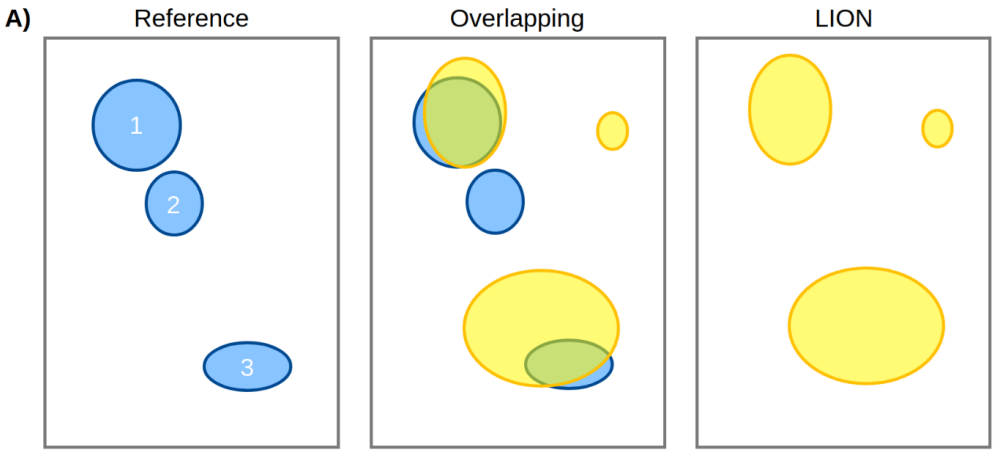
4. Segmentation Accuracy



Lesion 1 is detected
Lesion 2 is missed
Lesion 3 is detected
There is an extra lesion

Annexe: Evaluation methodology

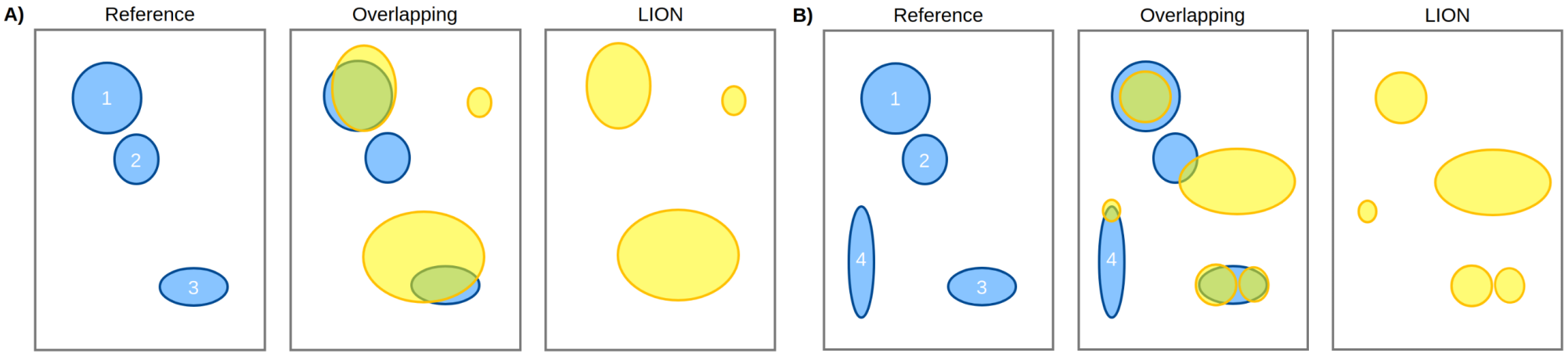
4. Segmentation Accuracy



*Lesion 1 is detected **and well segmented***
Lesion 2 is missed
*Lesion 3 is detected **and over-segmented***
There is an extra lesion

Annexe: Evaluation methodology

4. Segmentation Accuracy

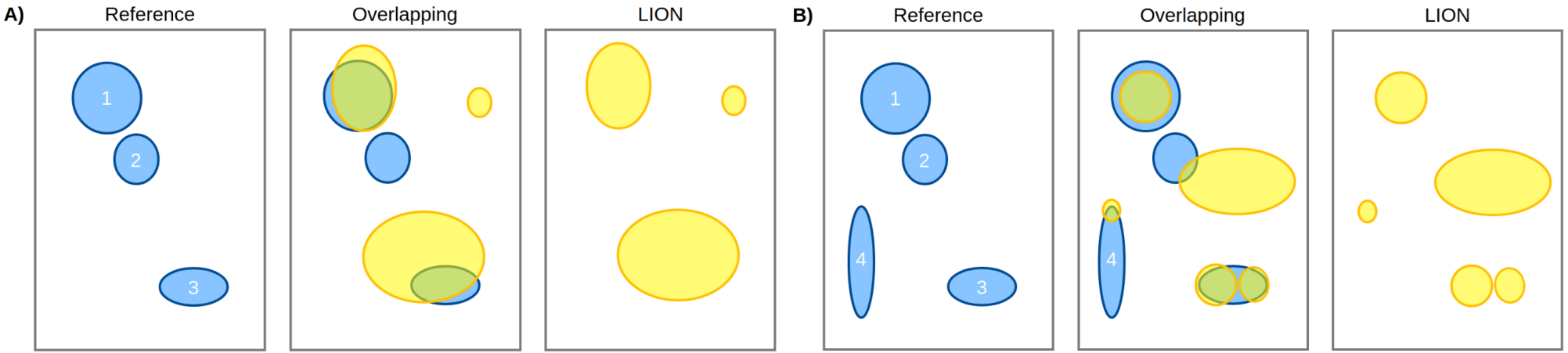


*Lesion 1 is detected **and well segmented***
Lesion 2 is missed
*Lesion 3 is detected **and over-segmented***
There is an extra lesion

Lesion 1 is detected
Lesion 2 is detected
Lesion 3 is detected
Lesion 4 is detected
There are no missed or extra lesions

Annexe: Evaluation methodology

4. Segmentation Accuracy



Lesion 1 is detected and well segmented
Lesion 2 is missed
Lesion 3 is detected and over-segmented
There is an extra lesion

*Lesion 1 is detected **and well segmented***
*Lesion 2 is detected **and under-segmented***
*Lesion 3 is detected **and well segmented***
*Lesion 4 is detected **and under-segmented***
There are no missed or extra lesions

Results: Segmentation

- **LION consistently under-segmented** the lesions.

	ALL	SINGLE-SITE	MULTI-SITE
Lesions well segmented	38%	46%	33%
Lesions under-segmented	50%	50%	49%
Lesions over-segmented	0%	0%	0%
Number of lesions (reference-LION)	570-817	73*-172	497-645
Number of patients	179	70	109

* 2 patients had multiple lesions in their reference segmentation because, during correction, the lesions were grouped together as a single lesion as they were very close and appeared to come from the same lesion.

Results: Segmentation

At compartments scale:

	Muscles	FatSC	FatV	Breast
Average % of lesions well segmented *	44%	47%	21%	42%
Average % of lesions under segmented *	38%	21%	24%	45%
Average % of lesions over segmented *	0%	0%	0%	0%
Number of patients with lesions in this tissue compartment	160	159	9	177

- In the **lung and bone compartments, LION tended to underestimate segmentation**, although segmentation quality was good in the rib region.

** Calculated only on patients with lesions in the specific tissue compartment (last row).*

Results: global PET Features and Dice Score

- Mean **Dice score** **of 0.84 for all groups**.
- **TMTV** **was well estimated** within 10% in 42% of the patients.
- **Dmax** **and** **TMTV** **had relatively low concordance rates (~40%).**

	ALL	SINGLE-SITE	MULTI-SITE
% of patients with correct **Dmax** within 10%	42%	49%	38%
Mean±std difference in the other patients	40±25 cm	53±23 cm	33±23 cm
Number of patients by group	180	70	110

	ALL	SINGLE-SITE	MULTI-SITE
% of patients with correct **TMTV** within 10%	42%	40%	43%
Mean±std difference in the other patients	8±13 mL	4±11 mL	11±14 mL
Number of patients by group	180	70	110