ARAMIS LAB
BRAIN DATA SCIENCE

ninon.burgos@cnrs.fr

Ensuring Robust AI in Medical Imaging:

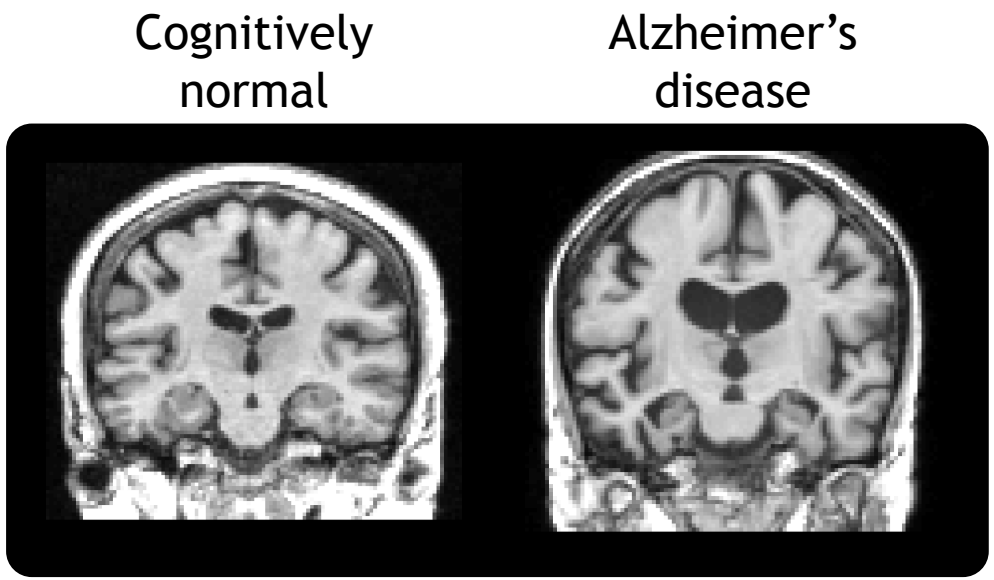
Validation on Real-World Data, Meaningful Metrics, and Rigorous Statistical Analysis

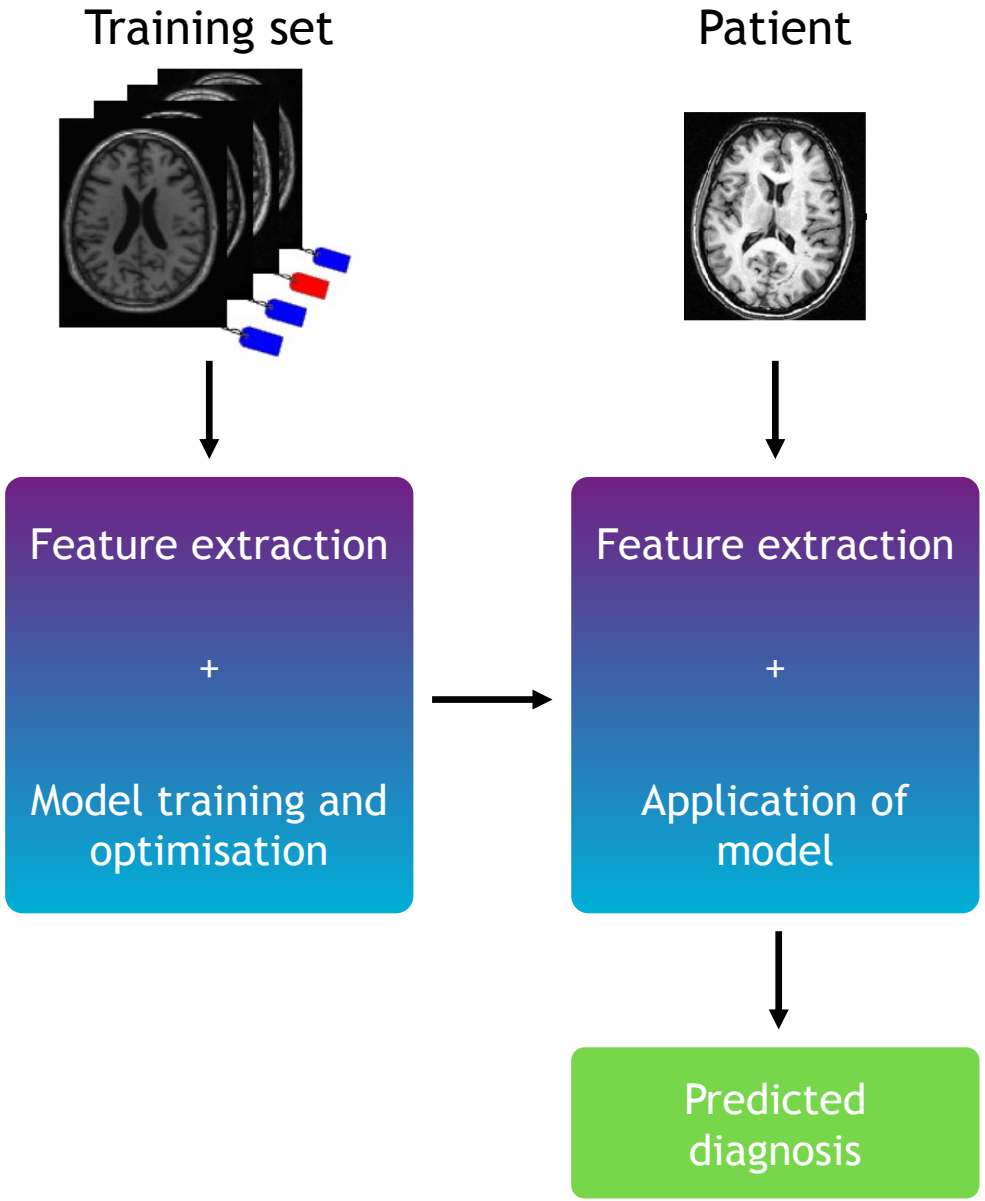
Ninon Burgos
CNRS researcher & PR[AI]RIE chair
ARAMIS Lab, Paris Brain Institute, France

Context

Computer-aided diagnosis

- Machine learning / deep learning task: classification
- Disorder targeted: dementia
- Imaging modality: T1-weighted MRI





Context

A (fake & slightly exaggerated) typical MICCAI conclusion

‘Our proposed deep learning framework achieves superior accuracy (86.6%) and robust generalizability in distinguishing Alzheimer’s disease patients from cognitively normal subjects, outperforming established methods on the ADNI dataset. The model provides anatomically interpretable results by highlighting disease-relevant regions such as the hippocampus, demonstrates high reproducibility and fast inference, and, by eliminating manual feature engineering while aligning with clinical biomarkers, is both practical and explainable—making it ready for immediate integration into routine diagnostic workflows.’

Ensuring Robust AI in Medical Imaging

- Validation on Real-World Data
- Meaningful Metrics
- Rigorous Statistical Analysis

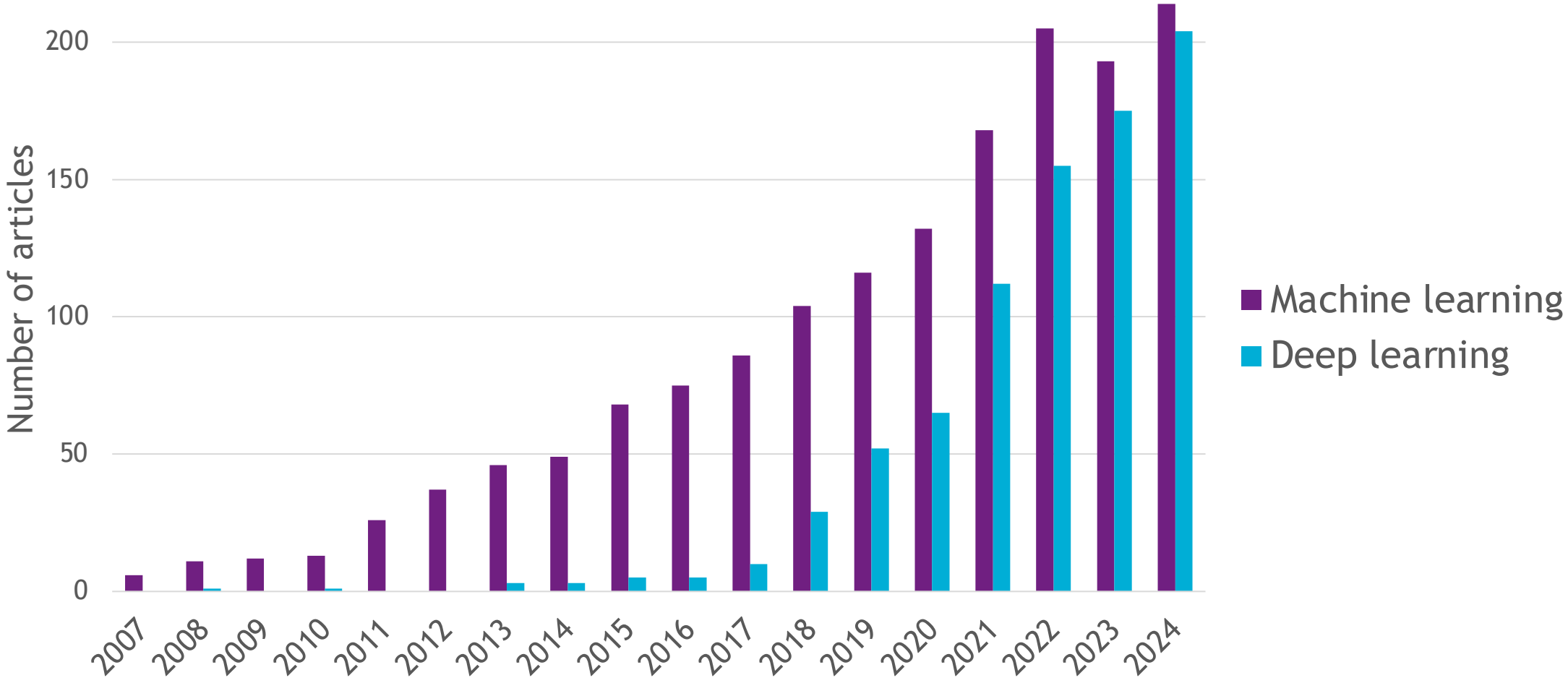
Ensuring Robust AI in Medical Imaging

- Validation on Real-World Data
- Meaningful Metrics
- Rigorous Statistical Analysis

Validation on Real-World Data

ML/DL for dementia diagnosis & prognosis

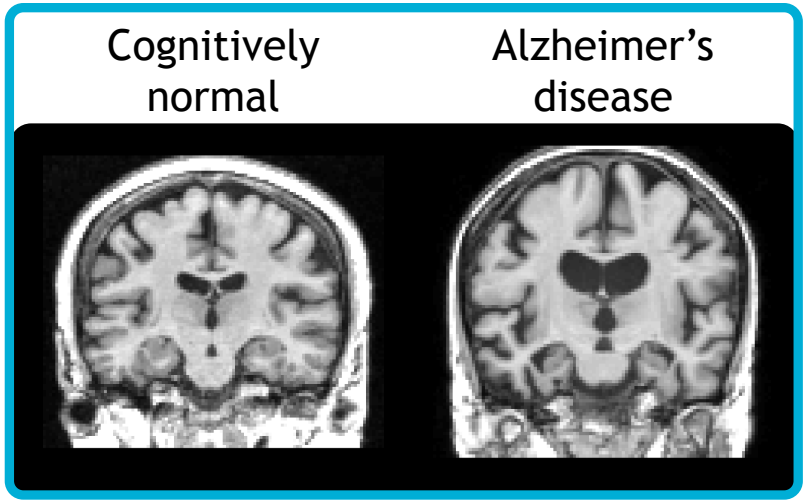
- A very active field of research



Validation on Real-World Data

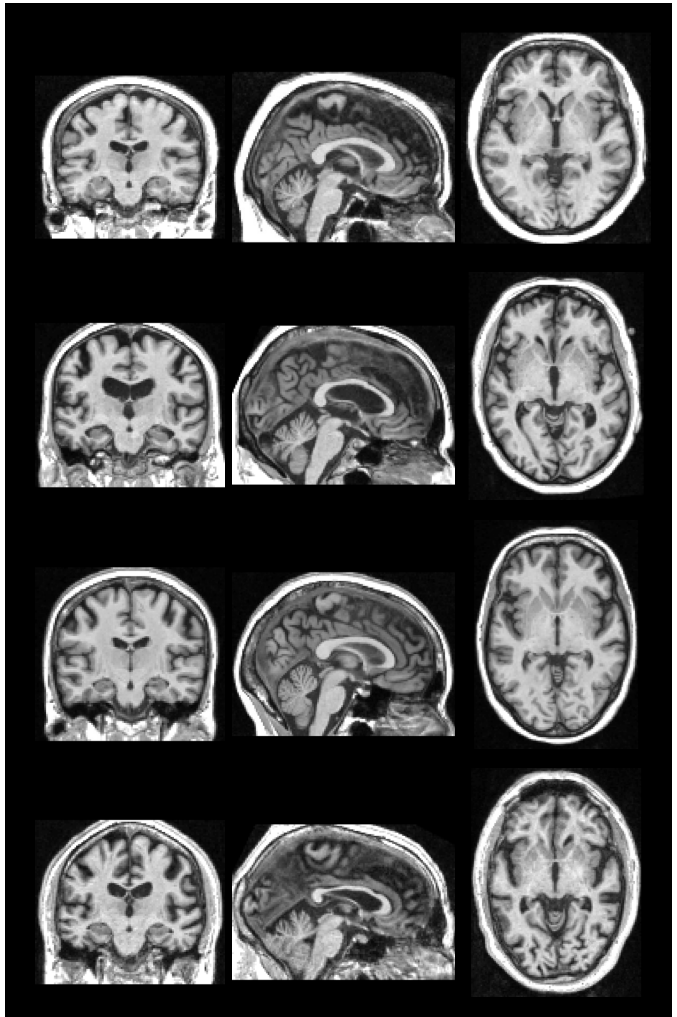
Motivation

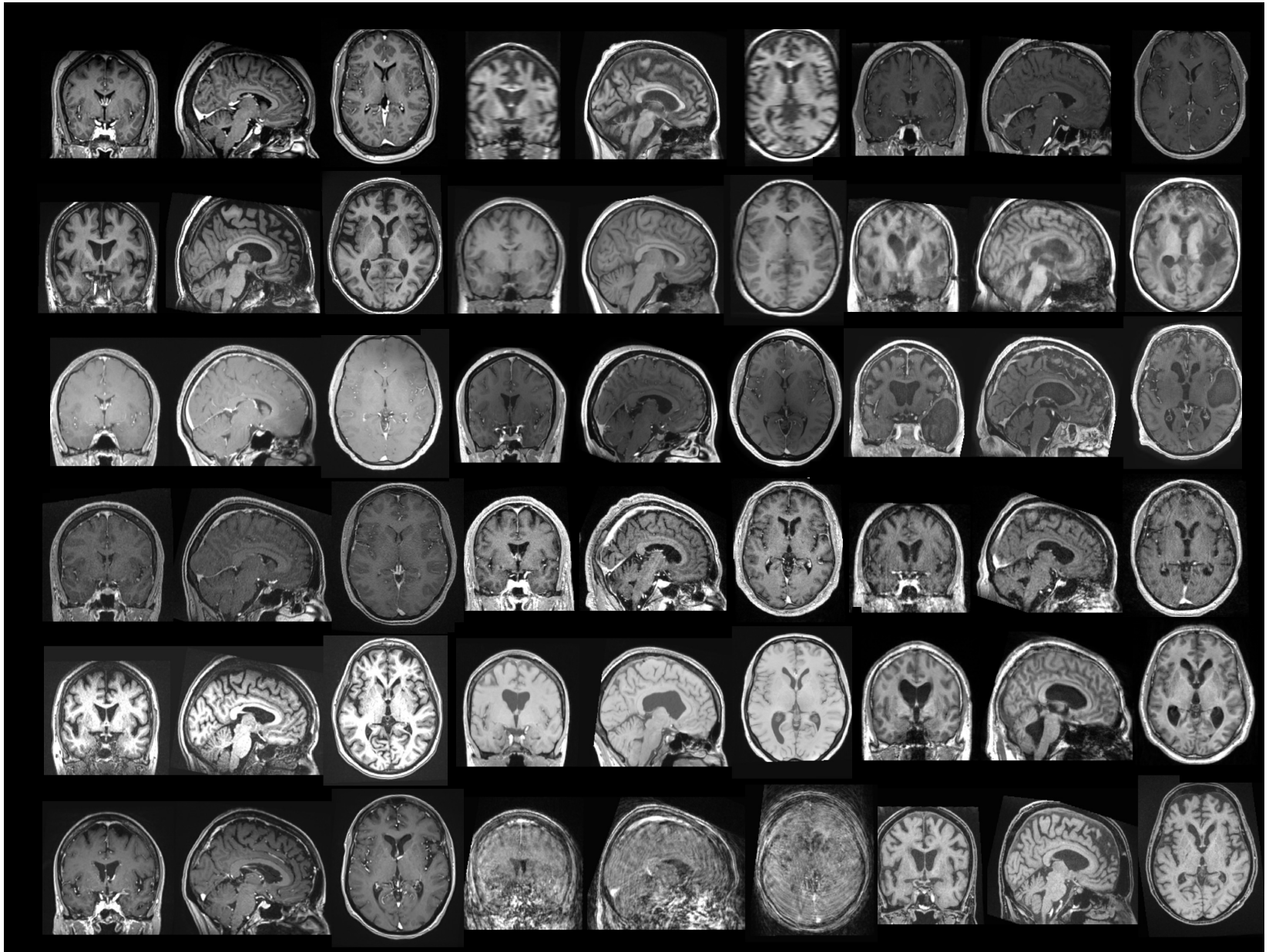
- Most machine learning algorithms for computer-aided diagnosis of dementia developed and applied on **research data**



Validation on Real-World Data

ADNI
Alzheimer's Disease Neuroimaging Initiative

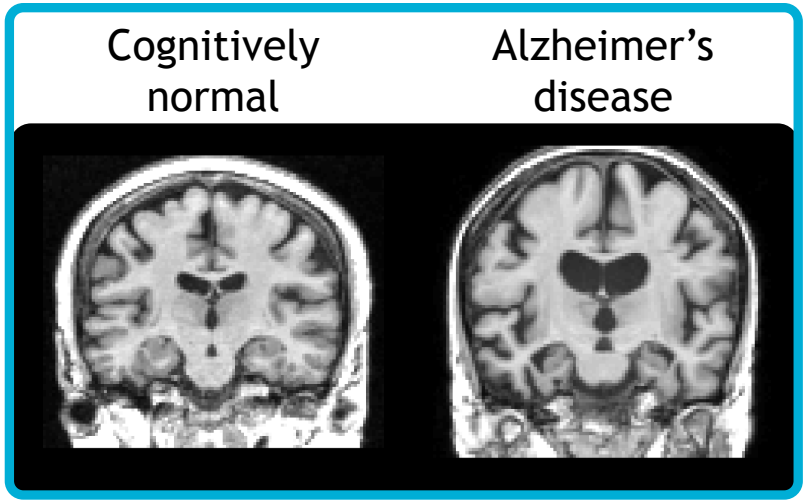




Validation on Real-World Data

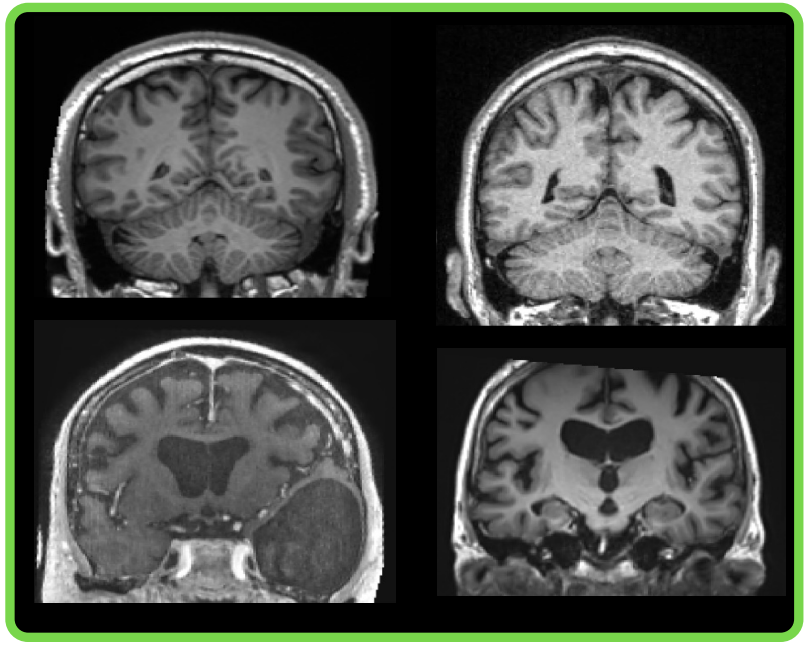
Motivation

- Most machine learning algorithms for computer-aided diagnosis of dementia developed and applied on **research data**



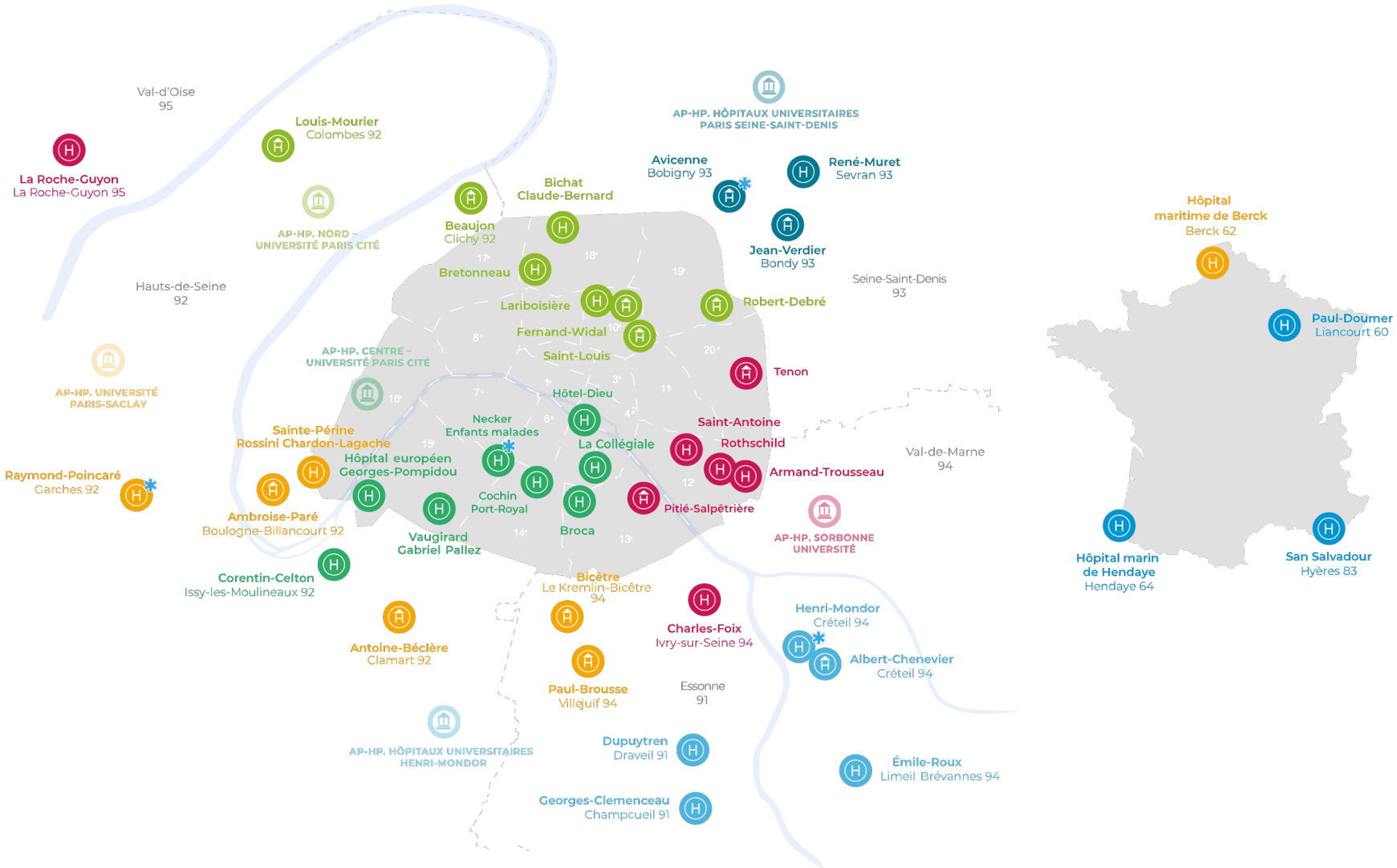
Objective

- Validate these algorithms on **clinical data**
- Develop robust algorithms able to handle clinical data



Validation on Real-World Data

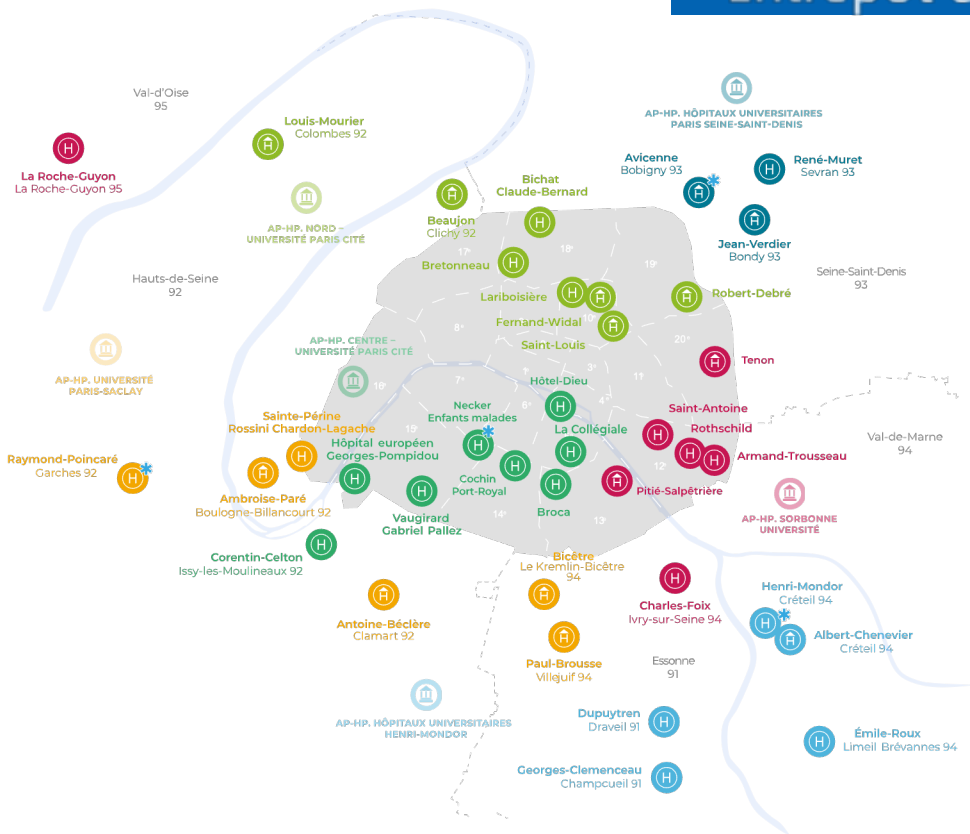
AP-HP: a network of 39 hospitals



Validation on Real-World Data

AP-HP clinical data warehouse

ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS
Entrepôt de Données de Santé



- 11 million patients
- 40 million medical reports
- 25 million imaging exams
- 325 million biological results

Validation on Real-World Data

Objective

- Identify patients with dementia among a clinical data warehouse from anatomical MRI using machine learning

Cohort definition

- **Inclusion criteria**
  - Age ≥18 years
  - At least one cerebral MRI scan including T1-weighted 3D acquisition
  - Clinical data available (available for ~18% of the subjects with a 3D T1 MRI)
- **Definition of the classes of interest using diagnosis (ICD-10) codes**
  - Dementia vs
    - No dementia and no lesions (NDNL)
    - No dementia with lesions (NDL)

Bottani et al., Medical Image Analysis, 2023

Validation on Real-World Data

Cohort

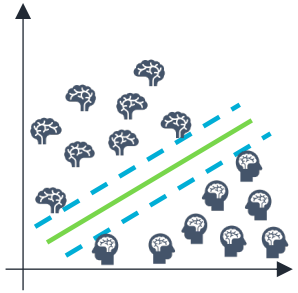
Category	N patients	N images	Age (mean ±std [range])	Sex (%F)
Dementia	756	887	71.17 ± 11.58 [18,90]	50.34%
NDNL	756	939	71.17 ± 11.58 [18,90]	50.34%
NDL	756	997	71.17 ± 11.58 [18,90]	50.34%
Total	**2268**	**2823**	**71.17 ± 11.58 [18,90]**	**50.34%**

Bottani et al., Medical Image Analysis, 2023

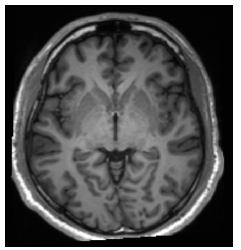
Validation on Real-World Data

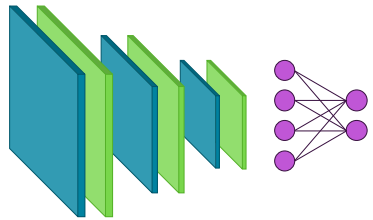
Classification algorithms

- Machine learning: support vector machine
  - Input: grey matter density map



- Deep learning: convolutional neural networks
  - Input: minimally processed MRI





Evaluation setting

- Test set of 152 patients/images per class
- 5-fold cross-validation

Bottani et al., Medical Image Analysis, 2023

Validation on Real-World Data

Results

Data set	Task	Classification strategy	
		SVM with grey matter maps	ResNet with minimally processed MRI
Clinical	D vs NDNL	68.75	73.62
	D vs NDL	73.09	72.24
Research	AD vs CN	86.80	85.30

Balanced accuracy (%)

Decrease of 15 percent points in balanced accuracy: clinical data set more heterogenous

Bottani et al., Medical Image Analysis, 2023

Validation on Real-World Data

Cohort

Category	N patients	N images	Age (mean ±std [range])	Sex (%F)	% Good/medium quality	% With gadolinium
Dementia	756	887	71.17 ± 11.58 [18,90]	50.34%	57.72%**	24.80%**
NDNL	756	939	71.17 ± 11.58 [18,90]	50.34%	36.42%**	66.13%**
NDL	756	997	71.17 ± 11.58 [18,90]	50.34%	52.25%	63.59%**
Total	**2268**	**2823**	**71.17 ± 11.58 [18,90]**	**50.34%**	**48.71%**	**52.24%**

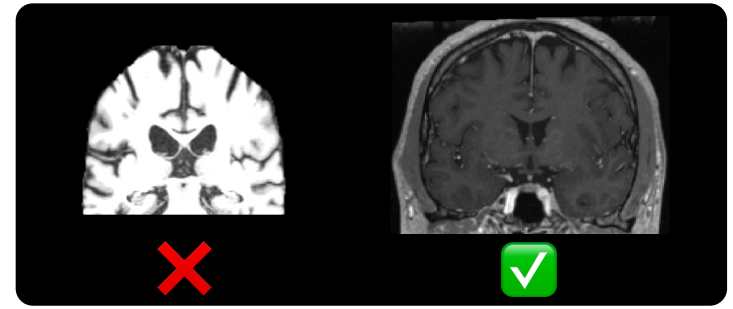
p-value corrected with Bonferroni < 0.05

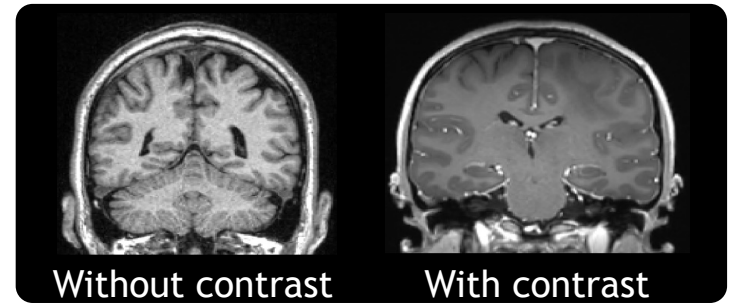
Bottani et al., Medical Image Analysis, 2023

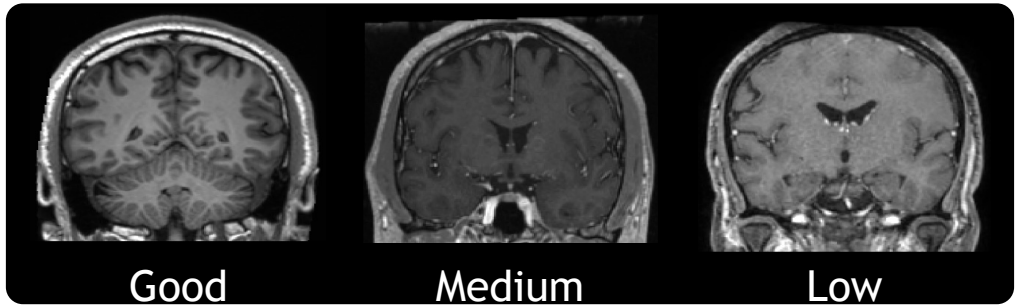
Validation on Real-World Data

T1-weighted image quality control

- **Objectives**
  - ➢ Reject images that are not proper T1-weighted anatomical MR images
  - ➢ Recognise images acquired after injection of a contrast agent (gadolinium)
  - ➢ Rate the overall image quality (good/medium/low)

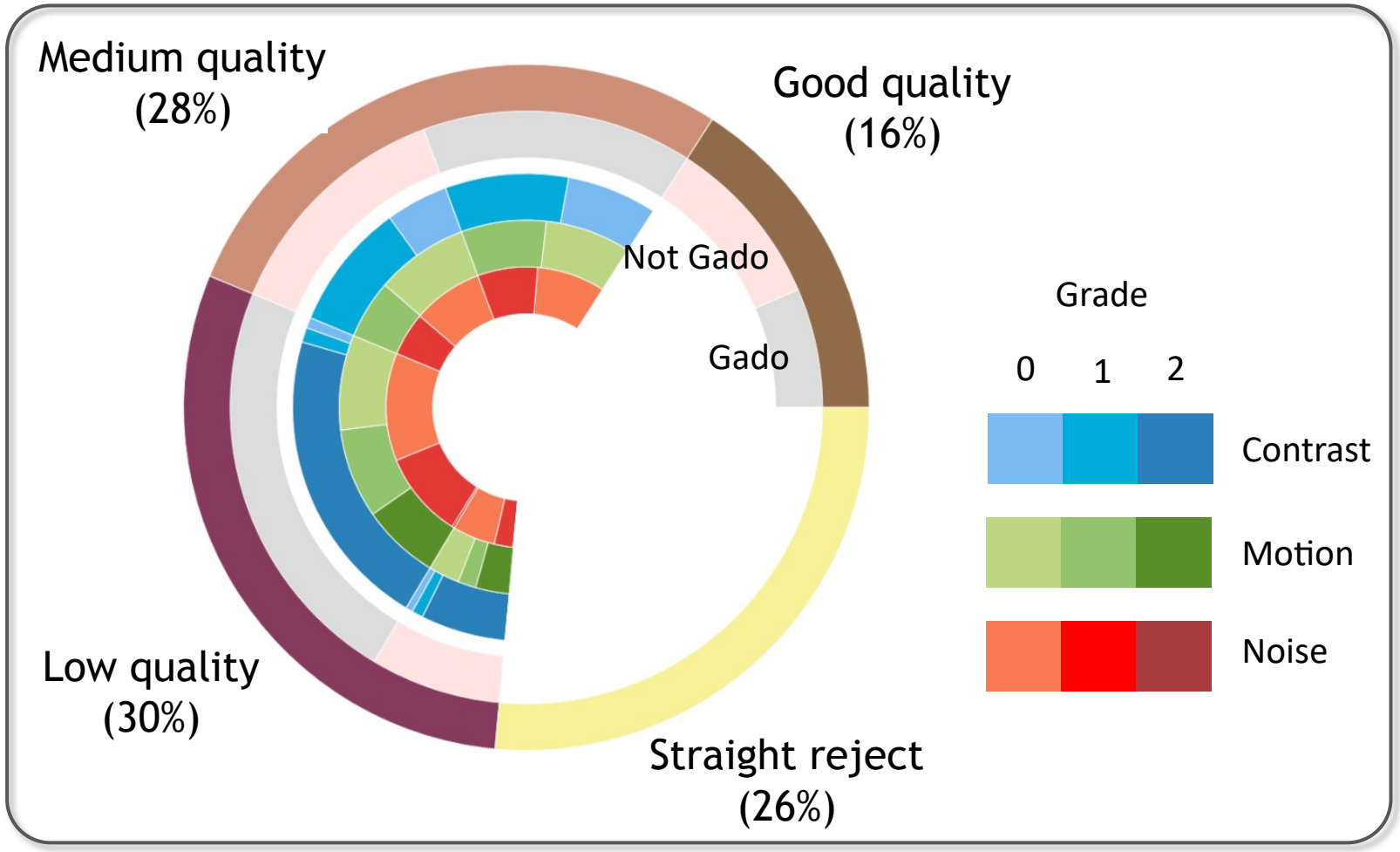






Validation on Real-World Data

T1-weighted image quality control



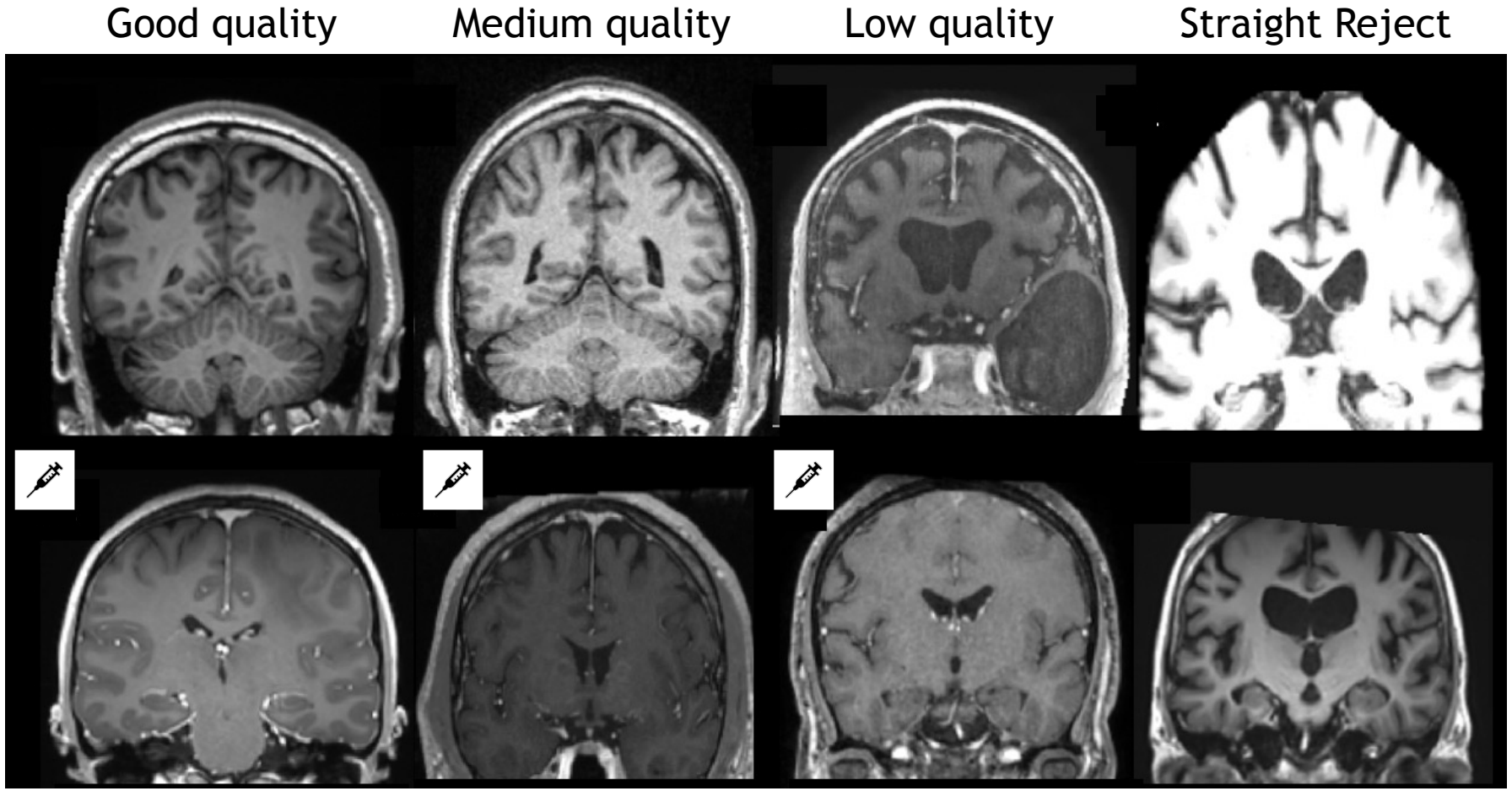
← Manual quality control

Bottani et al., Medical Image Analysis, 2022

Validation on Real-World Data

T1-weighted image quality control

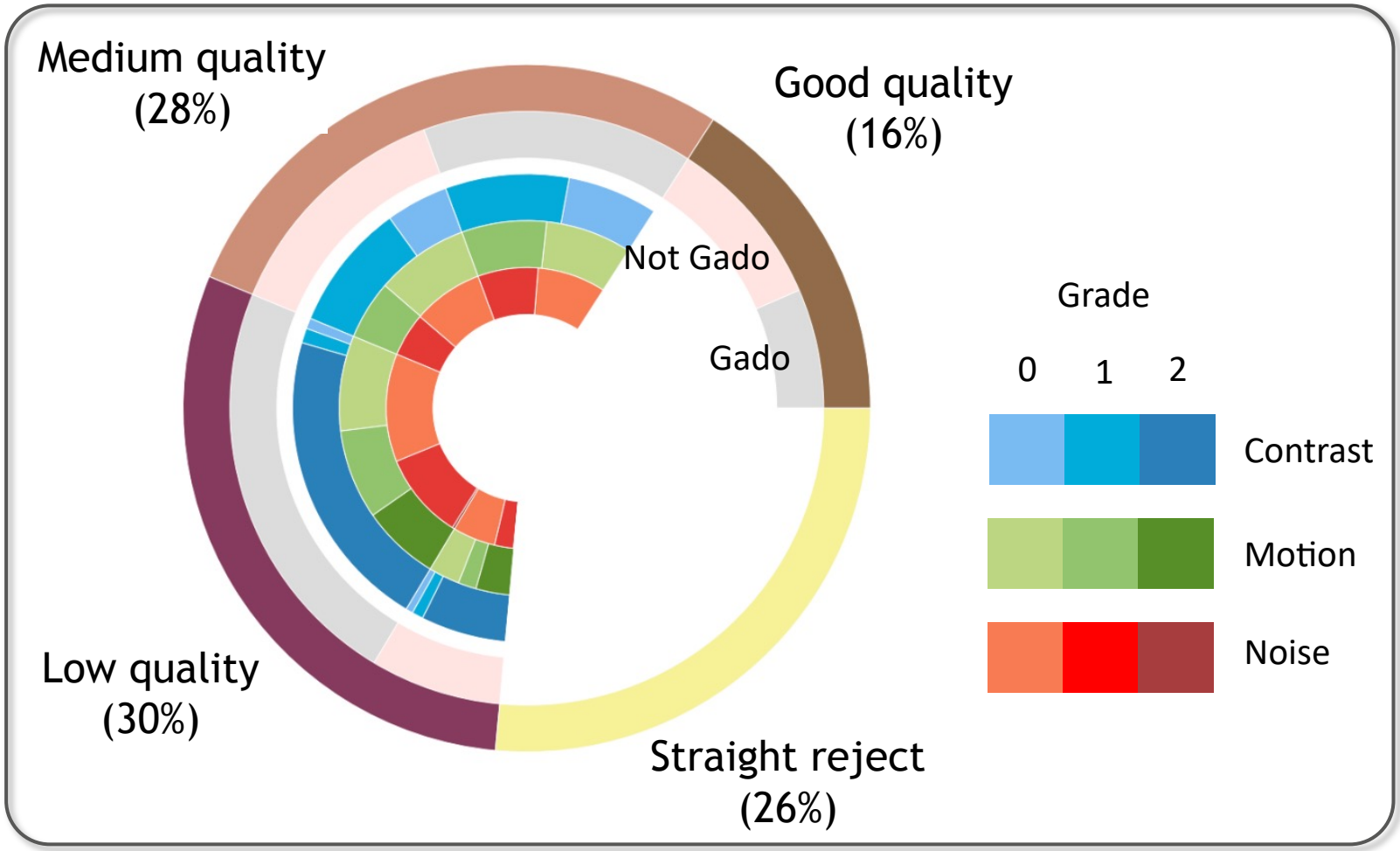
- Examples



Bottani et al., Medical Image Analysis, 2022

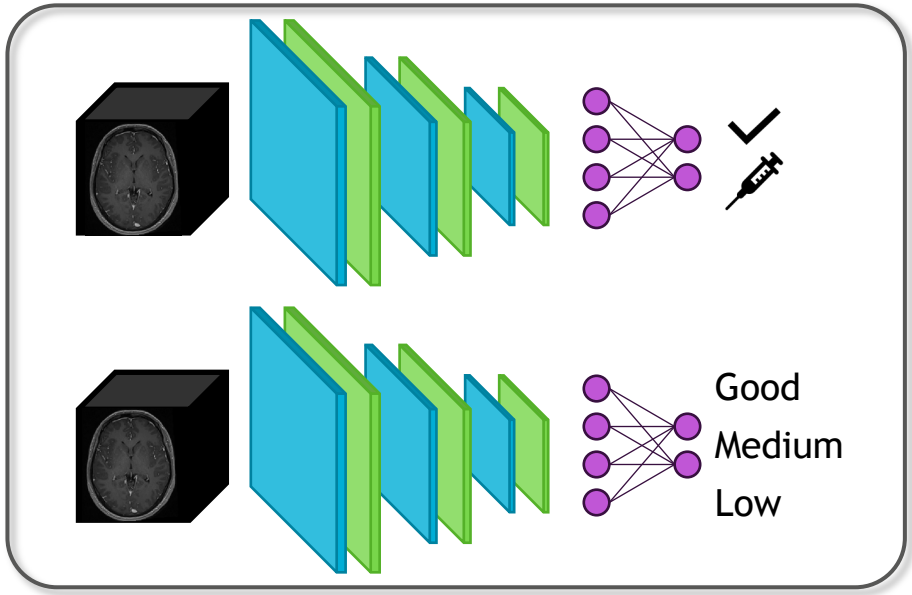
Validation on Real-World Data

T1-weighted image quality control



← Manual quality control

Automatic quality control ↓

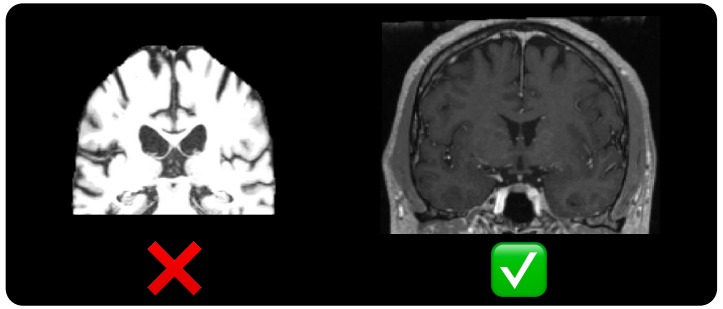


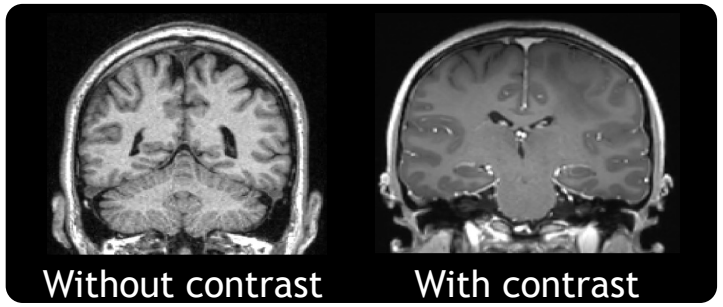
Bottani et al., Medical Image Analysis, 2022

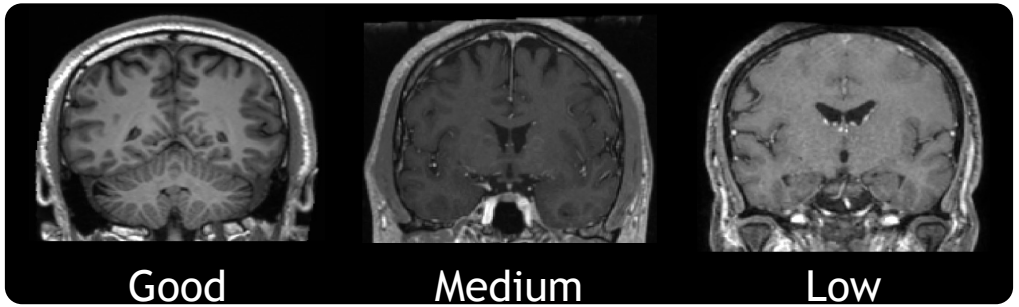
Validation on Real-World Data

T1-weighted image quality control

- **Results**
  - Reject images that are not proper T1-weighted anatomical MR images
    - Balanced accuracy > 90%
  - Recognise images acquired after contrast injection
    - Balanced accuracy > 95%
  - Rate the overall image quality
    - Good/medium vs low: balanced accuracy > 80%
    - Good vs medium: balanced accuracy > 70%







Bottani et al., Medical Image Analysis, 2022

Validation on Real-World Data

Cohort

Category	N patients	N images	Age (mean ±std [range])	Sex (%F)	% Good/medium quality	% With gadolinium
Dementia	756	887	71.17 ± 11.58 [18,90]	50.34%	57.72%**	24.80%**
NDNL	756	939	71.17 ± 11.58 [18,90]	50.34%	36.42%**	66.13%**
NDL	756	997	71.17 ± 11.58 [18,90]	50.34%	52.25%	63.59%**
Total	**2268**	**2823**	**71.17 ± 11.58 [18,90]**	**50.34%**	**48.71%**	**52.24%**

p-value corrected with Bonferroni < 0.05

Bottani et al., Medical Image Analysis, 2023

Diagnosis prediction from T1w MRI

- Results with training subsets

Task	Training set (n=88 per class)			
	Quality: good/medium + low Gadolinium: presence & absence	Quality: good/medium Gadolinium: presence & absence	Quality: good/medium Gadolinium: absence	Quality: good/medium Gadolinium: absence/synthetic
D vs NDNL	69.47	60.26	51.51	51.71
D vs NDL	73.03	68.29	50.00	54.08

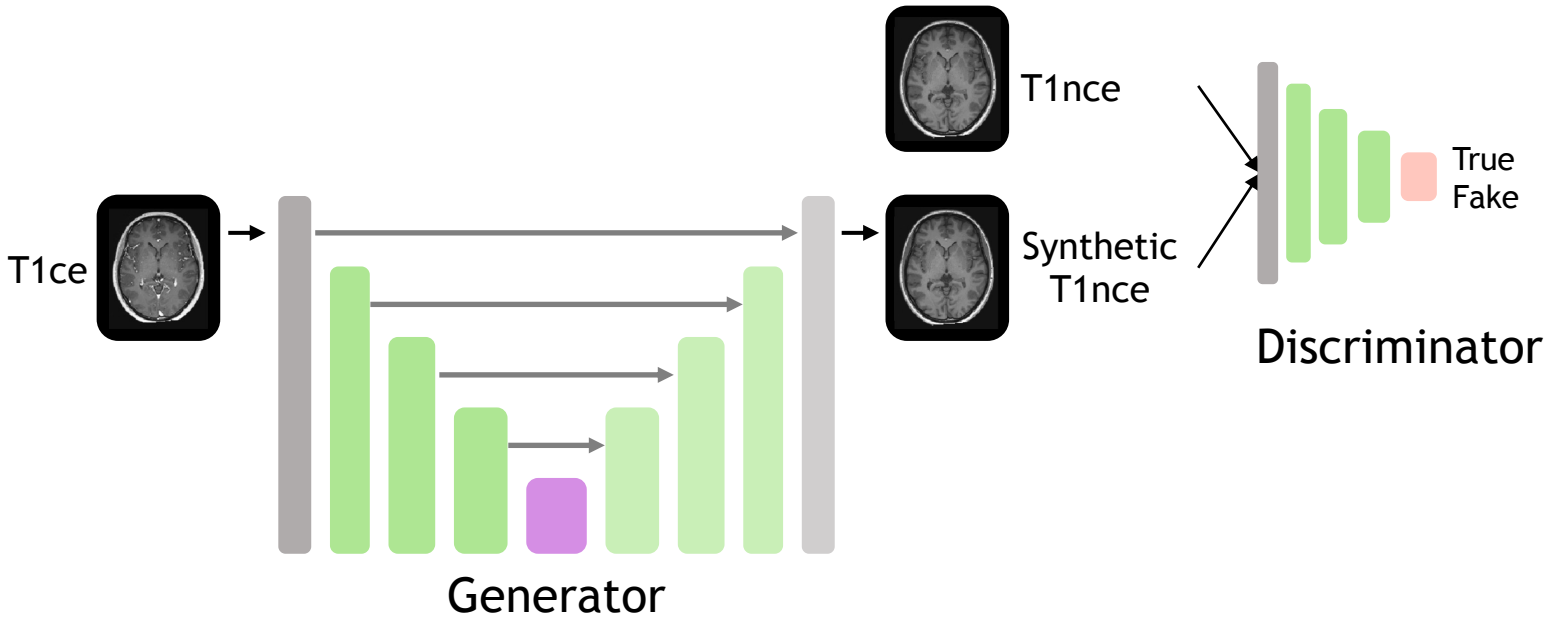
Balanced accuracy (%) - Classification strategy: SVM with grey matter maps

Bottani et al., Medical Image Analysis, 2023

Validation on Real-World Data

Contrast-enhanced to non-contrast-enhanced image translation

- **Methods**



3D U-Net like generators
- Residual connections
- Attention gates
- Transformer layers

Conditional GANs
- 3D U-Net like generators
- 3D patch discriminator

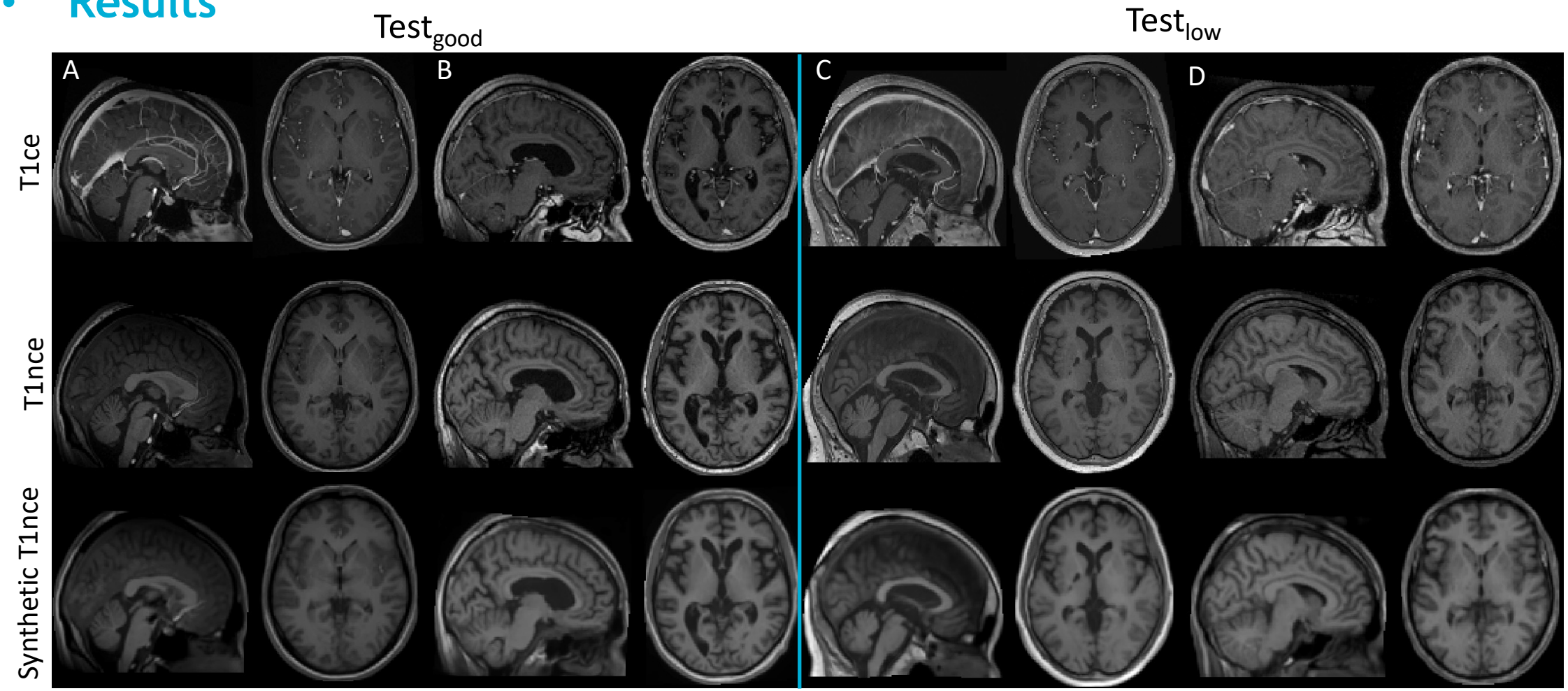
- Paired T1ce/T1nce MRI
- 230 pairs for training
- 77 pairs for testing

Bottani et al., BMC Medical Imaging, 2024

Validation on Real-World Data

Contrast-enhanced to non-contrast-enhanced image translation

- **Results**



Bottani et al., BMC Medical Imaging, 2024

Validation on Real-World Data

Contrast-enhanced to non-contrast-enhanced image translation

- **Results**
  1. Image synthesis accuracy
     - Mean absolute error
     - Peak signal to noise ratio
     - Structural similarity

Good synthesis accuracy

  2. Segmentation fidelity
     - Absolute volume difference
     - Volume difference

Reliable feature extraction

Grey matter map

Clinica

`t1-volume-tissue-segmentation`

Bottani et al., BMC Medical Imaging, 2024

Validation on Real-World Data

Diagnosis prediction from T1w MRI

- **Results with training subsets**

Task	Training set (n=88 per class)			
	Quality: good/medium + low Gadolinium: presence & absence	Quality: good/medium Gadolinium: presence & absence	Quality: good/medium Gadolinium: absence	Quality: good/medium Gadolinium: synthetic
D vs NDNL	69.47	60.26	51.51	51.71
D vs NDL	73.03	68.29	50.00	54.08

Balanced accuracy (%) - Classification strategy: SVM with grey matter maps

Classifier exploits biases of the training set

Image translation can be useful to limit bias due to the injection of gadolinium

Bottani et al., Medical Image Analysis, 2023

Validation on Real-World Data

Diagnosis prediction from T1w MRI

- Final results with unbiased training sets

Training set	Task	Classification strategy	
		SVM with grey matter maps	ResNet with minimally processed MRI
Research data set (ADNI Alzheimer's Disease Neuroimaging Initiative)	D vs NDNL	64.08	61.84
	D vs NDL	69.47	61.78
Clinical data set with synthetic images of good/medium quality	D vs NDNL	61.91	63.22
	D vs NDL	64.61	67.50

Balanced accuracy (%)

Balanced accuracy still ~20 percent points lower than when testing on research data set

Bottani et al., Medical Image Analysis, 2023

Validation on Real-World Data

Diagnosis prediction from T1w MRI

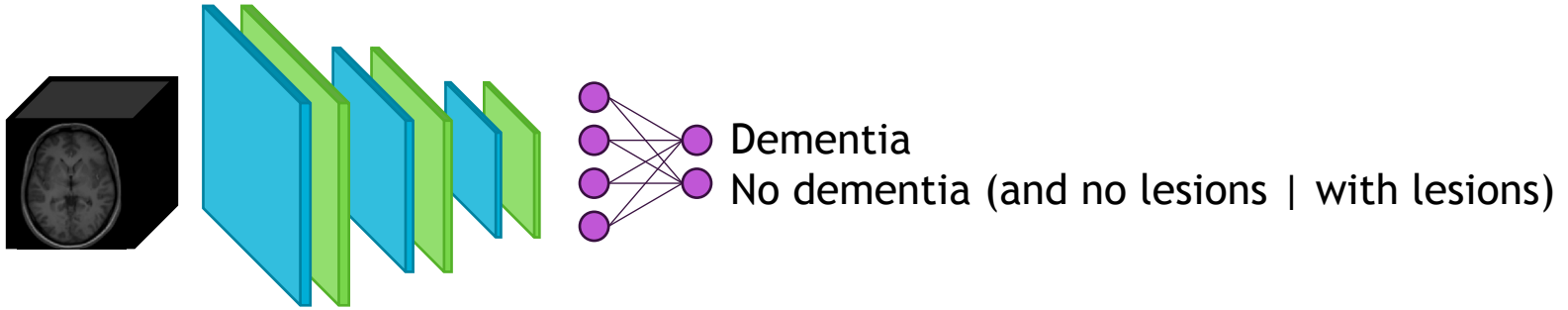
- **Summary of the results**
  - Balanced accuracy much lower than in research context
  - Difference in proportions of images with gadolinium and of medium/good quality can bias the results
    - **Solution**: training the algorithms using only research data or only clinical images of good/medium quality without gadolinium
- **Conclusion**
  - **Quality control** & **data set homogenisation** essential steps
- **Challenge remaining**
  - Lack of robustness of the labels

Bottani et al., Medical Image Analysis, 2023

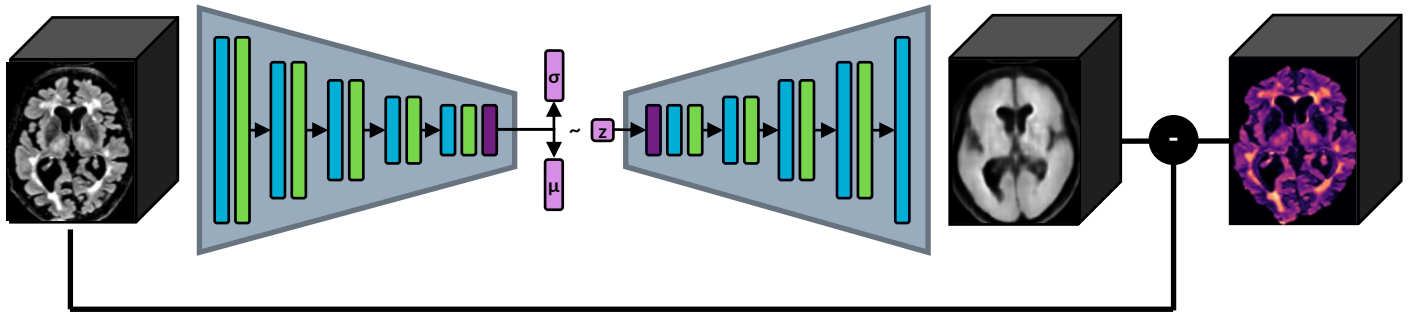
Validation on Real-World Data

Computer-aided diagnosis of neurodegenerative diseases

- **Diagnosis prediction**



- **Lesion detection**



Validation on Real-World Data

3D FLAIR cohort constitution

AP-HP Clinical Data Warehouse

IMAGING DATA

25 million imaging exams

CLINICAL DATA

Medical records of **26 million** patients

Inclusion Criteria

Age > 18 years

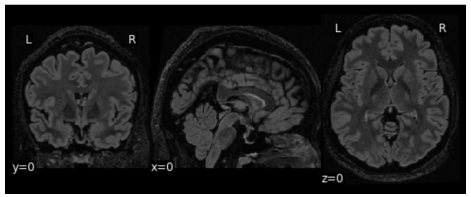
At least one cerebral 3D FLAIR MRI scan

Query on the CDW

~122 000 patients (clinical data available for ~67% of the patients)

Available for the APPRIMAGE project

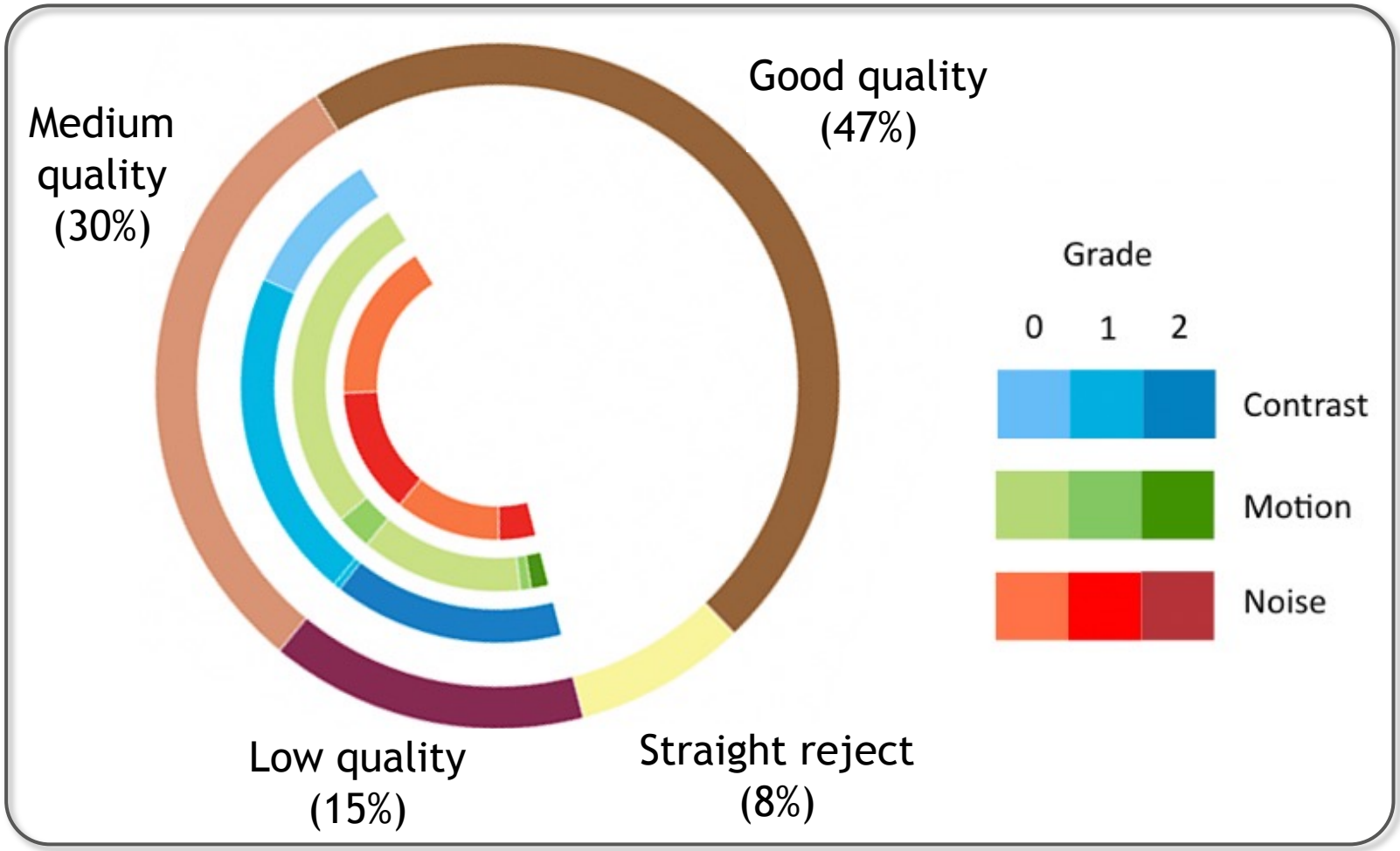
13 000 3D FLAIR MRIs



Loizillon et al., Medical Image Analysis, 2025

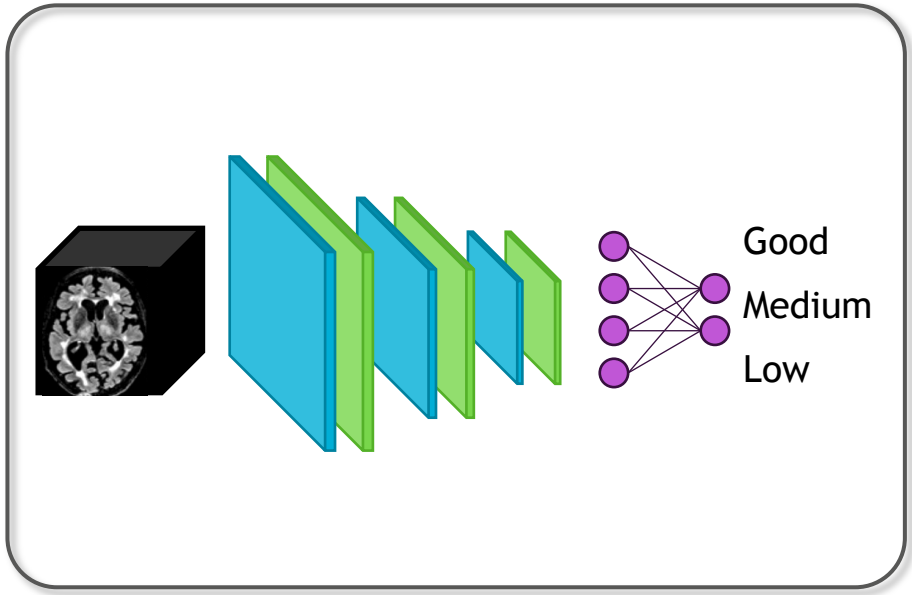
Validation on Real-World Data

3D FLAIR image quality distribution



← Manual quality control

Automatic quality control?



Loizillon et al., Medical Image Analysis, 2025

Validation on Real-World Data

3D FLAIR automatic quality control

- Annotated data available for training:
  - 358 3D FLAIR MRIs
  - 5000 3D T1w MRIs

Task: Low vs Medium/Good

	FLAIR	T1w
Annotators	86.54	91.56
Training with FLAIR	69.90 ± 2.38	48.20 ± 0.81
Training with T1w	50.06 ± 1.71	83.51 ± 0.93
Training with T1w + FLAIR	58.05 ± 4.45	82.37 ± 0.85

Balanced accuracy (%)

⇒ Domain gap

Loizillon et al., DART@MICCAI, 2024

Validation on Real-World Data

Domain adaptation for 3D FLAIR automatic quality control

Source domain

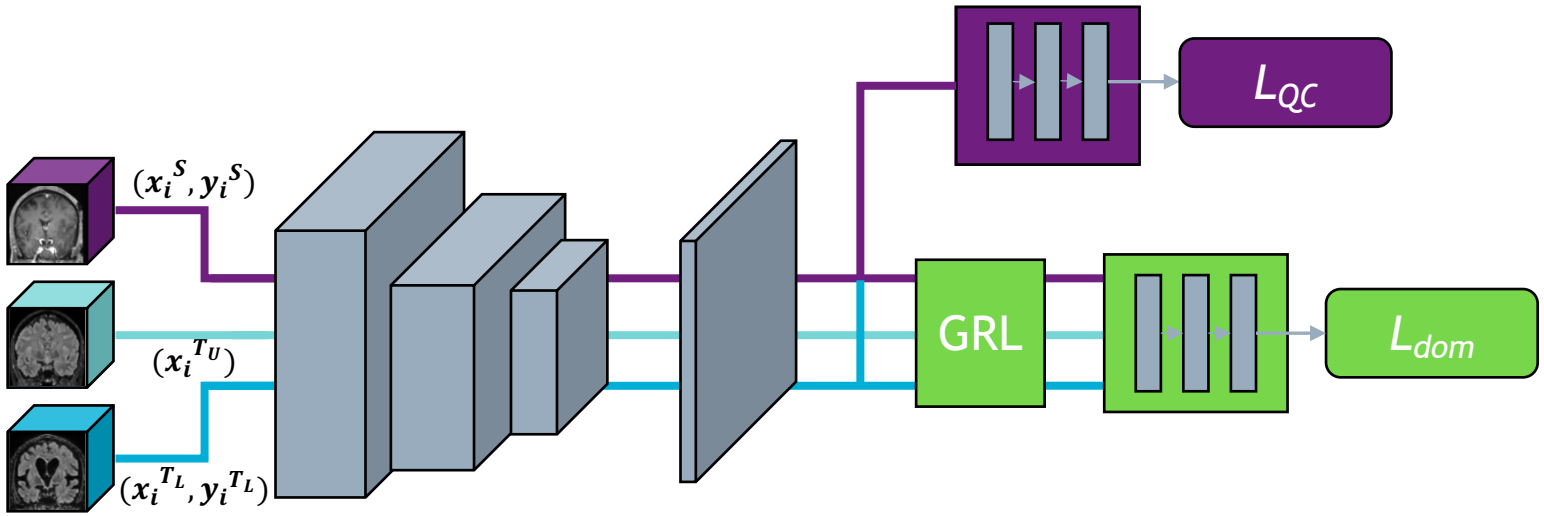
5000 3D T1w MRIs
manually annotated

$D_S = (x_i^S, y_i^S)^{N_S=5000}$

Target domain

5358 3D FLAIR MRIs, with
358 manually annotated

$D_{T_U} = (x_i^{T_U})^{N_{TU}=5000}$ $D_{T_L} = (x_i^{T_L}, y_i^{T_L})^{N_{TL}=358}$



Sundaresan et al., MedIA, 2021

Loizillon et al., Medical Image Analysis, 2025

Validation on Real-World Data

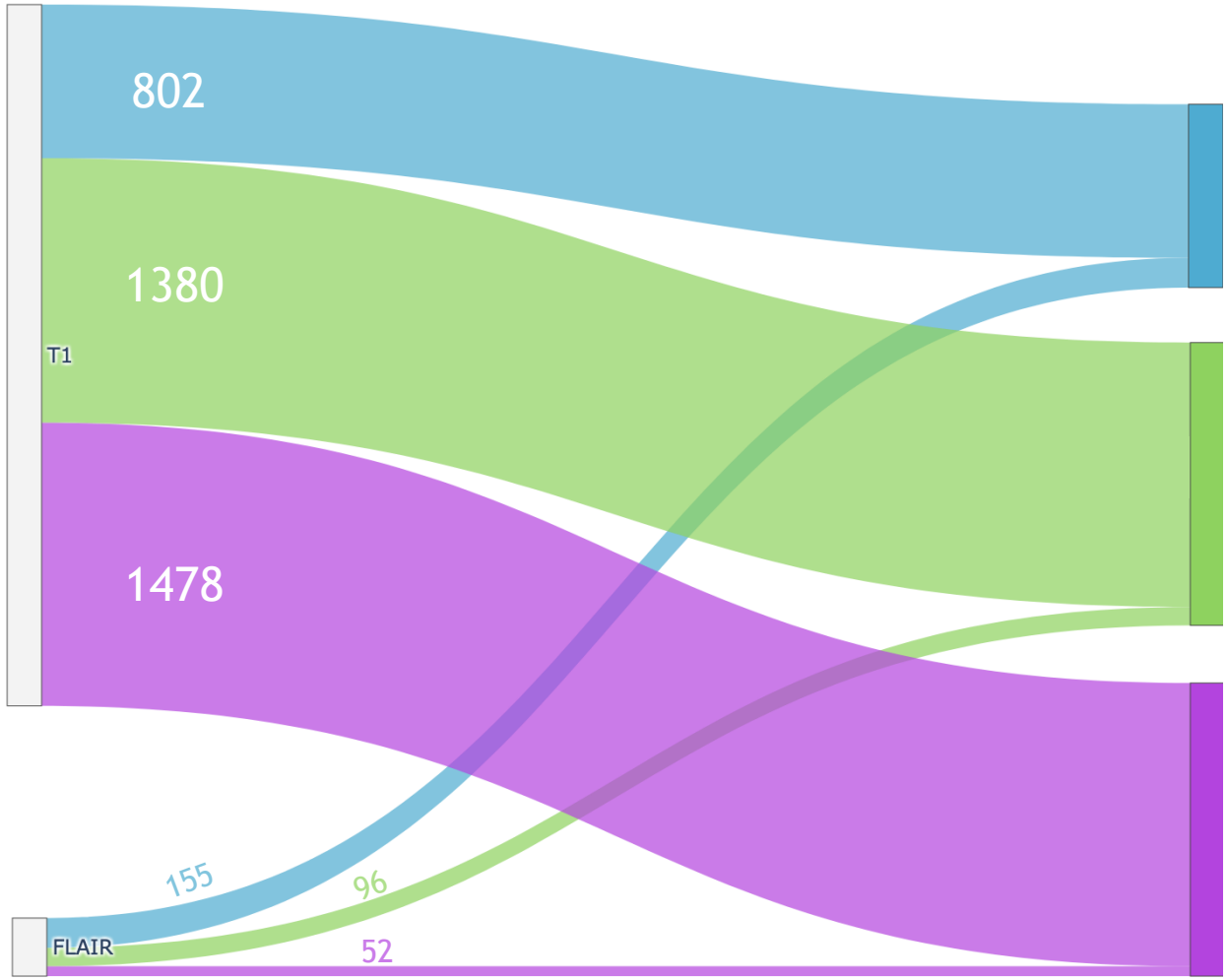
Domain adaptation for 3D FLAIR automatic quality control

Source domain

$$D_S = (x_i^S, y_i^S)^{N_S=5000}$$

Target domain

$$D_{T_L} = (x_i^{T_L}, y_i^{T_L})^{N_{TL}=303}$$



Loizillon et al., Medical Image Analysis, 2025

Validation on Real-World Data

Domain adaptation for 3D FLAIR automatic quality control

Source domain

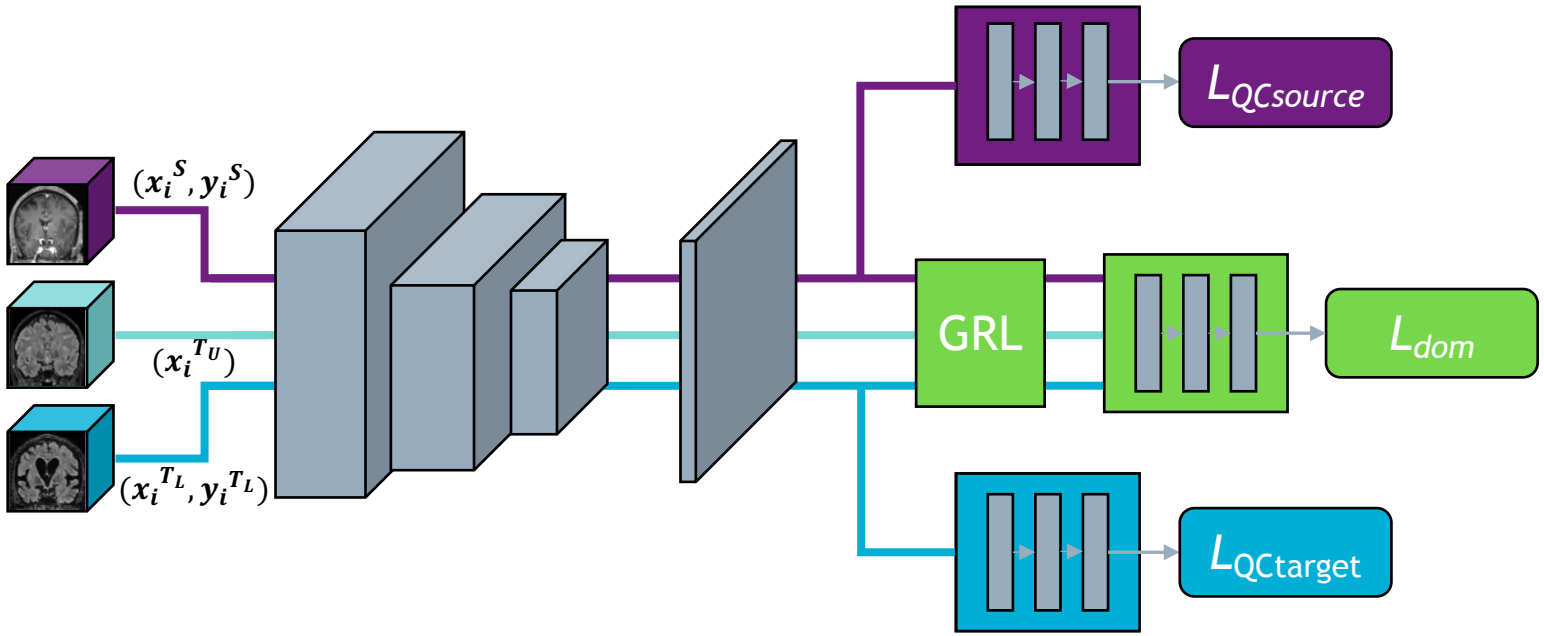
5000 3D T1w MRIs
manually annotated

$D_S = (x_i^S, y_i^S)^{N_S=5000}$

Target domain

5358 3D FLAIR MRIs, with
358 manually annotated

$D_{T_U} = (x_i^{T_U})^{N_{TU}=5000}$ $D_{T_L} = (x_i^{T_L}, y_i^{T_L})^{N_{TL}=358}$



Loizillon et al., Medical Image Analysis, 2025

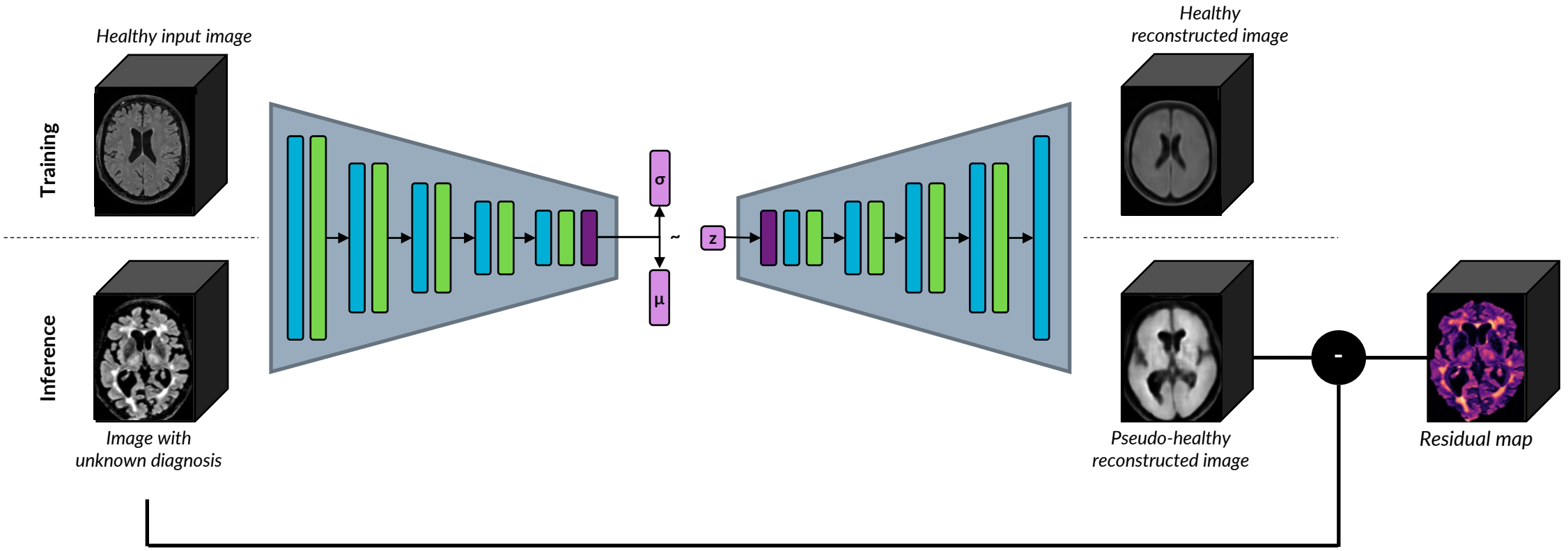
Domain adaptation for 3D FLAIR automatic quality control

	Straight reject (yes vs no)	Low vs Medium/Good	Medium vs Good
Annotators	94.67 [91.32,97.40]	86.31 [84.25,88.25]	84.43 [82.54,86.50]
Loizillon et al.	89.27 [79.69,97.08]	79.81 [74.65,84.71]	73.92 [70.73,77.22]

Balanced accuracy (%)

Loizillon et al., Medical Image Analysis, 2025

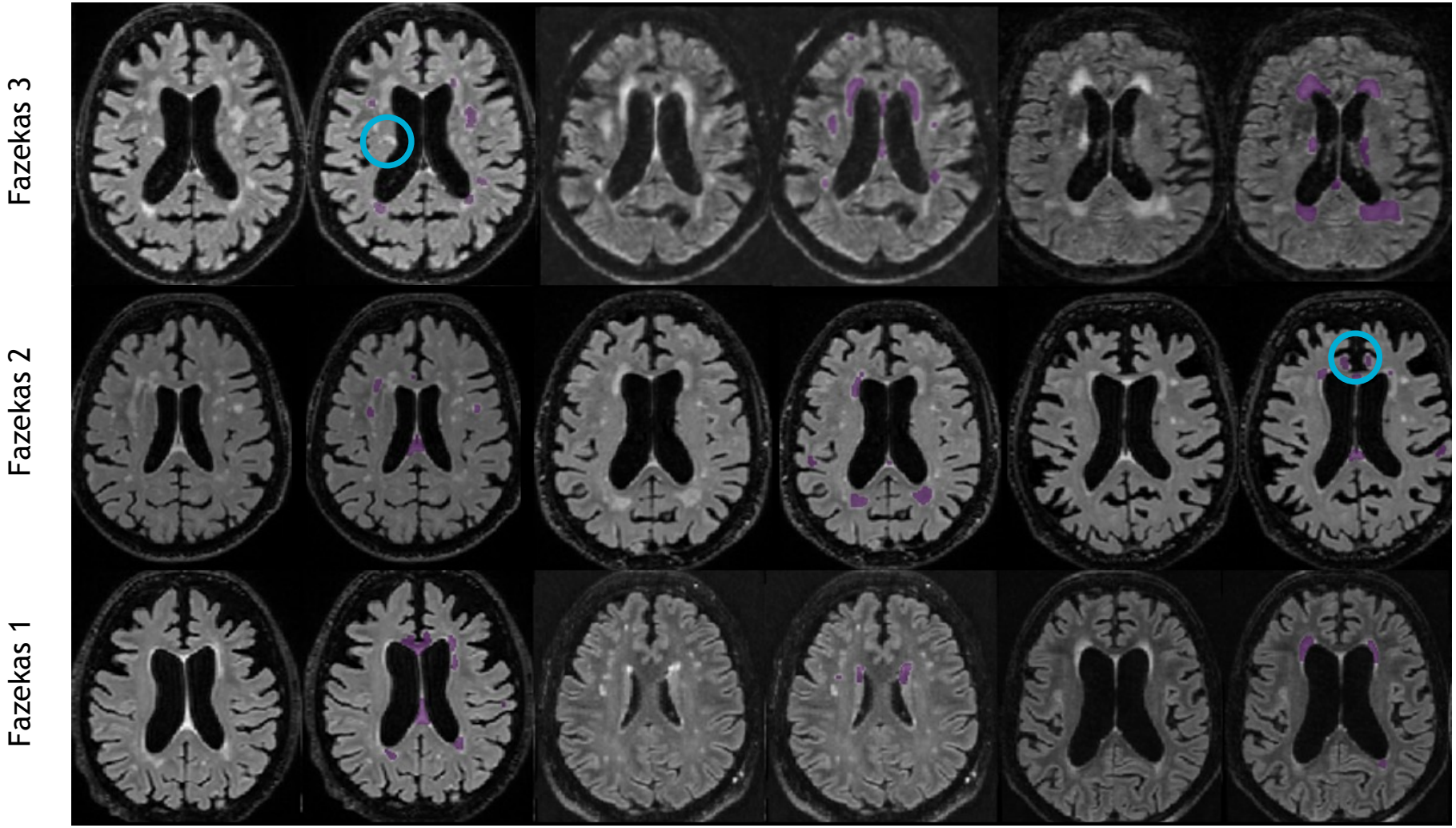
Detection of age-related white matter hyperintensities in 3D FLAIR MRI



Loizillon et al., MIDL, 2024

Validation on Real-World Data

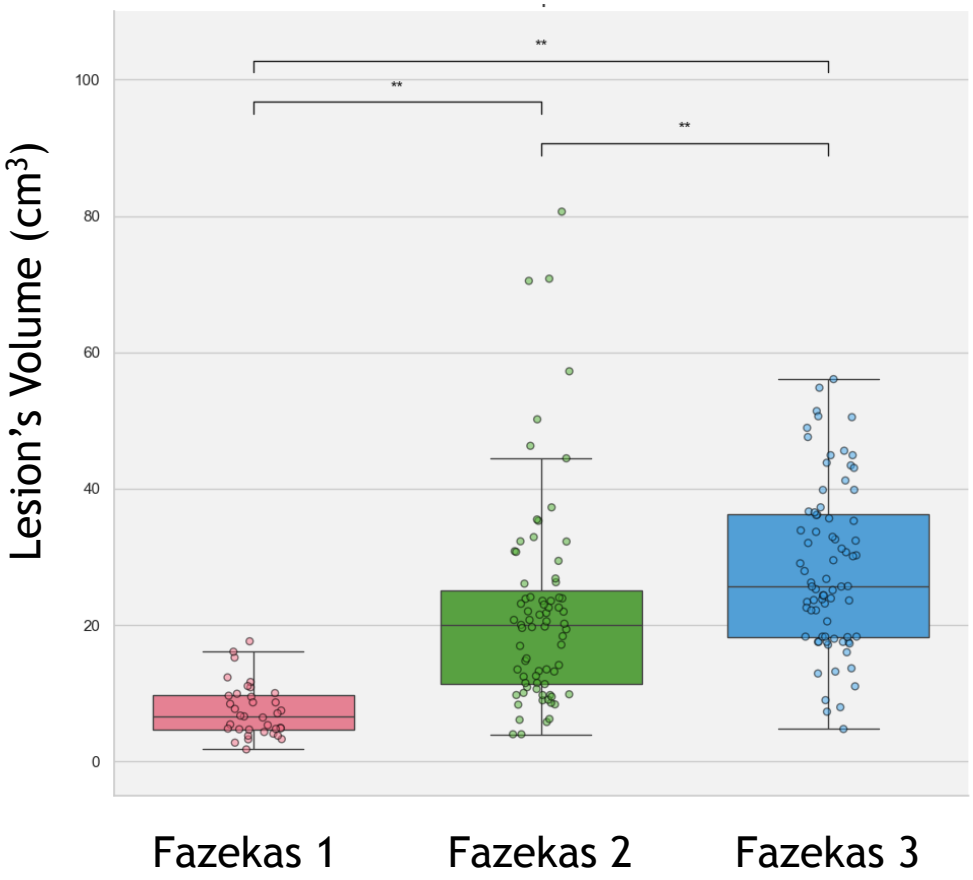
Detection of age-related white matter hyperintensities



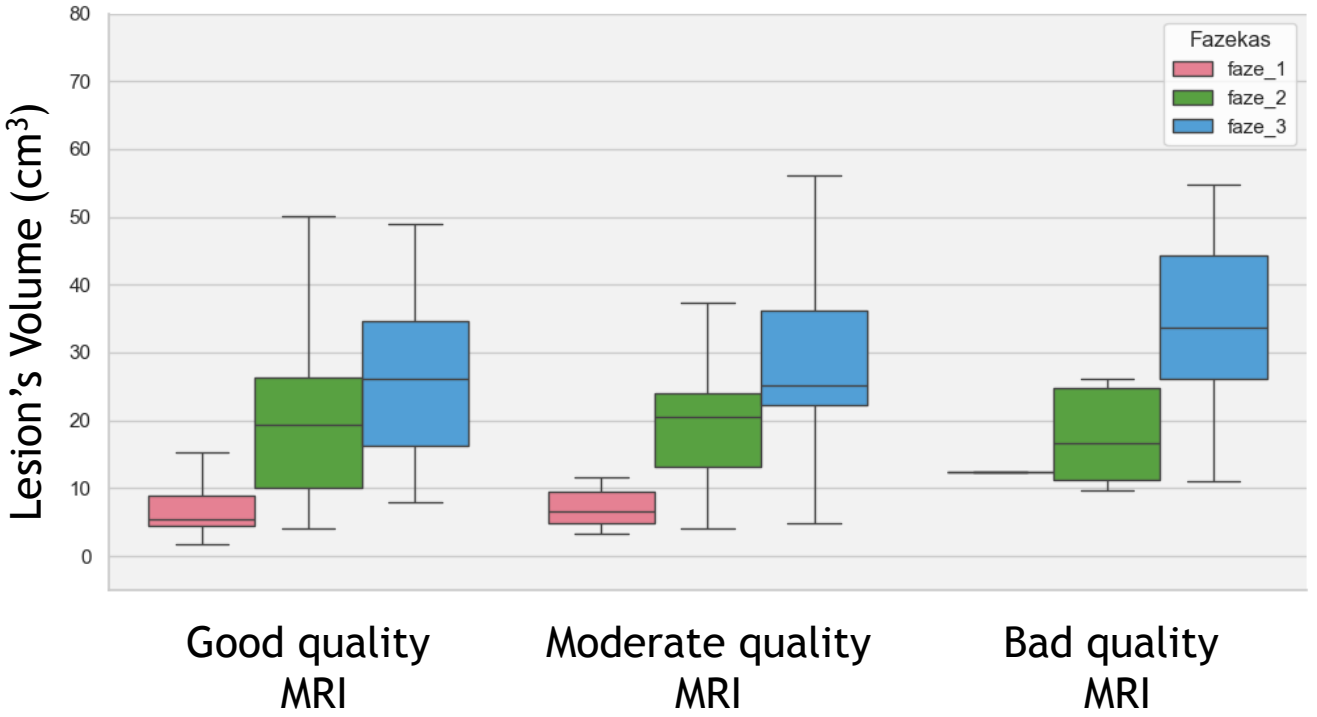
Loizillon et al., MIDL, 2024

Validation on Real-World Data

Detection of age-related white matter hyperintensities in 3D FLAIR MRI



- Robust to differences in image quality:



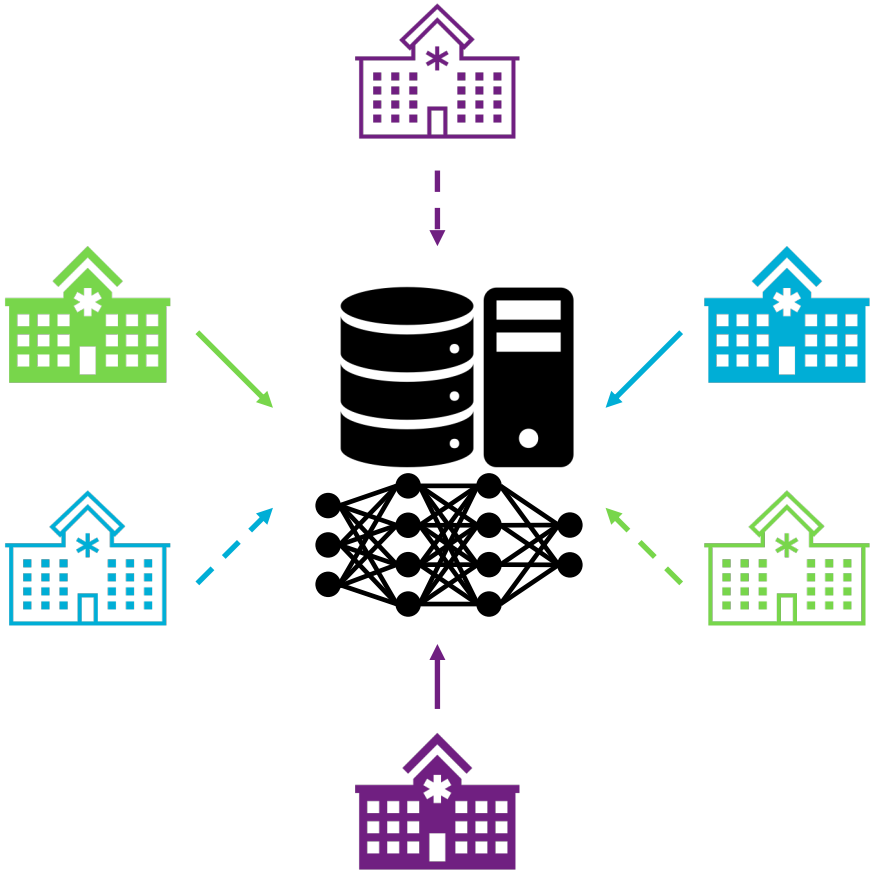
Loizillon et al., MIDL, 2024

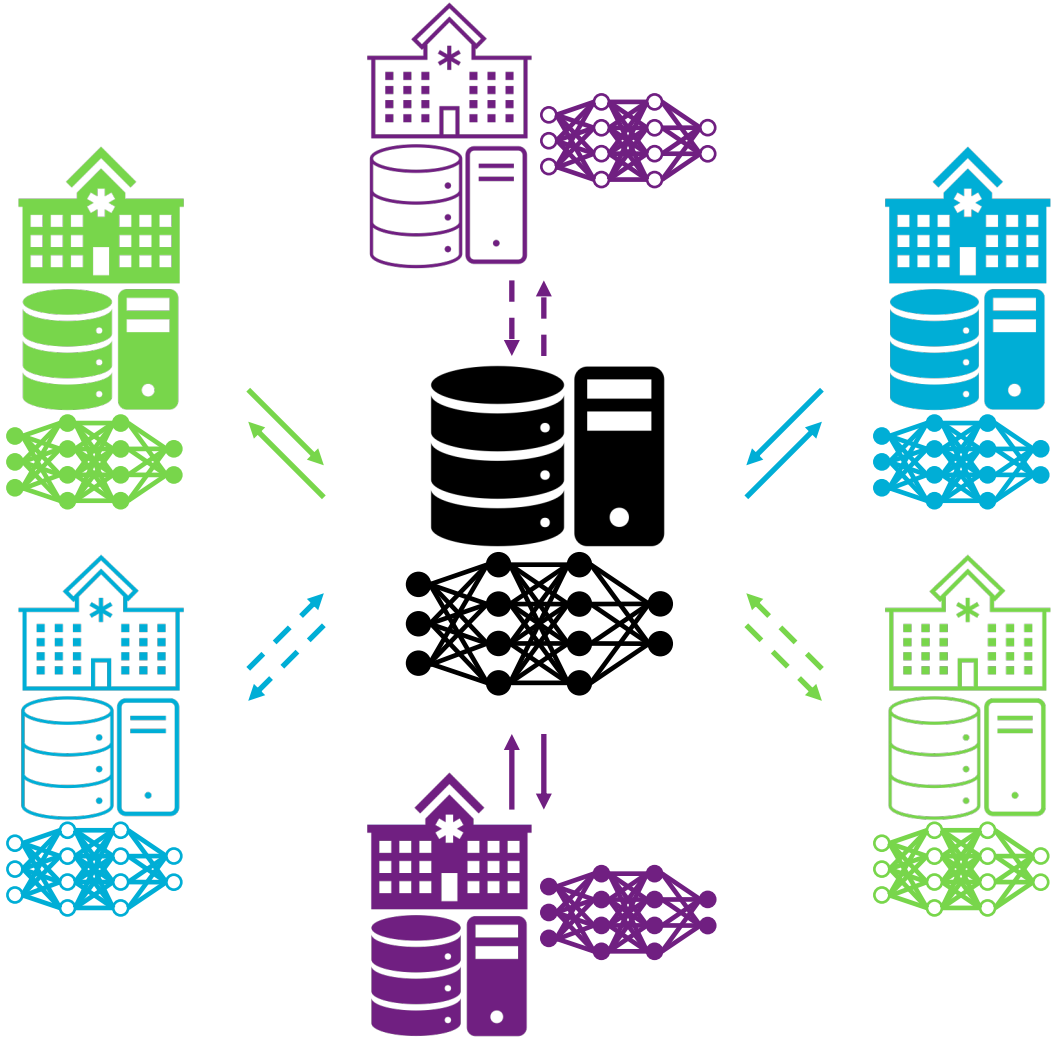
Validation on Real-World Data

Conclusions

- **Quality control** & **data set homogenisation** crucial to enable translation of computer-aided diagnosis tools to clinical practice
- Not all AI approaches are appropriate for use in clinical practice
- **Size** is not the only characteristic of data sets that matters

Centralised vs federated learning





Federated learning

Objective:

$$\min_w F(w), \qquad \text{where } F(w) := \sum_{k=1}^{m} p_k F_k(w)$$

m: total number of nodes

$p_k \geq 0$

$\sum_k p_k = 1$

F_k: local objective function for the k^{th} node

Node contribution in federated averaging framework (MacMahan et al., AISTATS, 2017):

$$p_k = \frac{n_k}{n}$$

n_k: number of datapoints in the node

n: total number of observations studied

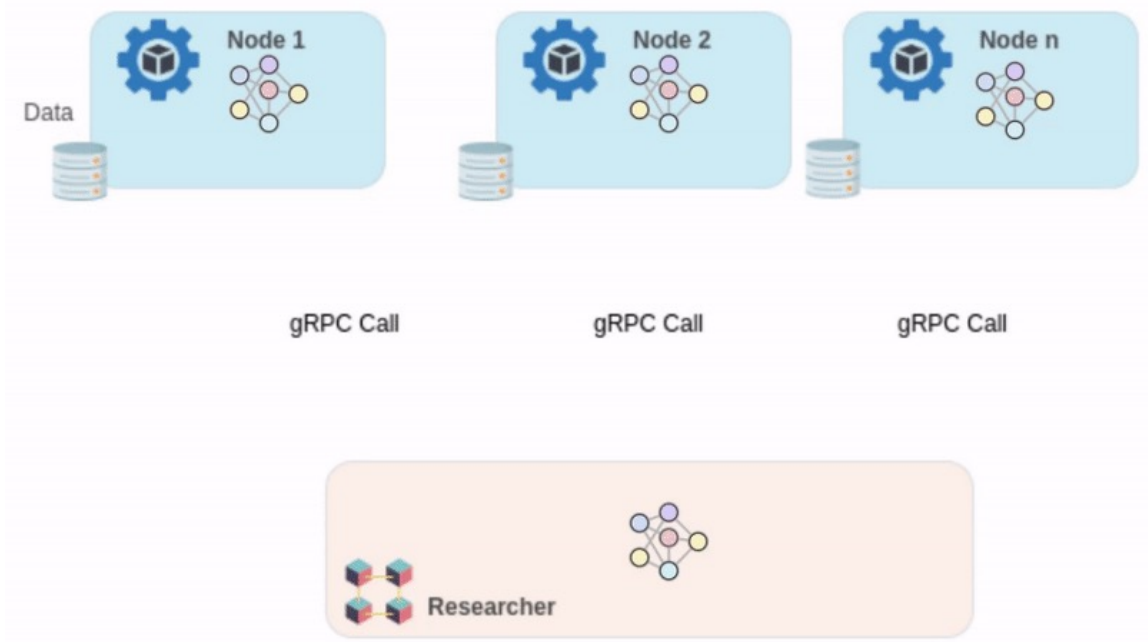
Validation on Real-World Data

Federated learning

Fed-BioMed

'Open, Transparent and Trusted Collaborative Learning for Real-World Healthcare Applications'

fedbiomed.org



Ensuring Robust AI in Medical Imaging

- Validation on Real-World Data
- **Meaningful Metrics**
- Rigorous Statistical Analysis

Meaningful Metrics

A (fake & slightly exaggerated) typical MICCAI conclusion

‘Our proposed deep learning framework achieves superior accuracy (86.6%) and robust generalizability in distinguishing Alzheimer’s disease patients from cognitively normal subjects, outperforming established methods on the ADNI dataset. The model provides anatomically interpretable results by highlighting disease-relevant regions such as the hippocampus, demonstrates high reproducibility and fast inference, and, by eliminating manual feature engineering while aligning with clinical biomarkers, is both practical and explainable—making it ready for immediate integration into routine diagnostic workflows.’

Meaningful Metrics

Image level binary classification

Confusion matrix:

True diagnosis \ Predicted diagnosis		
	True positive (TP)	False negative (FN)
	False positive (FP)	True negative (TN)

Accuracy:

$$ACC = \frac{TP + TN}{TP + FN + FP + TN}$$

Balanced accuracy:

$$\text{Bal. ACC} = 0.5 \times \left(\frac{TP}{TP + FN} + \frac{TN}{TN+FP}\right)$$

Scenario A

- 50 AD patients
- 50 CN subjects

	AD	CN
AD	45	5
CN	5	45

$$ACC = \frac{45 + 45}{45 + 5 + 5 + 45} = 0.9$$

$$\text{Bal. ACC} = 0.5 \times \left(\frac{45}{5 + 45} + \frac{45}{45+5}\right) = 0.9$$

Scenario B

- 50 AD patients
- 500 CN subjects

	AD	CN
AD	5	45
CN	10	490

$$ACC = \frac{5 + 490}{5 + 45 + 10 + 490} = 0.9$$

$$\text{Bal. ACC} = 0.5 \times \left(\frac{5}{5 + 45} + \frac{490}{490+10}\right) = 0.54$$

Meaningful Metrics

Brain tumour segmentation

Popular voxel-based metrics fail to capture clinical interest

Magnetic resonance imaging, same patient, different slices

Medical example: brain-tumor segmentation
A near-perfect voxel-level sensitivity hides information on missed lesions

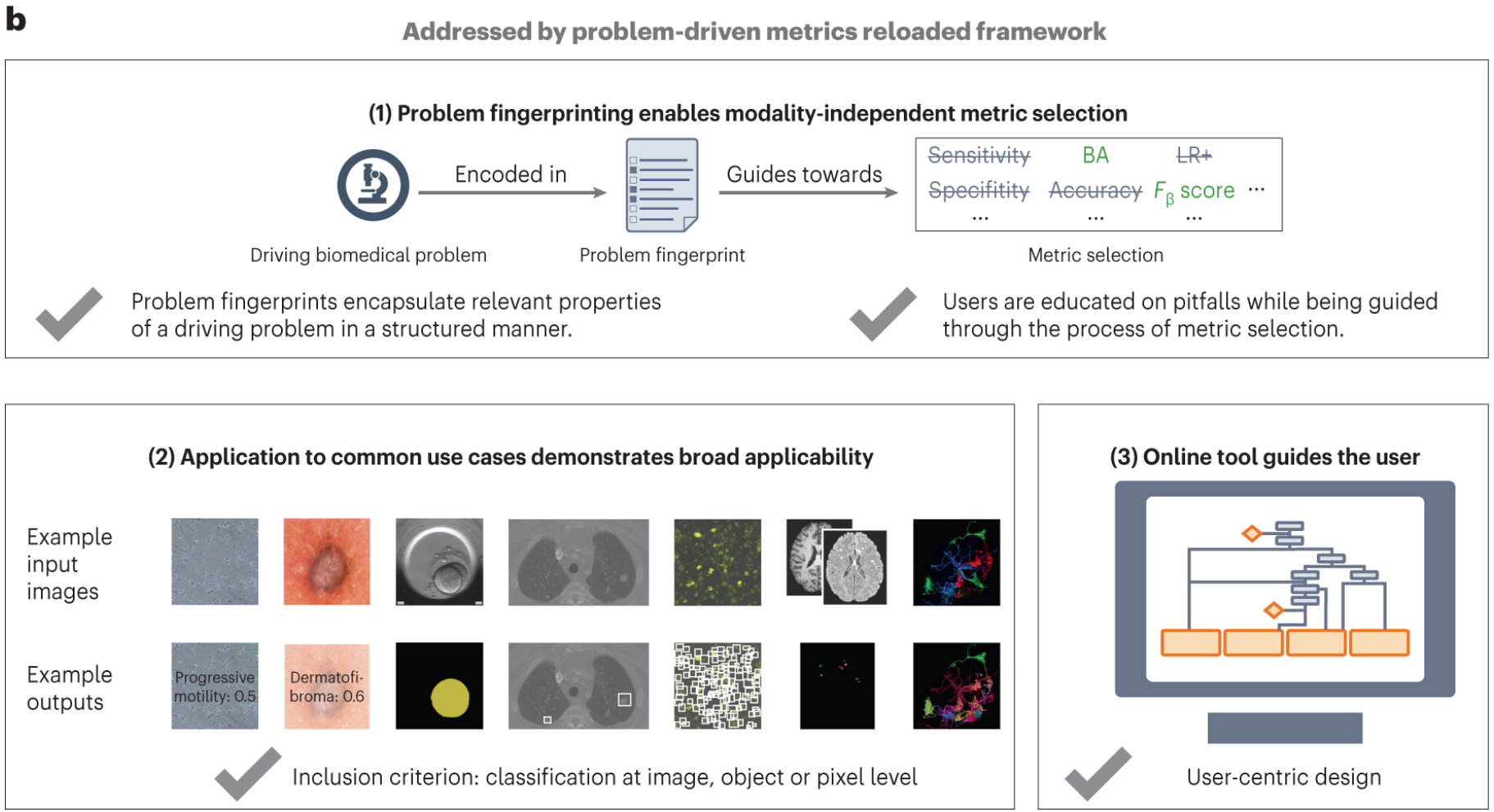
Reinke et al., Nature Methods, 2024

Meaningful Metrics

https://metrics-reloaded.dkfz.de

Meaningful Metrics

Metrics Reloaded



Maier-Hein et al., Nature Methods, 2024

Meaningful Metrics

Choose your tool

Start your selection

Problem Category Selection
Mapping a given research problem to the appropriate image processing task.

Metric Selection
Metric recommendation depending on the task type and specifications of the given problem.

Discover Metrics

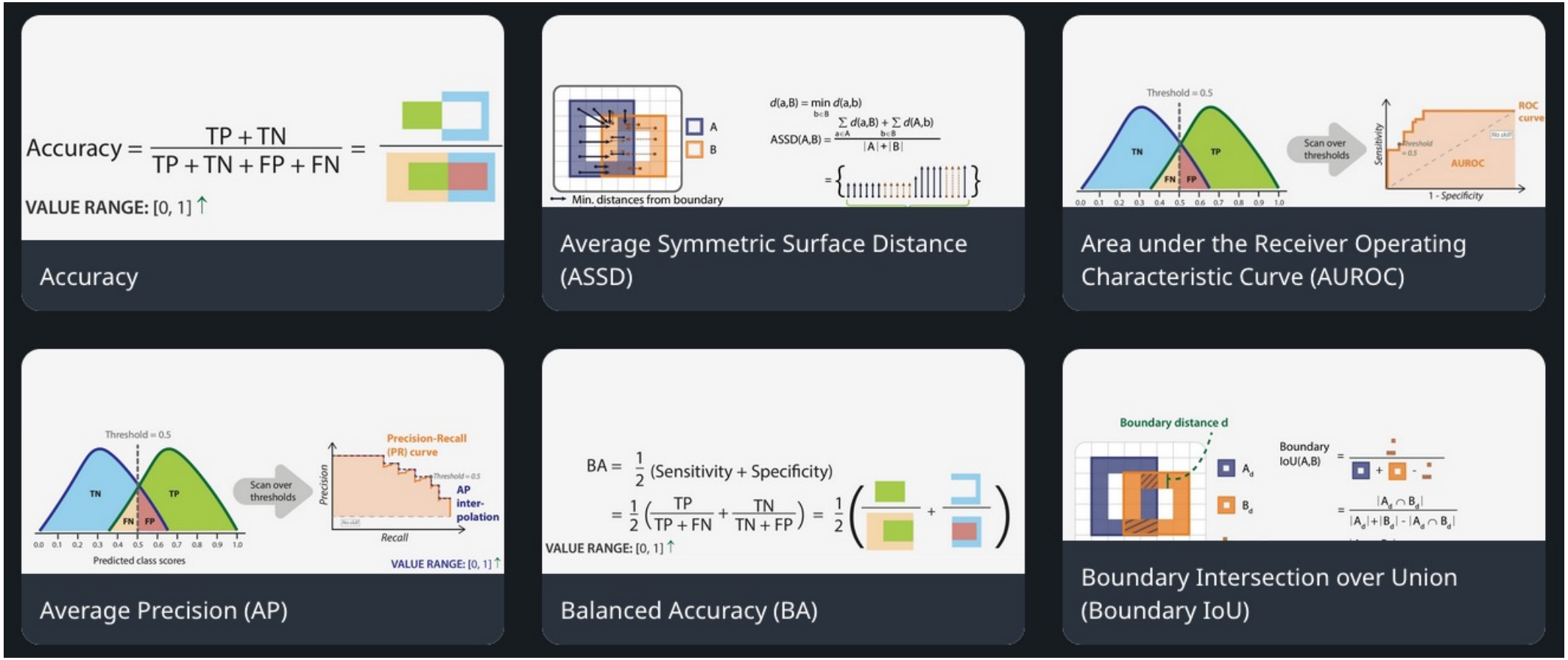
Metric Library
A collection of all metrics in our database including their definitions, references and restrictions.

https://metrics-reloaded.dkfz.de

Meaningful Metrics

Metrics Reloaded

- Metrics library



https://metrics-reloaded.dkfz.de

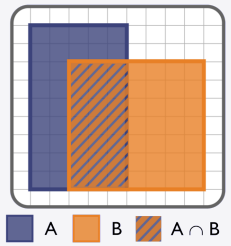
Meaningful Metrics

Metrics Reloaded

- Metrics library

DICE SIMILARITY COEFFICIENT (DSC)

Synonyms: Dice, Dice Coefficient, Sørensen–Dice Coefficient, F_1 Score, Balanced F Score



$$DSC(A,B) = \frac{2}{\square + \square} = \frac{2\,|A \cap B|}{|A| + |B|} = \frac{2\,PPV \cdot Sensitivity}{PPV + Sensitivity}$$

VALUE RANGE: [0, 1] ↑

DESCRIPTION
DSC measures the overlap between two structures.

RECOMMENDED FOR

ImLC	SemS	ObD	InS
●	●	●	●

CARDINALITIES

TP	FP	FN	TN
●	●	●	○

IMPORTANT RELATIONS
DSC is closely related to the IoU = Jaccard index: $DSC = \frac{2\,IoU}{1 + IoU}$

DSC is equal to the F_1 Score ($\beta = 1$ in F_β Score) at pixel level.

DEFINITION
[Dice, 1945]

METRIC FAMILY

Counting metric	Multi-threshold metric	Distance-based metric
●	○	○

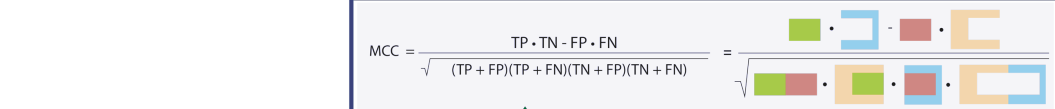
RELEVANT PITFALLS
- DSC is unaware of shapes, distances and centers (Figs. 4a, SN 2.5, SN 2.12, Extended Data Fig. 1b in [Reinke et al., 2023]).
- DSC penalizes missed pixels more in small objects (Figs. SN 2.10, SN 2.12, Extended Data Fig. 1a in [Reinke et al., 2023]).
- DSC is undefined if both the reference and prediction are empty, and ad hoc rules need to be established depending on the use case (Fig. SN 2.18 in [Reinke et al., 2023]).
- DSC treats oversegmentation and undersegmentation differently (Fig. SN 2.8 in [Reinke et al., 2023]).
- DSC does not compensate for inter-rater variability (Fig. SN 2.17 in [Reinke et al., 2023]).
- DSC behaves differently in 2D and 3D settings. In 3D settings, the additional z-dimension results in a cubical increase in erroneous pixels [Reinke et al., 2021].

RECOMMENDATIONS
- An overlap-based metric (by default the DSC or IoU) should be used in most cases of segmentation assessment. An exception is the case of consistently tiny structures along with a noisy reference.
- DSC should generally be used in combination with a boundary-based metric if boundaries are of interest.
- DSC should generally not be considered if…
  - … there is a high variability of structure sizes within an image or across images.
  - … inter-rater variability is requested to be compensated.
  - … over- and undersegmentation should be treated similarly.
- DSC should be considered as a metric in the medical community rather than in the computer vision and biology communities (where the almost identical IoU is preferred).

MATTHEWS CORRELATION COEFFICIENT (MCC)

Synonyms: Phi Coefficient

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$



VALUE RANGE: [-1, 1] ↑
A value of 0 refers to a prediction which is not better than random guessing.

DESCRIPTION
MCC measures the correlation between the actual and the predicted class.

DEFINITION
[Matthews, 1975]

RECOMMENDED FOR

ImLC	SemS	ObD	InS
●	○	○	○

PREVALENCE DEPENDENCY ●

CARDINALITIES

TP	FP	FN	TN
●	●	●	●

METRIC FAMILY

Counting metric	Multi-threshold metric	Distance-based metric
●	○	○

IMPORTANT RELATIONS
MCC can be rewritten as:

$$MCC = \sqrt{PPV \cdot Sensitivity \cdot Specificity \cdot NPV} - \sqrt{(1 - PPV) \cdot (1 - Sensitivity) \cdot (1 - Specificity) \cdot (1 - NPV)}$$

MCC is equivalent to the geometric mean of Markedness and Informedness.

MULTI-CLASS DEFINITION
For C classes, MCC can be defined as:

$$MCC = \frac{\sum_{i=1}^{C}\sum_{j=1}^{C}\sum_{k=1}^{C} n_{ii} \cdot n_{jk} - n_{ij} \cdot n_{ki}}{\sqrt{\sum_{i=1}^{C}\left(\sum_{j=1}^{C} n_{ij}\right)\left(\sum_{i'|i' \neq i}\sum_{j'=1}^{C} n_{i'j}\right)} \sqrt{\sum_{i=1}^{C}\left(\sum_{j=1}^{C} n_{ji}\right)\left(\sum_{i'|i' \neq i}\sum_{j'=1}^{C} n_{j'i'}\right)}}$$

n_{ij}: *entry of the confusion matrix for row i and column j, i.e., samples of actual class i that were predicted as class j*

RELEVANT PITFALLS
- MCC is prevalence-dependent, thus not comparable across data sets with different prevalences (Fig. SN 2.7 in [Reinke et al., 2023], [Reinke et al., 2021]).
- MCC does not allow for incorporating unequal severity of confusions across classes. This implies that the metric is not applicable to cost-benefit analyses (Fig. SN 2.9 in [Reinke et al., 2023]), or when target classes are related on an ordinal scale (Fig. 4b in [Reinke et al., 2023]).
- The theoretical lower bound of MCC (-1) may not always be achievable (Fig. SN 2.35 in [Reinke et al., 2023]).
- MCC is hard to interpret [Zhu 2020].
- Compared to other metrics like EC, MCC lacks a framework to identify and validate the decision rule applied to predicted class scores [Ferrer 2022].
- MCC depends on the definition of TN (undefined for ObD and InS).

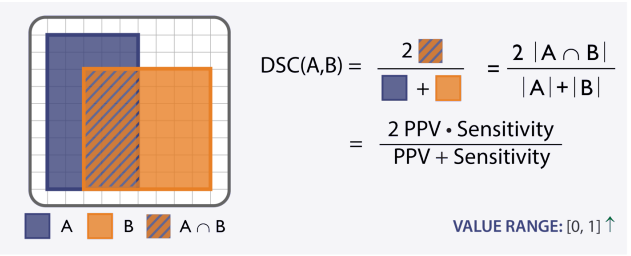
RECOMMENDATIONS
- MCC should not be used/used with care if…
  - … class confusions are of unequal severity (example: ordinal target classes).
  - … the provided class prevalences do not reflect the population of interest.
  - … there is a mismatch between class prevalences and class importance.
  - … compensation for class imbalance is not requested.
- Otherwise, MCC should be used as a multi-class metric specifically if all basic error rates (Sensitivity, Specificity, PPV, NPV) should be captured in one score.
- MCC scores should be carefully interpreted in the presence of class imbalance as the distribution becomes skewed [Zhu 2020].

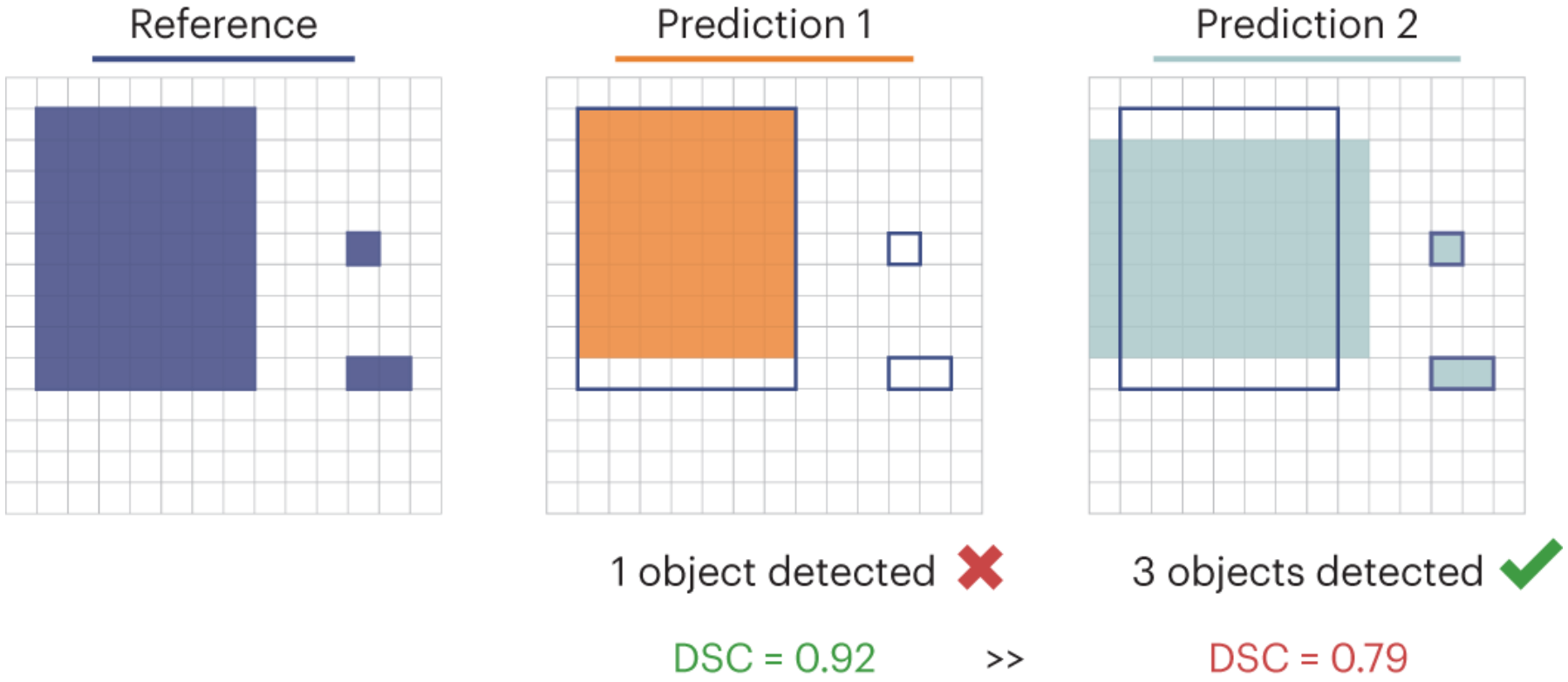
https://metrics-reloaded.dkfz.de

Meaningful Metrics

Metrics Reloaded

- Pitfalls related to the inadequate choice of the problem category





Reinke et al., Nature Methods, 2024

Meaningful Metrics

Metrics Reloaded

- Pitfalls related to poor metric selection
  - Disregard of the domain interest

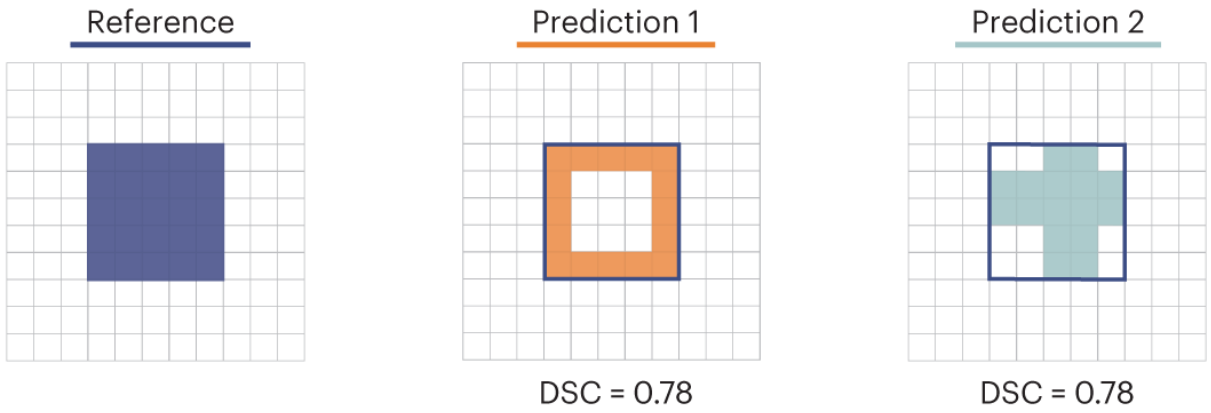
$$EC = w_{miss} \cdot \underbrace{\frac{FN}{TP+FN}}_{P_{miss}} \cdot \underbrace{\frac{TP+FN}{TP+TN+FP+FN}}_{P_{tar}} + w_{FA} \cdot \underbrace{\frac{FP}{TN+FP}}_{P_{FA}} \cdot \underbrace{1 - \frac{TP+FN}{TP+TN+FP+FN}}_{P_{tar}}$$

$$= w_{miss} \cdot \frac{\square}{\square} \cdot \frac{\square}{\square} + w_{FA} \cdot \frac{\square}{\square} \cdot 1 - \frac{\square}{\square}$$

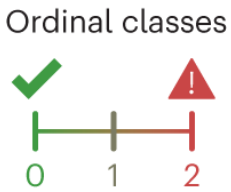
P_{miss}: FN (miss) rate, P_{FA}: FP (false alarm) rate
P_{tar}: prior probability (prevalence)
w_{miss}/w_{FA}: (estimation of) costs of the respective errors; can be adjusted as a weighting of them.

VALUE RANGE: $[0, \infty) \downarrow$
EC can be assumed to be positive if costs are non-negative, which can be done without loss of generality.

a Overlap-based metrics disregard structure boundaries



b Common multi-class metrics ignore ordinal grading



	Reference	Prediction 1	Prediction 2
Patient 1	Class 0	Class 0	Class 0
Patient 2	Class 1	Class 1	Class 1
Patient 3	Class 2	Class 0 ✗	Class 1 ✓

Accuracy = 0.67 = Accuracy = 0.67
EC = 0.83 >> EC = 0.33

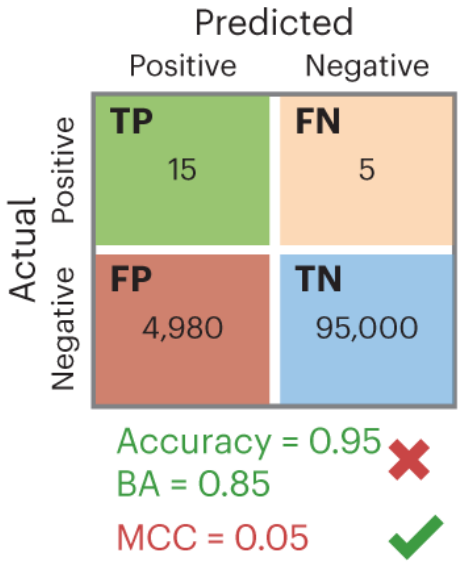
Reinke et al., Nature Methods, 2024

Meaningful Metrics

Metrics Reloaded

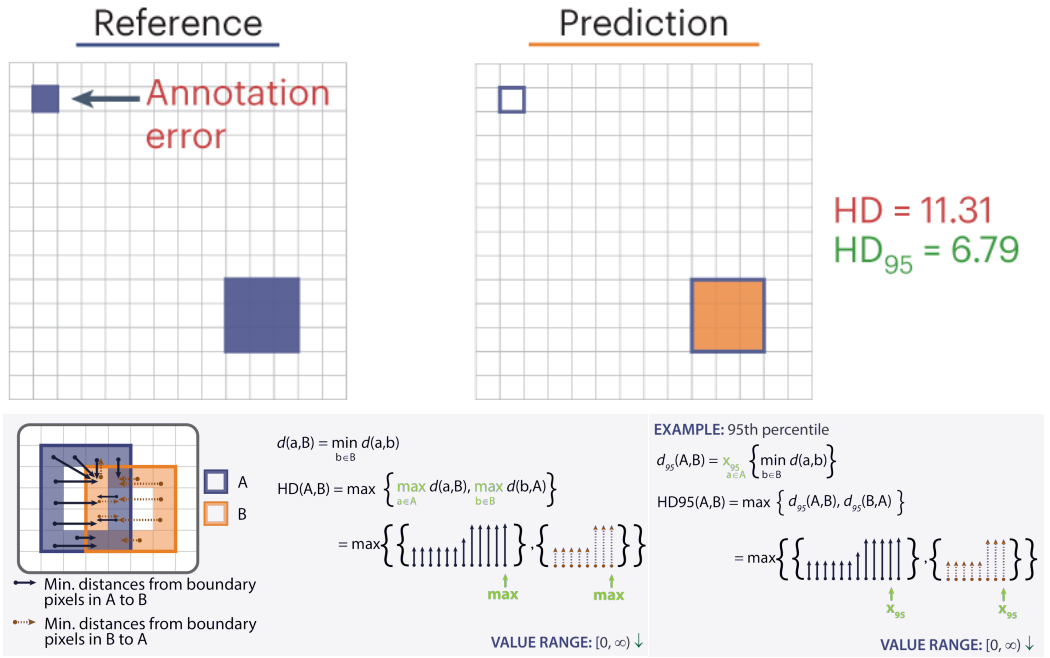
- Pitfalls related to poor metric selection
  - Disregard of the properties of the dataset

Common metrics yield implausible results in the presence of class imbalance



$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Annotation errors may have huge impact on metric scores

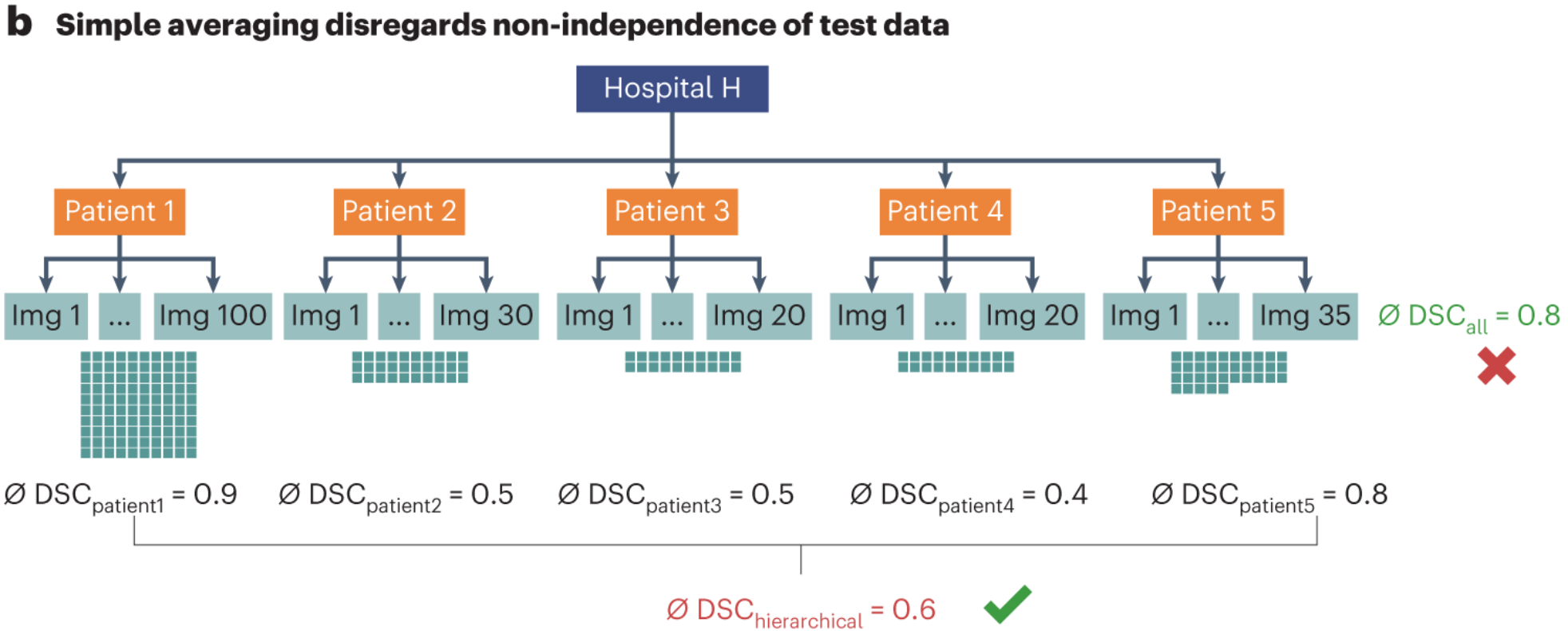


Reinke et al., Nature Methods, 2024

Meaningful Metrics

Metrics Reloaded

- Pitfalls related to poor metric application



Reinke et al., Nature Methods, 2024

Metrics Reloaded

Perspective | Published: 12 February 2024

Understanding metric-related pitfalls in image analysis validation

Annika Reinke, Minu D. Tizabi, Michael Baumgartner, Matthias Eisenmann, Doreen Heckmann-Nötzel, A. Emre Kavur, Tim Rädsch, Carole H. Sudre, Laura Acion, Michela Antonelli, Tal Arbel, Spyridon Bakas, Arriel Benis, Florian Buettner, M. Jorge Cardoso, Veronika Cheplygina, Jianxu Chen, Evangelia Christodoulou, Beth A. Cimini, Keyvan Farahani, Luciana Ferrer, Adrian Galdran, Bram van Ginneken, Ben Glocker, ... Lena Maier-Hein + Show authors

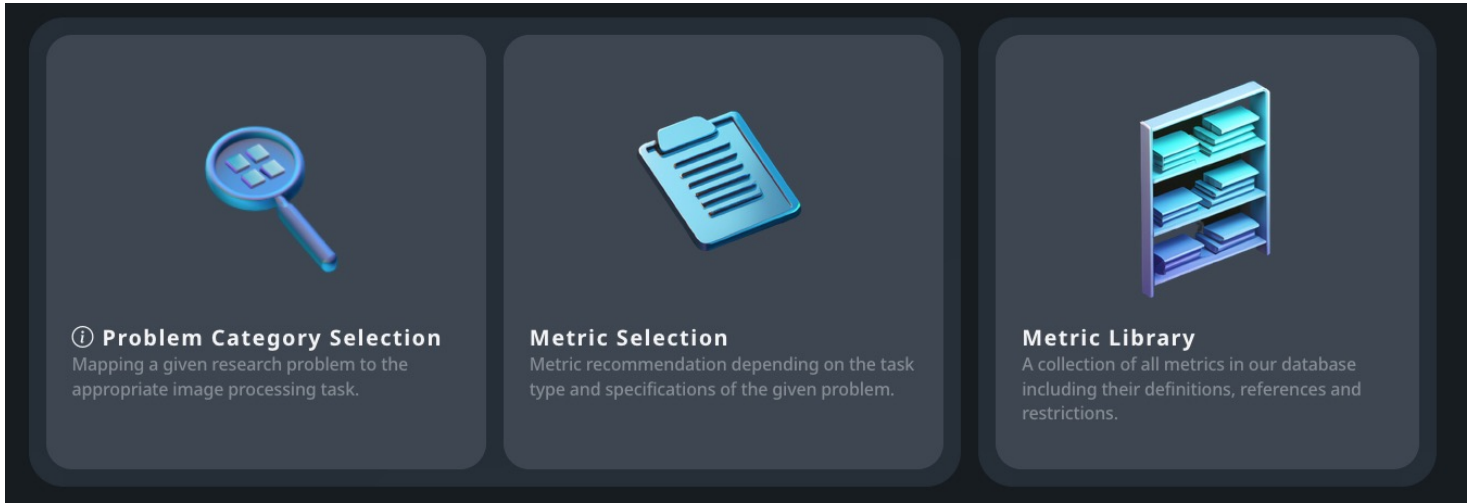
Nature Methods **21**, 182–194 (2024) | Cite this article

Perspective | Published: 12 February 2024

Metrics reloaded: recommendations for image analysis validation

Lena Maier-Hein, Annika Reinke, Patrick Godau, Minu D. Tizabi, Florian Buettner, Evangelia Christodoulou, Ben Glocker, Fabian Isensee, Jens Kleesiek, Michal Kozubek, Mauricio Reyes, Michael A. Riegler, Manuel Wiesenfarth, A. Emre Kavur, Carole H. Sudre, Michael Baumgartner, Matthias Eisenmann, Doreen Heckmann-Nötzel, Tim Rädsch, Laura Acion, Michela Antonelli, Tal Arbel, Spyridon Bakas, Arriel Benis, ... Paul F. Jäger + Show authors

Nature Methods **21**, 195–212 (2024) | Cite this article



Meaningful Metrics

Metrics Reloaded

- Metric implementation
  - Use reference implementations
  - Read metric-specific recommendations in the cheat sheets
- Aggregation
  - Address the potential correlation between classes when aggregating
  - Complement validation with multi-class metrics
  - Respect the hierarchical data structure when aggregating metrics
  - Leverage metadata to reveal potential algorithmic bias
  - Follow category-specific aggregation strategy
- Interpretation
  - Read metric-related recommendations to obtain awareness of the pitfall
  - Report on the quality of the reference (e.g. intra-rater and inter-rater variability).

Maier-Hein et al., Nature Methods, 2024

Ensuring Robust AI in Medical Imaging

- Validation on Real-World Data
- Meaningful Metrics
- Rigorous Statistical Analysis

A (fake & slightly exaggerated) typical MICCAI conclusion

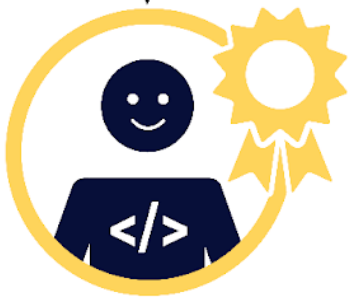
‘Our proposed deep learning framework achieves superior accuracy (86.6%) and robust generalizability in distinguishing Alzheimer’s disease patients from cognitively normal subjects, outperforming established methods on the ADNI dataset. The model provides anatomically interpretable results by highlighting disease-relevant regions such as the hippocampus, demonstrates high reproducibility and fast inference, and, by eliminating manual feature engineering while aligning with clinical biomarkers, is both practical and explainable—making it ready for immediate integration into routine diagnostic workflows.’

Rigorous Statistical Analysis

Common practice in medical imaging algorithm performance analysis

Commonly encountered results tables

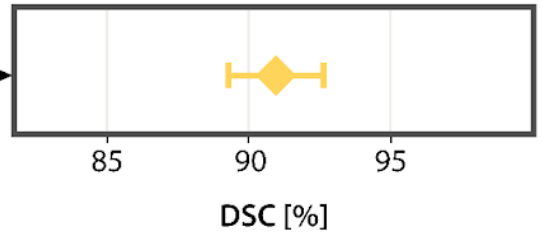
Methods	DSC	HD95
Method 1	86.82	43.22
Method 2	87.64	63.68
Method 3	90.67	67.34
Proposed method	**90.70**	**67.35**



?

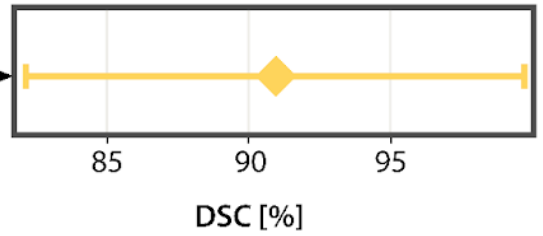
Unclear from results table

Scenario 1 (narrow CI; desired)



Mean DSC — CI

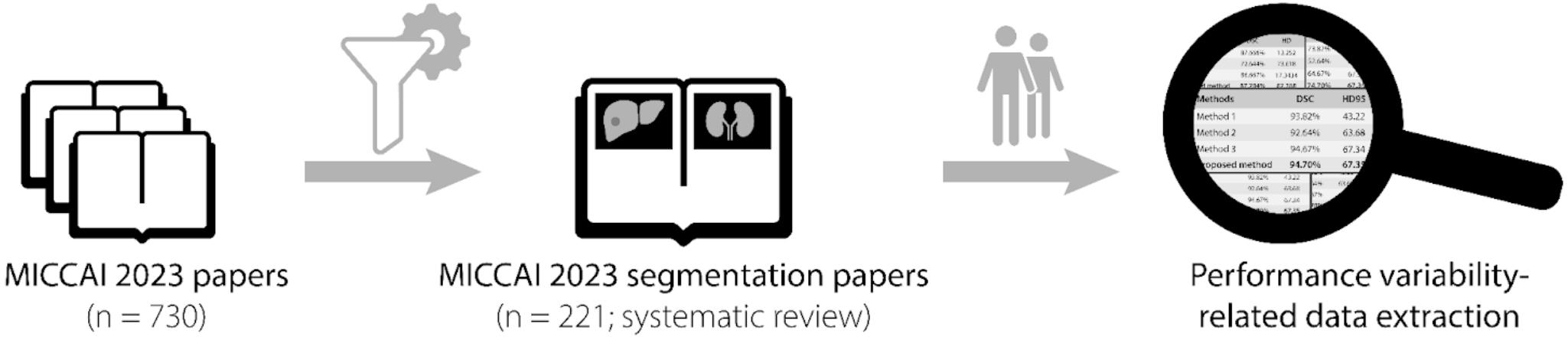
Scenario 2 (wide CI)



Christodoulou et al., MICCAI, 2024

Rigorous Statistical Analysis

Common practice in medical imaging algorithm performance analysis



Christodoulou et al., MICCAI, 2024

Rigorous Statistical Analysis

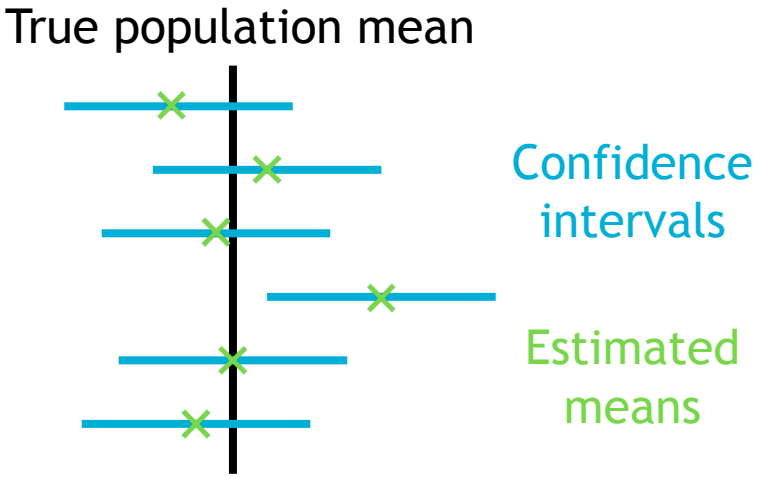
Standard deviation & confidence interval

- **Standard deviation (SD)**

  Measure of the dispersion or spread of data points from the mean value

- **Confidence interval (CI)**

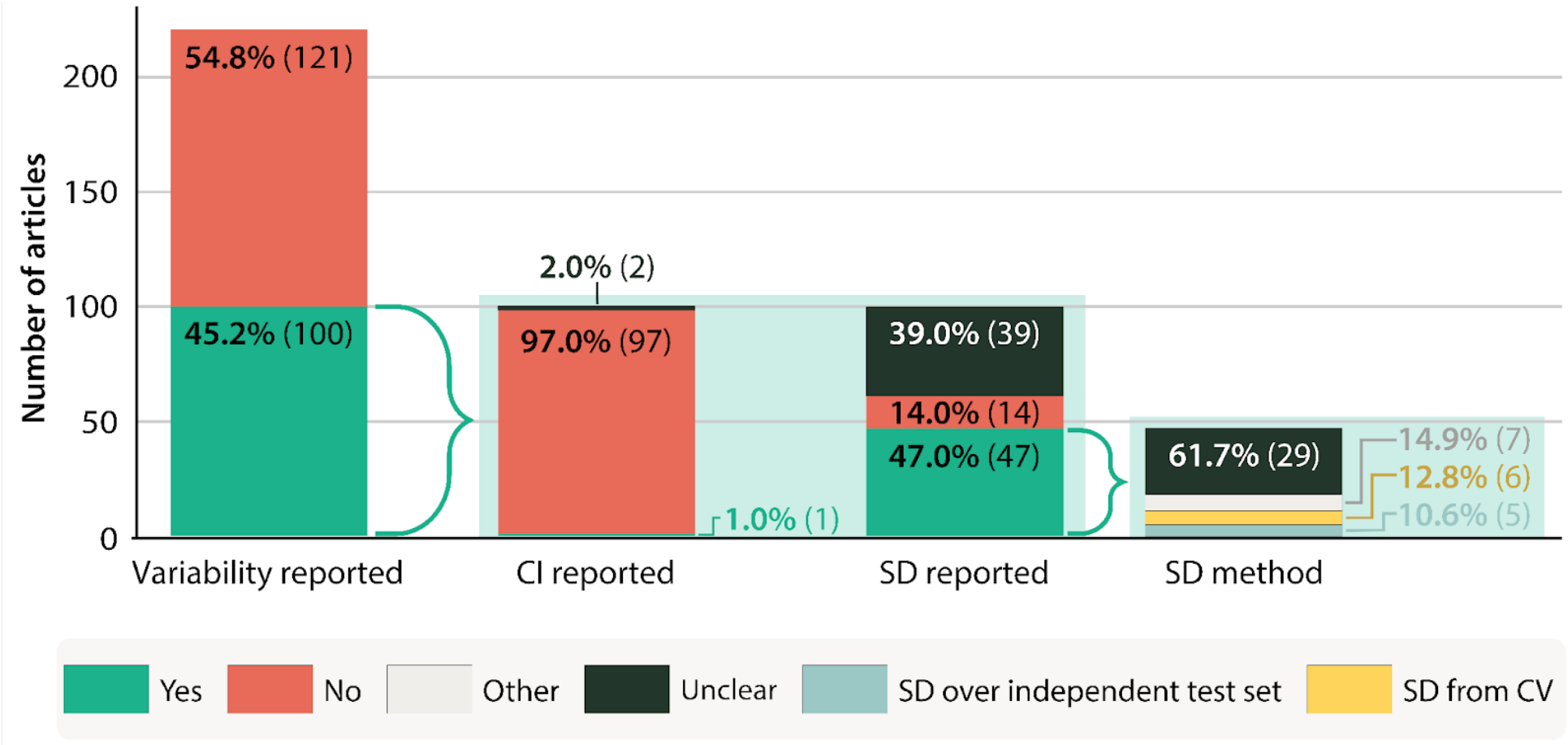
  Range within which a population parameter is expected to lie with a certain level of confidence.



Christodoulou et al., MICCAI, 2024

Rigorous Statistical Analysis

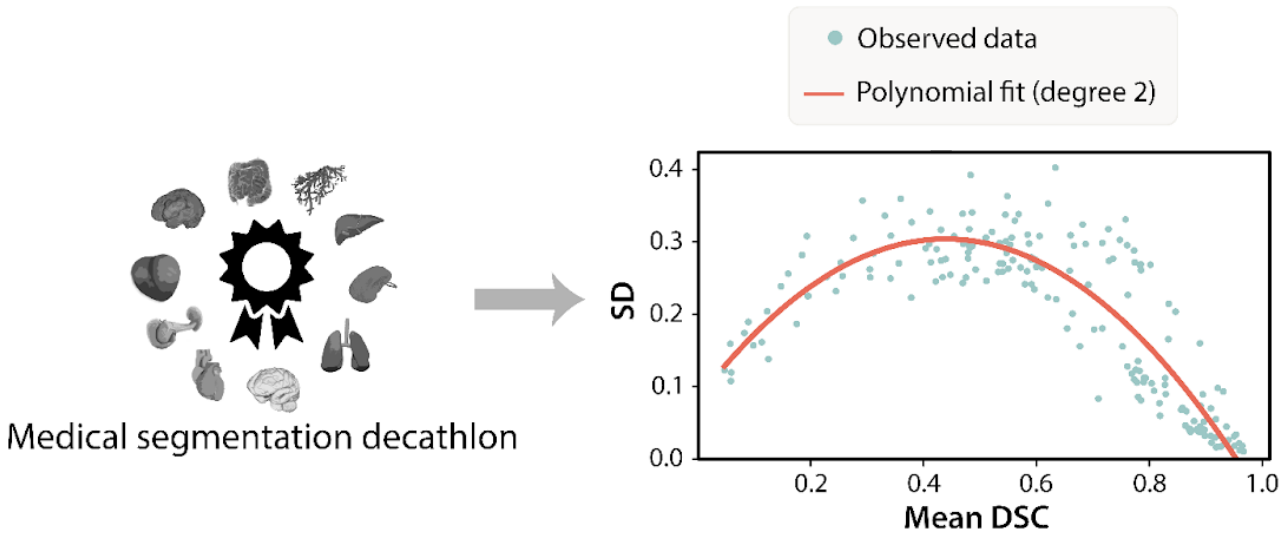
Common practice with respect to variability reporting

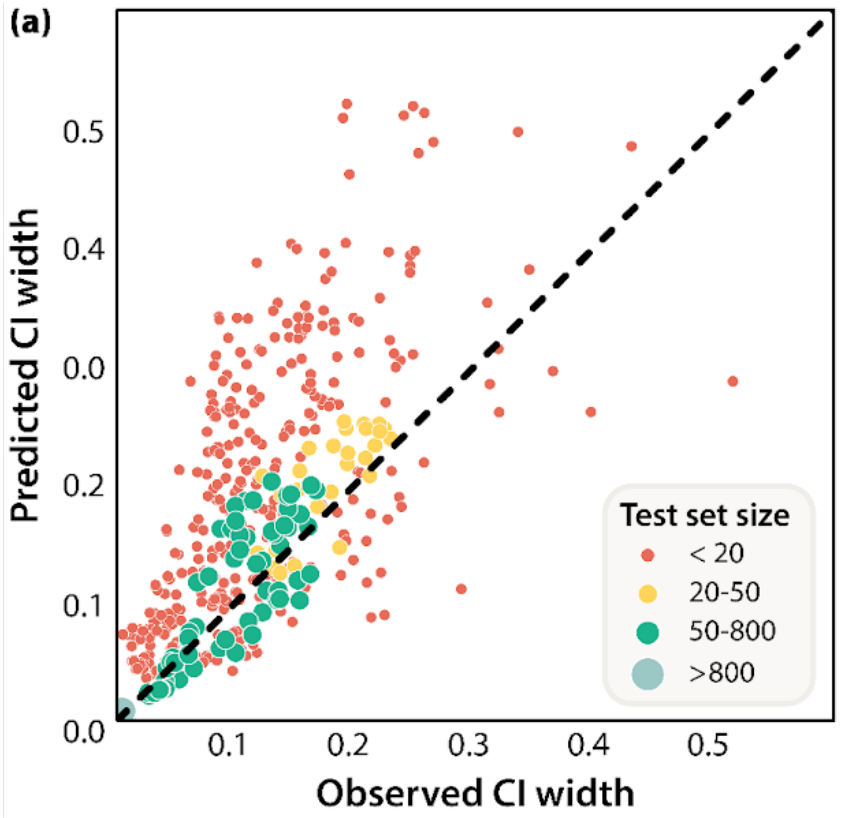


Christodoulou et al., MICCAI, 2024

Rigorous Statistical Analysis

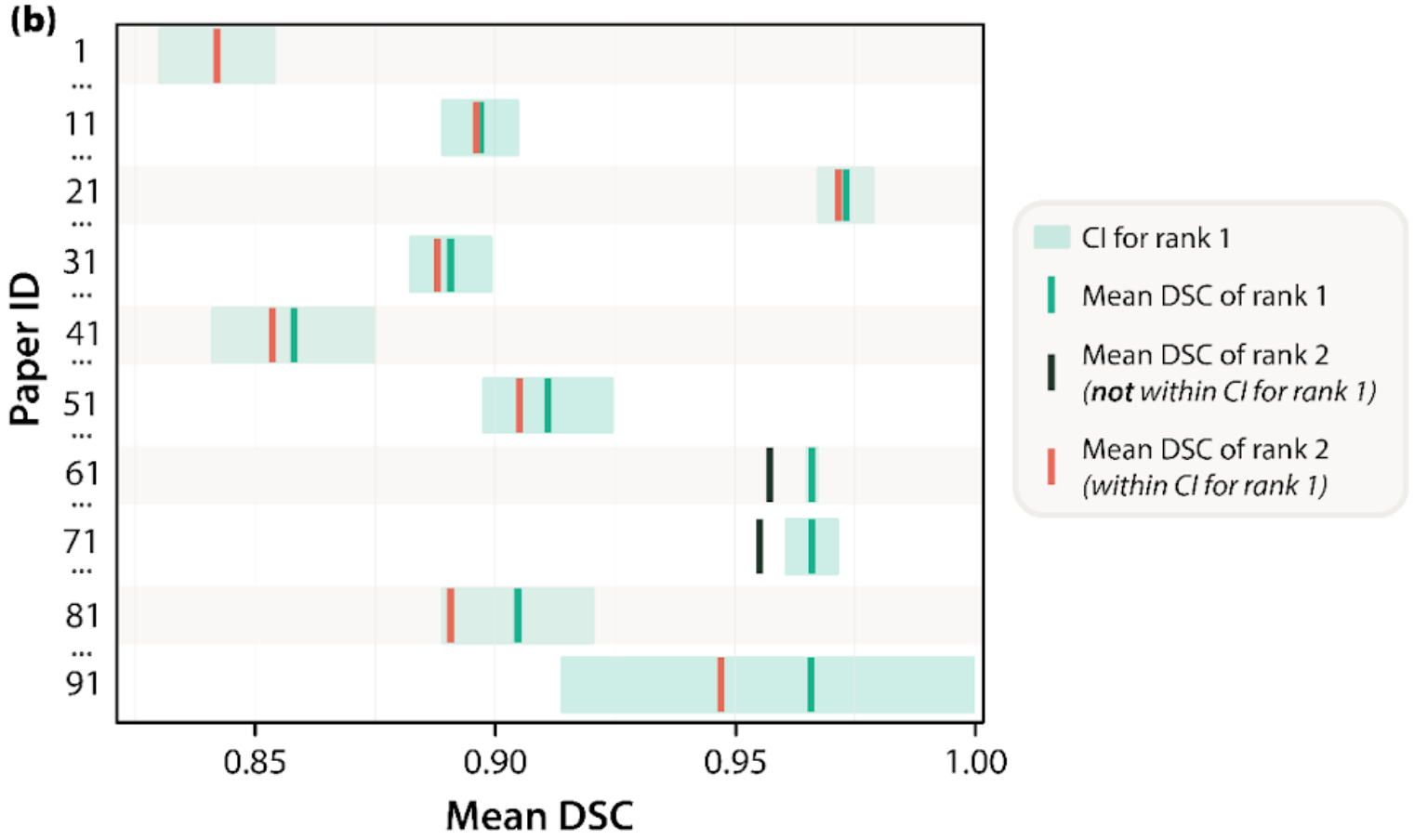
Approximation of the standard deviation





Christodoulou et al., MICCAI, 2024

Performance differences versus CI widths



Christodoulou et al., MICCAI, 2024

Rigorous Statistical Analysis

Conclusion

- Current publications typically do not provide sufficient evidence to support which models could potentially be translated into clinical practice

Recommendations

- Incorporate robust statistical analyses
- Report performance variability
- Use sufficiently large test sets to substantiate claims of outperformance

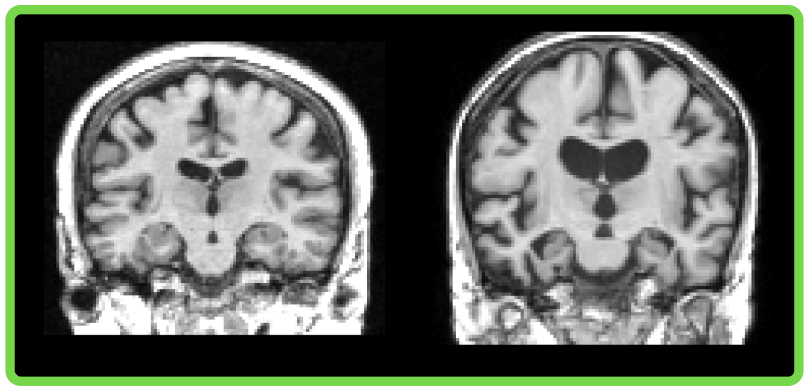
Christodoulou et al., MICCAI, 2024

Ensuring Robust AI in Medical Imaging:

Validation on Real-World Data,

Meaningful Metrics, and

Rigorous Statistical Analysis



Research vs clinical data