Failing forward: Rethinking the foundations of imaging AI

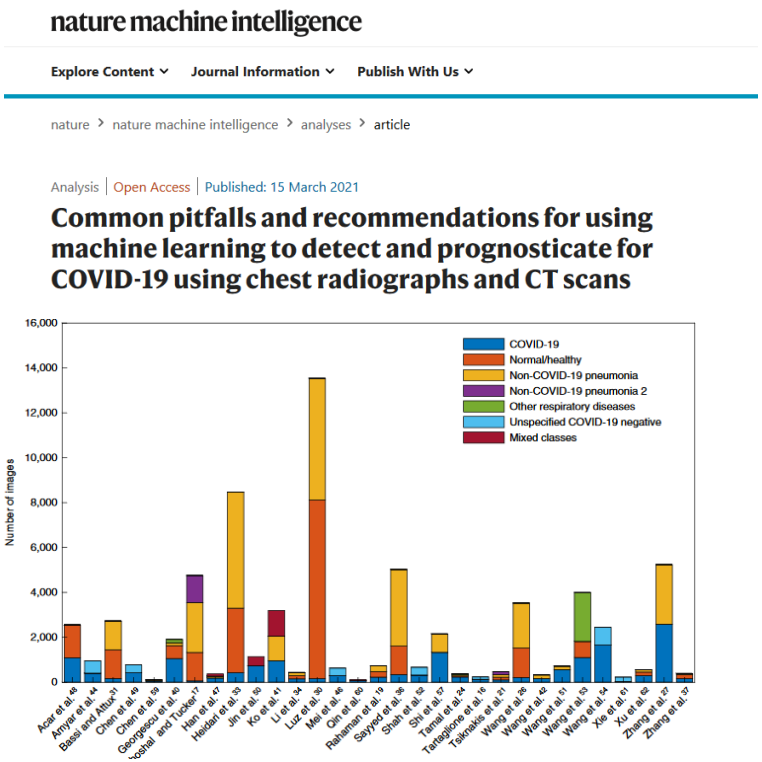
HELMHOLTZ IMAGING

ellis
European Laboratory for Learning and Intelligent Systems

erc European Research Council

Prof. Dr. Lena Maier-Hein
Head of Div. Intelligent Medical Systems (IMSY), DKFZ
Director National Center for Tumor Diseases (NCT) Heidelberg

AI in health: behind the scenes



nature machine intelligence

Explore Content | Journal Information | Publish With Us

nature > nature machine intelligence > analyses > article

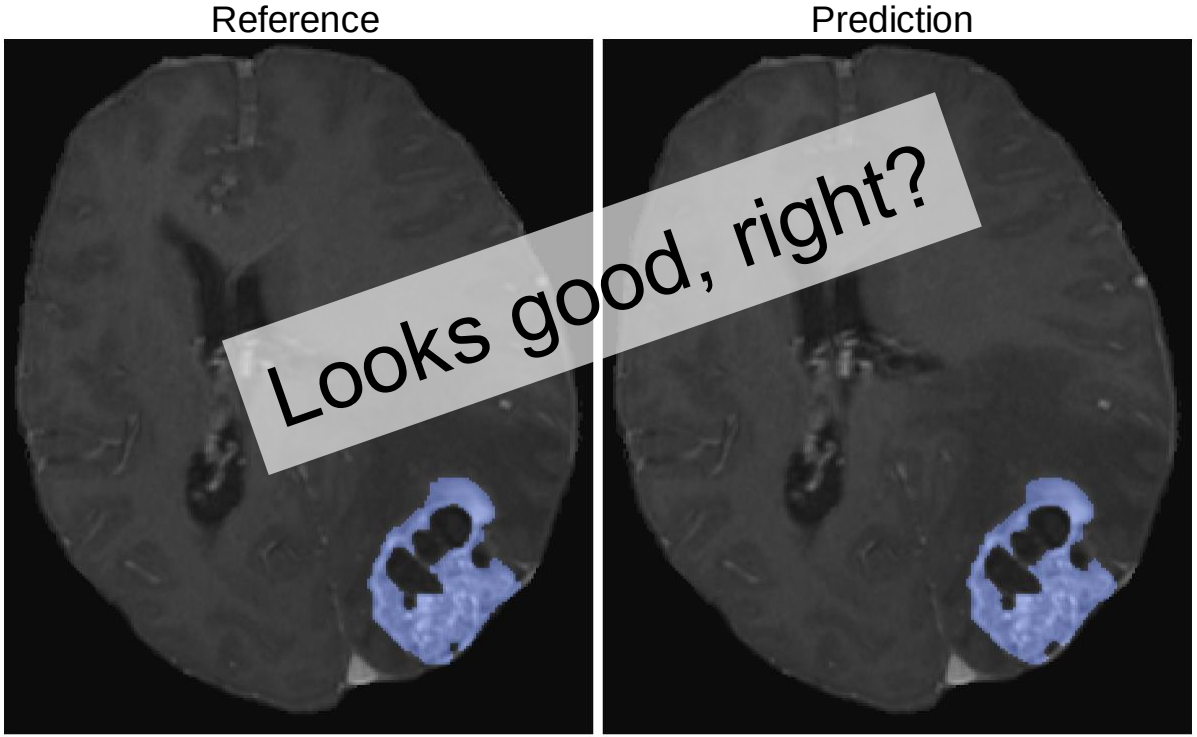
Analysis | Open Access | Published: 15 March 2021

Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans

Source: Pneumonia control group COVID-19 patients

www.siemens-healthineers.com/en-uk/news/mso-x-ray-imaging-for-covid-19.html

Metrics in (clinical) practice



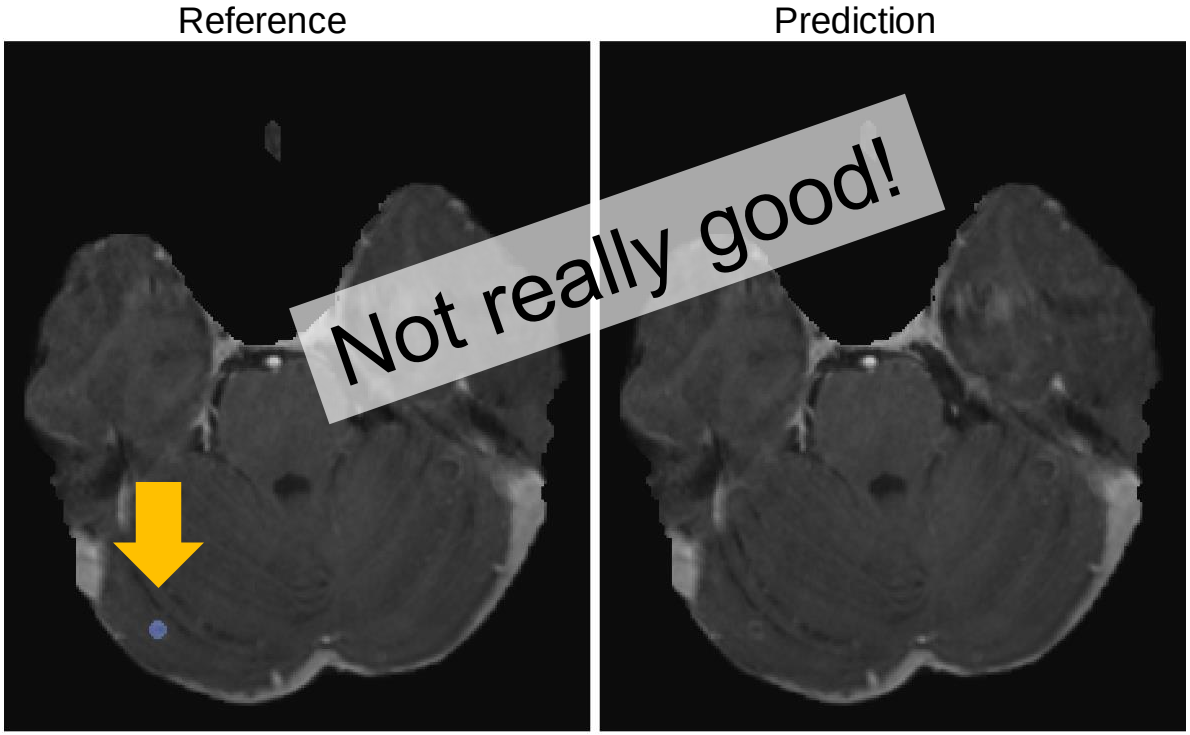
Courtesy Klaus Maier-Hein, DKFZ

Close to perfect DSC

Recall: 0.94

(voxel-level)

Metrics in (clinical) practice



Recall: 0.94

(voxel-level)

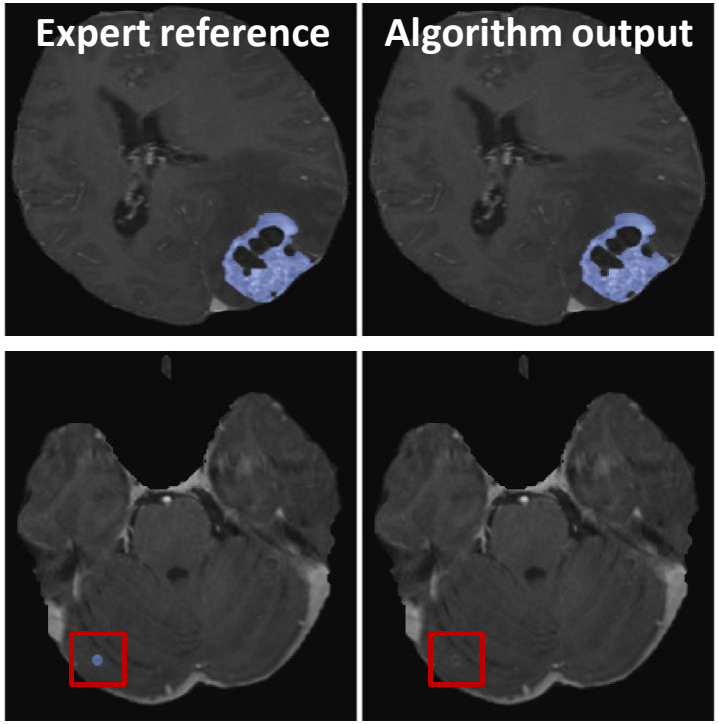
Recall: **0.5**

(object-level)

Courtesy Klaus Maier-Hein, DKFZ

Metrics in (clinical) practice

Algorithm with expert performance according to common validation metric DSC



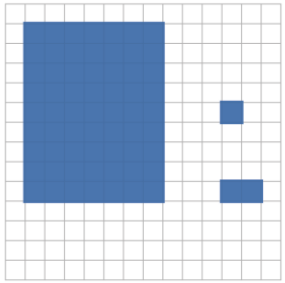
Most tumor pixels are detected…

… but the small (new) metastases are missed!

Instance progress not detected in ~1/3 of the cases!

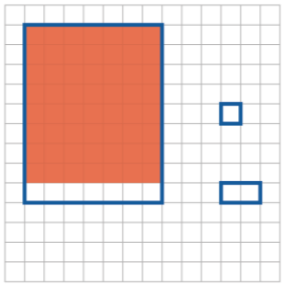
Reinke/Tizabi, …, Jäger/Maier-Hein. Understanding metric-related pitfalls in image analysis validation. **Nature Methods 2024**

Reference

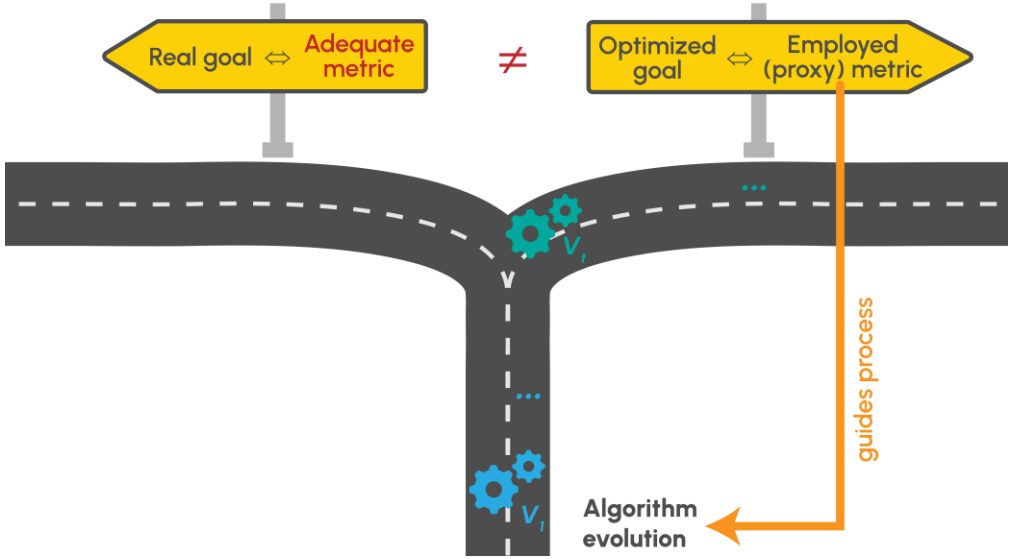


object-level metric

Algorithm 1

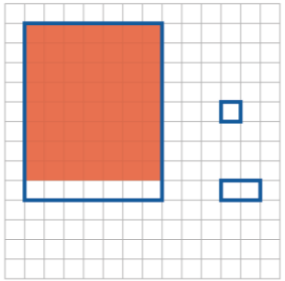


1/3 objects detected



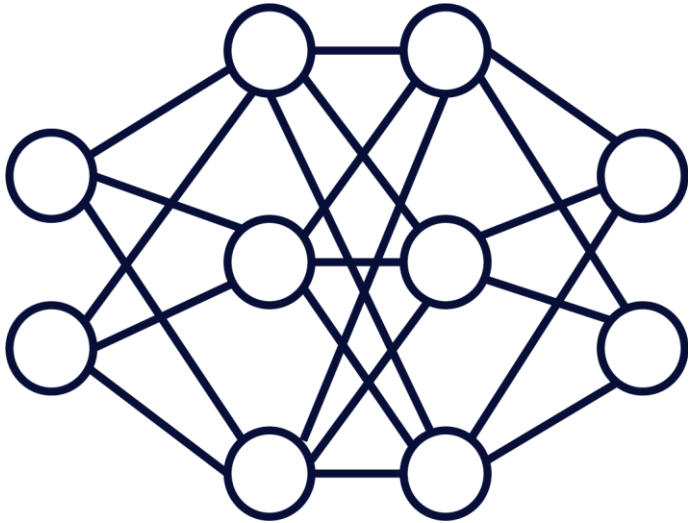
pixel-level metric

Algorithm 1



56/66 pixels detected

What to optimize?



How to optimize?

MICCAI Special Interest Group (SIG) for Challenges

SIG Board


Lena Maier-Hein (President)


Annika Reinke (Secretary)


Olivier Colliot (Treasurer)


Michal Kozubek


Bennett Landman


Nicholas Heller


Spyridon Bakas


Alexandros Karargyris


Annette Kopp-Schneider (Statistical Advisor)

SIG Members

IEEE Signal Processing Society
Gloria Menegaz

MICCAI challenge chairs
Nicholas Heller (2025)

ACR American College of Radiology
Kendall Schmidt

DREAM Challenges powered by Sage Bionetworks
Elise Blaese

International Symposium on Biomedical Imaging
Erik Meijering

Kitware
Stephen Aylward

NIH National Institutes of Health
Keyvan Farahani

SageBionetworks
Susheel Varma

kaggle
Maggie Demkin

RSNA Radiological Society of North America
Charles Kahn

Grand Challenge
Anne Mickan

MIDL
Bennett Landman

Impact of benchmarks / challenges

Algorithm	Rank
A1	**1**
A2	2
A3	3
A4	4
A5	5
A6	6
A7	7
A8	8

Up to €1 million price money

New state-of-the art method

Fame for researcher

…

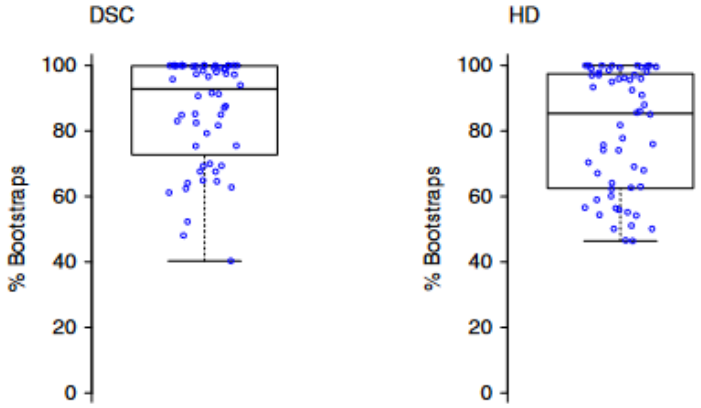

Technical University of Munich
UCL
TU/e Technische Universiteit Eindhoven University of Technology
UNIVERSITÄT ZU LÜBECK
INSTITUT FÜR MEDIZINISCHE INFORMATIK
MGH 1811
Hes·so VALAIS WALLIS
AIExplore
WARWICK
THE UNIVERSITY OF WARWICK
McGill
Centre for Intelligent Machines
UNIVERSITÄT BERN
UNIVERSITÉ DE RENNES 1
MIC
MEDICAL IMAGE COMPUTING
QUT
JOHNS HOPKINS UNIVERSITY
Heidelberg Collaboratory for Image Processing
NCT
NATIONAL CENTER FOR TUMOR DISEASES PARTNER SITE DRESDEN UNIVERSITY CANCER CENTER UCC
COMPLEXITY SCIENCE HUB VIENNA
VANDERBILT UNIVERSITY
Radboud Universiteit
Universität Rostock
Traditio et Innovatio
TU WIEN
TECHNISCHE UNIVERSITÄT WIEN
UNIVERSITÄT HEIDELBERG ZUKUNFT SEIT 1386
The University Of Sheffield.
Erasmus MC
University Medical Center Rotterdam
Instituts thématiques
Inserm
Institut national de la santé et de la recherche médicale
MEDICAL UNIVERSITY OF VIENNA
UNIVERSITY OF LEEDS

#Data (quantity)

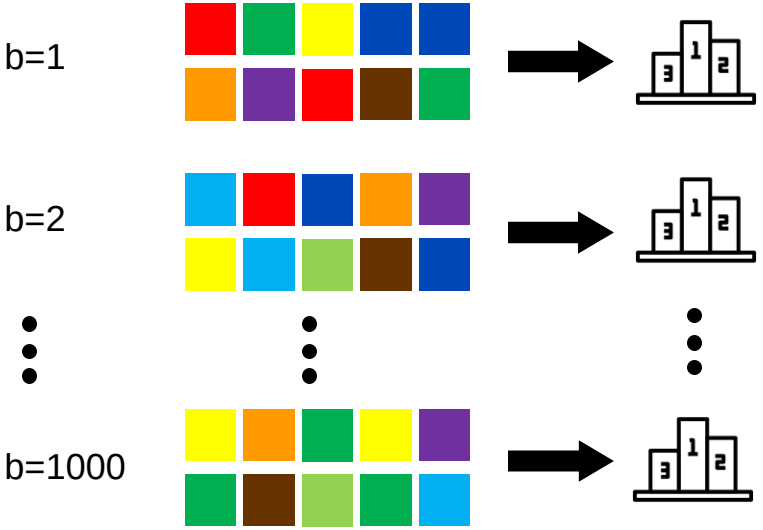
Pitfall: Data sparsity leads to instable rankings

n test cases

- **Instability of rankings** (analyzed over 56 biomedical image analysis challeges)
  -



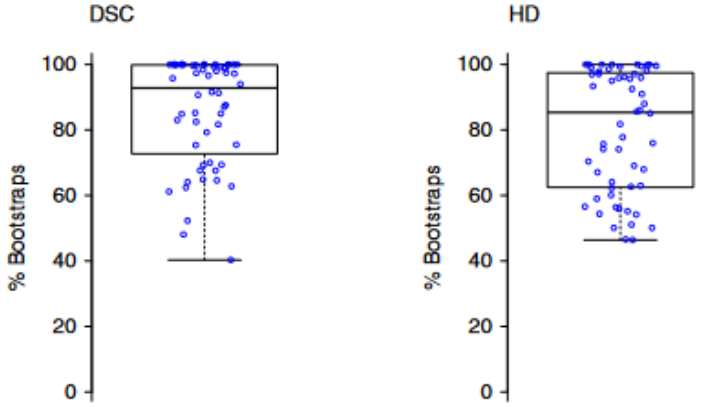
Percentage of cases in which the winner remains the winner when performing bootstrapping



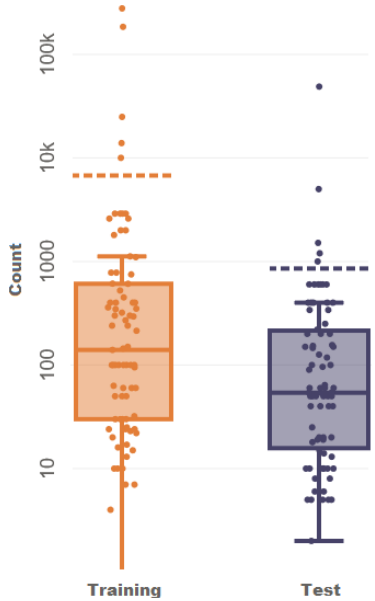
Maier-Hein et al. Why rankings of biomedical image analysiscompetitions should be interpreted with care **Nature Commun. 2018**

Pitfall: Data sparsity leads to instable rankings

- **Instability of rankings** (analyzed over 56 biomedical image analysis challeges)
  -



Percentage of cases in which the winner remains the winner when performing bootstrapping



Mean (n = 549)*:
231 training images
387 test images

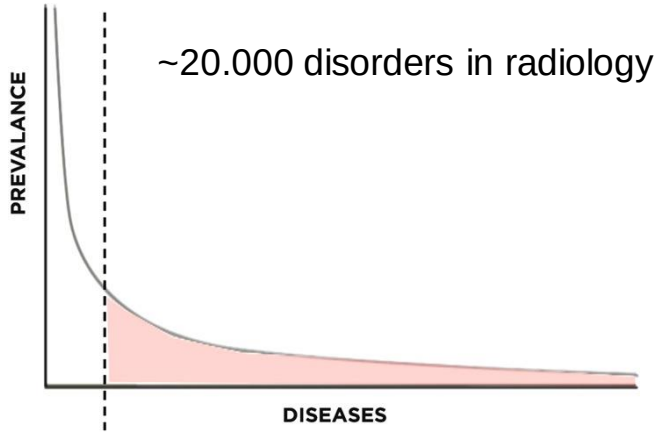
Maier-Hein et al. Why rankings of biomedical image analysiscompetitions should be interpreted with care **Nature Commun. 2018**

NEW Pitfall: Lack of *task* quantity and diversity

In the era of foundation models

- **Challenge:** Demonstrate broad capabilities through validation on diverse set of tasks
- **Example – Surgical FM Validation:**
  - No existing surgical FM validates beyond 6 downstream applications
  - Limited validation scope doesn't reflect real surgical complexity.

Extreme diversity of clinical disorders!



Shen, Tony, et al. "The long tail and rare disease research: the impact of next-generation sequencing for rare Mendelian disorders." *Genetics research* 97 (2015): e15.

C. P. Langlotz, "Will Artificial Intelligence Replace Radiologists?" Radiology: Ar- tificial Intelligence, vol. 1, no. 3, p. e190058, May 2019.

#DataSplitting

Data leakage

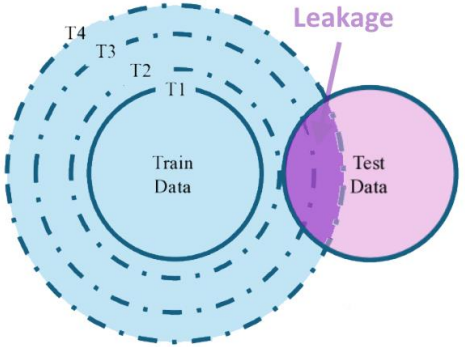
Field	Paper	Number of papers reviewed	Number of papers with pitfalls	[L1.1] No test set	[L1.2] Pre-proc. on train-test	[L1.3] Feature sel. on train-test	[L1.4] Duplicates	[L2] Illegitimate features	[L3.1] Temporal leakage	[L3.2] Non-ind. b/w train-test	[L3.3] Sampling bias	Comput. reproducibility issues	Data quality issues	Metric choice issues	Standard dataset used?
Medicine	Bouwmeester et al. (2012)	71	27	○								○			
Neuroimaging	Whelan & Garavan (2014)	–	14	○		○									
Bioinformatics	Blagus & Lusa (2015)	–	6		○										
Autism Diagnostics	Bone et al. (2015)	–	3				○				○		○	○	○
Nutrition Research	Ivanescu et al. (2016)	–	4	○									○	○	
Software Eng.	Tu et al. (2018)	58	11					○			○		○		○
Toxicology	Alves et al. (2019)	–	1				○				○		○		
Clinical Epidem.	Christodoulou et al. (2019)	71	48			○							○		
Satellite Imaging	Nalepa et al. (2019)	17	17						○				○		○
Tractography	Poulin et al. (2019)	4	2	○							○		○	○	○
Brain-computer Int.	Nakanishi et al. (2020)	–	1	○											○
Histopathology	Oner et al. (2020)	–	1						○						
Neuropsychiatry	Poldrack et al. (2020)	100	53	○	○								○	○	
Neuroimaging	Ahmed et al. (2021)	–	1						○						
Neuroimaging	Li et al. (2021)	122	18						○						
IT Operations	Lyu et al. (2021)	9	3					○							○
Medicine	Filho et al. (2021)	–	1				○								
Radiology	Roberts et al. (2021)	62	16	○			○				○	○			○
Neuropsychiatry	Shim et al. (2021)	–	1			○						○			
Medicine	Vandewiele et al. (2021)	24	21			○			○	○	○	○			○
Computer Security	Arp et al. (2022)	30	22	○	○	○		○	○	○	○		○	○	
Genomics	Barnett et al. (2022)	41	23			○							○		

Kapoor & Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. **Patterns 2023**

NEW pitfall: Leakage and accidental knowledge injection

- **Pitfall**: Scraping the internet for FM training/validation introduces unintended leakage
- **NEW challenge:** Live-benchmarks are needed to avoid benchmark overfitting and test set contamination [1]

Test set contamination over time



FM trained on older train sets crawled from the internet (T1, T2) experience test set contamination over time (T3, T4), as static test sets and public benchmarks are unintentionally included in the training, leading to unreliable performance reporting.
Figure adapted from [2].

[1] White et al. Livebench: A challenging, contamination-free llm benchmark. **ICLR 2025 spotlight**
[2] Shabtay et al. LiveXiv--A Multi-Modal Live Benchmark Based on Arxiv Papers Content. **arXiv preprint 2024**
[3] Oren et al. Proving Test Set Contamination in Black-Box Language Models, **ICLR 2025 (oral)**

#DataShifts

Medical Imaging AI in the wild

Refer to ophthalmologist

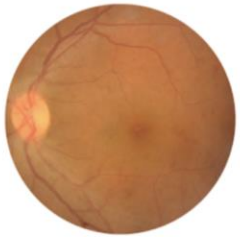
Right (OD)

DIABETIC RETINOPATHY (DR)

Severe NPDR

DIABETIC MACULAR EDEMA (DME)

DME detected



Left (OS)

DIABETIC RETINOPATHY (DR)

Mild NPDR

DIABETIC MACULAR EDEMA (DME)

No DME detected

A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy

Emma Beede
Google Health
Palo Alto, CA
embeede@google.com

Elizabeth Baylor
Google Health
Palo Alto, CA
ebaylor@google.com

Fred Hersch
Google Health
Singapore
fredhersch@google.com

Anna Iurchenko
Google Health

Lauren Wilcox
Google Health

Paisan Ruamviboonsuk
Rajavithi Hospital

Google deep learning algorithm: specialist-level accuracy (>90% sensitivity and specificity) for the detection of referable cases of diabetic retinopathy

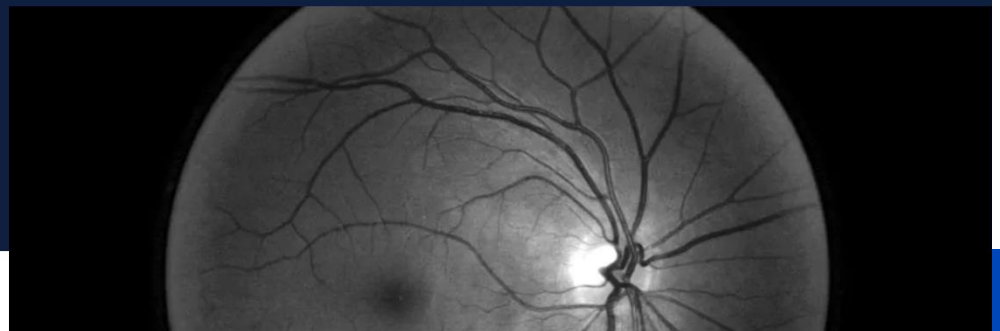
ARTIFICIAL INTELLIGENCE

Google's medical AI was super accurate in a lab. Real life was a different story.

If AI is really going to make a difference to patients we need to know how it works when real humans get their hands on it, in real situations.

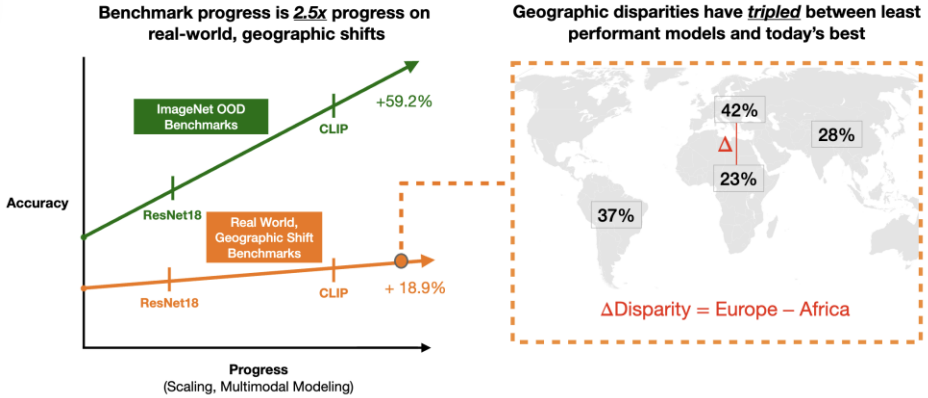
By Will Douglas Heaven

April 27, 2020



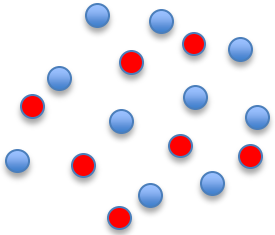
Imaging AI in the wild

- **Key Issue**: Models excel in benchmarks but fail in deployment due to distribution shifts.
- **Example**
  - Despite improvements in benchmarks, real-world disparities are increasing.
  - Performance gap has tripled between Europe & Africa.



Richards/Kirichenko et al. Does Progress On Object Recognition Benchmarks Improve Generalization on Crowdsourced, Global Data? **ICLR 2024**

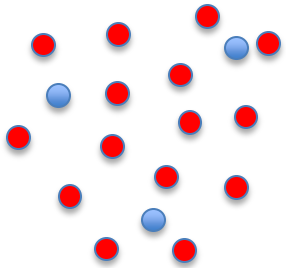
NEW: Shifts now also occur at *task* level; new shifts



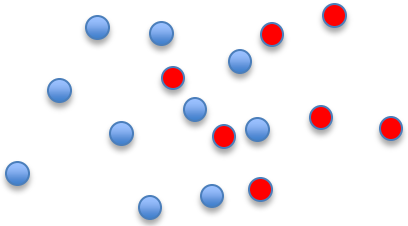
Ideal

Training tasks

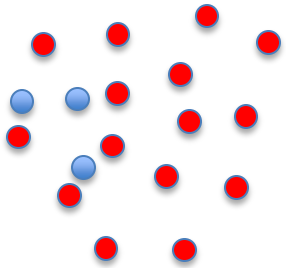
Test tasks



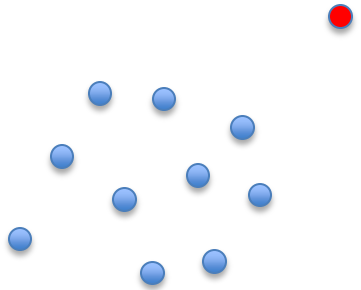
Sparsity



Domain gap



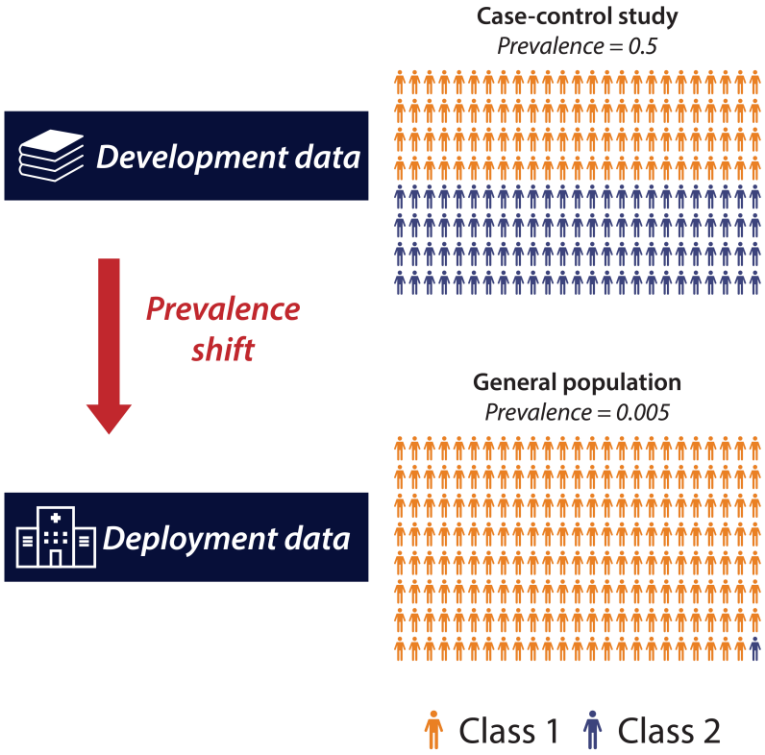
Selection bias

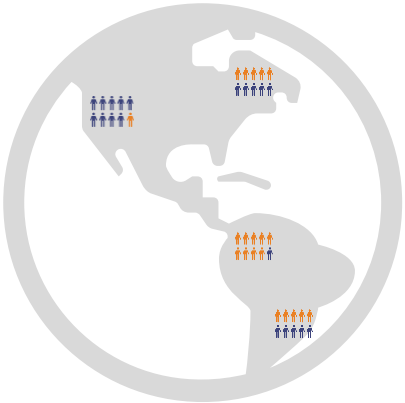


NEW: New types of shifts, e.g. unintended use, prompt shifts…

#Metrics

Example II: Prevalence shifts - a real-world challenge



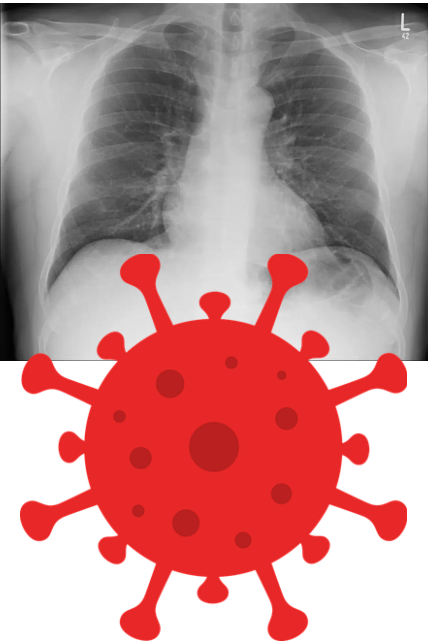


Patrick Godau

Piotr Kalinowski

Reinke/Tizabi…Jäger/Maier-Hein: Understanding metric-related pitfalls in image analysis validation. **Nature Methods 2023** (cond. accept)

Example II: COVID-19 classification



Development data:
Train with balanced sampling (50/50)
Test on balanced dataset (50/50)

→

Deployment data:
Prevalence = 0.01

Inherent properties of the method: *Sensitivity* = 0.90, *Specificity* = 0.70

		PREDICTED P	PREDICTED N
ACTUAL	P	3600	400
	N	1200	2800

Accuracy = 0.80
F1 Score = 0.82

?

Godau/Kalinowski, ..., Maier-Hein. Deployment of Image Analysis Algorithms under Prevalence Shifts, **MICCAI 2023**
Image source: www.siemens-healthineers.com/en-uk/news/mso-x-ray-imaging-for-covid-19.html

A so far overlooked metric: Expected Cost (EC)

$$\text{Accuracy} = 1 - \text{Error Rate} = 1 - \frac{\text{FN} + \text{FP}}{\text{N}}$$

$$\text{Error Rate} = \frac{\text{FN} + \text{FP}}{\text{N}} = \frac{\text{FN}}{\text{TP} + \text{FN}} \cdot \frac{\text{TP} + \text{FN}}{\text{N}} + \frac{\text{FP}}{\text{FP} + \text{TN}} \cdot \frac{\text{FP} + \text{TN}}{\text{N}}$$

R_{12} Prevalence

$$= R_{12} \cdot \text{Prevalence} + R_{21} \cdot (1 - \text{Prev}$$

Replace test-set prevalence by estimated real-world prevalence!

$$\text{EC} = \sum_k \text{Prevalence}(k) \sum_j c_{kj} R_{kj}$$

binary

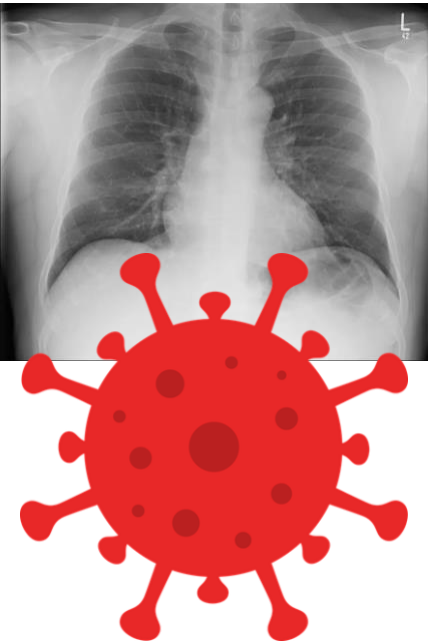
$c_{11} = c_{22} = 0$
$c_{12} = c_{21} = 1$

$$= \text{Prevalence} \cdot (c_{11}R_{11} + c_{12}R_{12}) + (1 - \text{Prevalence}) \cdot (c_{21}R_{21} + c_{22}R_{22})$$

$$= \text{Prevalence} \cdot R_{12} + (1 - \text{Prevalence}) \cdot R_{21}$$

$$= \text{Error Rate} = 1 - \text{Accuracy}$$

Example II: COVID-19 classification



Development data:
Train with balanced sampling (50/50)
Test on balanced dataset (50/50)

Deployment data:
Prevalence = 0.01

Inherent properties of the method: *Sensitivity* = 0.90, *Specificity* = 0.70

		PREDICTED P	PREDICTED N
ACTUAL	P	3600	400
ACTUAL	N	1200	2800

Accuracy = 0.80
F1 Score = 0.82

EC* = 0.30

* Prevalence-corrected

Δ Accuracy = 0.10
Δ F1 Score = 0.76

		PREDICTED P	PREDICTED N
ACTUAL	P	72	8
ACTUAL	N	2375	5545

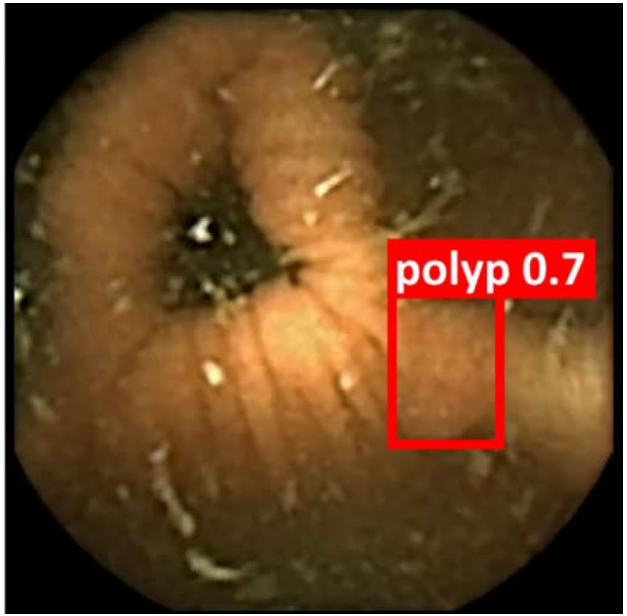
Accuracy = 0.70
F1 Score = 0.06

EC* = 0.30

* Prevalence-corrected

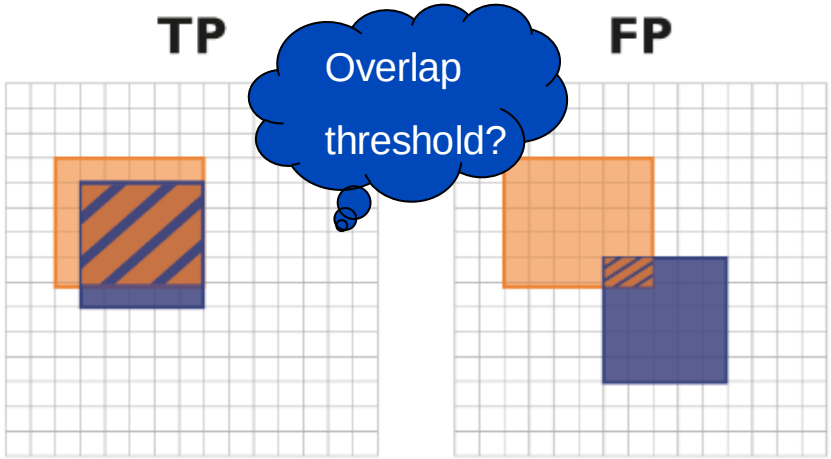
Godau/Kalinowski, ..., Maier-Hein. Deployment of Image Analysis Algorithms under Prevalence Shifts, **MICCAI 2023** (best paper nom.)

Example III: Metric configuration matters as well



Amine Yamlahi

Nuong Tran



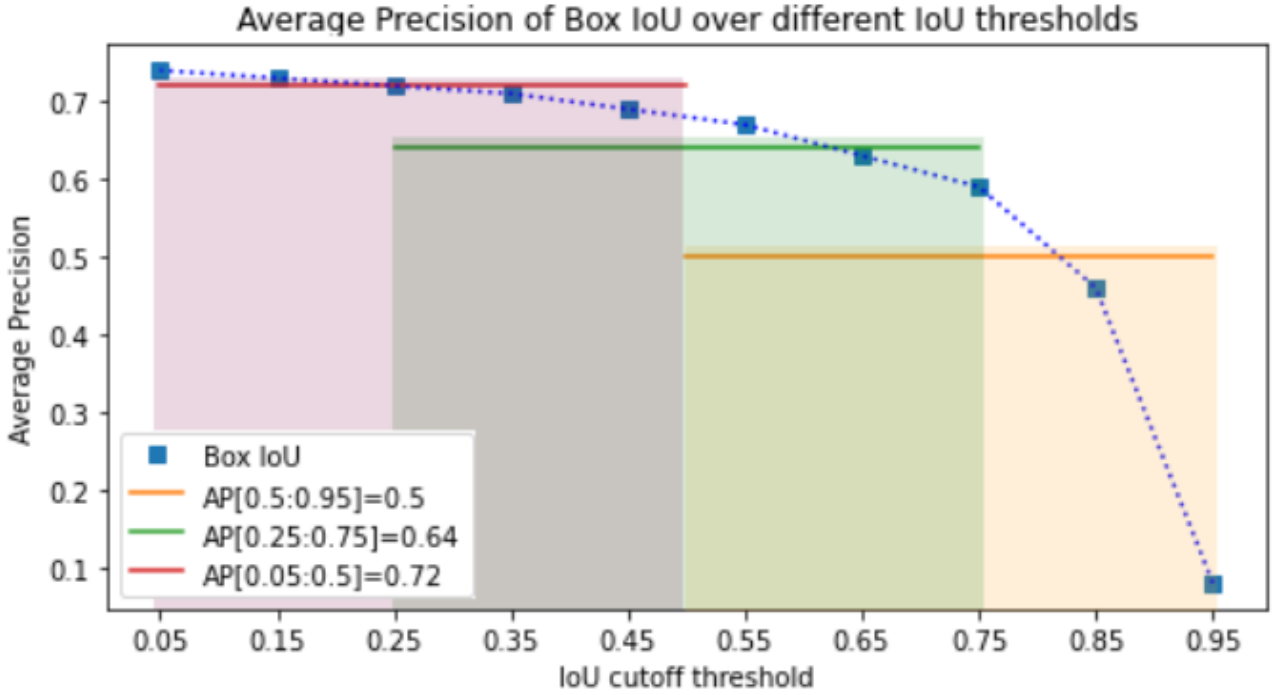
How to define a TP/FP?

Reinke/Tizabi, …, Jäger/Maier-Hein. Understanding metric-related pitfalls in image analysis validation. **Nature Methods 2024**
Tran, …, Maier-Hein. Sources of performance variability in deep learning-based polyp detection. **IPCAI 2023**

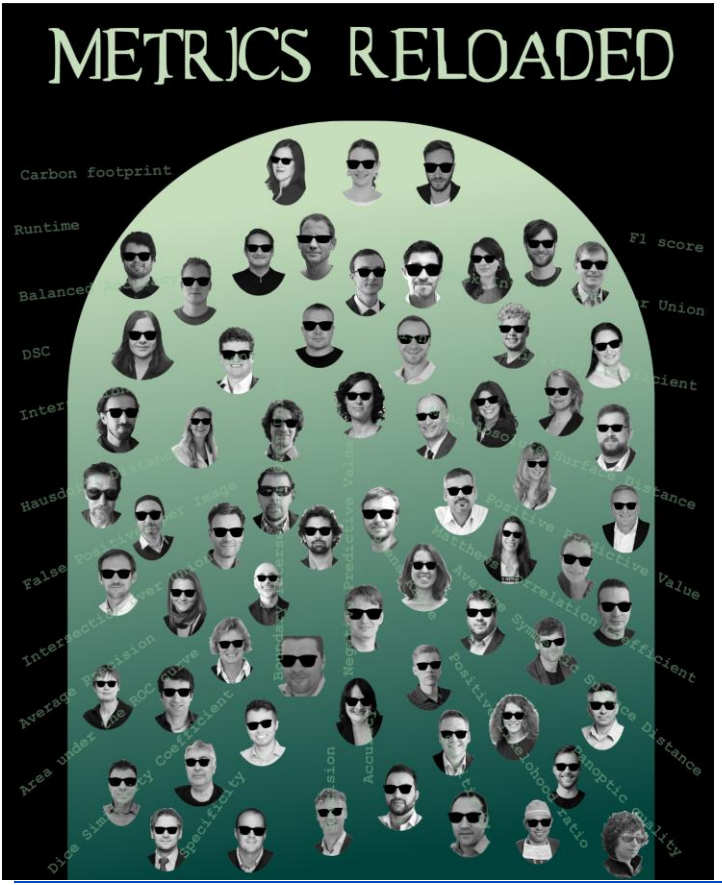
Standard solution does not reflect the clinical need!

Clinicians‘ preference

ML community



Tran, …, Maier-Hein. Sources of performance variability in deep learning-based polyp detection. **IPCAI 2023**



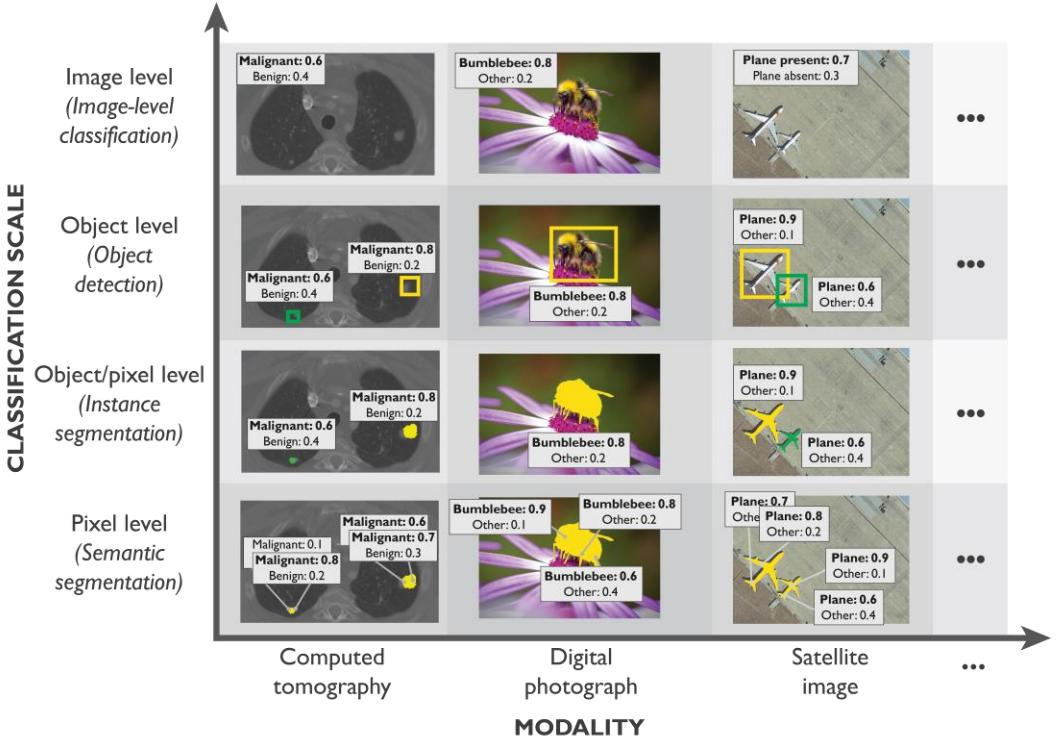
nature methods

Rita Strack, *Nature Methods Editor*

"It's really hard for me to state in words how important I think these two [papers] are for the future of bioimage analysis."

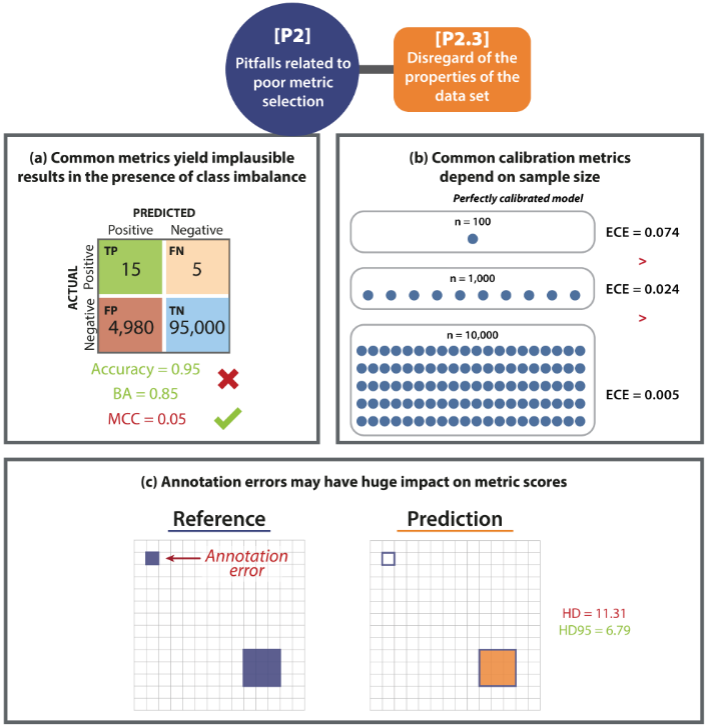
Maier-Hein/Reinke et al. Metrics reloaded: Recommendations for image analysis validation. **Nature Methods 2024**
Reinke/Tizabi, …, Maier-Hein. Understanding metric-related pitfalls in image analysis validation. **Nature Methods 2024**

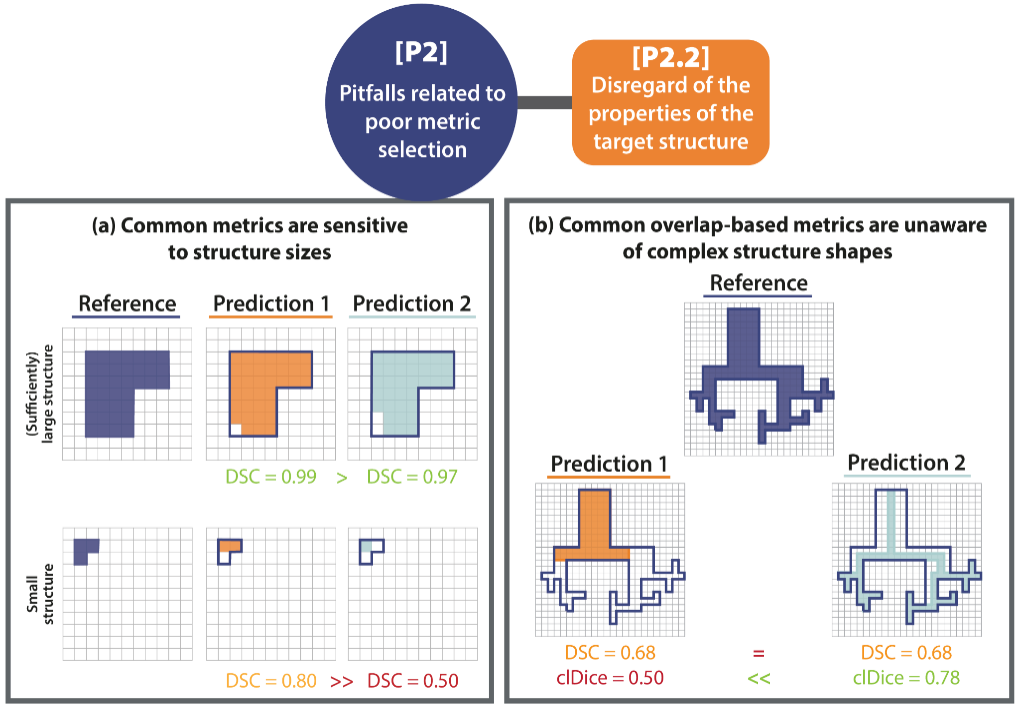
Scope of (first set of) recommendations



Maier-Hein/Reinke et al. Metrics reloaded: Recommendations for image analysis validation. **Nature Methods 2024**

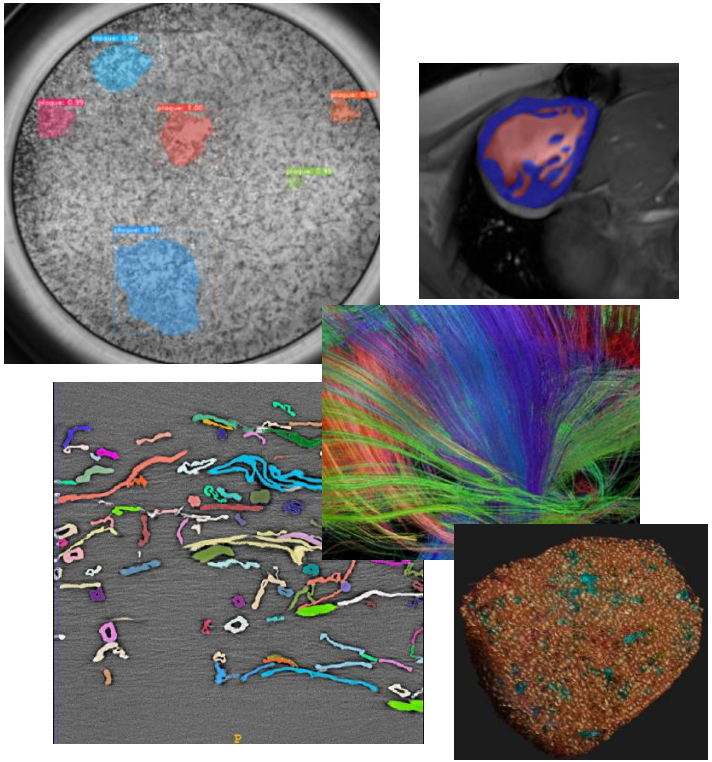
Step 1: Collection of pitfalls





Reinke/Tizabi, …, Maier-Hein. Understanding metric-related pitfalls in image analysis validation. **Nature Methods 2024**

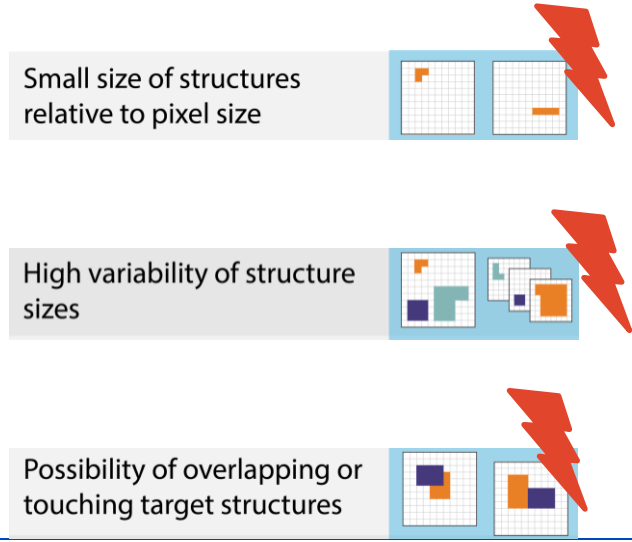
Step 2: Abstracting pitfalls from modality/domain



Annika Reinke

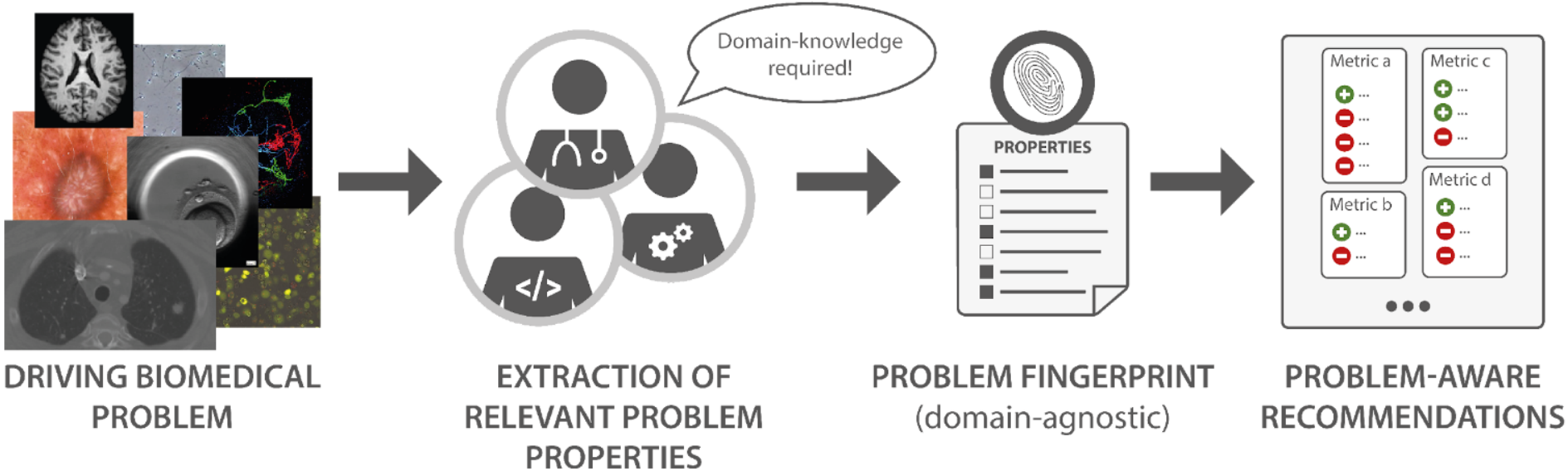
Minu Tizabi

Abstraction via fingerprinting

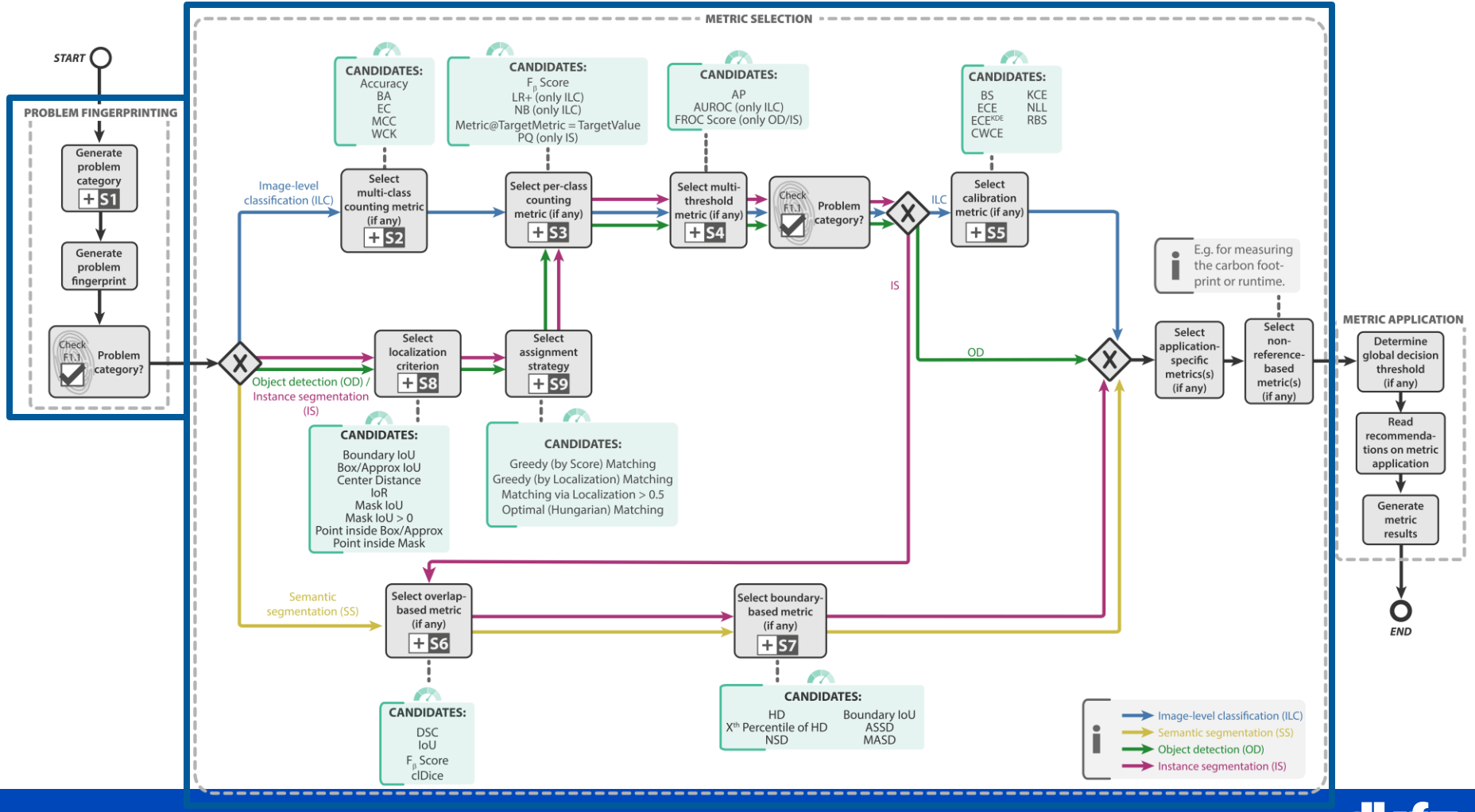


Maier-Hein/Reinke et al. Metrics reloaded: Recommendations for image analysis validation. **Nature Methods 2024**

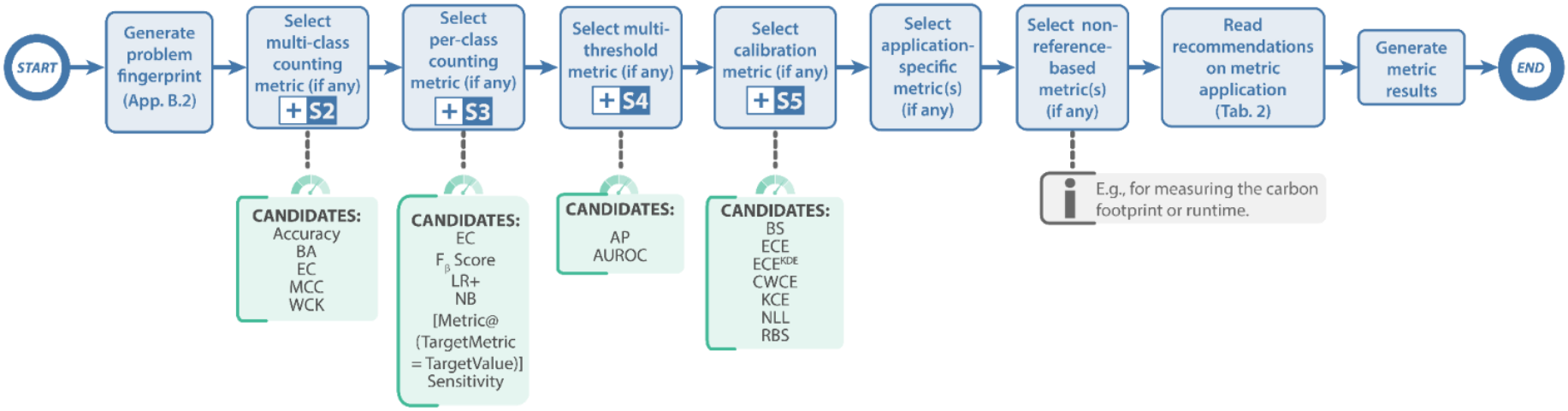
Metrics Reloaded approach



Maier-Hein/Reinke et al. Metrics reloaded: Recommendations for image analysis validation. **Nature Methods 2024**


START
PROBLEM FINGERPRINTING
Generate problem category
+ S1
Generate problem fingerprint
Check F1.1
Problem category?
METRIC SELECTION
CANDIDATES:
Accuracy
BA
EC
MCC
WCK
CANDIDATES:
F_β Score
LR+ (only ILC)
NB (only ILC)
Metric@TargetMetric = TargetValue
PQ (only IS)
CANDIDATES:
AP
AUROC (only ILC)
FROC Score (only OD/IS)
CANDIDATES:
BS
ECE
ECE^{KDE}
CWCE
KCE
NLL
RBS
Image-level classification (ILC)
Select multi-class counting metric (if any)
+ S2
Select per-class counting metric (if any)
+ S3
Select multi-threshold metric (if any)
+ S4
Check F1.1
Problem category?
ILC
IS
Select calibration metric (if any)
+ S5
E.g. for measuring the carbon foot-print or runtime.
OD
Object detection (OD) / Instance segmentation (IS)
Select localization criterion
+ S8
Select assignment strategy
+ S9
CANDIDATES:
Boundary IoU
Box/Approx IoU
Center Distance
IoR
Mask IoU
Mask IoU > 0
Point inside Box/Approx
Point inside Mask
CANDIDATES:
Greedy (by Score) Matching
Greedy (by Localization) Matching
Matching via Localization > 0.5
Optimal (Hungarian) Matching
Select application-specific metrics(s) (if any)
Select non-reference-based metric(s) (if any)
METRIC APPLICATION
Determine global decision threshold (if any)
Read recommenda-tions on metric application
Generate metric results
END
Semantic segmentation (SS)
Select overlap-based metric (if any)
+ S6
Select boundary-based metric (if any)
+ S7
CANDIDATES:
DSC
IoU
F_β Score
clDice
CANDIDATES:
HD
X^{th} Percentile of HD
NSD
Boundary IoU
ASSD
MASD
Image-level classification (ILC)
Semantic segmentation (SS)
Object detection (OD)
Instance segmentation (IS)

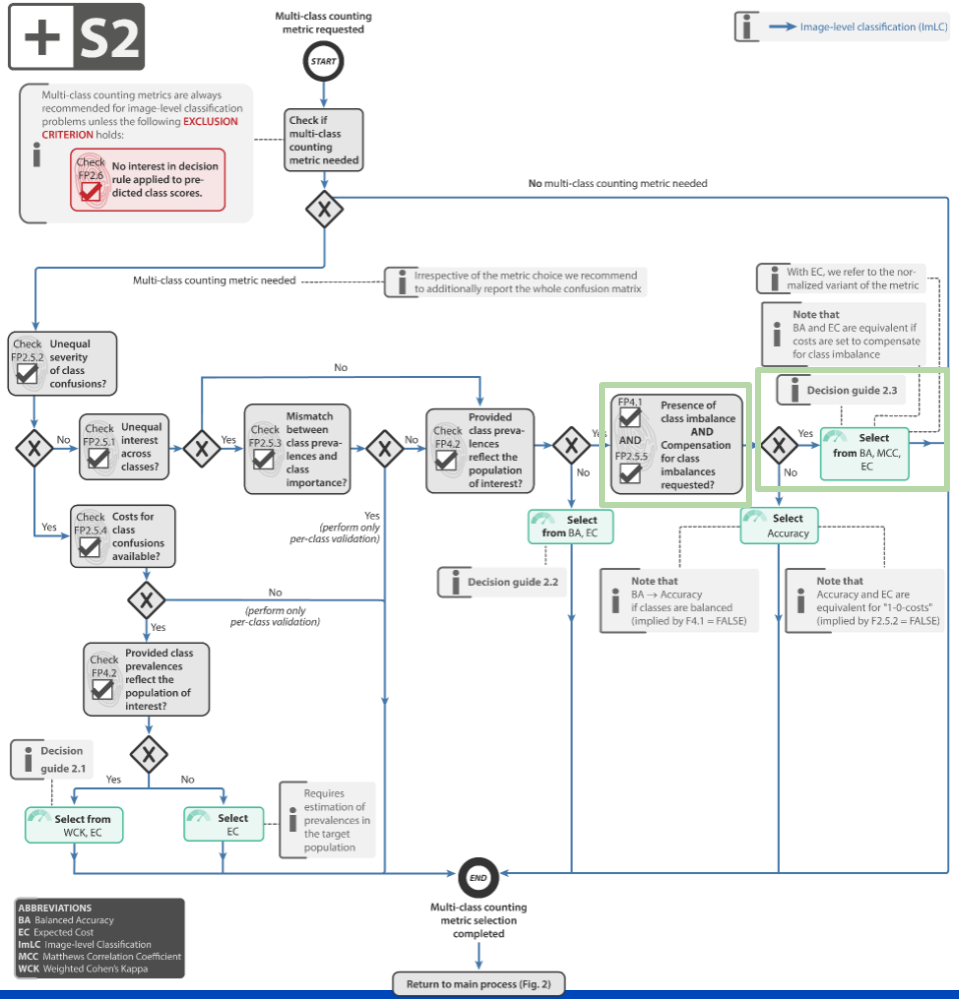
Metric selection for image-level classification



ABBREVIATIONS

AP Average Precision
AUROC Area Under the Receiver Operating Characteristic Curve
BA Balanced Accuracy
BS Brier Score
CWCE Class-wise Calibration Error
EC Expected Cost
ECE Expected Calibration Error
ECE^{KDE} Expected Calibration Error Kernel Density Estimate
KCE Kernel Calibration Error
LR+ Positive Likelihood Ratio
MCC Matthews Correlation Coefficient
NB Net Benefit
NLL Negative log likelihood
RBS Root Brier Score
WCK Weighted Cohen's Kappa

Multi-class counting metrics



Decision guides

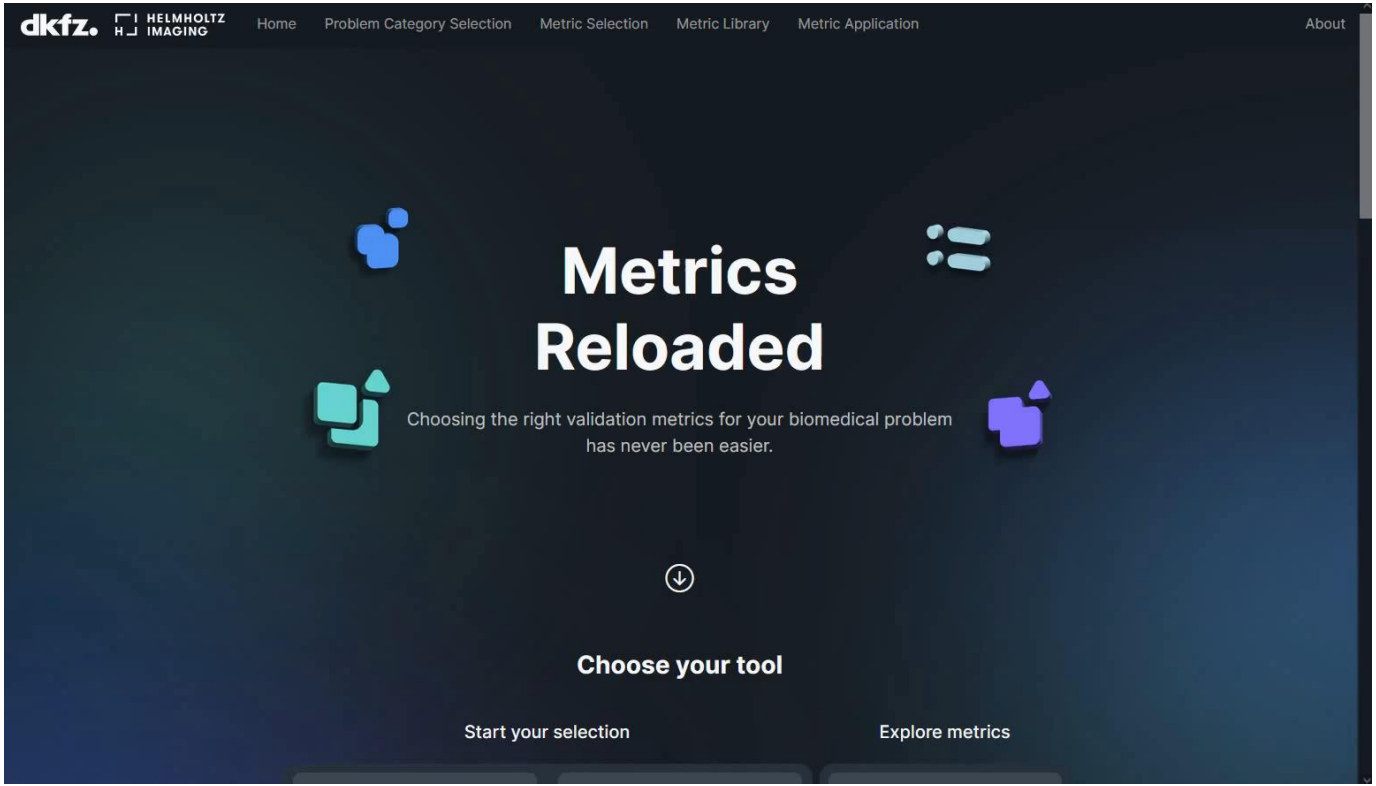
Summary of DG2.3: BA versus MCC versus ECN

BA	MCC	ECN
⊕ Inherent interpretability with respect to naive classifier	⊕ Inherent interpretability with respect to naive classifier	⊕ Inherent interpretability with respect to naive classifier
⊖ Implication of equal class contribution	⊖ Implication of equal class contribution	⊖ No establishment of equal class contribution
⊖ Insensitive to predictive values (Positive Predictive Value (PPV) and Negative Predictive Value (NPV))	⊖ High scores ensure high predictive values (PPV and NPV)	⊖ Limited sensitivity to predictive values (PPV and NPV)
⊕ Availability of framework to identify and validate decision rule applied to class scores	⊖ Lack of framework to identify and validate the decision rule applied to class scores	⊕ Availability of framework to identify and validate the decision rule applied to class scores
⊕ Good interpretability	⊖ Limited interpretability	⊕ Good interpretability
⊕ Widely used	⊖ Fairly well-known but not much used	⊖ Not known or used in the biomedical imaging domain although based on well-studied statistical concepts

Table 6. Comparison of Balanced Accuracy (BA) to Matthews Correlation Coefficient (MCC) to normalized EC (ECN) in the context of the decision guide DG2.3 for Subprocess S2. Context: Equal severity of class confusions (FP2.5.2 = FALSE), either (1) unequal interest across classes (FP2.5.1 = TRUE) and no mismatch between class prevalences and class importance (FP2.5.3 = FALSE) or (2) equal interest across classes (FP2.5.1 = FALSE), provided class prevalences reflect the population of interest (FP4.2 = TRUE), presence of class imbalance (FP4.1 = TRUE) and compensation for class imbalances requested (FP2.5.5 = TRUE).

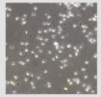
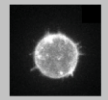
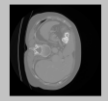
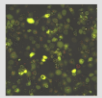
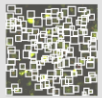
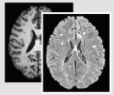
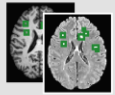
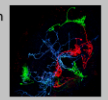
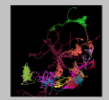
Making it accessible to the community: Metrics Reloaded toolkit

Emre Kavur

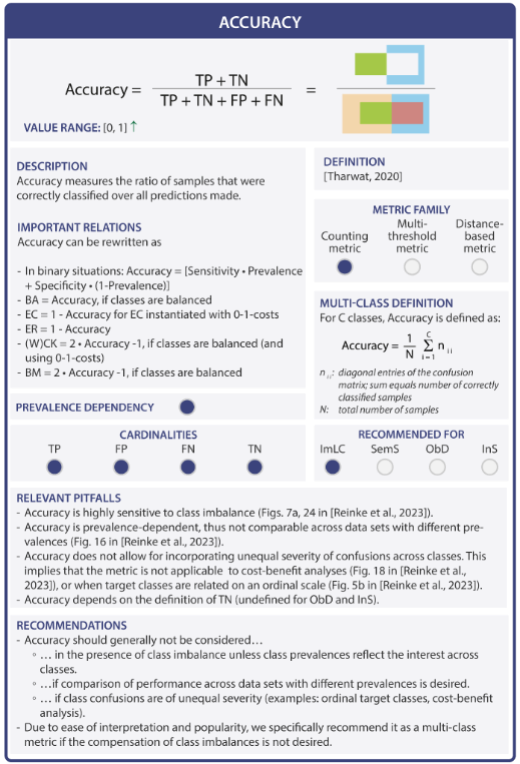


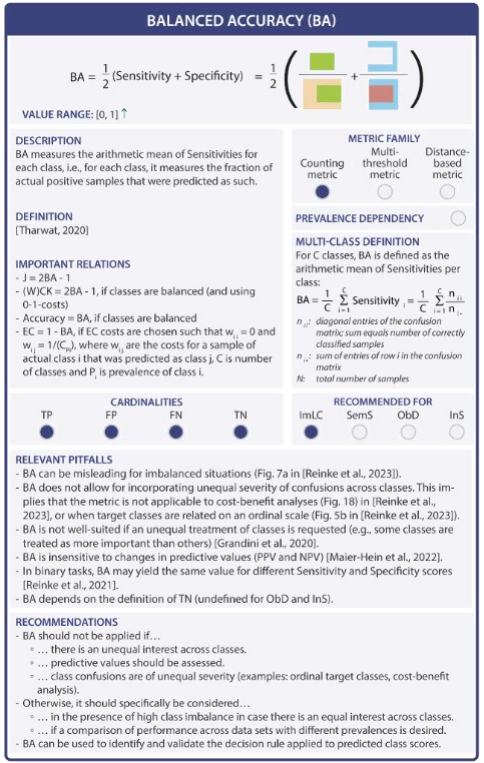
https:// metrics-reloaded.helmholtz-imaging.de

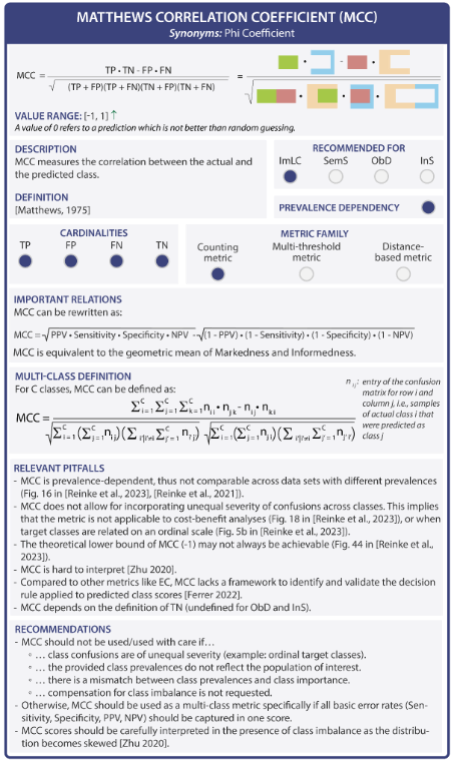
Instantiation for common biomedical use cases

DESCRIPTION OF PROBLEM	SCENARIO	SAMPLE INPUT IMAGE	RECOMMENDED OUPUT IMAGE	RECOMMENDATION
Classification of images	Frame-based sperm motility classification based on microscopy time-lapse video containing human spermatozoa			**Problem category:** Image-level classification **Multi-class counting metric (S2):** Balanced Accuracy (BA) **Multi-threshold metric (S3):** Area under the Receiver Operating Characteristc Curve (AUROC) **Output calibration:** Expected Calibration Error (ECE) **Per-class counting metric (S4):** Positive Likelihood Ratio (LR+)
	Disease classification in dermoscopic images			
Segmentation of large objects	Lung cancer cell segmentation from microscopy images			**Problem category:** Semantic segmentation **Overlap-based metric (S7):** Dice Similarity Coefficient (DSC) **Boundary-based metric (S8):** Normalized Surface Distance (NSD) **Specific property-related metric:** Liver segmentation: Absolute Volume Difference
	Liver segmentation in computed tomography (CT) images			
Detection of multiple and arbitrary located objects	Cell detection and tracking during the autophagy process in time-lapse microscopy			**Problem category:** Object detection **Localization criterion (S5):** Box Intersection over Union (Box IoU) **Assignment strategy (S6):** Greedy (by Score) Matching, set double assignments to False Positives (FP) **Multi-threshold metric (S3):** Free-Response Receiver Operating Characteristic (FROC) Score **Output calibration:** MS lesion detection: Proper Scoring Rules (PSR) **Per-class counting metric (S4):** FP per Image (FPPI)@Sensitivity
	MS Lesion detection in multi-modal brain MRI images			
Segmentation and distinction of tubular objects	Instance segmentation of neurons from the fruit fly in 3D multi-color light microscopy images			**Problem category:** Instance segmentation **Localization criterion (S5):** Neuron segmentation: Mask IoU Instrument segmentation: Boundary IoU **Assignment strategy (S6):** Greedy (by Score) Matching, set double assignments to FP **Multi-threshold metric (S3):** AP **Per-class counting metric (S4):** F_β Score **Overlap-based metric (S7):** Center line Dice Similarity Coefficient (clDice) **Boundary-based metric (S8):** NSD
	Surgical instrument instance segmentation in colonoscopy videos			

Making it accessible to the community: Metric cheat sheets







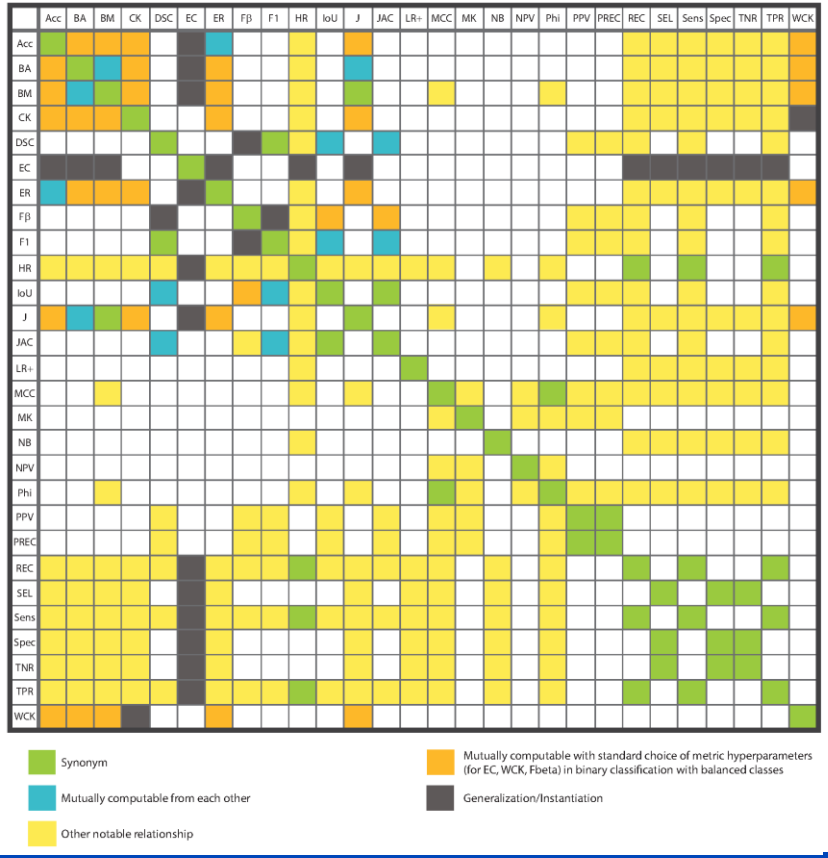
...

Anecdote…

Reviewer #3

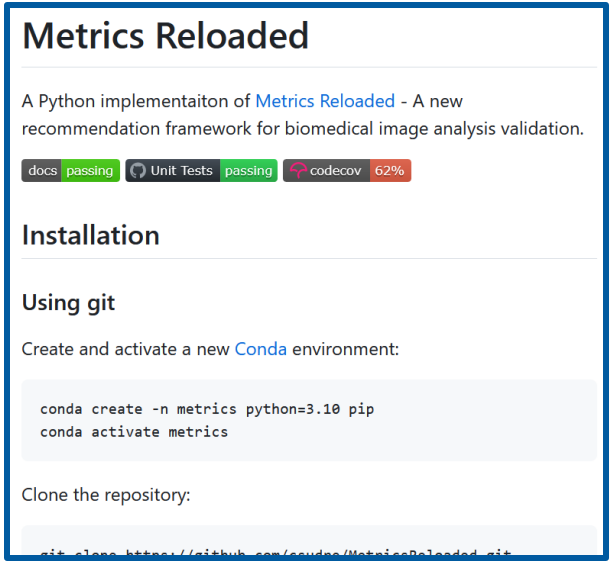
7. Why have only DSC and NSD scores been used for result comparison? Other metrices (F1 score, accuracy, precision) are commonly used to calculate segmentation performance of a network. Any particular reason why they have not been compared?

Making it accessible to the community: Metric relations



Making it accessible to the community: Standard implementation

Carole Sudre



Metrics Reloaded

A Python implementaiton of Metrics Reloaded - A new recommendation framework for biomedical image analysis validation.

docs passing | Unit Tests passing | codecov 62%

Installation

Using git

Create and activate a new Conda environment:

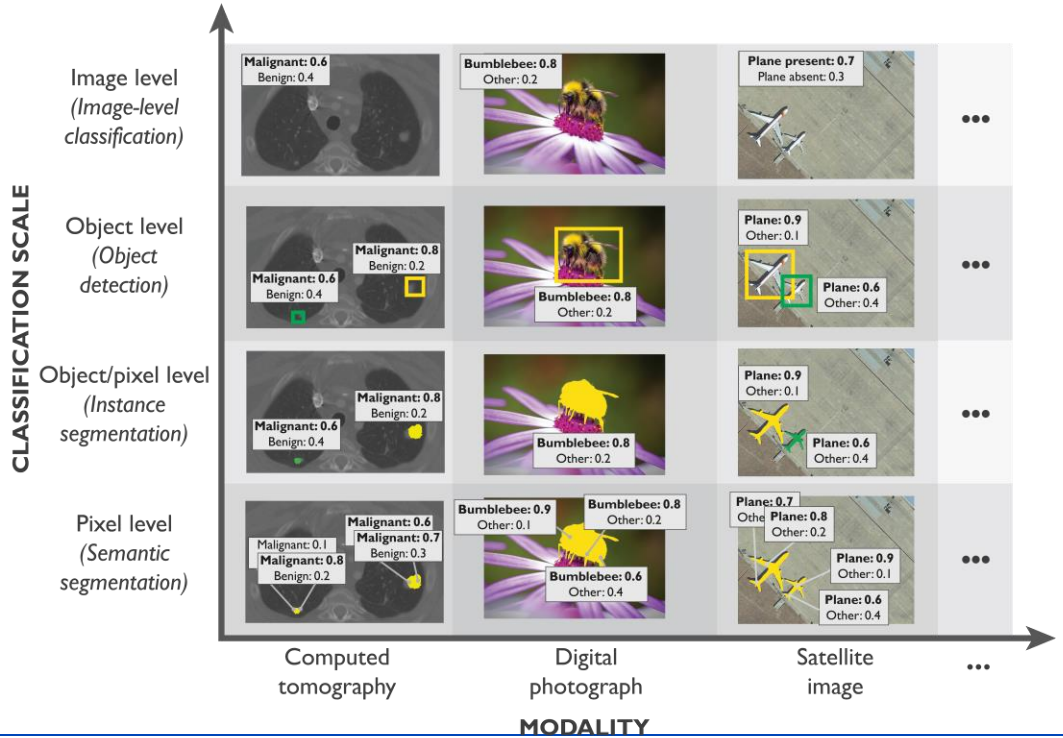
```
conda create -n metrics python=3.10 pip
conda activate metrics
```

Clone the repository:

```
git clone https://github.com/csudre/MetricsReloaded.git
```

https://github.com/Project-MONAI/MetricsReloaded/tree/main/MetricsReloaded

Scope of (first set of) recommendations



Annika Reinke

Minu Tizabi

In the pipeline:

- Metrics Reloaded **Reconstruction**
- Metrics Reloaded **Video**
- Metrics Reloaded **Foundation Models**
- …

(contact me if you want to get involved)

Maier-Hein/Reinke et al. Metrics reloaded: Recommendations for image analysis validation. **Nature Methods 2024**

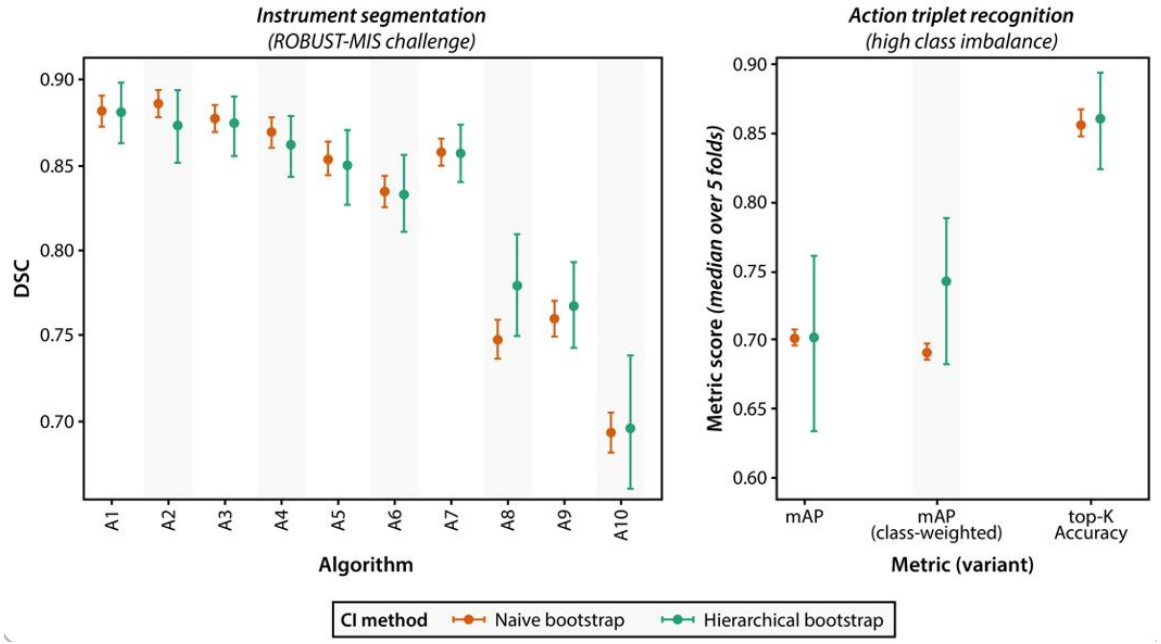
Example video pitfall: Non-independence within the test data

Work in progress

(a) Ignoring hierarchical structure leads to overconfident results

Confidence intervals (CIs) may be underestimated by 2x to over 10x, especially for imbalanced data.

(b) Experimental evidence

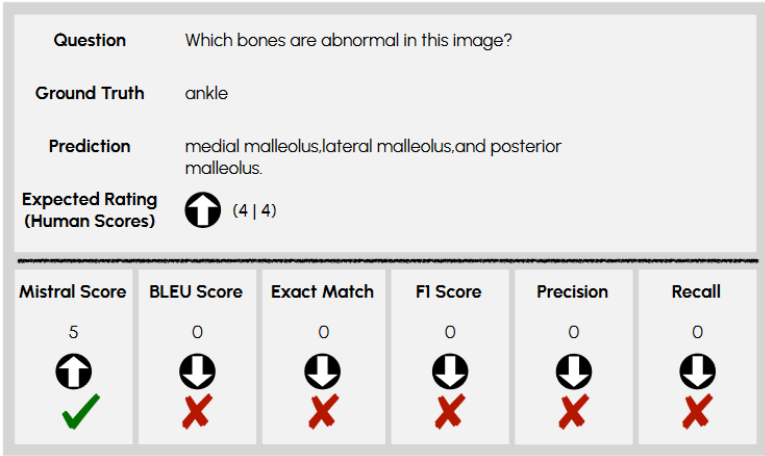


Reinke et al. Current validation practice undermines surgical AI development, Positive pre-submission request Nature Biomedical Engineering 2025

NEW challenges in the context of foundation models

- How to measure important **aspects beyond accuracy**, e.g.
  - Memorization versus true reasoning?
  - Cross-modal consistency?
  - Fairness?
  - Uncertainty awareness? Especially in open-ended generation?
  - Explainability?
  - Generalizability?
  - Energy-efficiency?
  - Long-term alignment?

- **Rating of open-ended questions?**
- **Heterogeneity of metrics across tasks:** How to aggregate over tasks with multiple metrics?



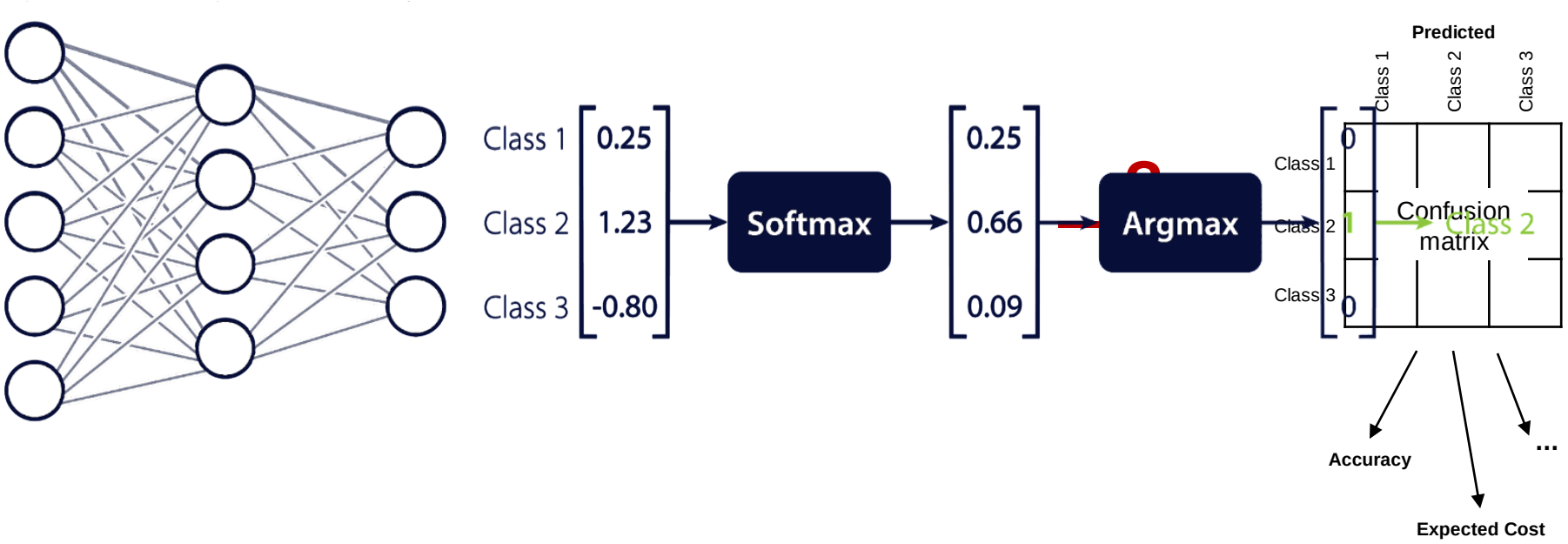
Kahl/Erkan, Maier-Hein, Jäger: SURE-VQA: Systematic understanding of robustness evaluation in medical VQA tasks, **arXiv 2024**

Abbasian et al. Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative AI, **npj Digital Medicine 2024**

#Decision Threshold

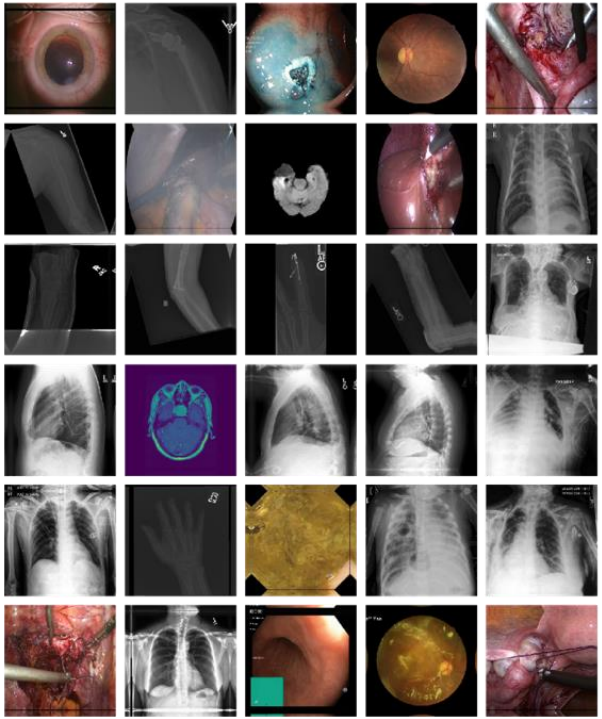
How to convert predicted class scores to classes?

Neural Network **Unnormalized output** **Class Scores**



Godau/Kalinowski, ..., Maier-Hein. Navigating prevalence shifts in image analysis algorithm deployment, **MICCAI 2023**
Medical Image Analysis MICCAI 2023 Best Paper Award

Let‘s look at a variety of tasks to draw solid conclusions...



# of samples: 1,200 - 121,583

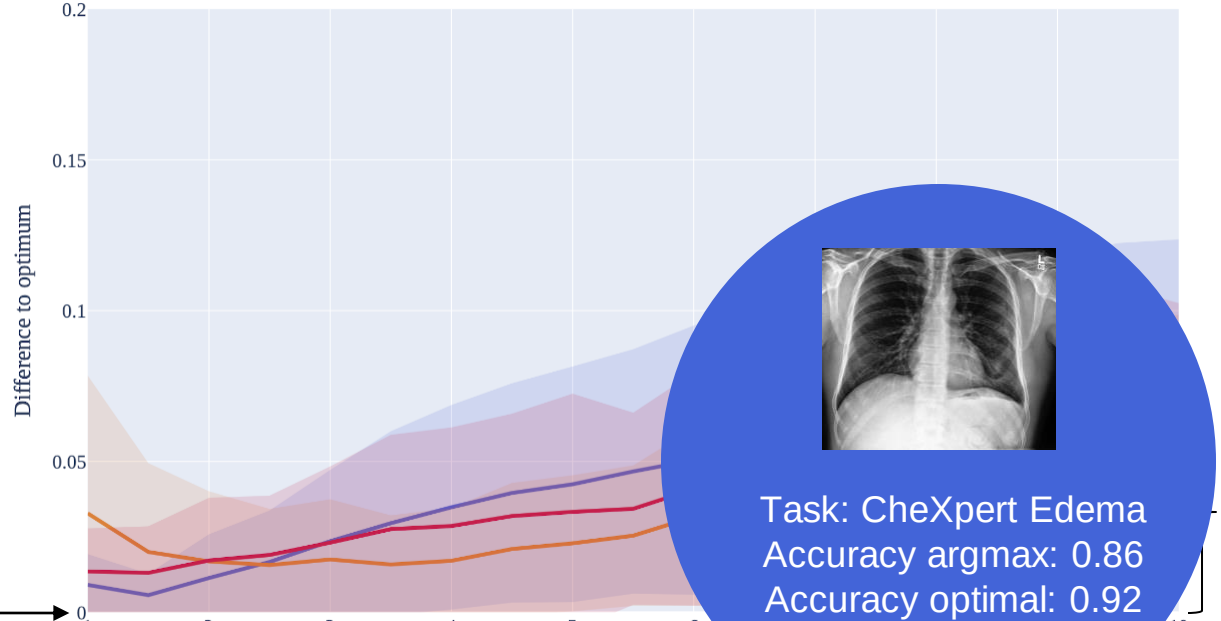
# of classes: 2-8

Imbalance ratio: 1-10.9

Assumption for this talk: Deployment prevalences (approximately) known
Paper shows: Robustness to uncertainty

Godau/Kalinowski, ..., Maier-Hein. Navigating prevalence shifts in image analysis algorithm deployment, **MICCAI 2023**
Medical Image Analysis MICCAI 2023 Best Paper Award

Argmax – what‘s going on?



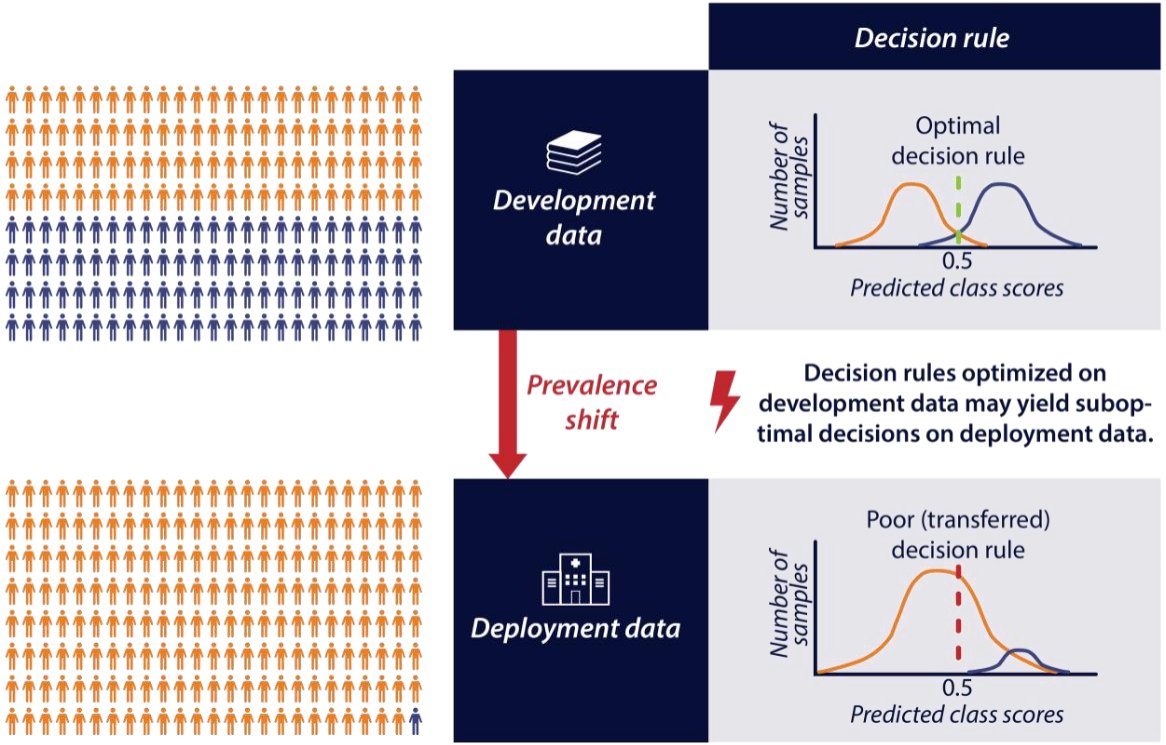
Best cutoff optimized on grid search

Task: CheXpert Edema
Accuracy argmax: 0.86
Accuracy optimal: 0.92

Large deviation from optimal performance!

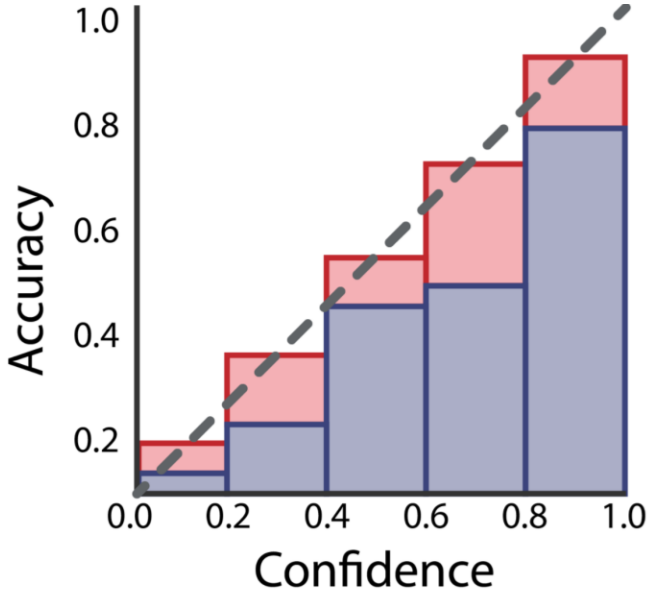
Godau/Kalinowski, ..., Maier-Hein. Navigating prevalence shifts in image analysis algorithm deployment, **MICCAI 2023**
Medical Image Analysis MICCAI 2023 Best Paper Award

That‘s why…



Godau/Kalinowski, ..., Maier-Hein. Navigating prevalence shifts in image analysis algorithm deployment, **MICCAI 2023**
Medical Image Analysis MICCAI 2023 Best Paper Award

Common calibration metric: Expected Calibration Error



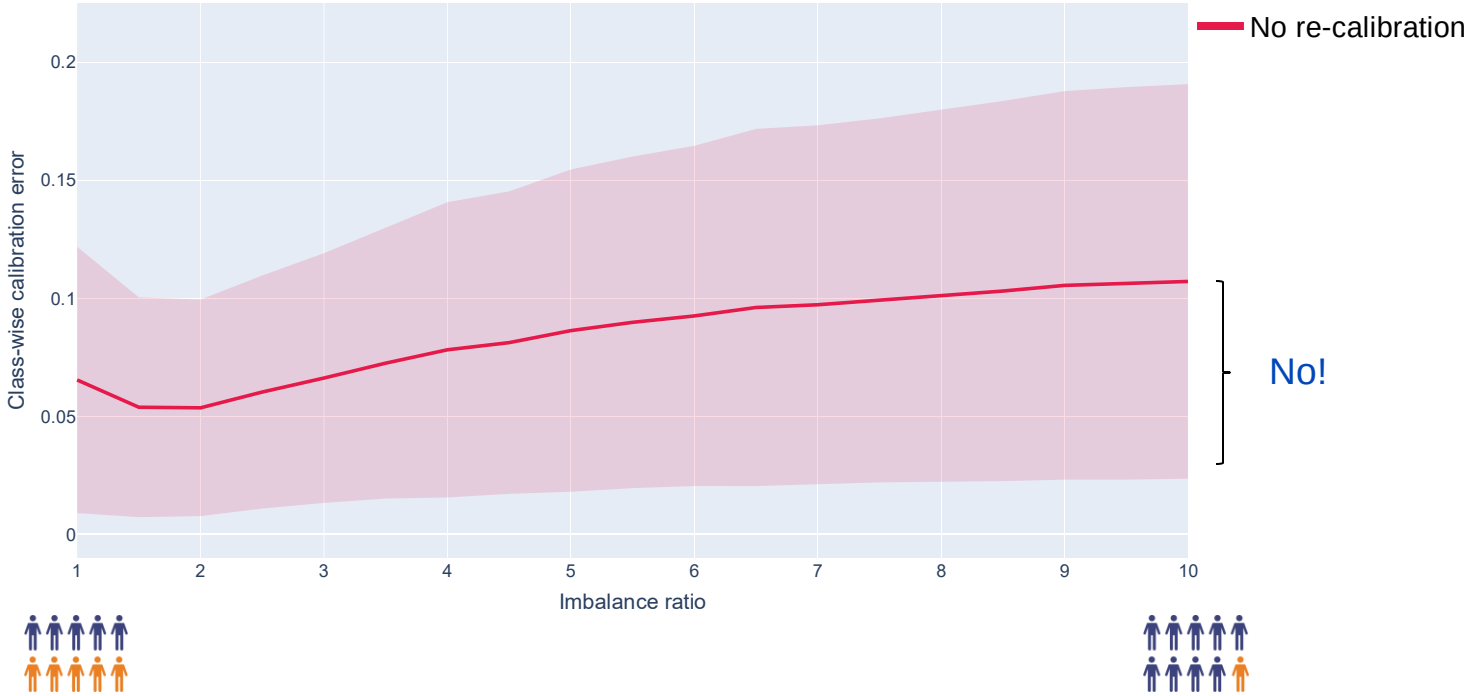
$$ECE = \sum_{m=1}^{n} \frac{|B_m|}{n} |Accuracy(B_m) - Confidence(B_m)|$$

= Weighted Average { , , , , }

VALUE RANGE: [0, 1] ↑

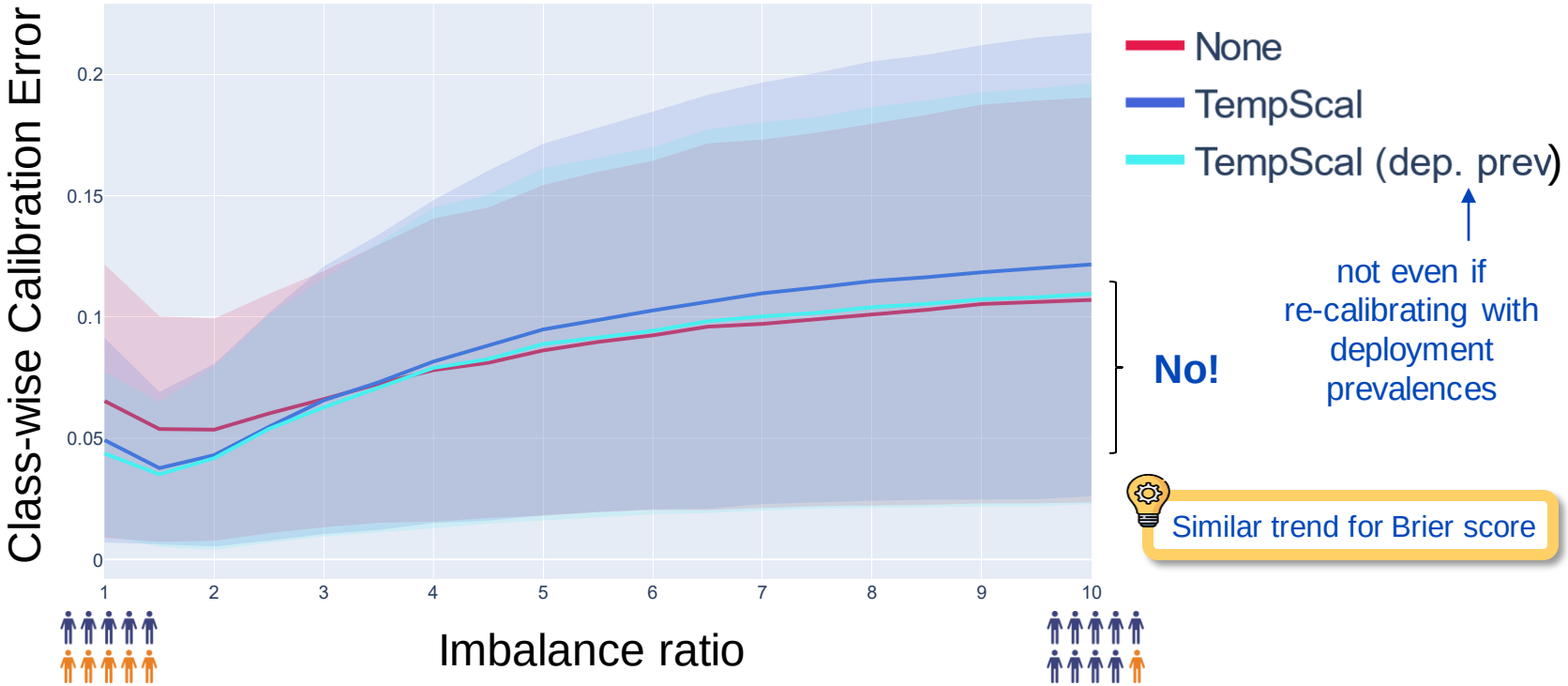
Godau/Kalinowski, ..., Maier-Hein. Navigating prevalence shifts in image analysis algorithm deployment, **MICCAI 2023**
Medical Image Analysis MICCAI 2023 Best Paper Award

Are the scores calibrated?



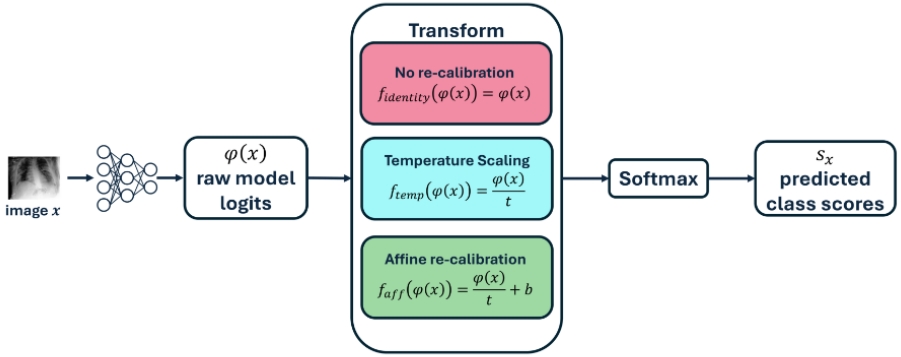
Godau/Kalinowski, ..., Maier-Hein. Navigating prevalence shifts in image analysis algorithm deployment, **MICCAI 2023**
Medical Image Analysis MICCAI 2023 Best Paper Award

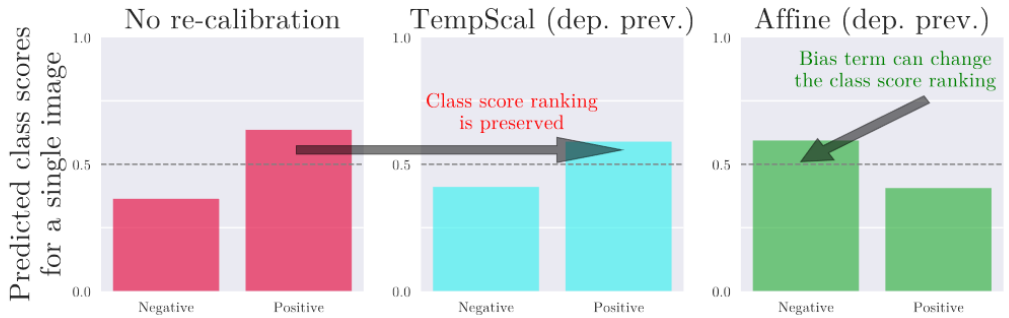
Does temperature scaling fix our problem?



Godau/Kalinowski, ..., Maier-Hein. Navigating prevalence shifts in image analysis algorithm deployment, **MICCAI 2023**
Medical Image Analysis MICCAI 2023 Best Paper Award

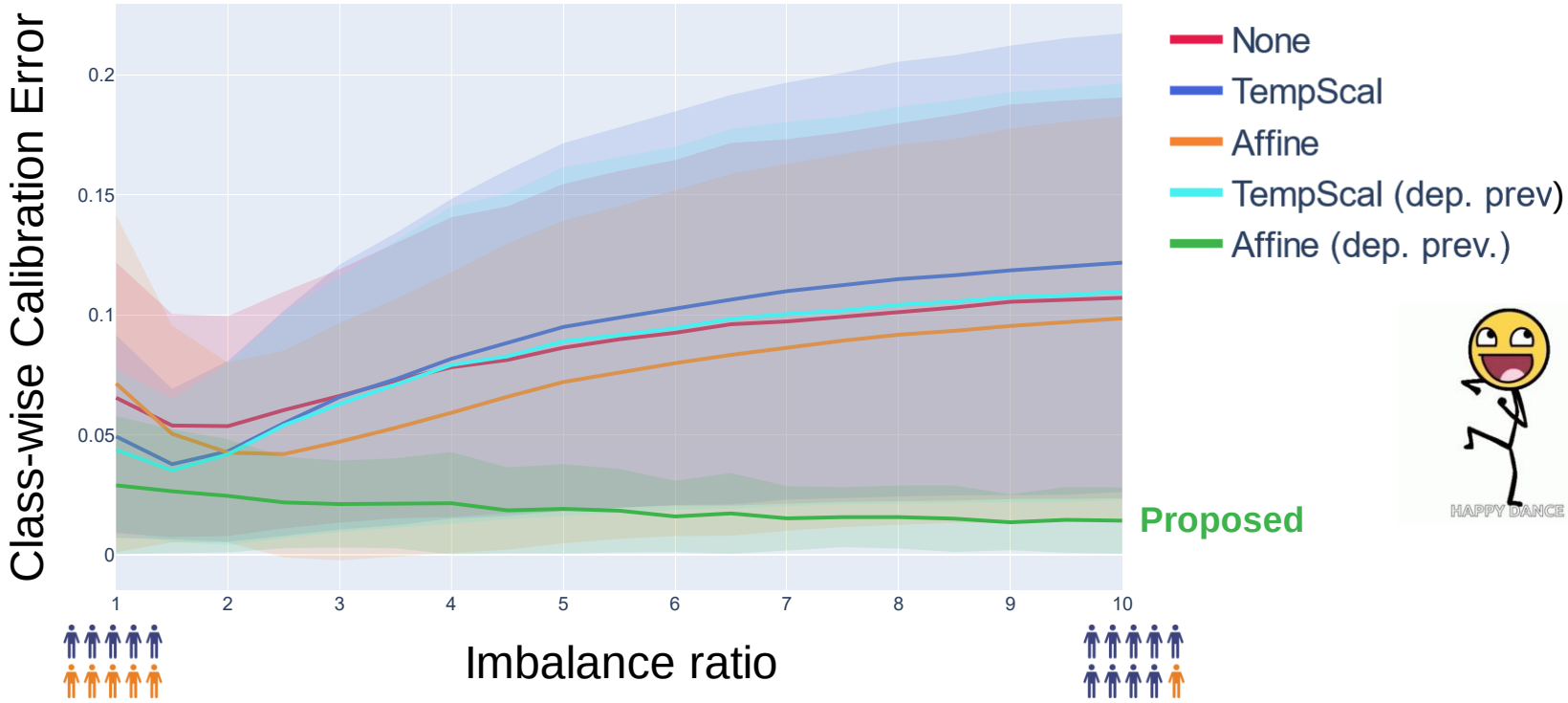
Why not? Because we need a bias term





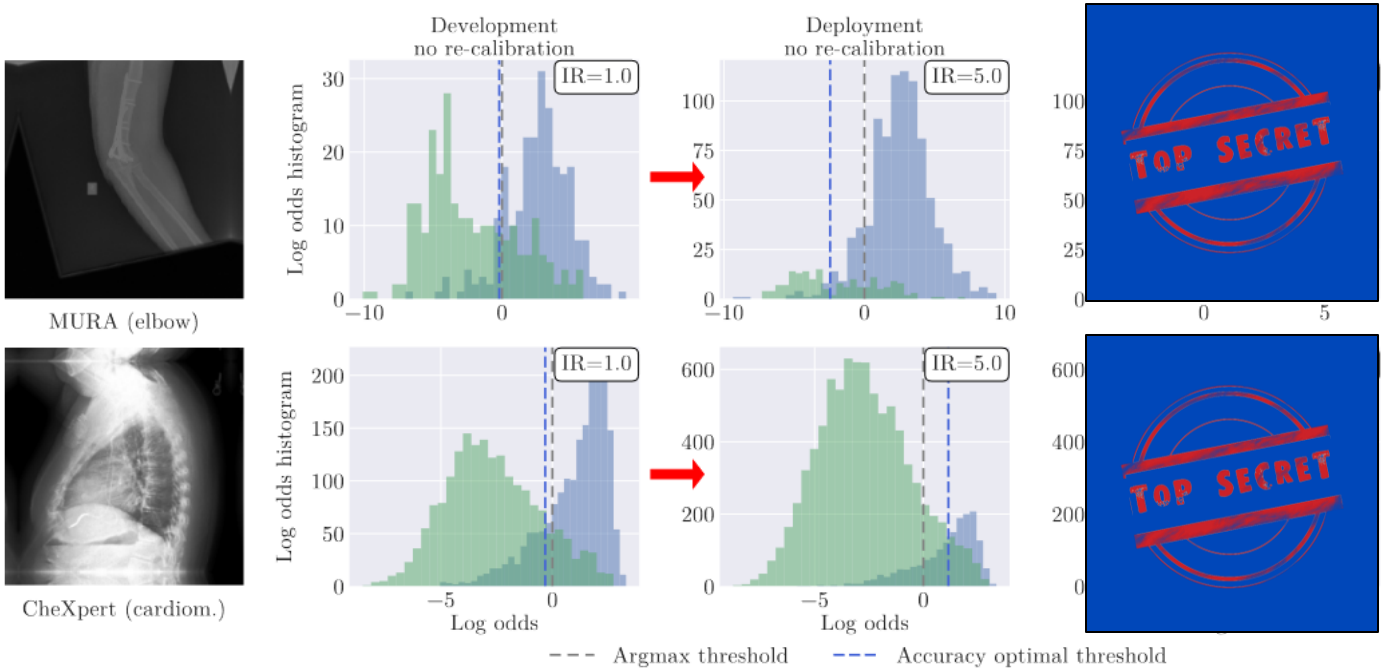
Godau/Kalinowski, ..., Maier-Hein. Navigating prevalence shifts in image analysis algorithm deployment, **MICCAI 2023**
Medical Image Analysis MICCAI 2023 Best Paper Award

Affine re-calibration solves the calibration issue

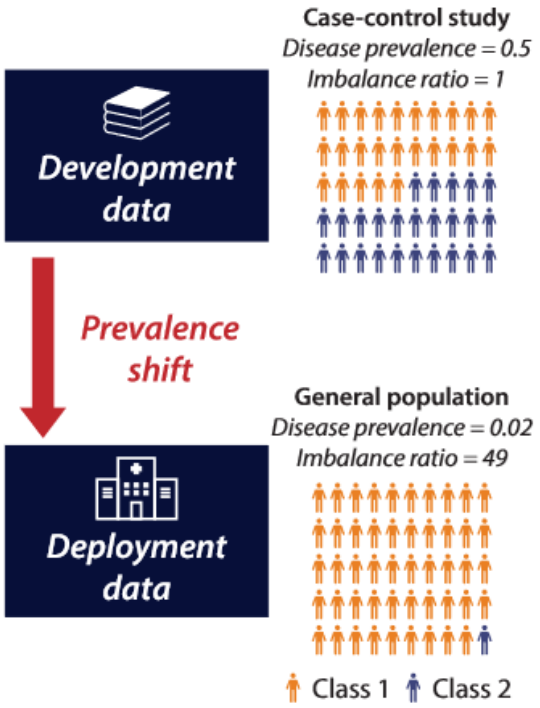


Godau/Kalinowski, ..., Maier-Hein. Navigating prevalence shifts in image analysis algorithm deployment, **MICCAI 2023**
Medical Image Analysis MICCAI 2023 Best Paper Award

Some real examples…



Godau/Kalinowski, ..., Maier-Hein. Navigating prevalence shifts in image analysis algorithm deployment, **MICCAI 2023**
Medical Image Analysis MICCAI 2023 Best Paper Award



RQ1: How well can we estimate prevalences from unlabeled deployment data?

RQ2: What is the effect of prevalence shifts on the quality of

a) calibration,

b) decision rules,

c) performance assessment?

RQ3: How to best compensate for prevalence shifts?

Godau/Kalinowski, ..., Maier-Hein. Navigating prevalence shifts in image analysis algorithm deployment, **MICCAI 2023**
Medical Image Analysis MICCAI 2023 Best Paper Award

(a) PROPOSED WORKFLOW

I. ESTIMATE DEPLOYMENT PREVALENCES

Based on medical records, epidemiological research, or using a data-driven approach.

CC EMQ PACC KDEyCS BBSE ACC HDy KDEyML KDEyHD

II. PERFORM PREVALENCE-AWARE RE-CALIBRATION

Transform model logits with Affine re-calibration. Parameters are optimized with the cross-entropy loss and prevalence dependent class weights.

$$f_{aff}(\varphi(x)) = \frac{\varphi(x)}{t} + b,\ t \in \mathbb{R},\ b \in \mathbb{R}^C$$

key to success

$\varphi(x)$: model output logits for input x

III. CONFIGURE VALIDATION METRIC WITH DEPLOYMENT PREVALENCES

Adjust class prevalences in metric formulas to evaluate algorithm performance in realistic settings. Configure Expected Cost (EC) with deployment prevalences P_{dep}:

$$EC = \sum_k P_{dep}(k) \sum_j c_{kj} R_{kj}$$

c_{kj}: "costs" assigned to classifying a sample of class k as j.

R_{kj}: fraction of all samples with reference class k that have been predicted as j.

IV. SET PREVALENCE-AWARE DECISION RULE

Prevalences and validation metrics determine optimal decision rules. Knowing these enables deriving theoretically-optimal decisions for deployment.

$$d(s_x) = \text{argmin}_k \sum_j c_{jk} s_x(j)$$

s_x: predicted class scores for input x after f_{aff} and softmax operator.

Setting $c_{kk} = 0$ for all k and $c_{kj} = 1$ for $k \neq j$: $EC = 1 - Accuracy$ and d simplifies to $\text{argmax}(s_x)$

V. EXTERNAL VALIDATION

Additional validation on the actual data of the new environment is indispensable for monitoring!

(b) KEY INSIGHTS

(partially experimental proof of theory)

Data-driven estimation of deployment prevalences can be achieved with high accuracy *(see Fig. 5)*.

In contrast to common Temperature Scaling, **weight-adjusted Affine re-calibration compensates for miscalibration** due to prevalence shifts *(see Fig. 6)*.

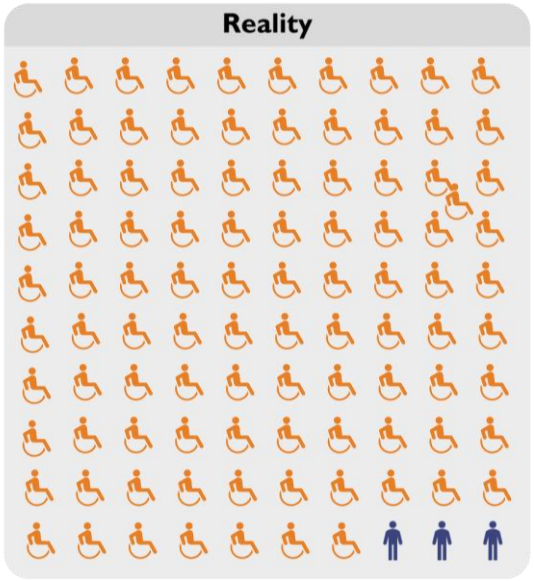
Performance estimations based on development data may be inadequate under prevalence shifts. EC, with adjusted prevalences, can be a robust solution in these scenarios *(see Fig. 10)*.

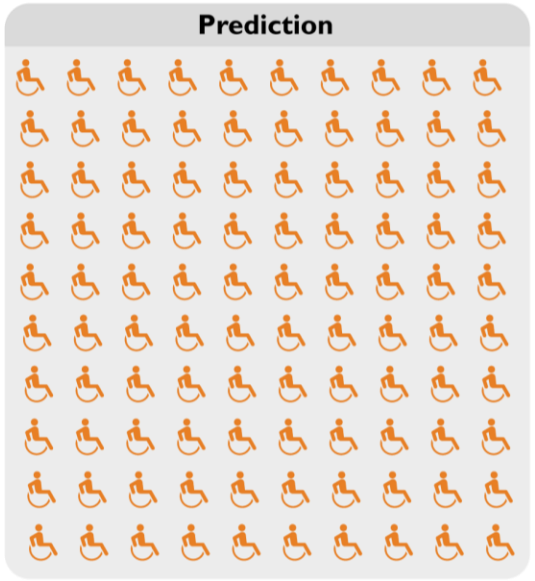
Argmax should not be used indiscriminately as a decision rule. However, it is the optimal decision rule if predicted class scores are calibrated and Accuracy (or EC) is the target metric *(see Figs. 7 & 8)*.

A decision rule, tuned on the development set without accounting for prevalence shifts, may not generalize to the deployment setting *(see Fig. 9)*.

#Baselines

Pitfall: No random baseline



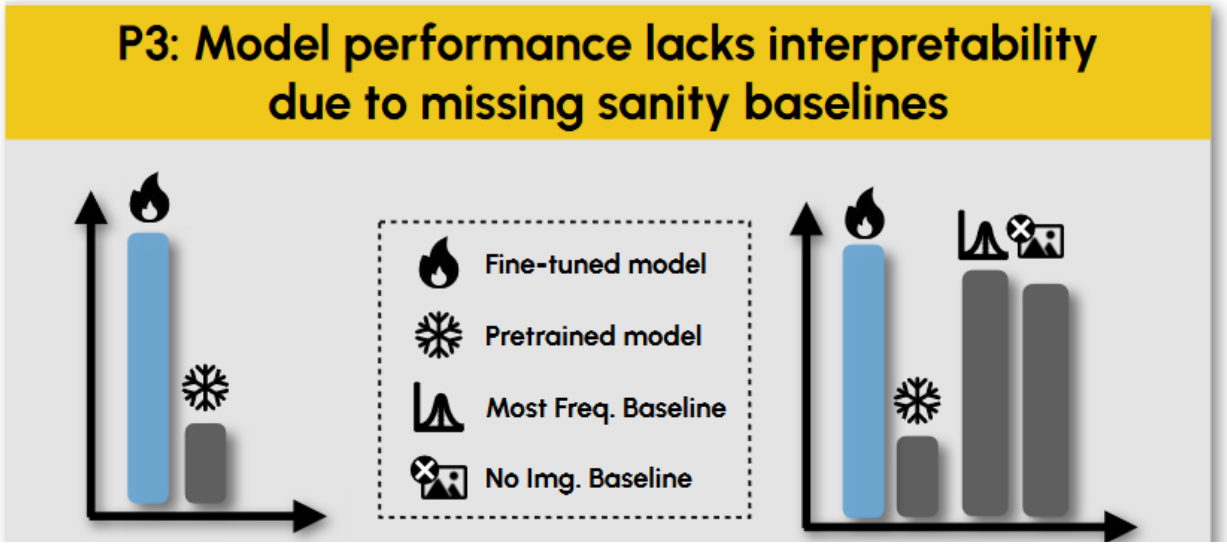


Accuracy: 97%

Baseline most frequent class: 97%

Reinke/Tizabi, …, Maier-Hein. Understanding metric-related pitfalls in image analysis validation. **Nature Methods 2024**
Reinke et al. Common Limitations of Image Processing Metrics: A Picture Story. **arXiv 2023**

NEW pitfall: New random baselines required



Kim-Celine Kahl

Selen Erkan

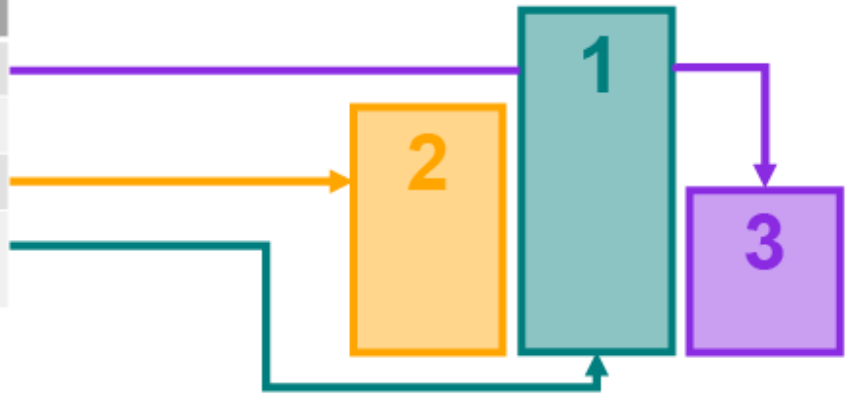
Kahl/Erkan, Maier-Hein, Jäger: SURE-VQA: Systematic understanding of robustness evaluation in medical VQA tasks, **arXiv 2024**

#Rankings

Challenging common practice

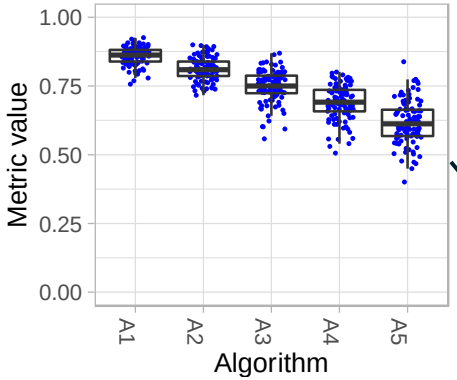
Commonly encountered results tables

Methods	Accuracy	AUC
Method 1	0.83	0.91
Method 2	0.80	0.89
Method 3	0.83	0.92
Proposed method	**0.84**	**0.92**

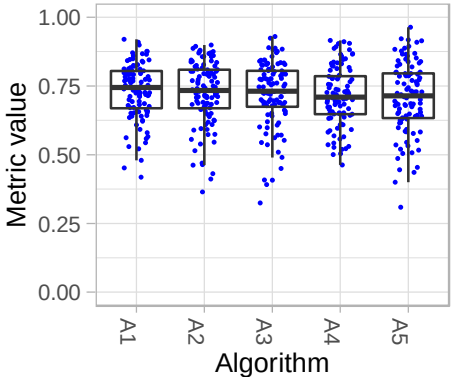


“As shown in Table 1, our method outperforms all previously proposed state-of-the-art methods”

Why result analysis and visualization is critical: Example



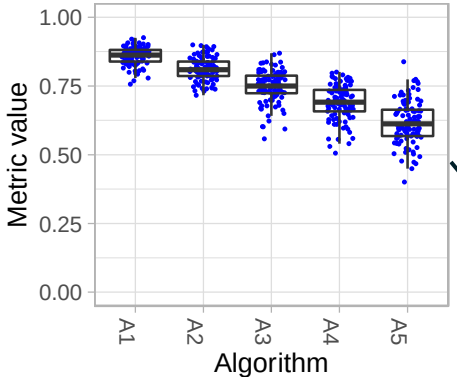
Same ranking

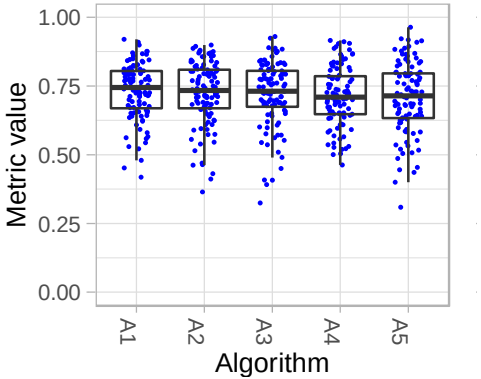


Rank	Algorithm
1	A1
2	A4
3	A3
4	A2
5	A5

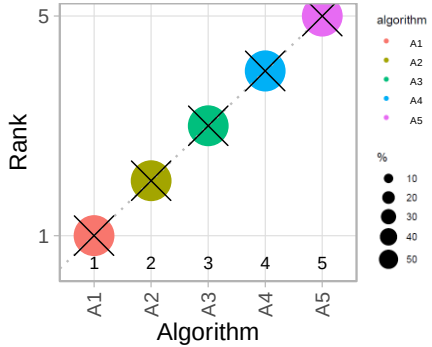
Wiesenfarth, … Maier-Hein/Kopp-Schneider. Methods and open-source toolkit for analyzing and visualizing challenge results, **Scientific Reports 2021**

Why result analysis and visualization is critical: Example

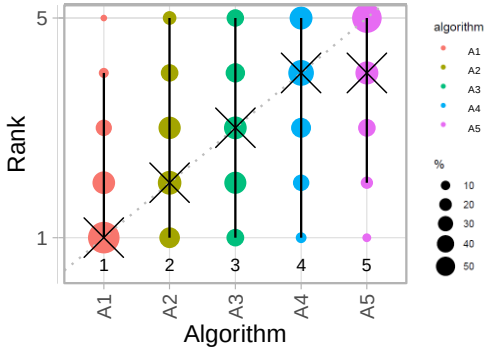




Same ranking



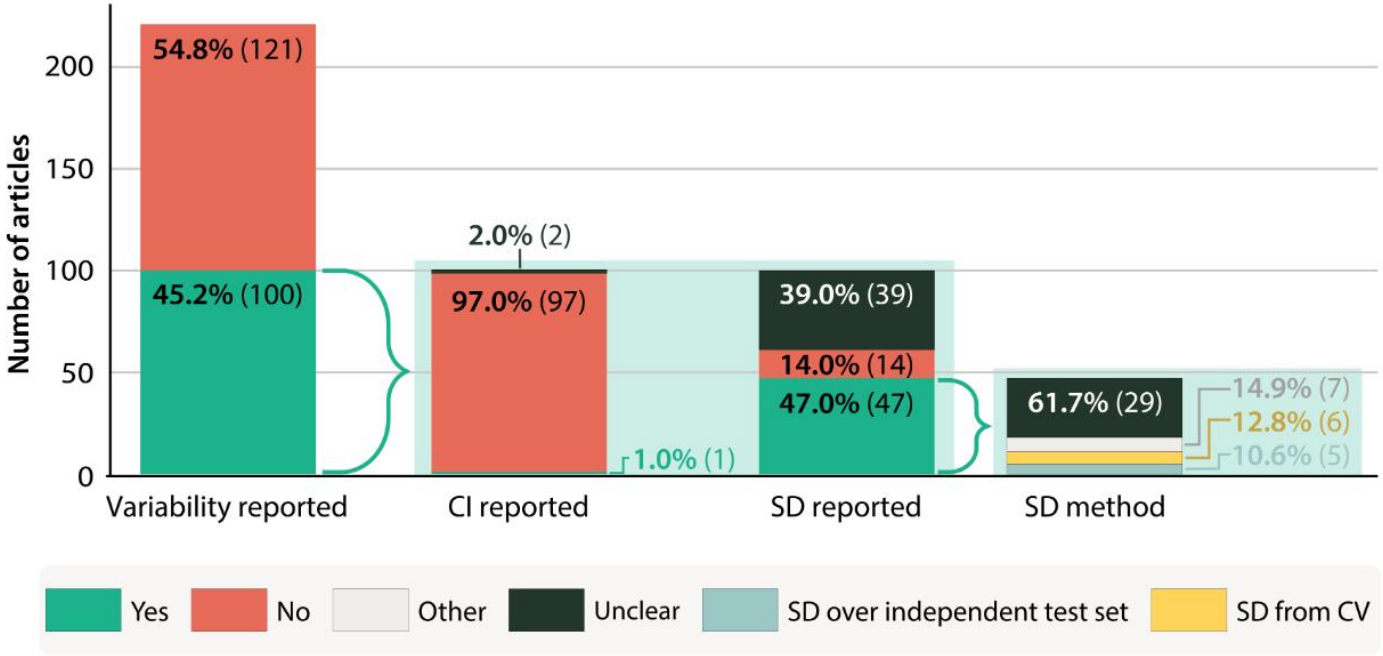
Rank	Algorithm
1	A1
2	A4
3	A3
4	A2
5	A5



Wiesenfarth, … Maier-Hein/Kopp-Schneider. Methods and open-source toolkit for analyzing and visualizing challenge results, **Scientific Reports 2021**

Do we have a problem in practice?

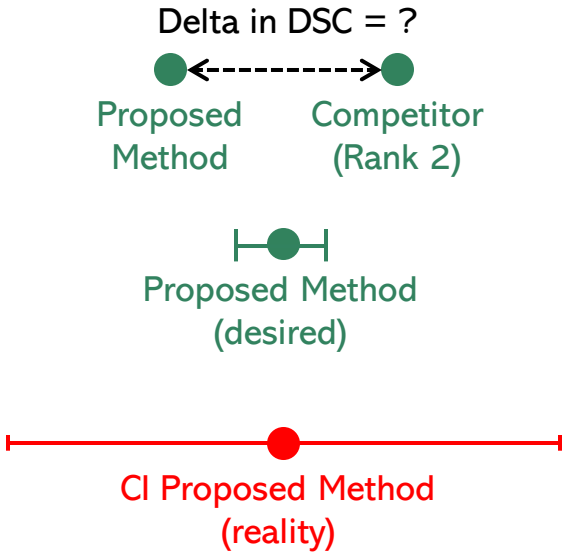
(analysis of all >200 MICCAI 2023 segmentation papers)



One 1(!) paper reported the CI

Christodoulou/Reinke et al. Confidence intervals uncovered: Are we ready for real-world medical imaging AI? MICCAI 2024.

By which margin does a proposed method beat the strongest competitor?



Median improvement in DSC: 0.01
(all MICCAI 2023 segmentation papers)

Confidence interval width should be smaller …

Reality: Median CI width: ~0.03

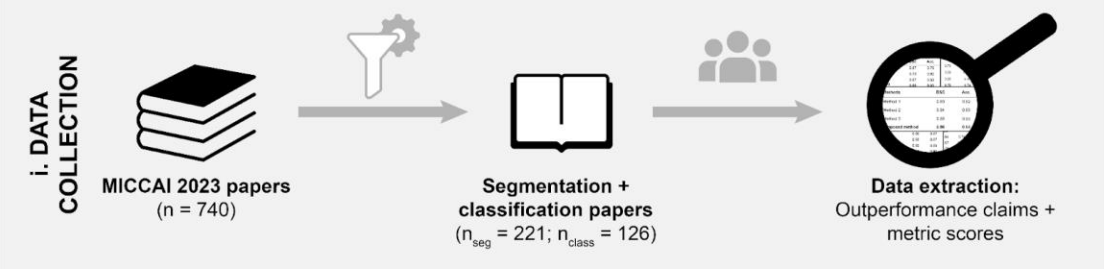
Eva Christodoulou

Annika Reinke

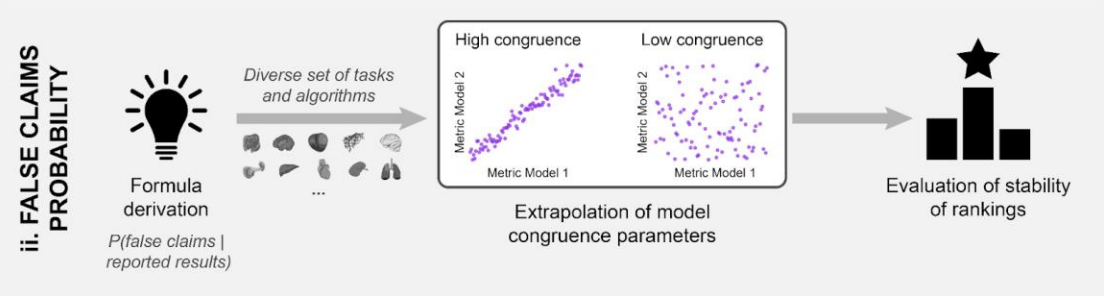
Close collaboration with Gael Varoquaux and Olivier Colliot

Christodoulou/Reinke et al. Confidence intervals uncovered: Are we ready for real-world medical imaging AI? **MICCAI 2024**

False outperformance claims?



RQ: Are common claims of outperformance in medical imaging AI well-substantiated?



Eva Christodoulou

Annika Reinke

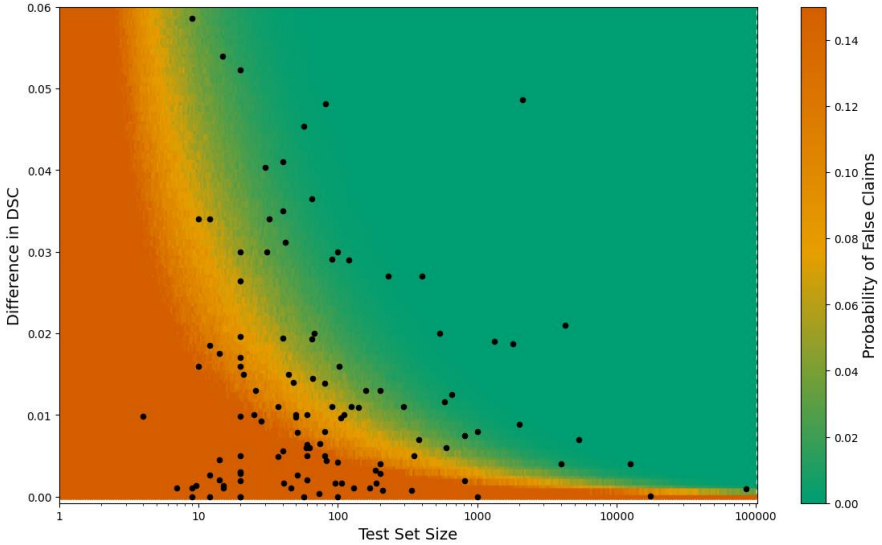
Close collaboration with Gael Varoquaux and Olivier Colliot

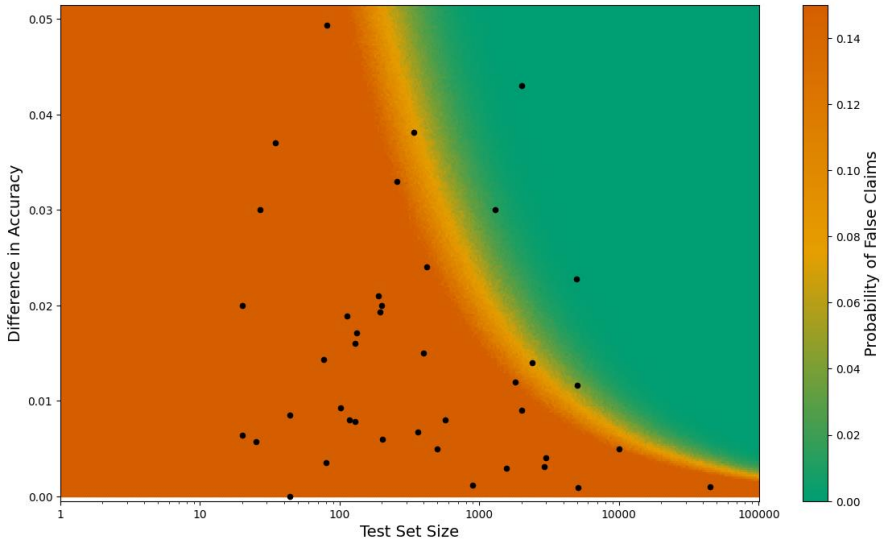
Christodoulou, ..., Varoquaux/Colliot/Maier-Hein. False Promises in Medical Imaging AI? Assessing Validity of Outperformance Claims, **arXiv 2025**

Outperformance claims are not well-substantiated

High probability (>5%) of false outperformance claims in

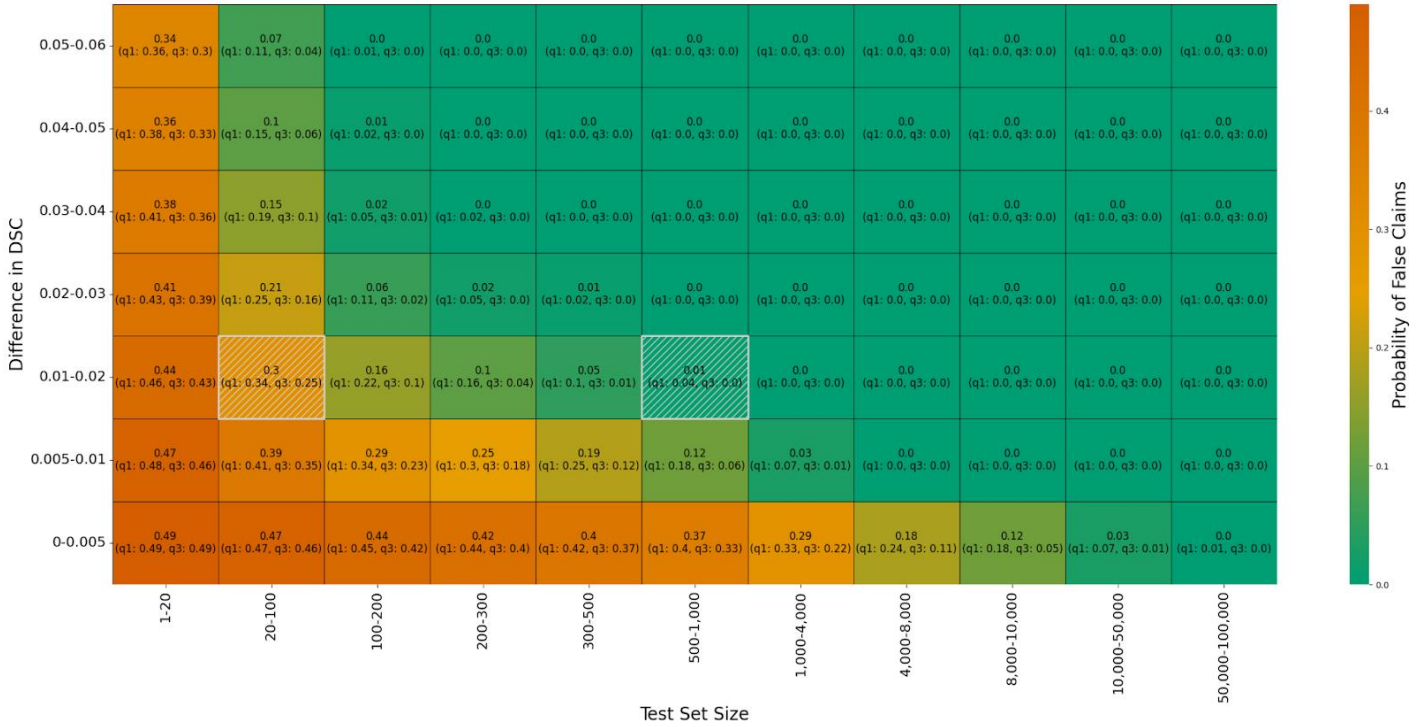
- 86% of classification papers and
- 53% of segmentation papers





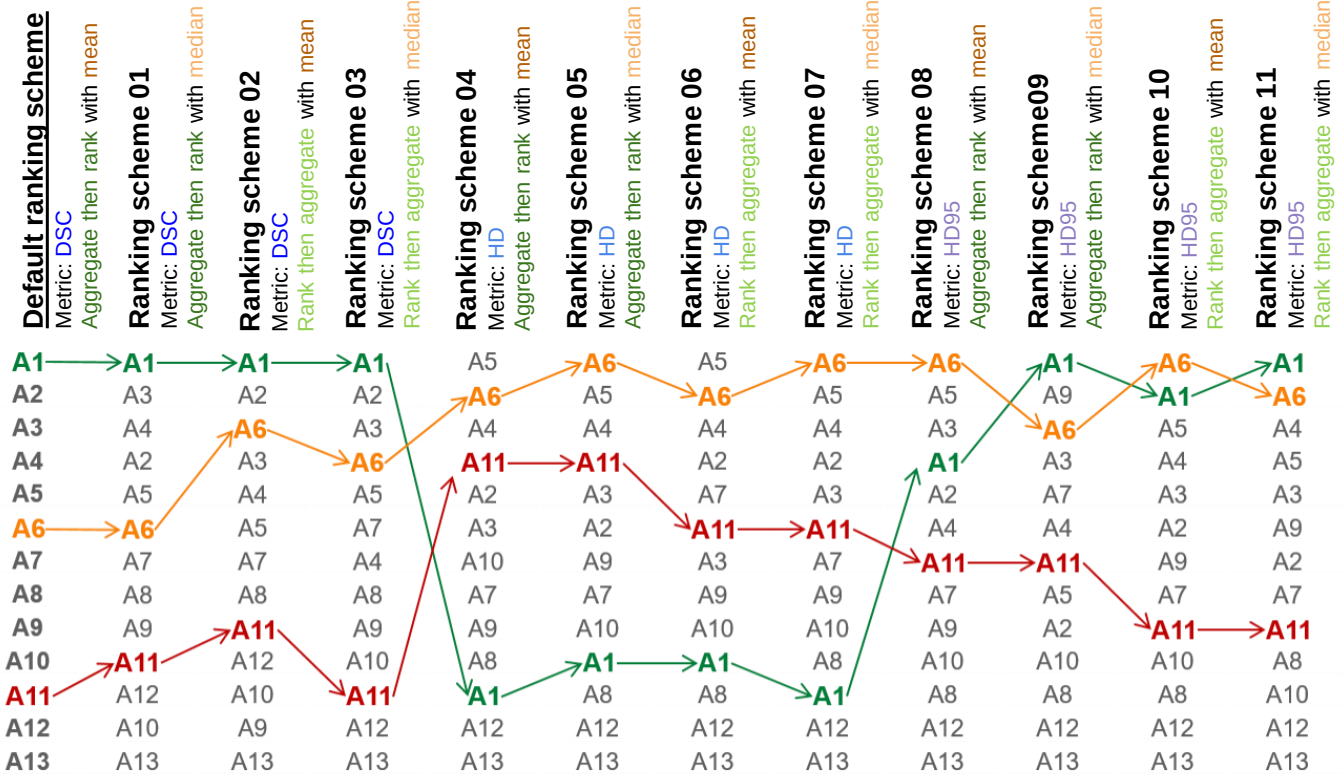
Christodoulou, ..., Varoquaux/Colliot/Maier-Hein. False Promises in Medical Imaging AI? Assessing Validity of Outperformance Claims, **arXiv 2025**

We need larger sample sizes



Christodoulou, ..., Varoquaux/Colliot/Maier-Hein. False Promises in Medical Imaging AI? Assessing Validity of Outperformance Claims, **arXiv 2025**

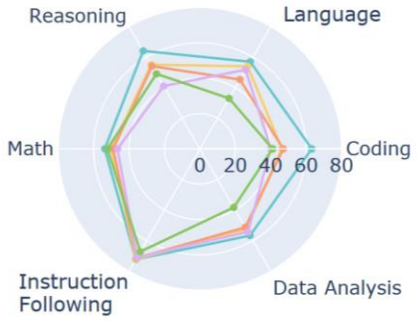
Pitfalls beyond sample size: Sensitivity to ranking scheme



Maier-Hein et al. Why rankings of biomedical image analysis competitions should be interpreted with care. **Nature Commun. 2018**
Reinke et al. How to Exploit Weaknesses in Biomedical Challenge Design and Organization. **MICCAI 2018**

NEW pitfalls in rankings

- More sources of variability, e.g. prompt sensitivity
- More assessment categories
- More relevant complementary aspects, e.g. human-AI-collaboration
- How to incorporate relevance?
- Arena-style validation
  - *Relative* performance testing
  - Lack of relieable reference
  - Overfitting to arena-preferred examples
  - Reasons of failure?

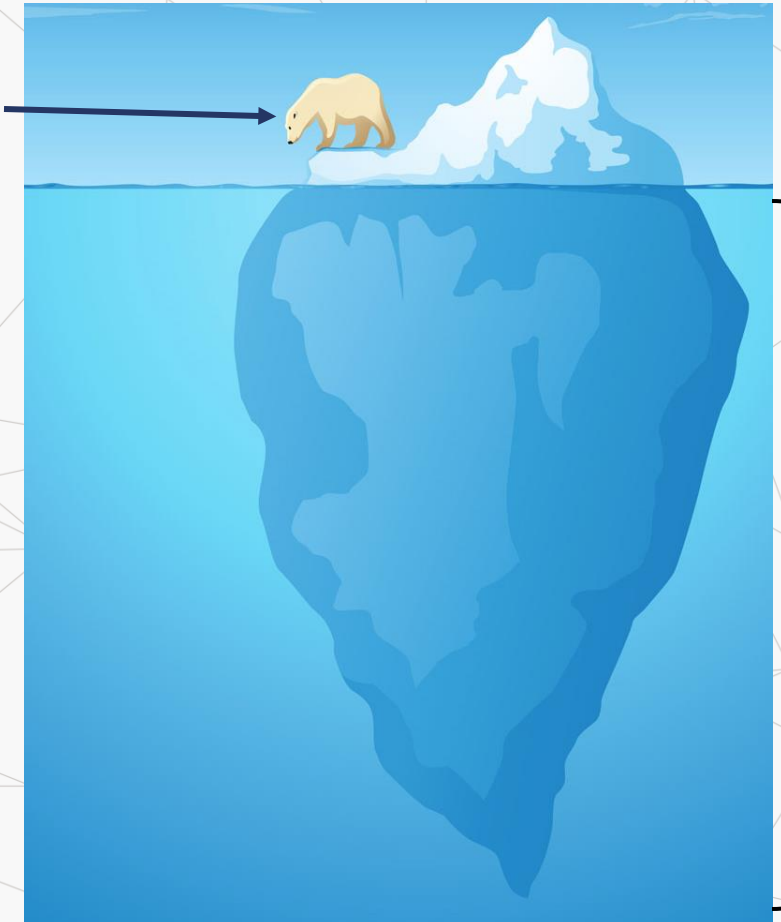


Dataset	Flan-PaLM 540B with non-medical prompt	Flan-PaLM 540B with medical prompt
MedQA 4 options	65.4	67.6
MedMCQA	55.2	57.6
PubMedQA	77.2	75.2

White C., …, Goldblum M. Livebench: A challenging, contamination-free LLM benchmark. **ICLR 2025 spotlight**
Singhal et al. Large Language Models Encode Clinical Knowledge. **Nature 2023**

Summary: Proper evaluation requires attention to detail

A(G)I developer



Data splitting

Data shifts

Metrics

Rankings

Cheating

Reporting

…

Image source: Regulatory Affairs Professional Society

The role of academia in AI?

nature

Explore content | About the journal | Publish with us | Subscribe

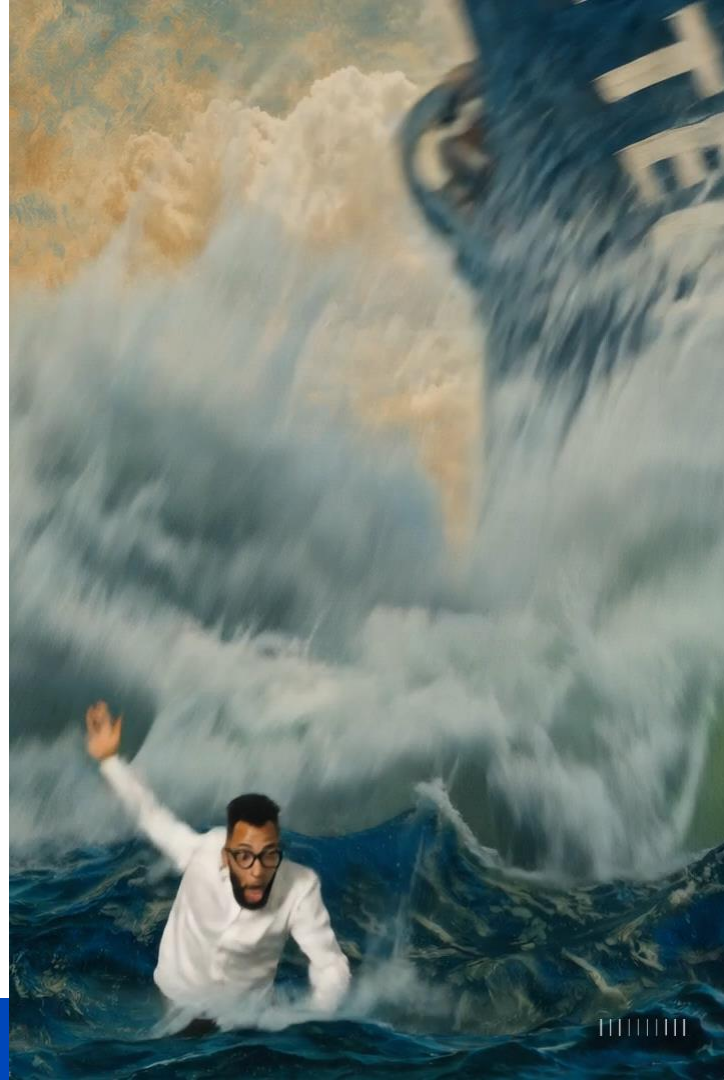
nature > nature index > article

NATURE INDEX | 18 September 2024

Rage against machine learning driven by profit

Industry research funding is vastly eclipsing academia's spend, but healthy development demands broad input.

"Academia is the only place where researchers still have the ability to work without an obvious roadmap to profit."



Intelligent Medical Systems, DKFZ

@lena_maierhein

@DKFZ_IMSY_lab

New: @lena-maier-hein.bsky.social

HELMHOLTZ IMAGING

NCT

INTELLIGENT SYSTEMS IN SURGICAL ONCOLOGY

erc European Research Council