

这份PDF《Failing forward: Rethinking the foundations of imaging AI》是由德国癌症研究中心（DKFZ）智能医学系统负责人 Prof. Dr. Lena Maier-Hein 在2025年7月的一个演讲中展示的内容，重点讨论了当前医学影像AI发展中存在的评估方法缺陷与研究实践问题，并提出了改善建议。以下是该演讲的核心内容总结：

核心主题：医学影像AI发展中“评估指标”和“挑战设计”的反思与改进

一、常见陷阱与问题（Pitfalls）

1. ! 评价指标误导

- 像素级指标（如 DSC）往往过于乐观，无法反映小病灶或新发病灶检测能力；
- 目标级指标（如 recall）较低但被忽视，导致AI表现被高估；
- 临床中重要的“实例进展”未被现有指标捕捉到。

2. ! 挑战赛评估机制存在问题

- 挑战赛排名在数据不足时不稳定；
- 由于数据稀疏（如仅200+测试样本），Bootstrap分析表明第一名经常不是“真正最优”的模型；
- 缺乏多样性任务，导致模型泛化能力无法评估。

3. ! 数据泄露与“测试集污染”

- 大模型训练过程中可能无意中包含测试数据，破坏公正性；
- 强调需要LiveBench等动态测试集替代静态benchmark。

4. ! 数据分布偏移（Data Shift）

- 训练与部署场景不同（如国家间/医院间差异），模型性能下降；
- 示例：Google糖尿病视网膜病变检测模型，在现实临床中表现不佳。

5. ! 指标定义与配置敏感

- 如多聚焦/多病灶任务中如何定义TP/FP、Overlap阈值等，影响结果很大；
- AI社区和临床医生对“好结果”的标准可能不同。

二、“Metrics Reloaded”项目：推动指标重构

🔍 目标：

构建更适用于医学影像的评估体系，包括：

- 任务抽象化与指纹化（Fingerprinting）；
- 指标选择决策树与推荐系统；
- 提供工具包、标准实现与cheat sheet：

🔗 工具网站：<https://metrics-reloaded.helmholtz-imaging.de>

🔗 GitHub实现：[MONAI/MetricsReloaded](#)

三、其他关键议题

✓ 预估代价指标（Expected Cost, EC）

- 提议使用更接近真实部署环境的评估方式；
- 替代静态accuracy，加入疾病先验率（prevalence）建模。

✓ 校准问题与偏差矫正

- 温度缩放无法应对类别不平衡，应使用Affine重新校准；
- 多篇MICCAI论文中存在过拟合与无信度区间的问题。

排名系统敏感性

- 不同聚合策略会极大影响算法排名；
- 引入了“竞技场验证（Arena-style validation）”与相对性能评估概念；
- 强调排名结果的可视化与统计置信度报告。

四、未来挑战与研究方向

- 多模态大模型评估：记忆vs推理？一致性？公平性？解释性？
- 如何定义和评价生成式任务（如开放式问答）的“正确性”？
- 如何整合跨任务、跨指标的评估？
- 面向长期部署，提出了新的稳健性、节能性与协作性指标需求。

总结建议

医学影像AI的可靠评估，是推动临床应用的基石。当前社区应：

1. 反思现有评估方法的不适用性与误导性；
2. 推动指标、验证框架与排名机制的系统化改进；
3. 倡导公开工具、标准化报告与跨领域合作（如AI开发者与临床专家的深入结合）。

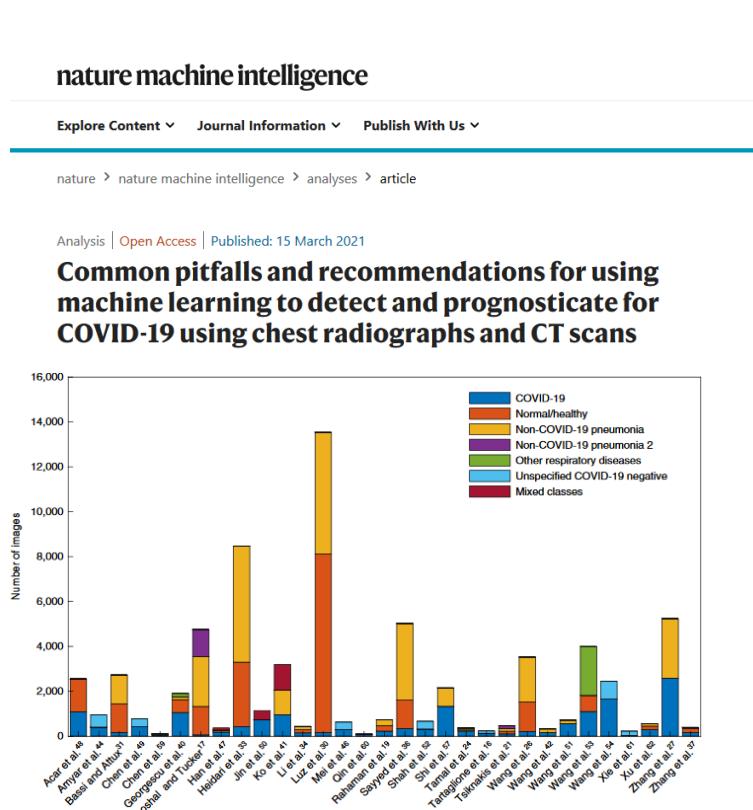
Failing forward: Rethinking the foundations of imaging AI



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

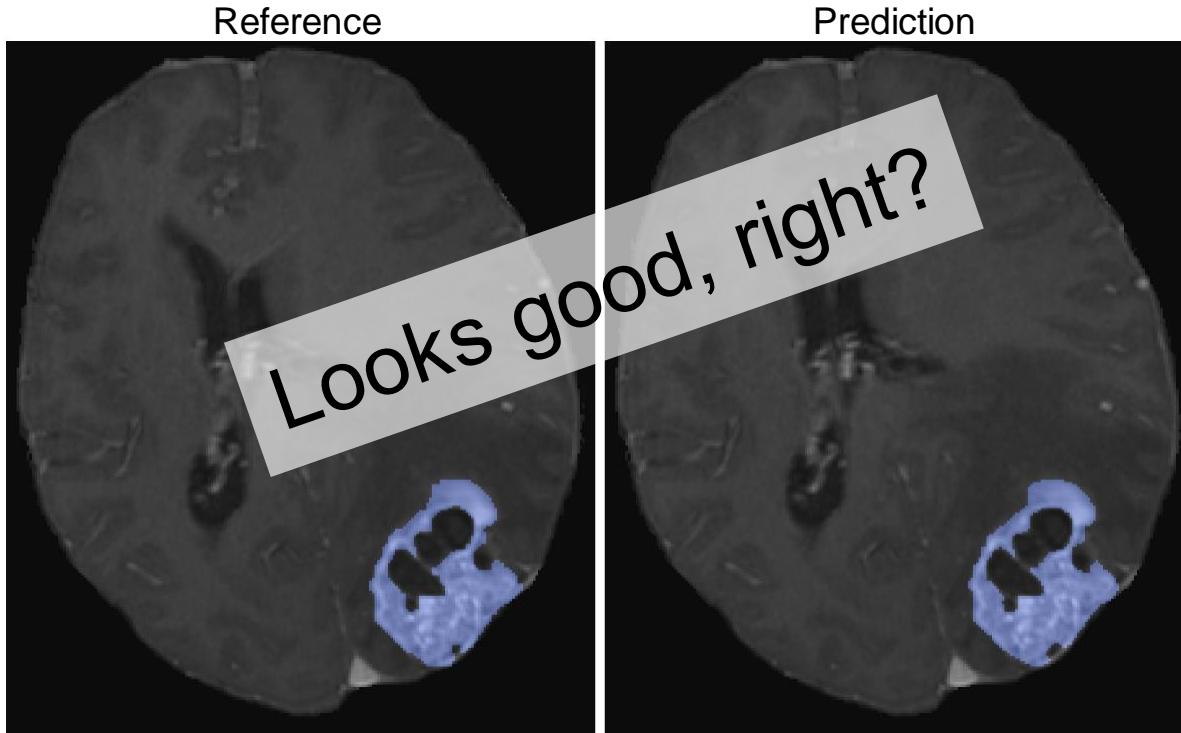


Source: [Pneumonia control group](#)

COVID-19 patients

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

Metrics in (clinical) practice

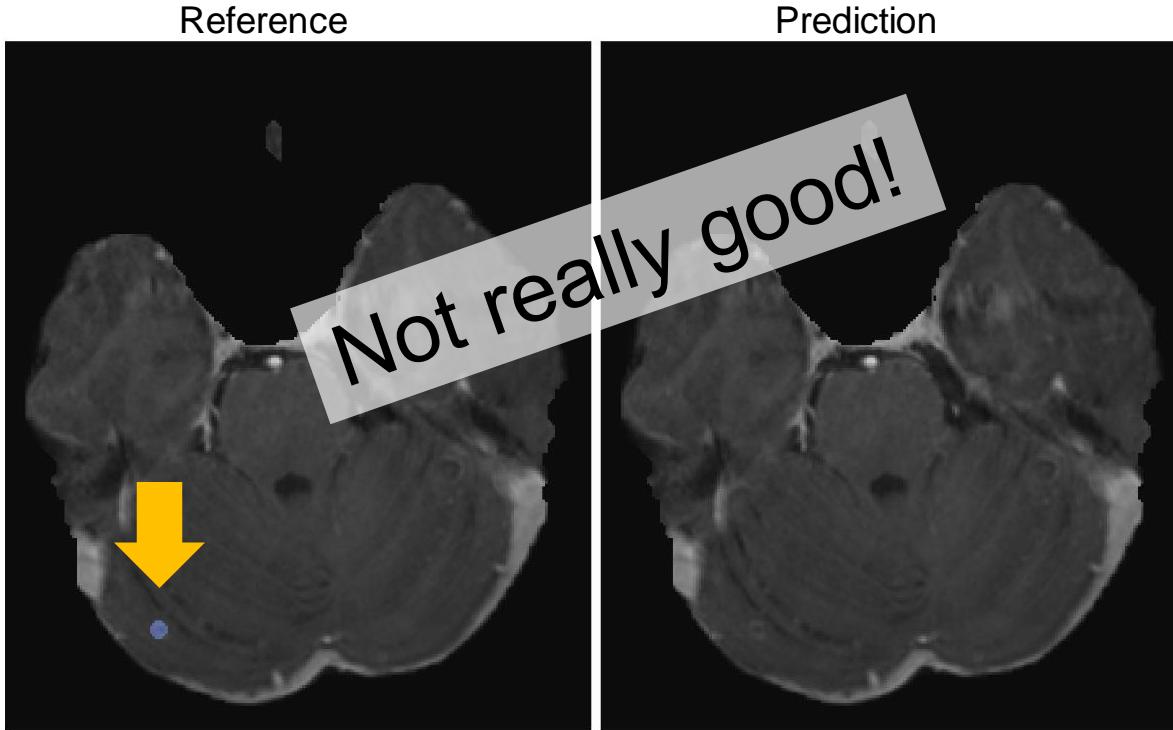


Close to perfect DSC

Recall: 0.94
(voxel-level)

Courtesy Klaus Maier-Hein, DKFZ

Metrics in (clinical) practice



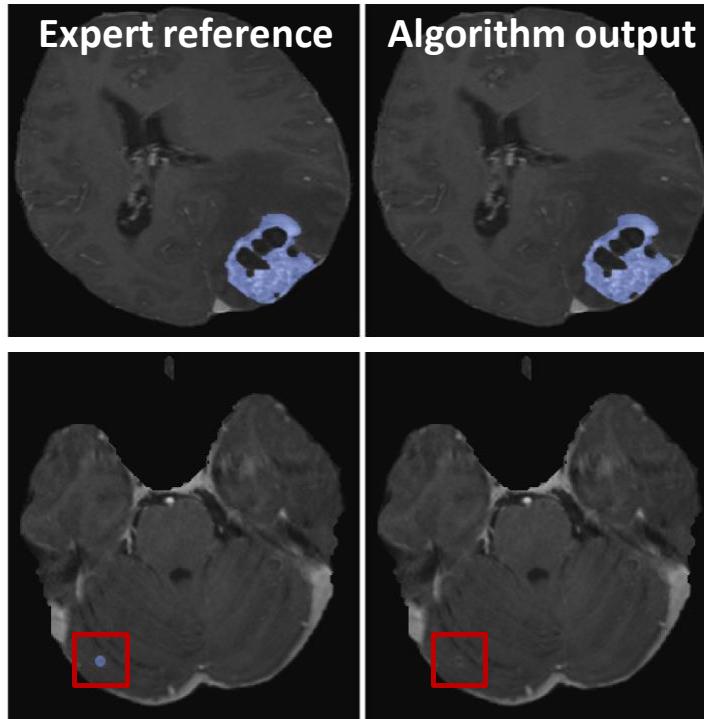
Recall: 0.94
(voxel-level)

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!

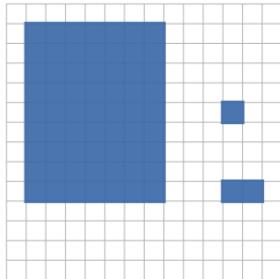


Instance progress not detected in ~1/3 of the cases!



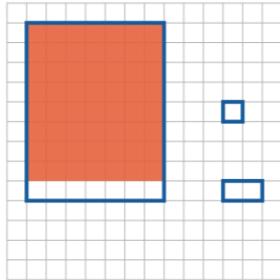
Reinke/Tizabi, ..., Jäger/Maier-Hein. Understanding metric-related pitfalls in image analysis validation. **Nature Methods 2024**

Reference

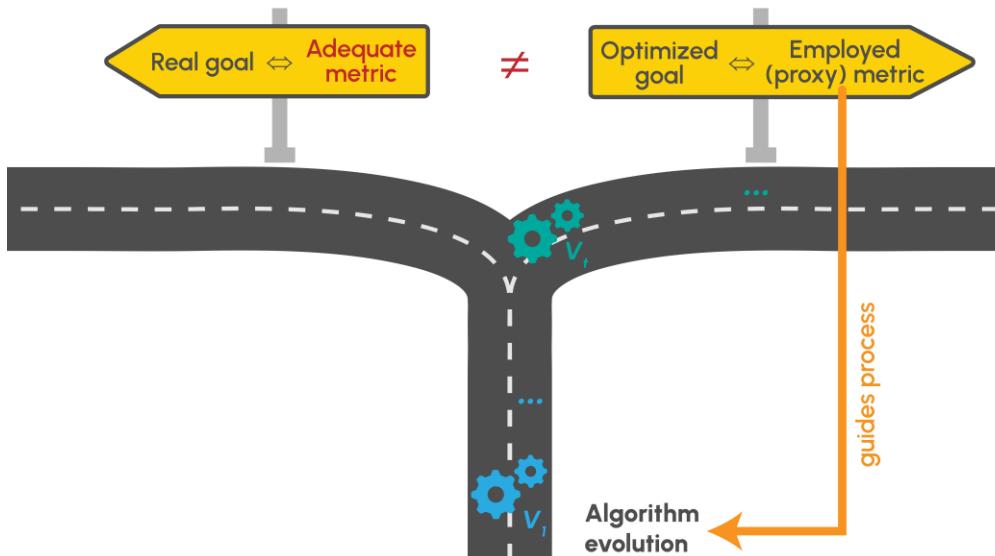


object-level
metric

Algorithm 1

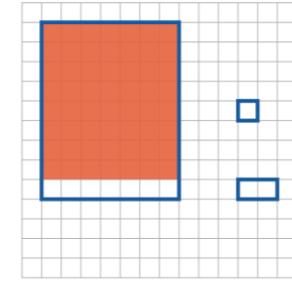


1/3 objects
detected



pixel-level
metric

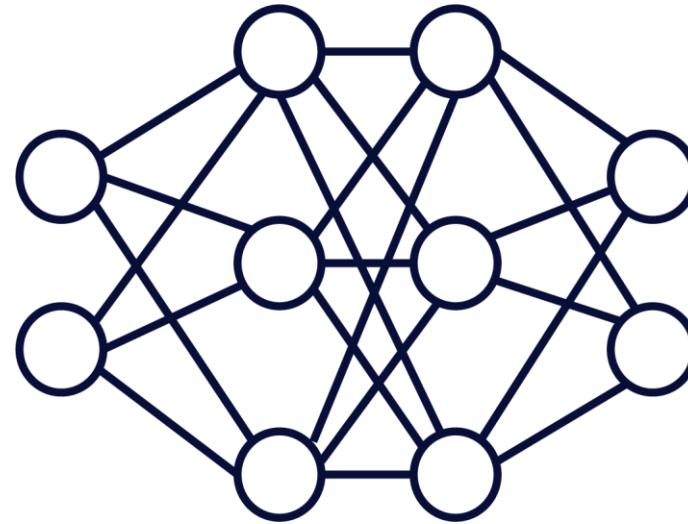
Algorithm 1



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



Gloria Menegaz



Nicholas Heller (2025)



Kendall Schmidt



Elise Blaese



Erik Meijering



Stephen Aylward



Keyyan Farahani



Susheel Varma



Maggie Demkin



Charles Kahn



Anne Mickan



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



Technische Universiteit
Eindhoven
University of Technology



UNIVERSITÄT ZU LÜBECK
INSTITUT FÜR MEDIZINISCHE INFORMATIK



AIExplore



WARWICK
THE UNIVERSITY OF WARWICK



McGill
Centre for
Intelligent Machines



UNIVERSITÄT
BERN



Heidelberg Collaboratory

UNIVERSITÉ DE
RENNES 1



NATIONAL CENTER
FOR TUMOR DISEASES
PARTNER SITE DRESDEN
UNIVERSITY CANCER CENTER UCC

Supported by:
Helmholtz Research Center
University Hospital Carl Gustav Carus Dresden
University Center of Medicine, TU Dresden
Hannover Medical School Hannover



COMPLEXITY
SCIENCE
HUB
VIENNA



VANDERBILT
UNIVERSITY

Radboud Universiteit



Universität
Rostock



Traditio et Innovatio



TECHNISCHE
UNIVERSITÄT
WIEN



UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386



The
University
Of
Sheffield.



Instituts
thématiques

Inserm

Institut national
de la santé et de la recherche médicale



MEDICAL UNIVERSITY
OF VIENNA

UNIVERSITY OF LEEDS

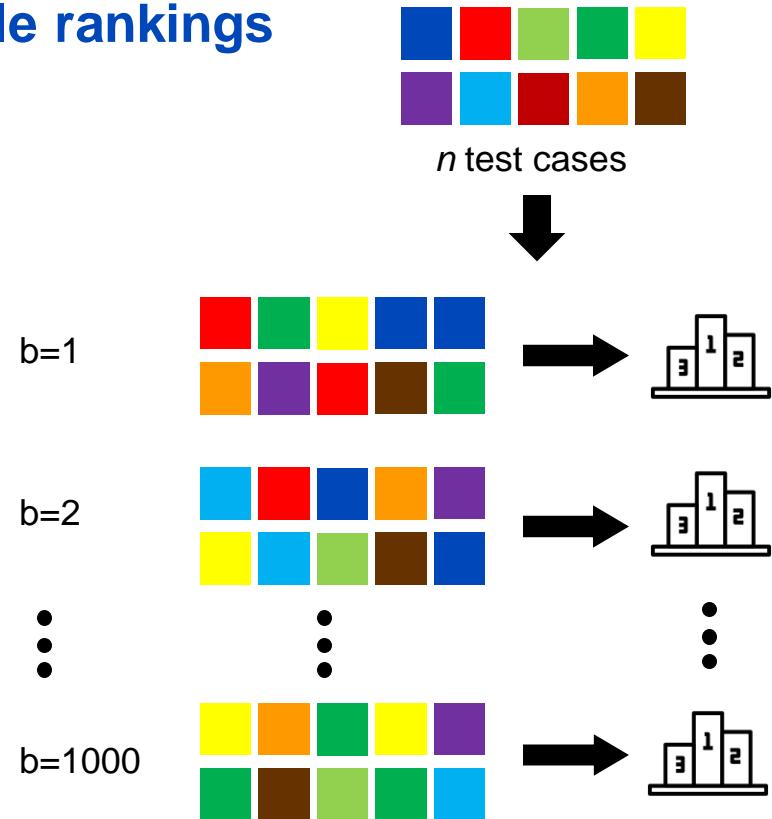
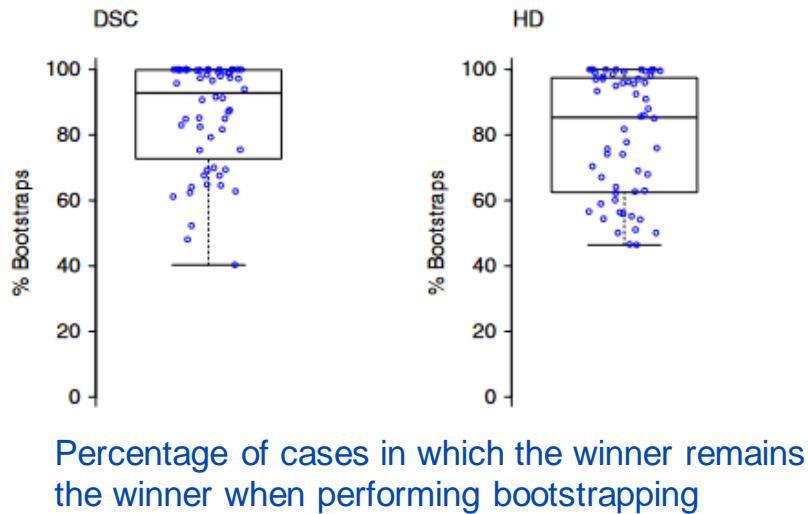




#Data (quantity)

Pitfall: Data sparsity leads to instable rankings

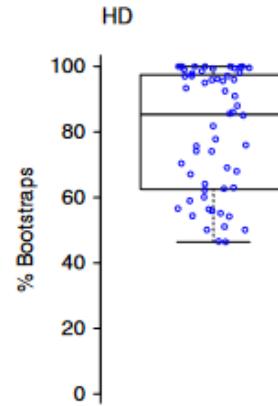
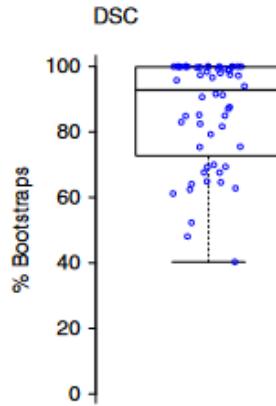
- **Instability of rankings** (analyzed over 56 biomedical image analysis challenges)
 -



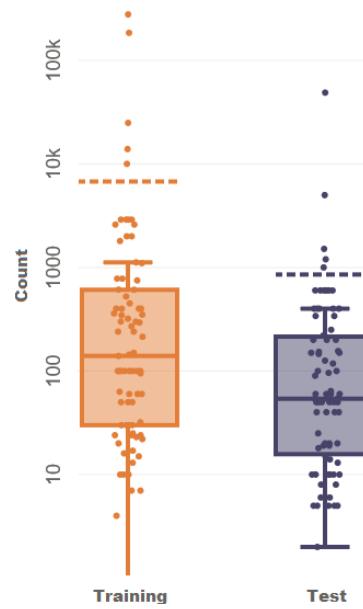
Maier-Hein et al. Why rankings of biomedical image analysis competitions should be interpreted with care *Nature Commun.* 2018

Pitfall: Data sparsity leads to instable rankings

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



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



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

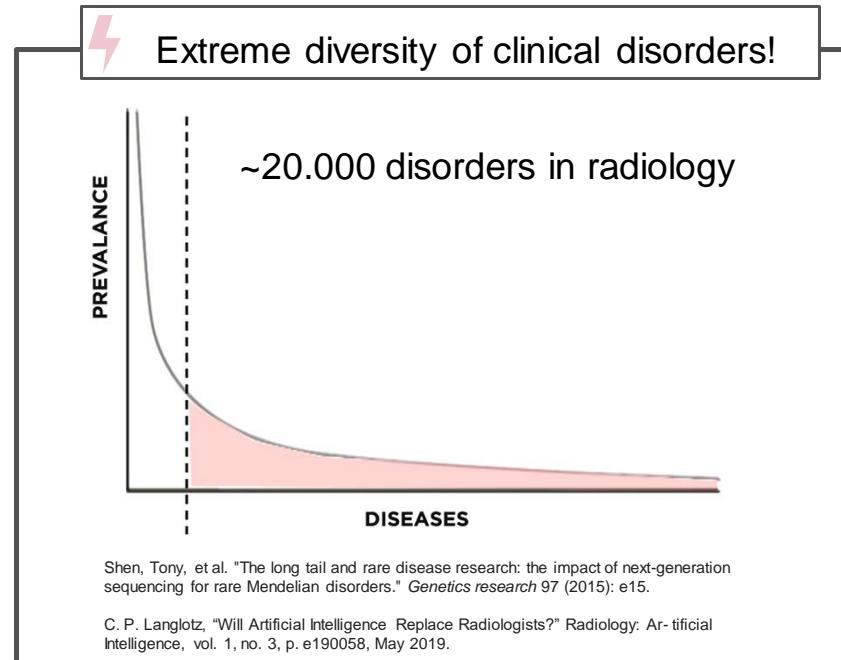


Maier-Hein et al. Why rankings of biomedical image analysis competitions should be interpreted with care *Nature Commun.* 2018

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.





#DataSplitting

Data leakage

Field	Paper	Number of papers reviewed	Number of papers with pitfalls	[I1.1] No test set	[I1.2] Pre-proc. on train-test	[I1.3] Feature sel. on train-test	[I1.4] Duplicates	[L2] Illegitimate features	[L3.1] Temporal leakage	[L3.2] Non-ind. b/w train-test	[L3.3] Sampling bias	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		○						○			

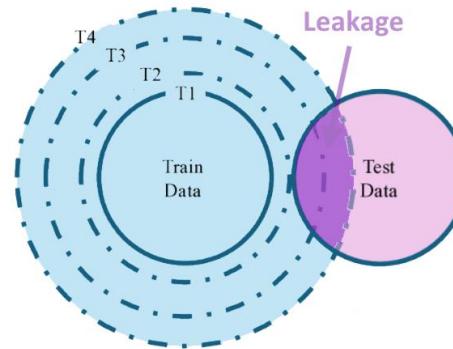


Kapoor & Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. Patterns 2023

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



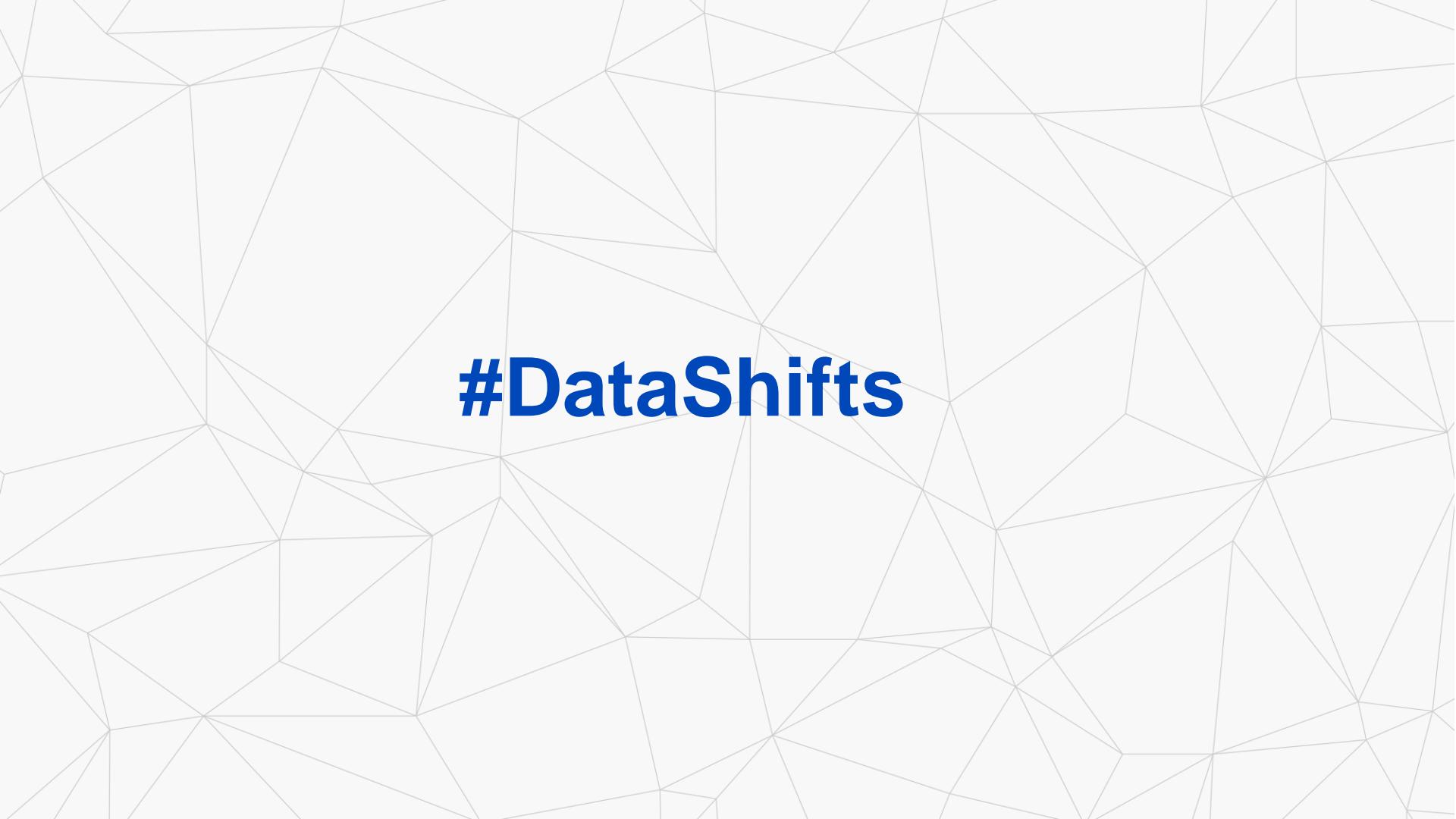
FM trained on older train sets crawled from the internet (T1, T2) experience test set contamination over time (T3, T4), as static test sets and public benchmarks are unintentionally included in the training, leading to unreliable performance reporting.
Figure adapted from [2].



[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



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



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

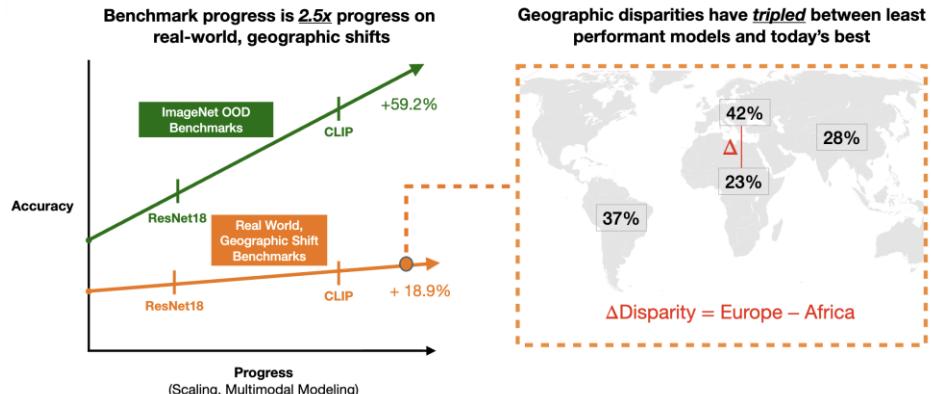
Anna Iurchenko
Google Health

Lauren Wilcox
Google Health

Paisan Ruamviboonsuk
Rajavithi Hospital

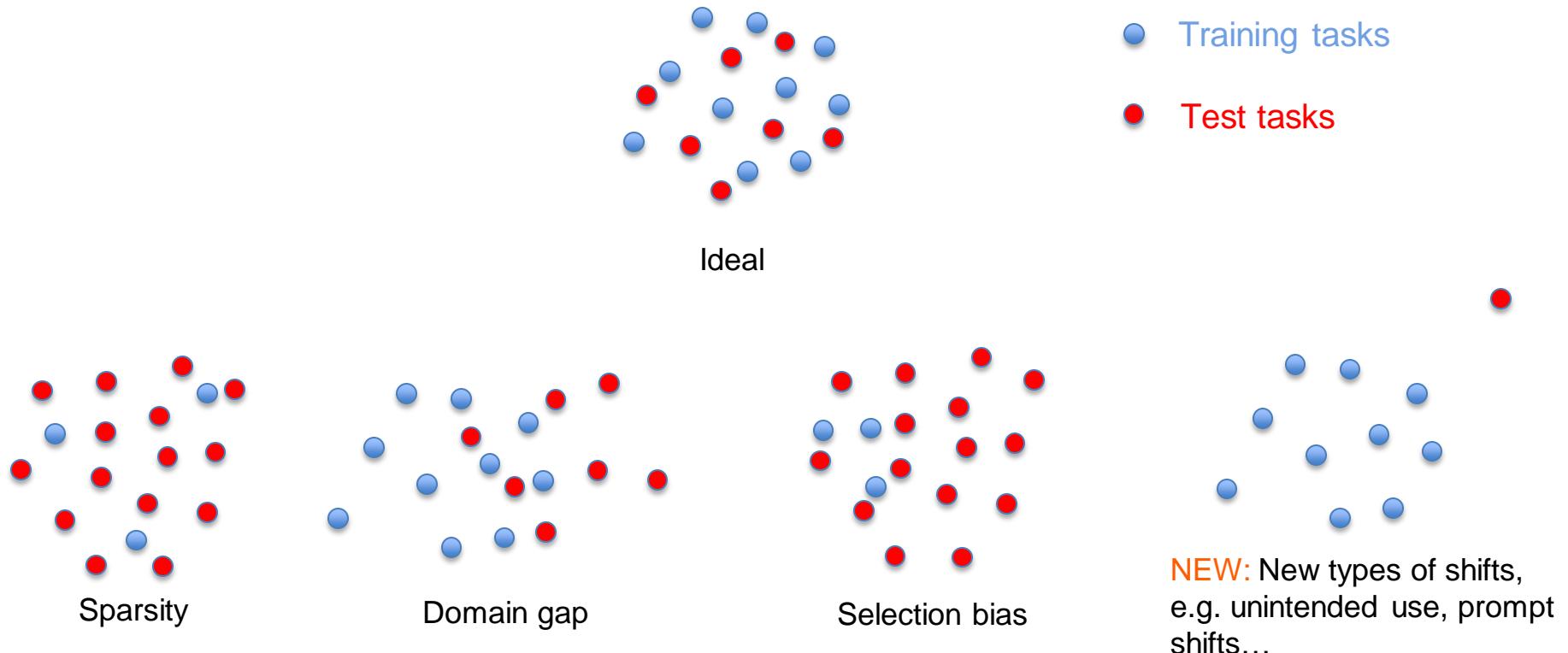
Imaging AI in the wild

- **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



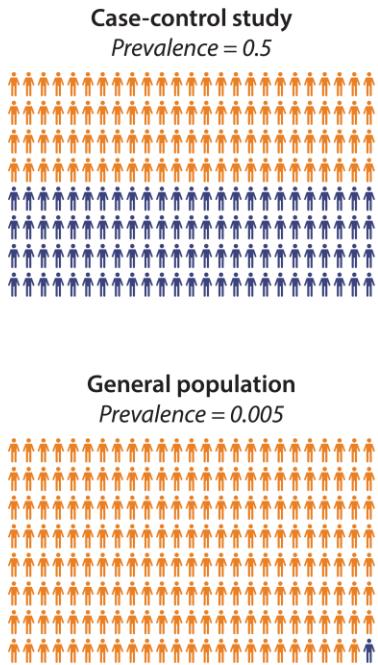


#Metrics

Example II: Prevalence shifts - a real-world challenge



↓
Prevalence shift



Class 1 Class 2



Patrick Godau

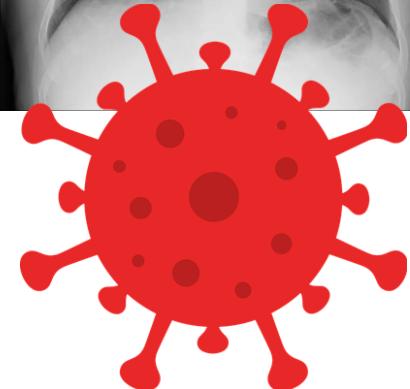
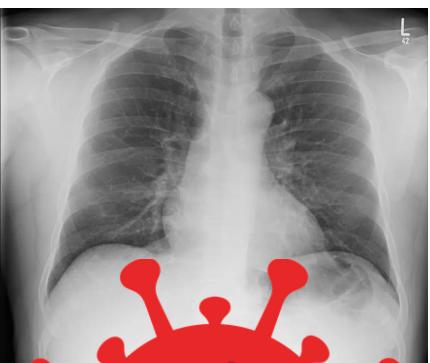


Piotr Kalinowski



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	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)

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

$$\begin{aligned}\text{Error Rate} &= \frac{\text{FN} + \text{FP}}{N} = \frac{\text{FN}}{\text{TP} + \text{FN}} \cdot \frac{\text{TP} + \text{FN}}{N} + \frac{\text{FP}}{\text{FP} + \text{TN}} \cdot \frac{\text{FP} + \text{TN}}{N} \\ &\quad R_{12} \quad \text{Prevalence} \\ &= R_{12} \cdot \text{Prevalence} + R_{21} \cdot (1 - \text{Prevalence})\end{aligned}$$



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

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

binary

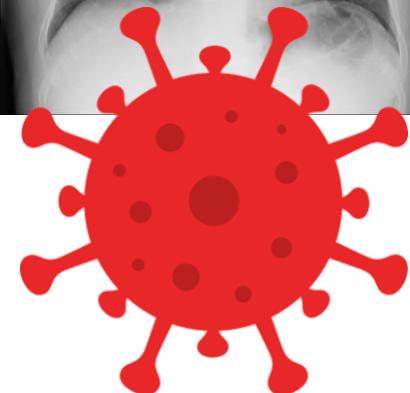
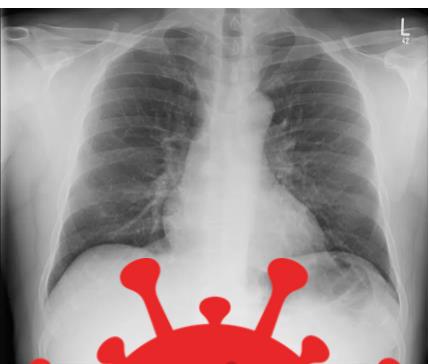
$$c_{11} = c_{22} = 0 \quad = \text{Prevalence} \cdot (c_{11}R_{11} + c_{12}R_{12}) + (1 - \text{Prevalence}) \cdot (c_{21}R_{21} + c_{22}R_{22})$$

$$c_{12} = c_{21} = 1$$

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

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

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	N
ACTUAL	P	3600	400
	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	N
ACTUAL	P	72	8
	N	2375	5545

Accuracy = 0.70

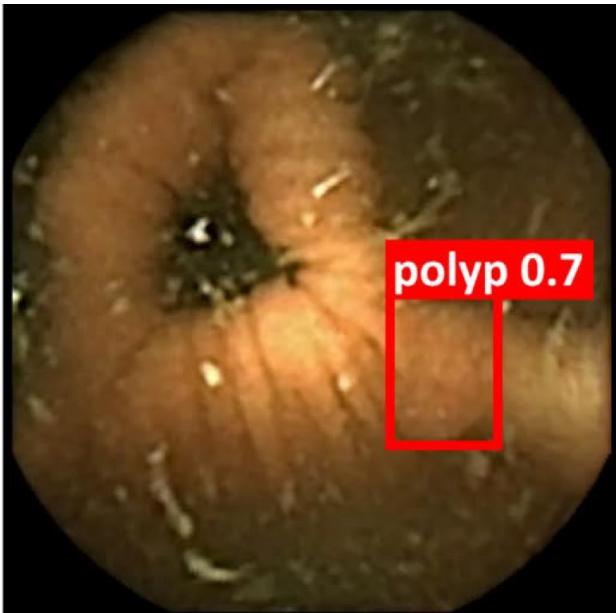
F1 Score = 0.06

EC* = 0.30

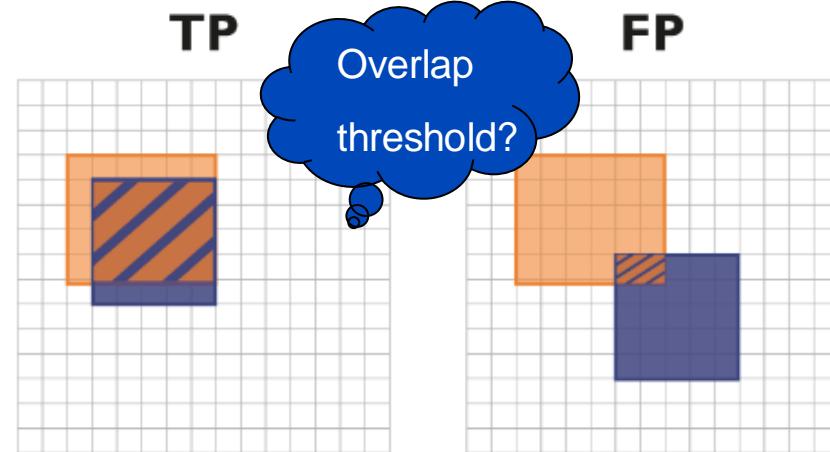
* Prevalence-corrected



Example III: Metric configuration matters as well



Amine Yamlahi Nuong Tran



How to define a TP/FP?

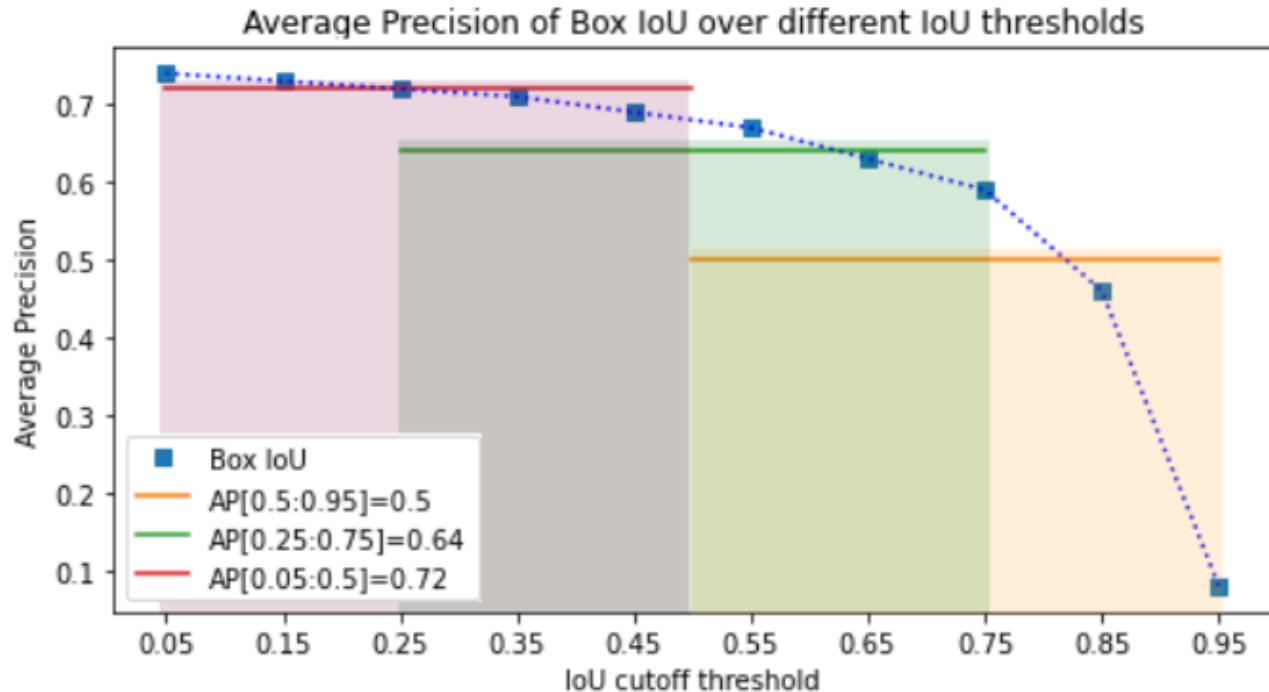


Reinke/Tizabi, ..., Jäger/Maier-Hein. Understanding metric-related pitfalls in image analysis validation. **Nature Methods 2024**

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

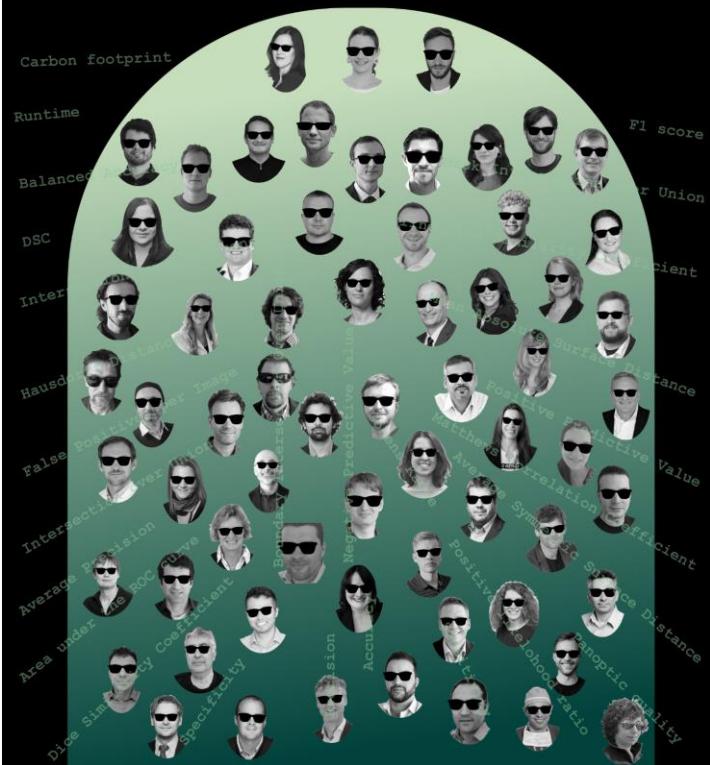
Standard solution does not reflect the clinical need!

Clinicians' preference
ML community



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

METRICS RELOADED



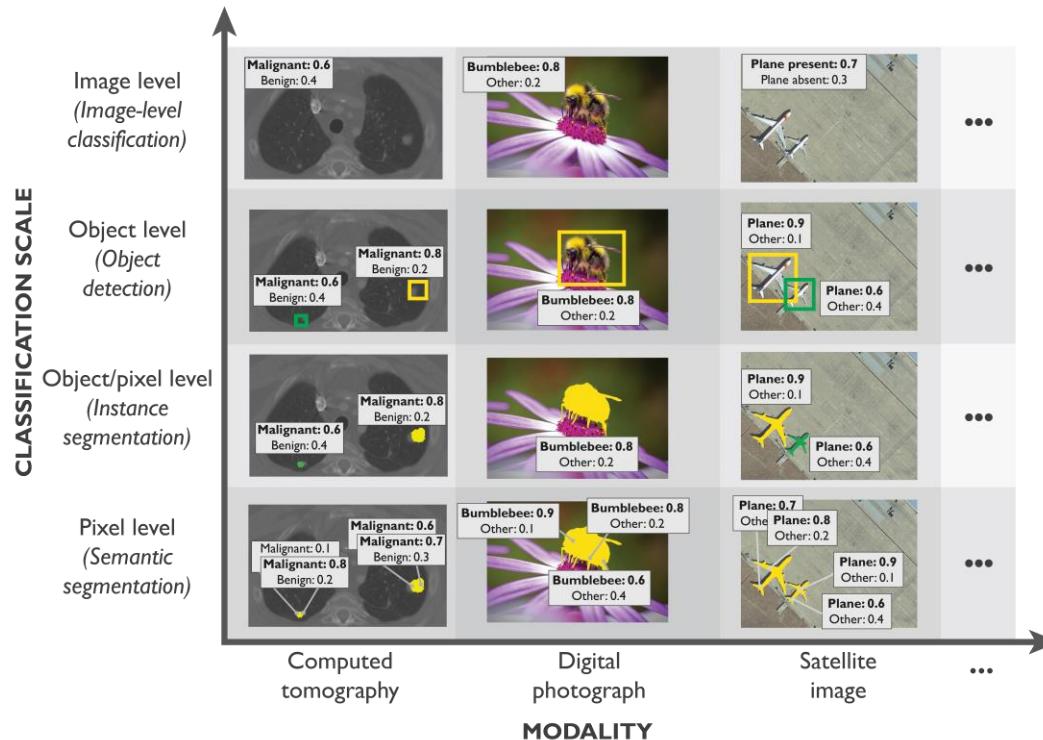
Rita Strack,
Nature Methods
Editor

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



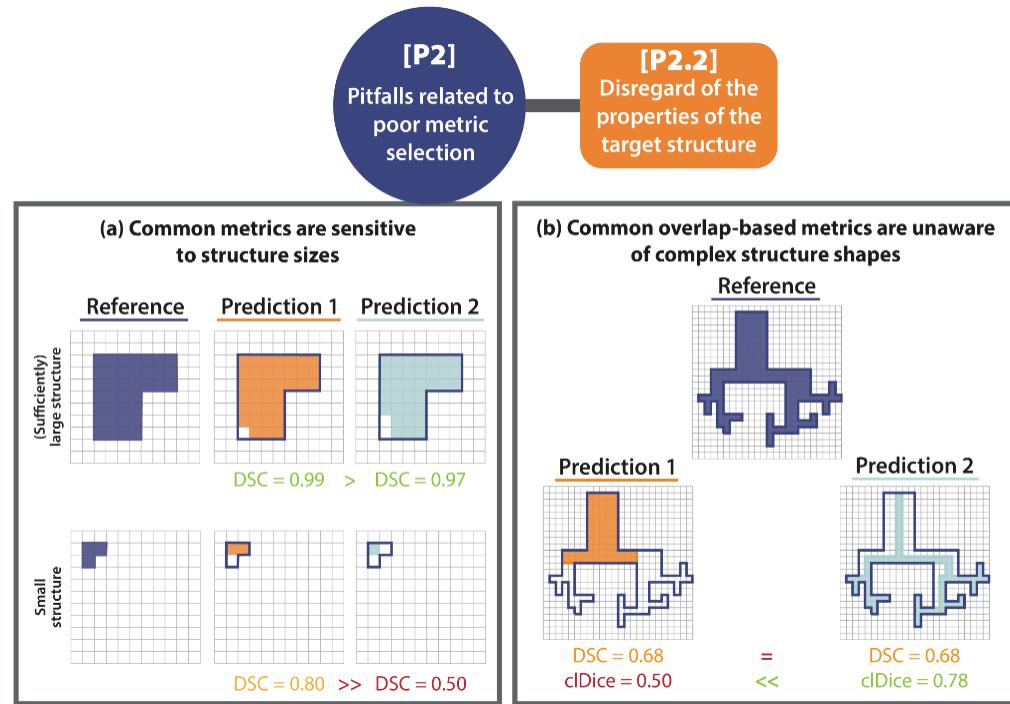
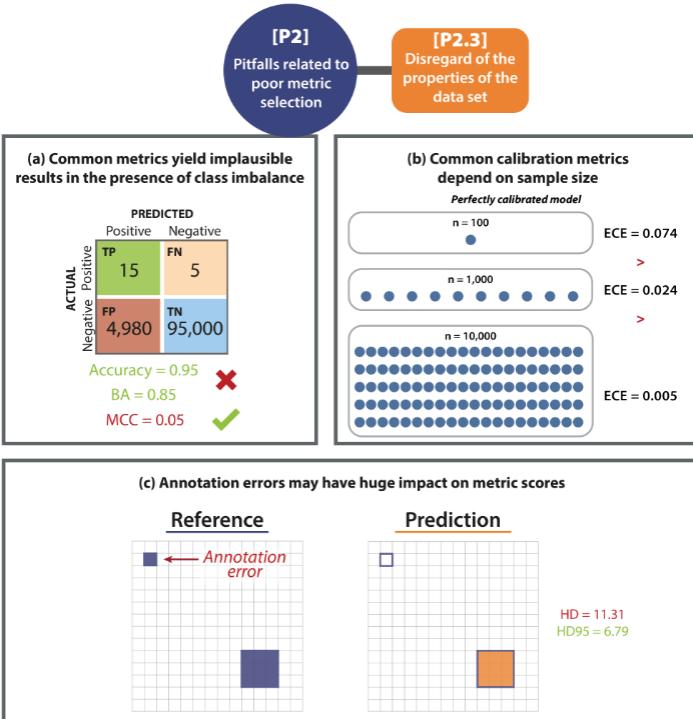
Maier-Hein/Reinke et al. Metrics reloaded: Recommendations for image analysis validation. **Nature Methods 2024**
Reinke/Tizabi, ..., Maier-Hein. Understanding metric-related pitfalls in image analysis validation. **Nature Methods 2024**

Scope of (first set of) recommendations



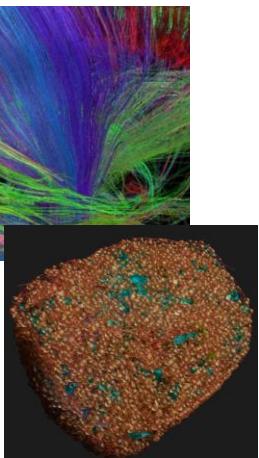
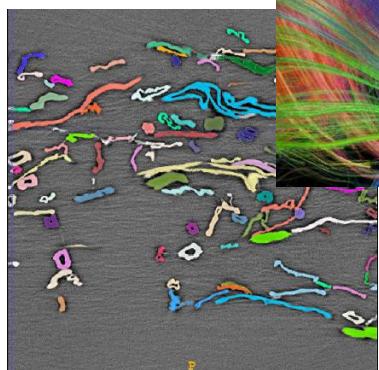
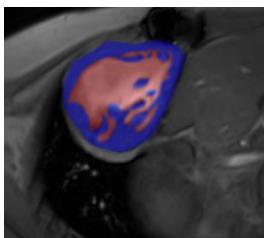
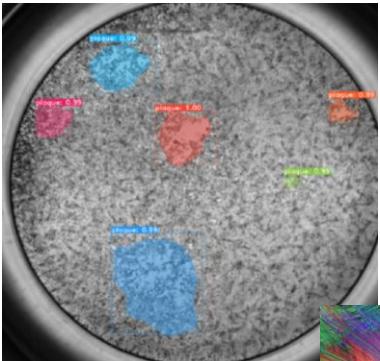
Maier-Hein/Reinke et al. Metrics reloaded: Recommendations for image analysis validation. *Nature Methods* 2024

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



Abstraction via
fingerprinting

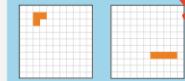


Annika Reinke

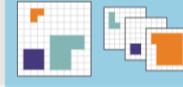


Minu Tizabi

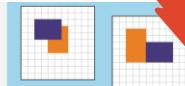
Small size of structures
relative to pixel size



High variability of structure
sizes

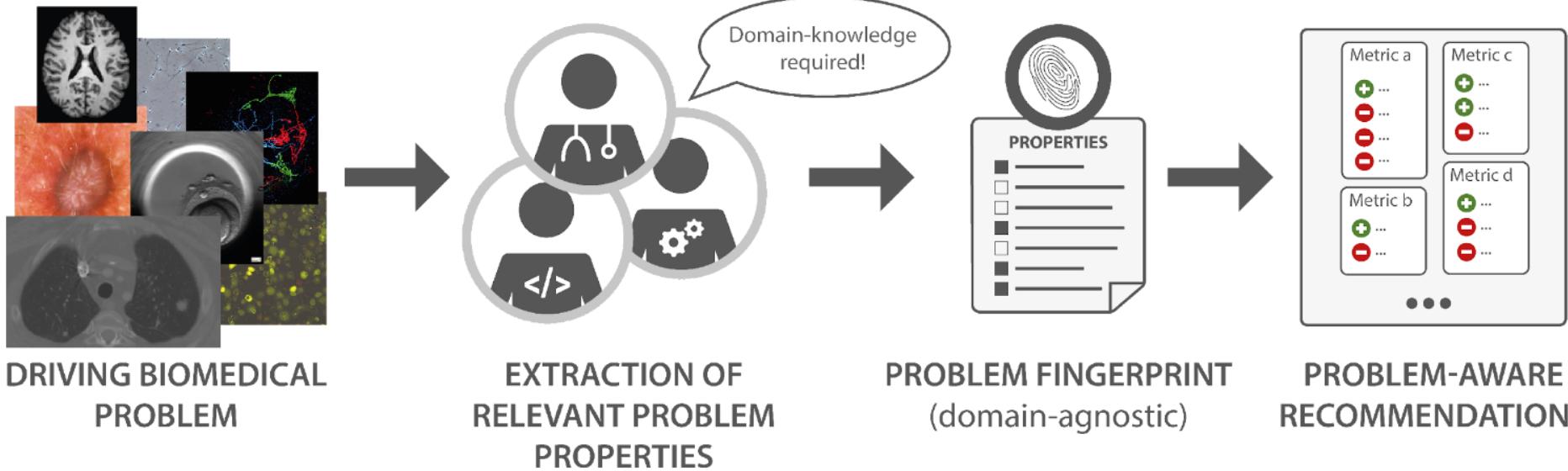


Possibility of overlapping or
touching target structures

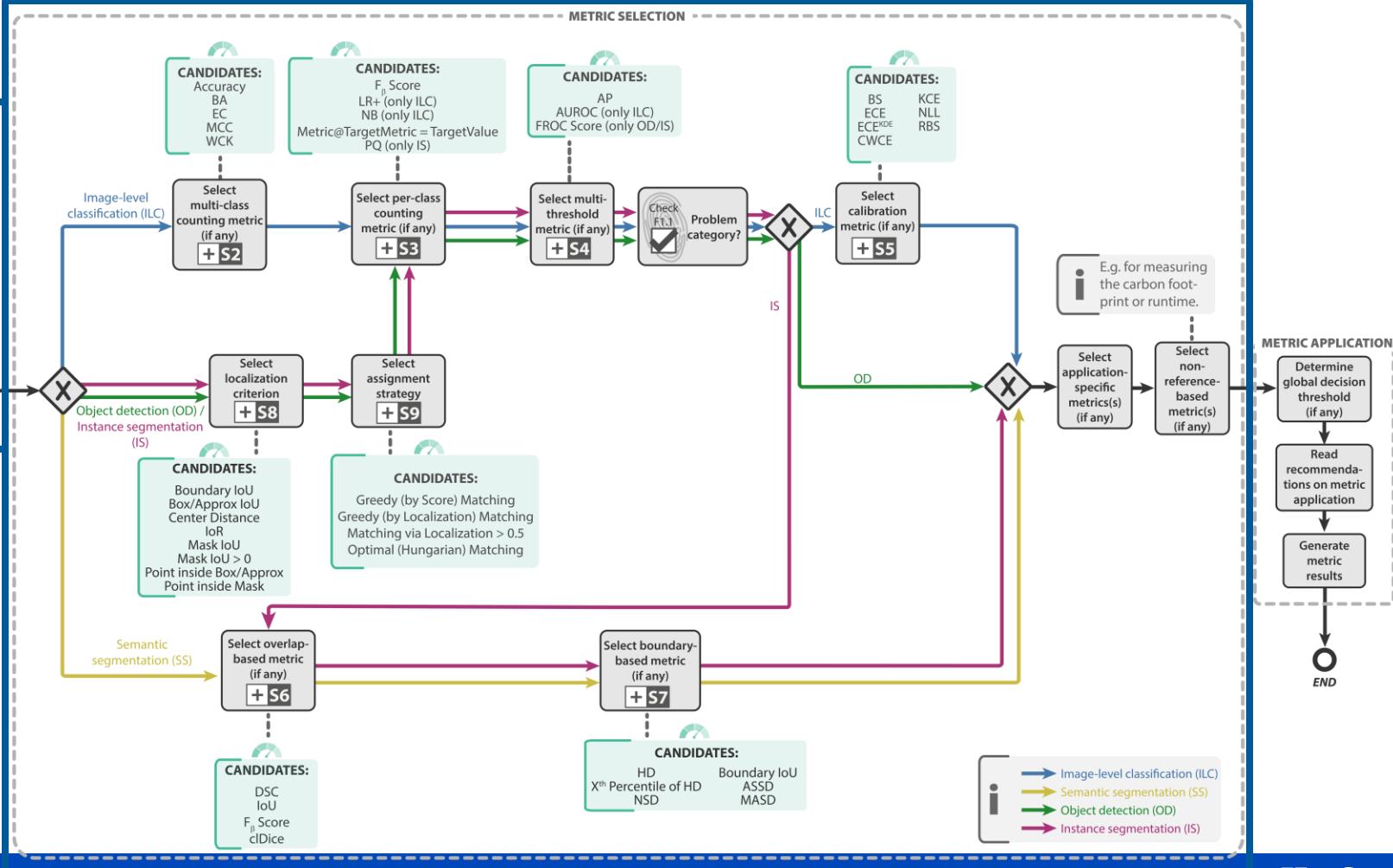
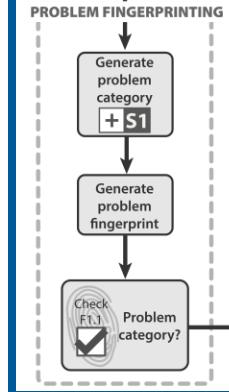


Maier-Hein/Reinke et al. Metrics reloaded: Recommendations for image analysis validation. *Nature Methods* 2024

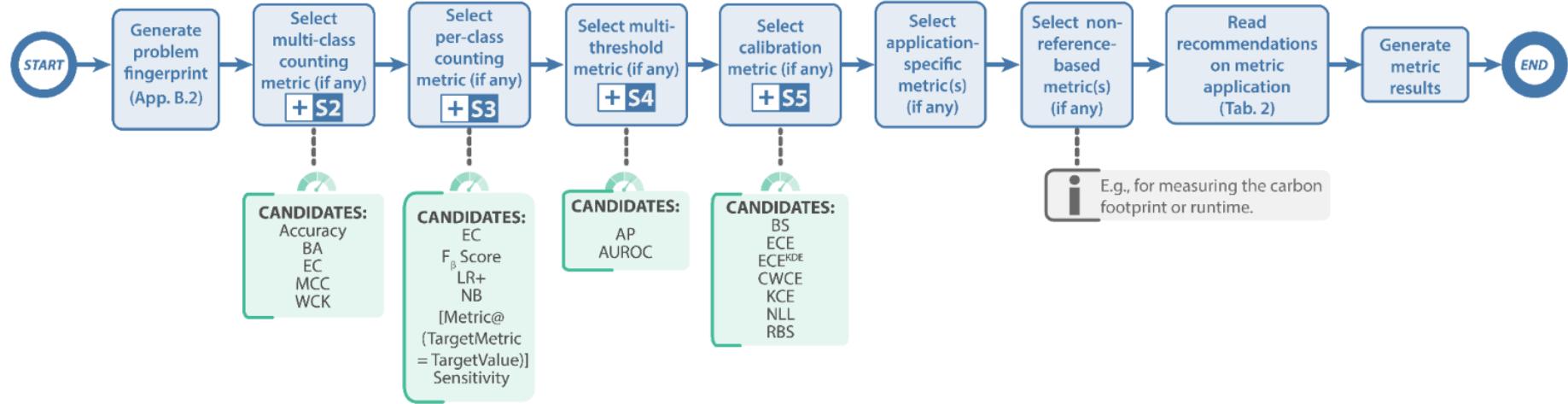
Metrics Reloaded approach



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



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

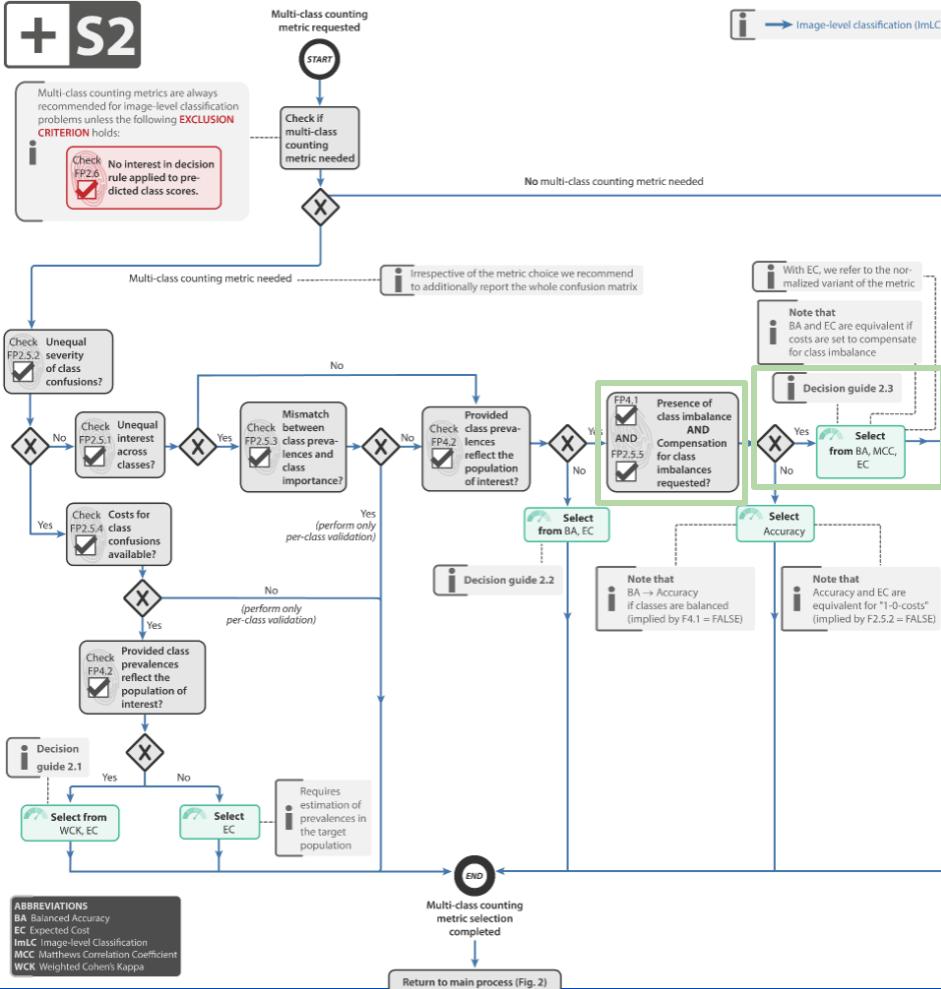
NB Net Benefit

NLL Negative log likelihood

RBS Root Brier Score

WCK Weighted Cohen's Kappa

Multi-class counting metrics

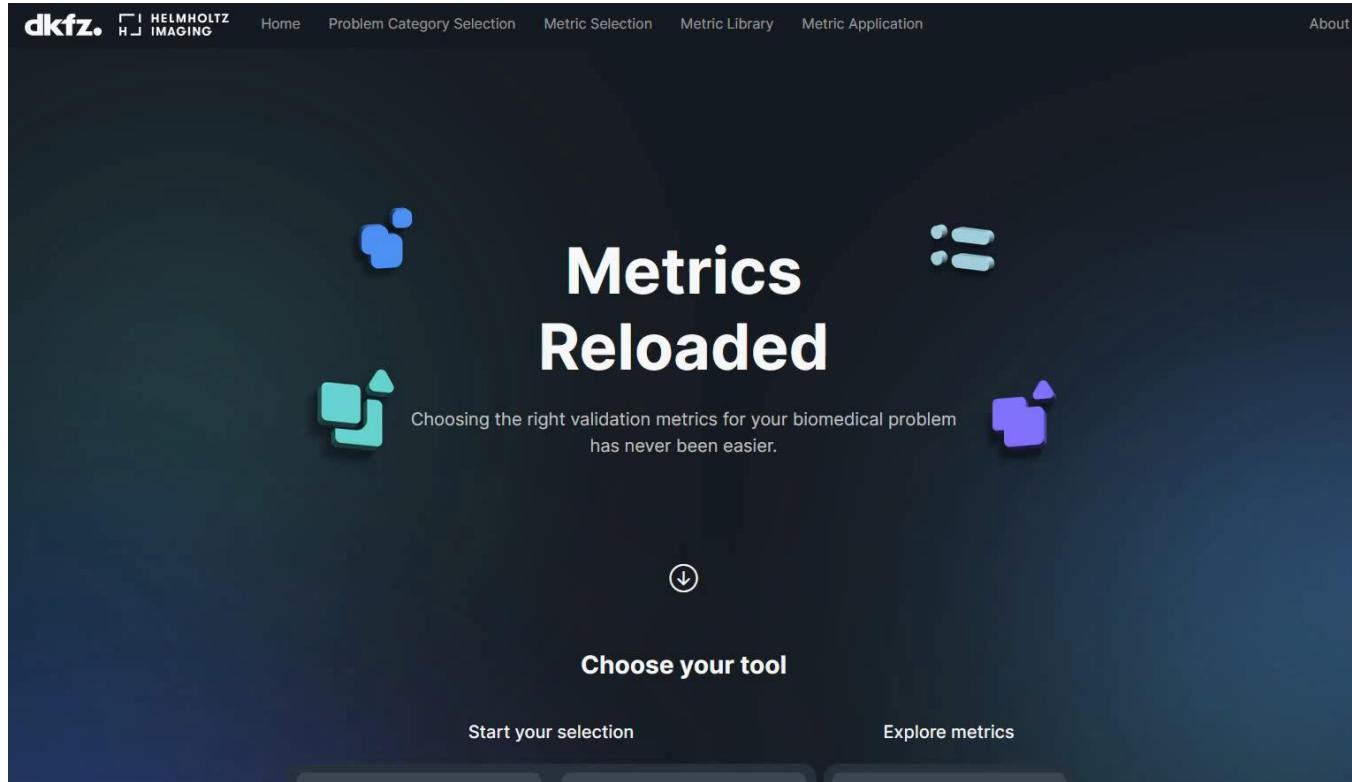


Decision guides

Summary of DG2.3: BA versus MCC versus ECN		
BA	MCC	ECN
<ul style="list-style-type: none">⊕ Inherent interpretability with respect to naive classifier⊖ Implication of equal class contribution⊖ Insensitive to predictive values (Positive Predictive Value (PPV) and Negative Predictive Value (NPV))⊕ Availability of framework to identify and validate decision rule applied to class scores⊕ Good interpretability⊕ Widely used	<ul style="list-style-type: none">⊕ Inherent interpretability with respect to naive classifier⊖ Implication of equal class contribution⊖ High scores ensure high predictive values (PPV and NPV)⊖ Lack of framework to identify and validate the decision rule applied to class scores⊖ Limited interpretability⊖ Fairly well-known but not much used	<ul style="list-style-type: none">⊕ Inherent interpretability with respect to naive classifier⊖ No establishment of equal class contribution⊖ Limited sensitivity to predictive values (PPV and NPV)⊕ Availability of framework to identify and validate the decision rule applied to class scores⊕ Good interpretability⊖ 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 ($FP2.5.2 = \text{FALSE}$), either (1) unequal interest across classes ($FP2.5.1 = \text{TRUE}$) and no mismatch between class prevalences and class importance ($FP2.5.3 = \text{FALSE}$) or (2) equal interest across classes ($FP2.5.1 = \text{FALSE}$), provided class prevalences reflect the population of interest ($FP4.2 = \text{TRUE}$), presence of class imbalance ($FP4.1 = \text{TRUE}$) and compensation for class imbalances requested ($FP2.5.5 = \text{TRUE}$).

Making it accessible to the community: Metrics Reloaded toolkit



The screenshot shows the homepage of the Metrics Reloaded toolkit. At the top, there is a navigation bar with the dkfz. logo, the text "HELMHOLTZ IMAGING", and links for Home, Problem Category Selection, Metric Selection, Metric Library, Metric Application, and About. The main title "Metrics Reloaded" is prominently displayed in the center, with a subtitle below it: "Choosing the right validation metrics for your biomedical problem has never been easier." Below the title, there is a large downward arrow icon. At the bottom of the page, there are two buttons: "Start your selection" and "Explore metrics".



Emre Kavur

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



Instantiation for common biomedical use cases

DESCRIPTION OF PROBLEM	SCENARIO	SAMPLE INPUT IMAGE	RECOMMENDED OUTPUT IMAGE	RECOMMENDATION
Classification of images	Frame-based sperm motility classification based on microscopy time-lapse video containing human spermatozoa		 Progressive motility: 0.5 Non-progressive motility: 0.4 Immotile: 0.1	Problem category: Image-level classification Multi-class counting metric (S2): Balanced Accuracy (BA)
	Disease classification in dermoscopic images		 Dermatofibroma: 0.6 Seborrheic keratosis: 0.2 Mole: 0.1 Basal cell carcinoma: 0.0 Actinic keratosis: 0.0 Vascular lesion: 0.1	Multi-threshold metric (S3): Area under the Receiver Operating Characteristic Curve (AUROC) Output calibration: Expected Calibration Error (ECE) Per-class counting metric (S4): Positive Likelihood Ratio (LR+)
Segmentation of large objects	Lung cancer cell segmentation from microscopy images			Problem category: Semantic segmentation Overlap-based metric (S5): Dice Similarity Coefficient (DSC)
	Liver segmentation in computed tomography (CT) images			Boundary-based metric (S6): Normalized Surface Distance (NSD) Specific property-related metric: Liver segmentation: Absolute Volume Difference
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)
	MS Lesion detection in multi-modal brain MRI images			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 (FPI)@Sensitivity
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
	Surgical instrument instance segmentation in colonoscopy videos			Multi-threshold metric (S3): AP Per-class counting metric (S4): F_1 Score Overlap-based metric (S5): Center line Dice Similarity Coefficient (cIDice) Boundary-based metric (S6): NSD

Making it accessible to the community: Metric cheat sheets

ACCURACY

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

VALUE RANGE: [0, 1] ↑

DESCRIPTION
Accuracy measures the ratio of samples that were correctly classified over all predictions made.

IMPORTANT RELATIONS
Accuracy can be rewritten as

- In binary situations: Accuracy = [Sensitivity + Prevalence + Specificity - (1 - Prevalence)]
- BA = Accuracy, if classes are balanced
- EC = 1 - Accuracy for EC instantiated with 0-1 costs
- ER = 1 - Accuracy
- (W/C)K = 2 * Accuracy - 1, if classes are balanced (and using 0-1 costs)
- BM = 2 * Accuracy - 1, if classes are balanced

PREVALENCE DEPENDENCY ●

TP	CARDINALITIES	FP	FN	TN
●	ImLC	●	●	●
●	SemS	○	○	○
●	ObD	○	○	○
●	InS	○	○	○

RELEVANT PITFALLS

- Accuracy is highly sensitive to class imbalance (Figs. 7a, 24 in [Reinke et al., 2023]).
- Accuracy is prevalence-dependent, thus not comparable across data sets with different prevalences (Fig. 16 in [Reinke et al., 2023]).
- Accuracy does not allow for incorporating unequal severity of confusions across classes. This implies that the metric is not applicable to cost-benefit analyses (Fig. 18 in [Reinke et al., 2023]), or when target classes are related on an ordinal scale (Fig. 5b in [Reinke et al., 2023]).
- Accuracy depends on the definition of TN (undefined for ObD and InS).

RECOMMENDATIONS

- Accuracy should generally not be considered...
 - ... in the presence of class imbalance unless class prevalences reflect the interest across classes.
 - ... if comparison of performance across data sets with different prevalences is desired.
 - ... if class confusions are of unequal severity (examples: ordinal target classes, cost-benefit analysis).
- Due to ease of interpretation and popularity, we specifically recommend it as a multi-class metric if the compensation of class imbalances is not desired.

BALANCED ACCURACY (BA)

$BA = \frac{1}{2} (\text{Sensitivity} + \text{Specificity}) = \frac{1}{2} \left(\frac{\text{green}}{\text{green} + \text{orange}} + \frac{\text{blue}}{\text{blue} + \text{red}} \right)$

VALUE RANGE: [0, 1] ↑

DESCRIPTION
BA measures the arithmetic mean of Sensitivities for each class, i.e., for each class, it measures the fraction of actual positive samples that were predicted as such.

DEFINITION
[Tharwat, 2020]

IMPORTANT RELATIONS

- $J = 2BA - 1$
- $(W/C)K = 2BA - 1$, if classes are balanced (and using 0-1-costs)
- Accuracy = BA, if classes are balanced
- EC = 1 - BA, if EC costs are chosen such that $w_{ij} = 0$ and $w_{ii} = 1/w_{ij}$, where w_{ij} are the costs for a sample of actual class i that was predicted as class j . C is number of classes and P_i is prevalence of class i .

PREVALENCE DEPENDENCY ○

MULTI-CLASS DEFINITION
For C classes, BA is defined as the arithmetic mean of Sensitivities per class:

$$BA = \frac{1}{C} \sum_{i=1}^C \text{Sensitivity}_i = \frac{1}{C} \sum_{i=1}^C \frac{n_{ii}}{n_{..}}$$

$n_{..}$: diagonal entries of the confusion matrix; sum equals number of correctly classified samples
 $n_{..}$: sum of entries of row i in the confusion matrix
 N : total number of samples

TP	CARDINALITIES	FP	FN	TN
●	ImLC	●	●	●
●	SemS	○	○	○
●	ObD	○	○	○
●	InS	○	○	○

RELEVANT PITFALLS

- BA can be misleading for imbalanced situations (Fig. 7a in [Reinke et al., 2023]).
- BA does not allow for incorporating unequal severity of confusions across classes. This implies that the metric is not applicable to cost-benefit analyses (Fig. 18) in [Reinke et al., 2023], or when target classes are related on an ordinal scale (Fig. 5b in [Reinke et al., 2023]).
- BA is not well-suited if an unequal treatment of classes is requested (e.g., some classes are treated as more important than others) [Grandini et al., 2020].
- BA is insensitive to changes in predictive values (PPV and NPV) [Maier-Hein et al., 2022].
- In binary tasks, BA may yield the same value for different Sensitivity and Specificity scores [Reinke et al., 2021].
- BA depends on the definition of TN (undefined for ObD and InS).

RECOMMENDATIONS

- BA should not be applied if...
 - ... there is unequal interest across classes.
 - ... predictive values should be assessed.
 - ... class confusions are of unequal severity (examples: ordinal target classes, cost-benefit analysis).
- Otherwise, it should specifically be considered...
 - ... in the presence of high class imbalance in case there is an equal interest across classes.
 - ... if a comparison of performance across data sets with different prevalences is desired.
- BA can be used to identify and validate the decision rule applied to predicted class scores.

MATTHEWS CORRELATION COEFFICIENT (MCC)

Synonyms: Phi Coefficient

$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} = \frac{\text{green} \cdot \text{blue} - \text{orange} \cdot \text{red}}{\sqrt{\text{green} \cdot \text{orange} \cdot \text{blue} \cdot \text{red}}}$

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]

PREVALENCE DEPENDENCY ●

TP	CARDINALITIES	FP	FN	TN
●	Counting metric	●	●	●
●	ImLC	○	○	○
●	SemS	○	○	○
●	ObD	○	○	○
●	InS	○	○	○

IMPORTANT RELATIONS
MCC can be rewritten as:

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

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

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

$$MCC = \frac{\sum_{i=1}^C \sum_{j=1}^C \sum_{k=1}^C n_{ik} \cdot n_{jk} - \sum_{i=1}^C n_{ii} \cdot \sum_{j=1}^C n_{jj}}{\sqrt{\sum_{i=1}^C (\sum_{j=1}^C n_{ij}) (\sum_{j=1}^C \sum_{k=1}^C n_{jk})} \cdot \sqrt{\sum_{j=1}^C (\sum_{i=1}^C n_{ij}) (\sum_{i=1}^C \sum_{k=1}^C n_{ik})}}$$

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. 16 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. 18 in [Reinke et al., 2023]), or when target classes are related on an ordinal scale (Fig. 5b in [Reinke et al., 2023]).
- The theoretical lower bound of MCC (-1) may not always be achievable (Fig. 44 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).

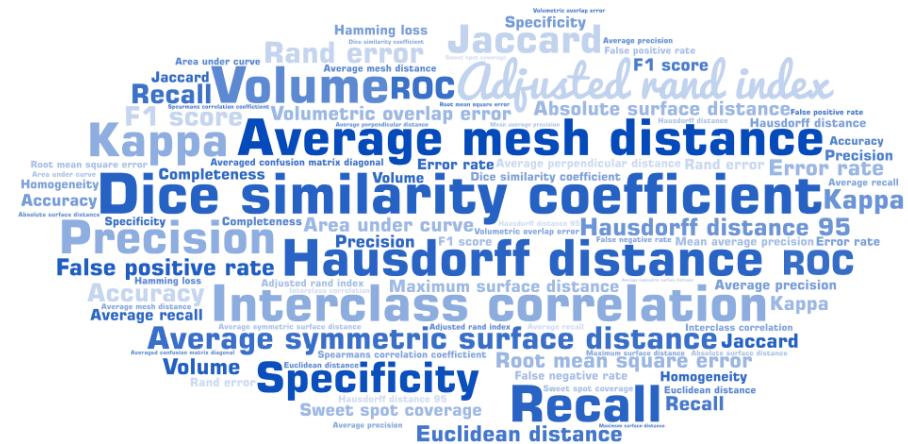
RECOMMENDATIONS

- 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.
 - ... computational 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].

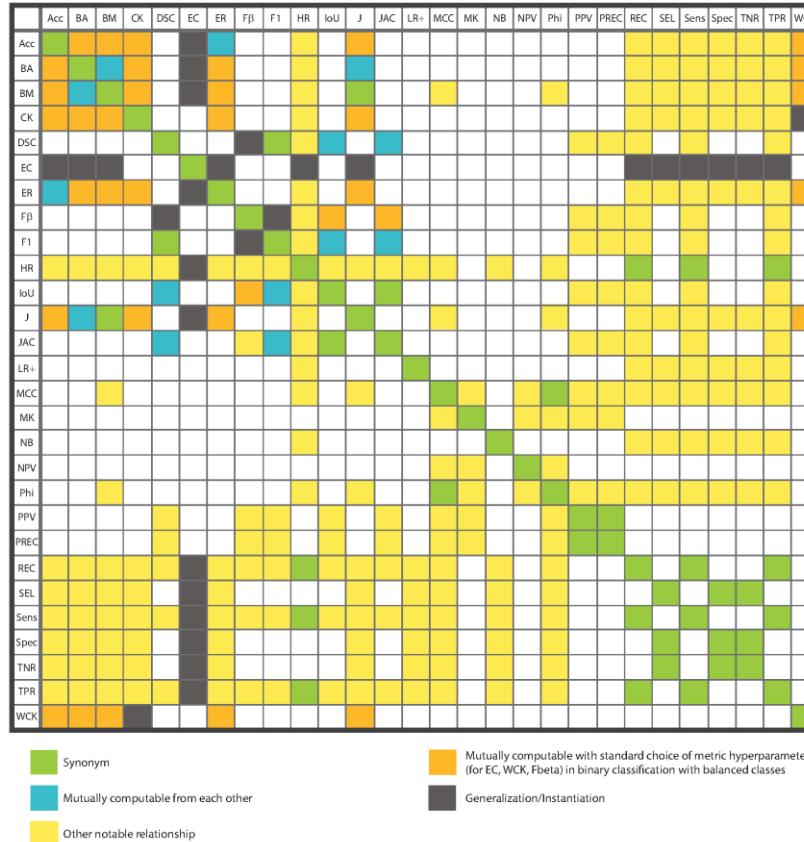
Anecdote...

Reviewer #3

7. Why have only DSC and NSD scores been used for result comparison? Other metrics (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
```

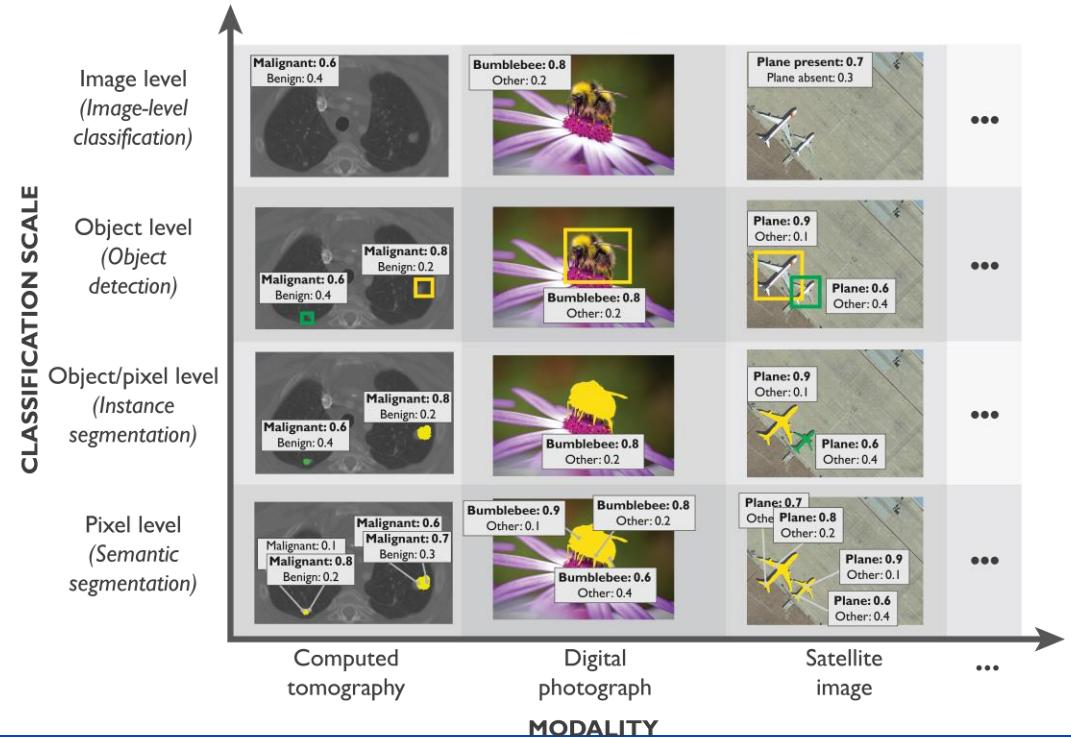
Clone the repository:

```
git clone https://github.com/c Sudre/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

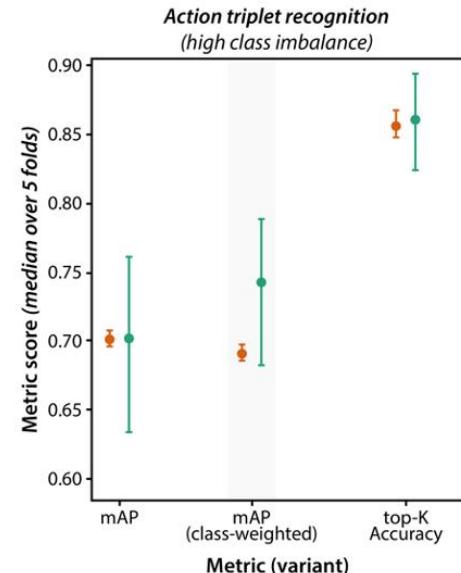
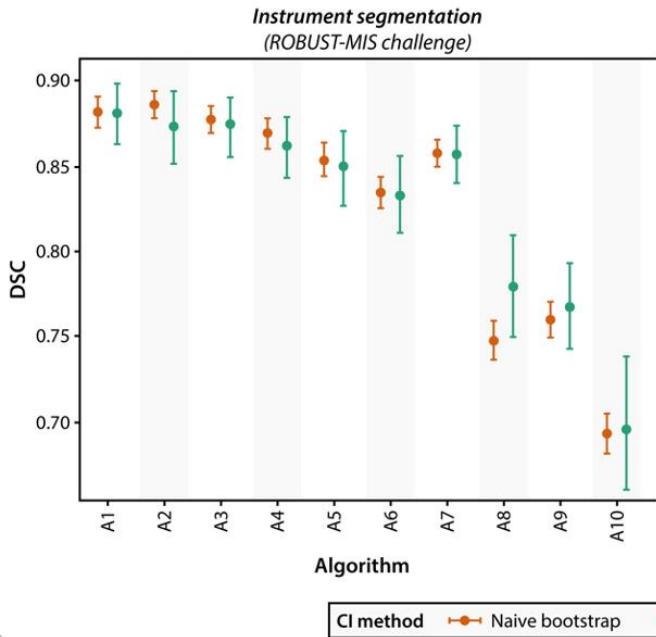
Work in progress

(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



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?

Question	Which bones are abnormal in this image?					
Ground Truth	ankle					
Prediction	medial malleolus,lateral malleolus, and posterior malleolus.					
Expected Rating (Human Scores)	 (4 4)					
Mistral Score	BLEU Score	Exact Match	F1 Score	Precision	Recall	
5 	0 	0 	0 	0 	0 	



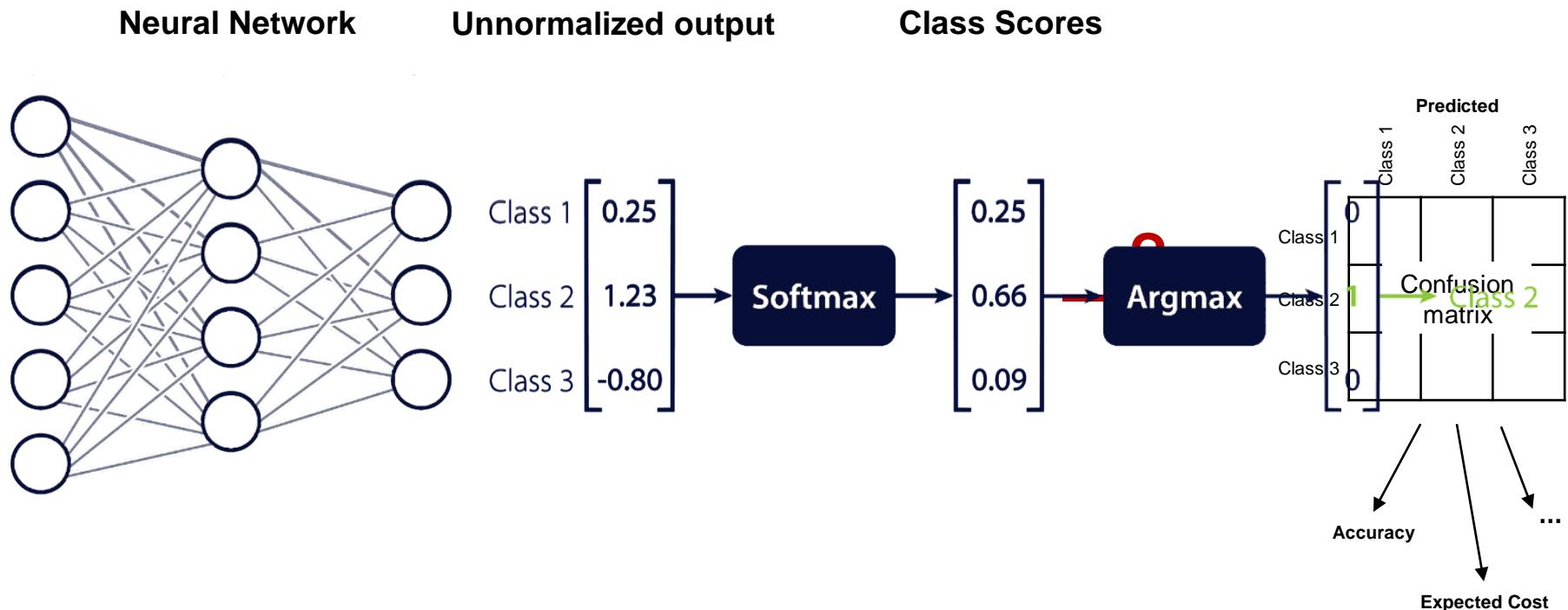
Kahl/Erkan, Maier-Hein, Jäger: SURE-VQA: Systematic understanding of robustness evaluation in medical VQA tasks, [arXiv 2024](#)

Abbasian et al. Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative AI, [npj Digital Medicine 2024](#)



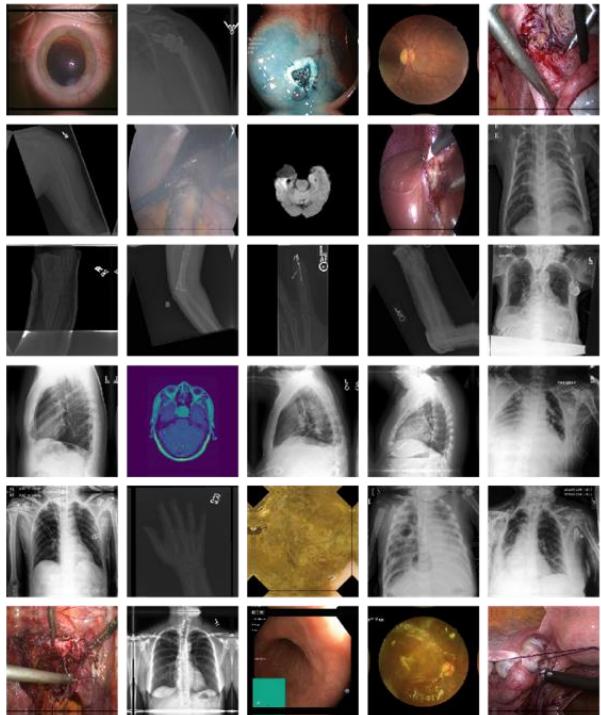
#Decision Threshold

How to convert predicted class scores to classes?



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

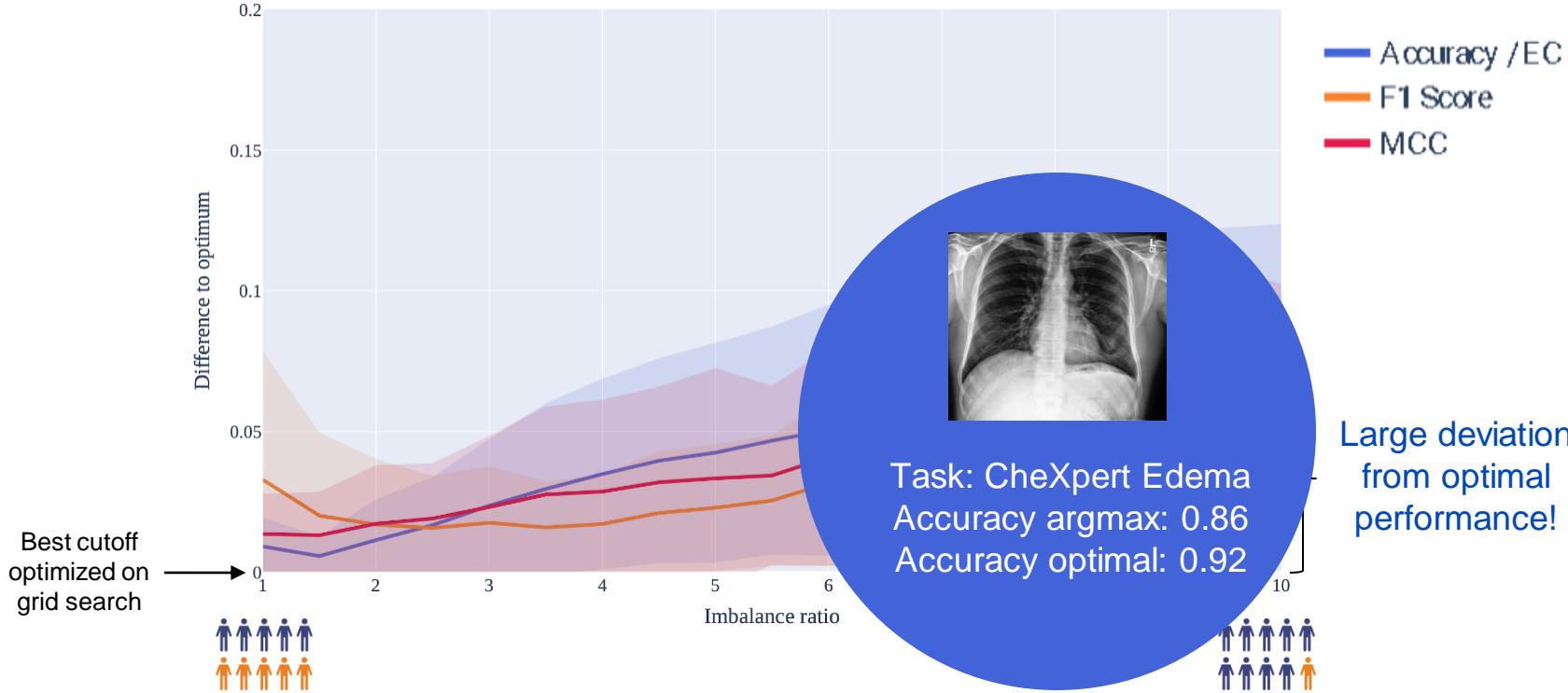
**Assumption for this talk: Deployment prevalences
(approximately) known**

Paper shows: Robustness to uncertainty



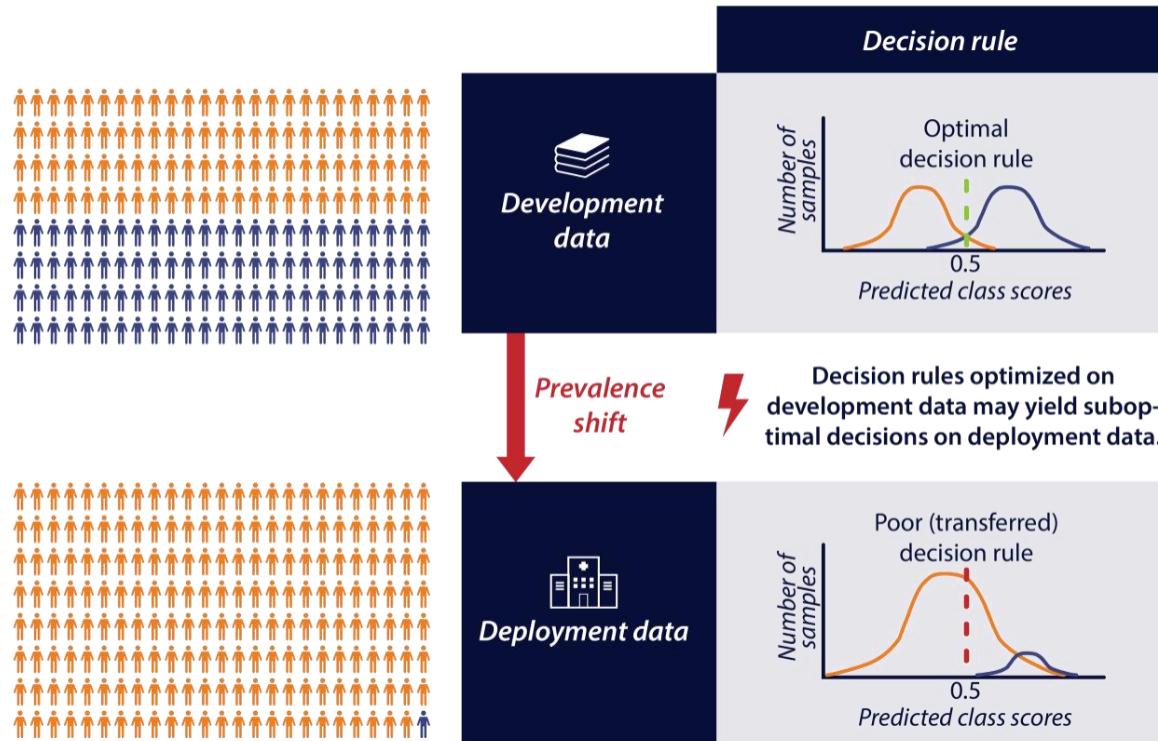
Godau/Kalinowski, ..., Maier-Hein. Navigating prevalence shifts in image analysis algorithm deployment, **MICCAI 2023**
Medical Image Analysis MICCAI 2023 Best Paper Award

Argmax – what's going on?



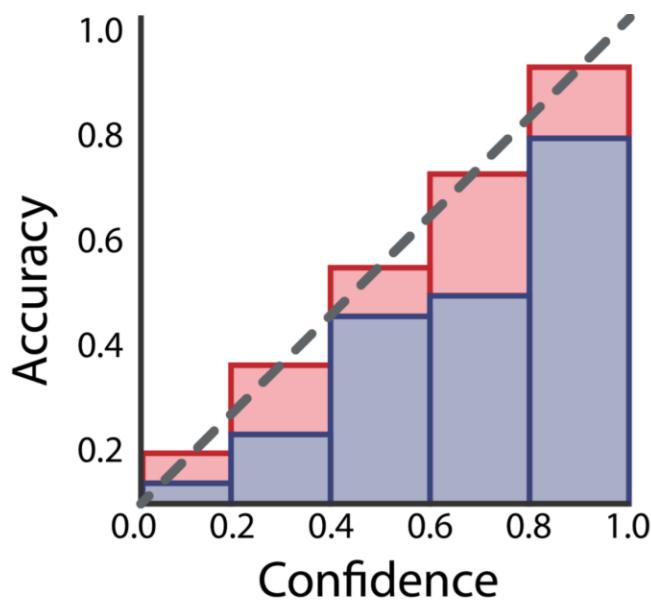
Godau/Kalinowski, ..., Maier-Hein. Navigating prevalence shifts in image analysis algorithm deployment, **MICCAI 2023**
Medical Image Analysis MICCAI 2023 Best Paper Award

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 {

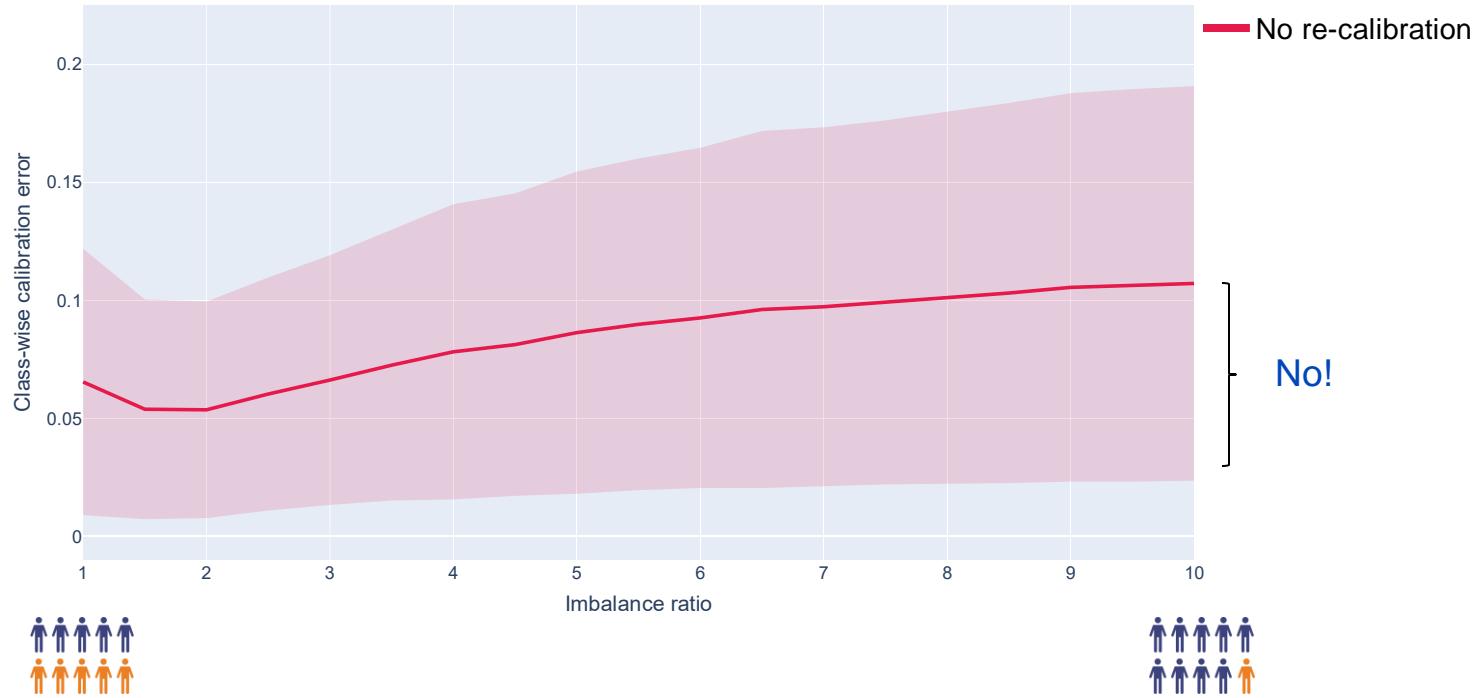


VALUE RANGE: [0, 1] ↑



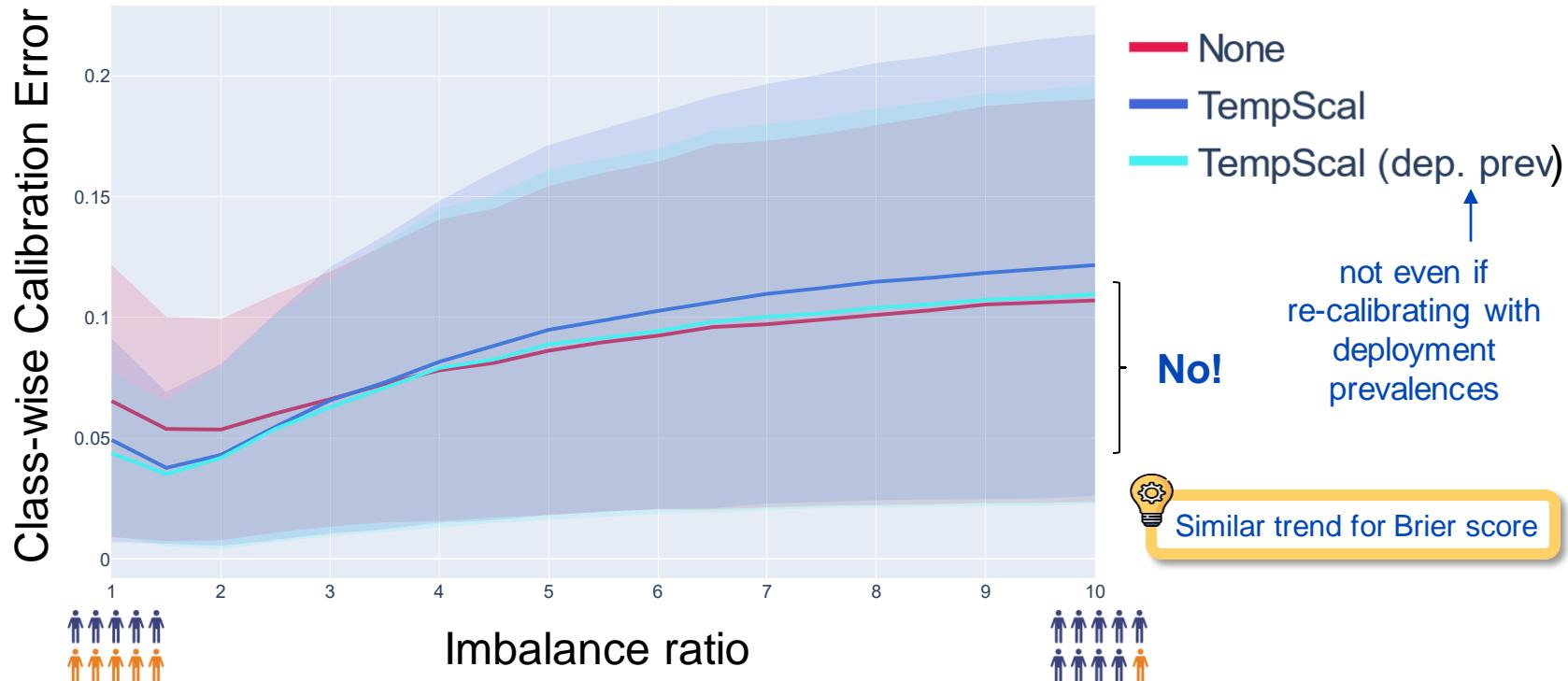
Godau/Kalinowski, ..., Maier-Hein. Navigating prevalence shifts in image analysis algorithm deployment, **MICCAI 2023**
Medical Image Analysis MICCAI 2023 Best Paper Award

Are the scores calibrated?



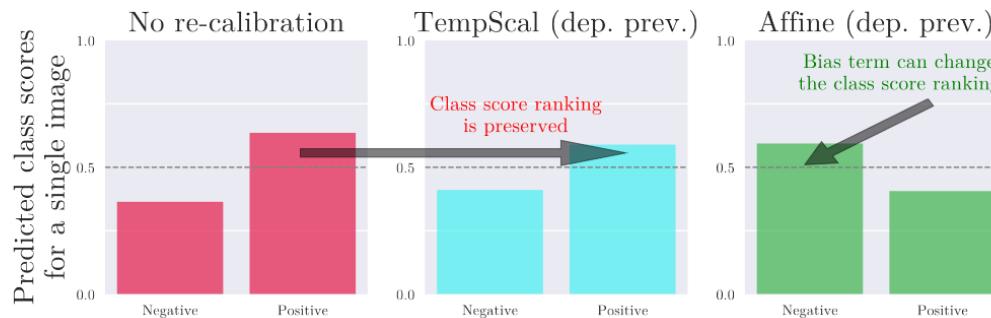
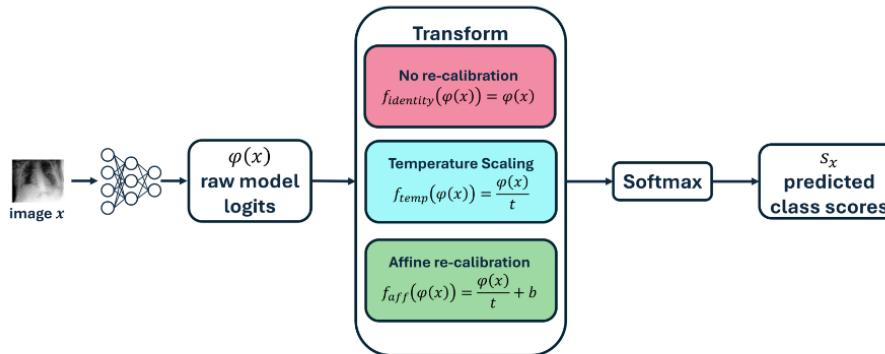
Godau/Kalinowski, ..., Maier-Hein. Navigating prevalence shifts in image analysis algorithm deployment, **MICCAI 2023**
Medical Image Analysis MICCAI 2023 Best Paper Award

Does temperature scaling fix our problem?



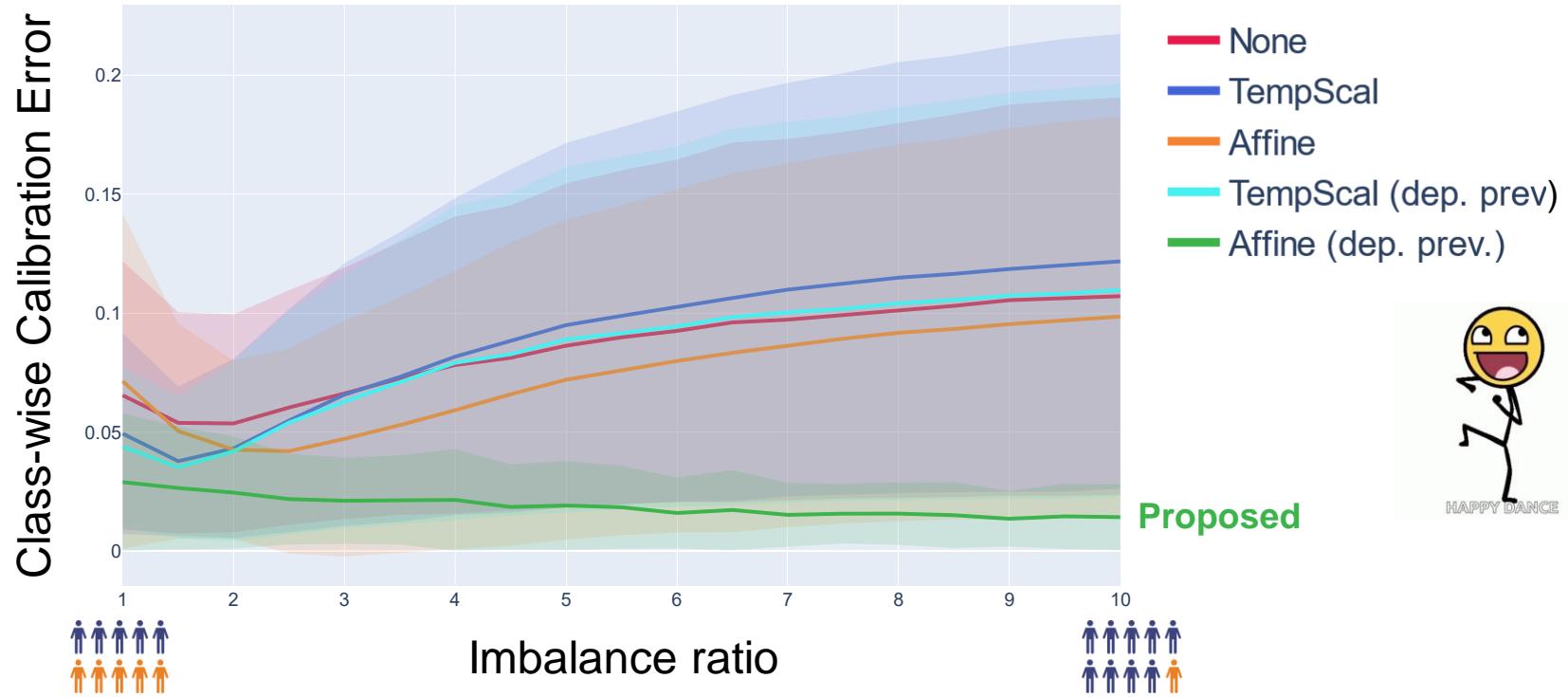
Godau/Kalinowski, ..., Maier-Hein. Navigating prevalence shifts in image analysis algorithm deployment, **MICCAI 2023**
Medical Image Analysis MICCAI 2023 Best Paper Award

Why not? Because we need a bias term



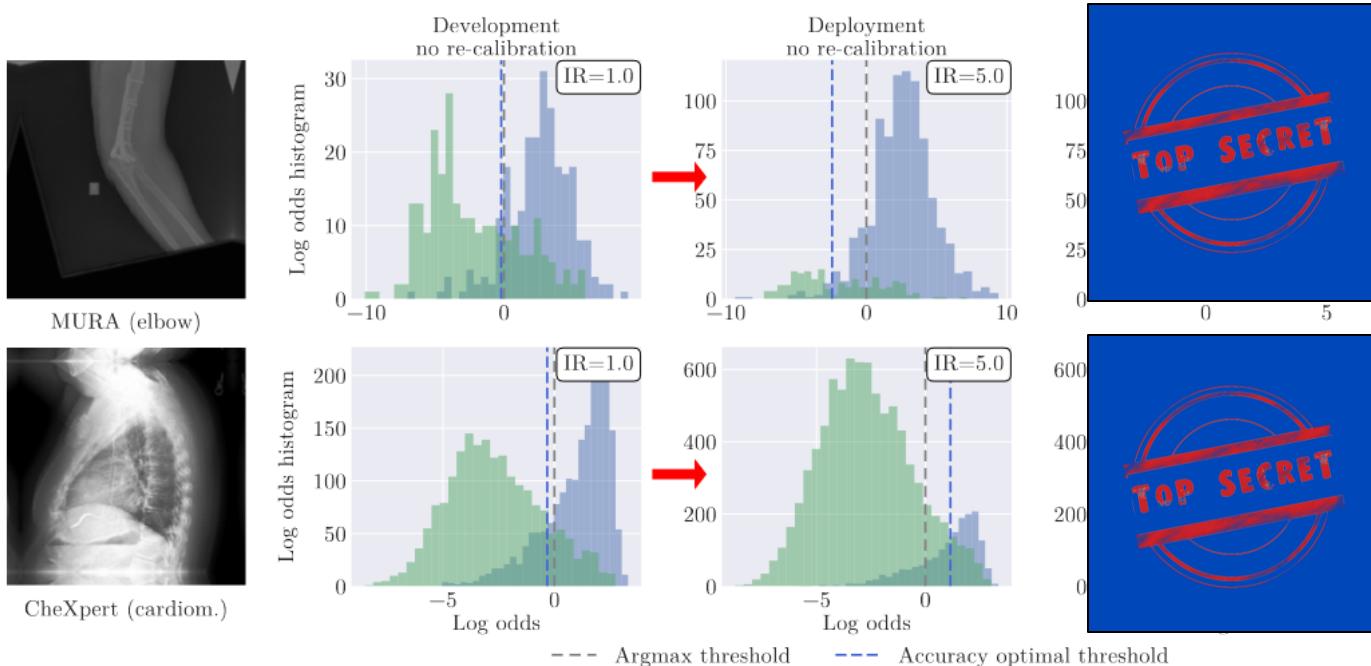
Godau/Kalinowski, ..., Maier-Hein. Navigating prevalence shifts in image analysis algorithm deployment, **MICCAI 2023**
Medical Image Analysis MICCAI 2023 Best Paper Award

Affine re-calibration solves the calibration issue

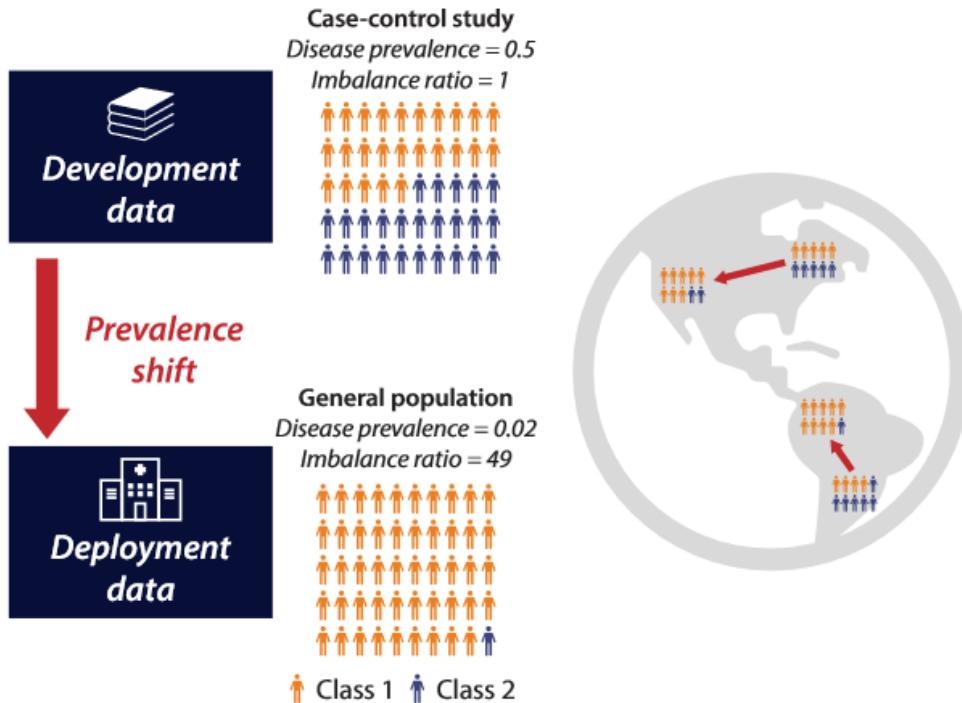


Godau/Kalinowski, ..., Maier-Hein. Navigating prevalence shifts in image analysis algorithm deployment, **MICCAI 2023**
Medical Image Analysis MICCAI 2023 Best Paper Award

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

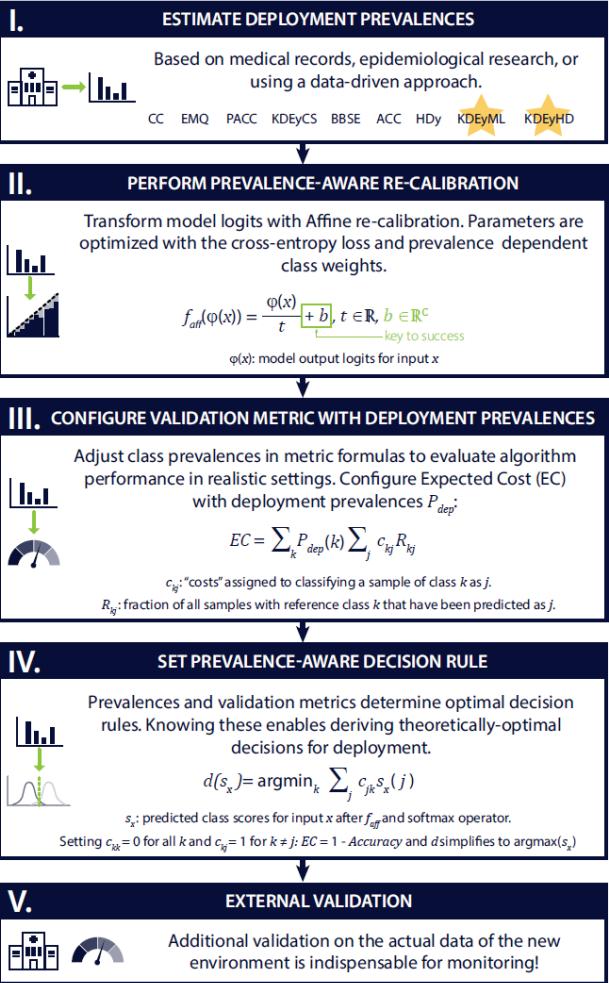
- calibration,
- decision rules,
- performance assessment?

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



(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).

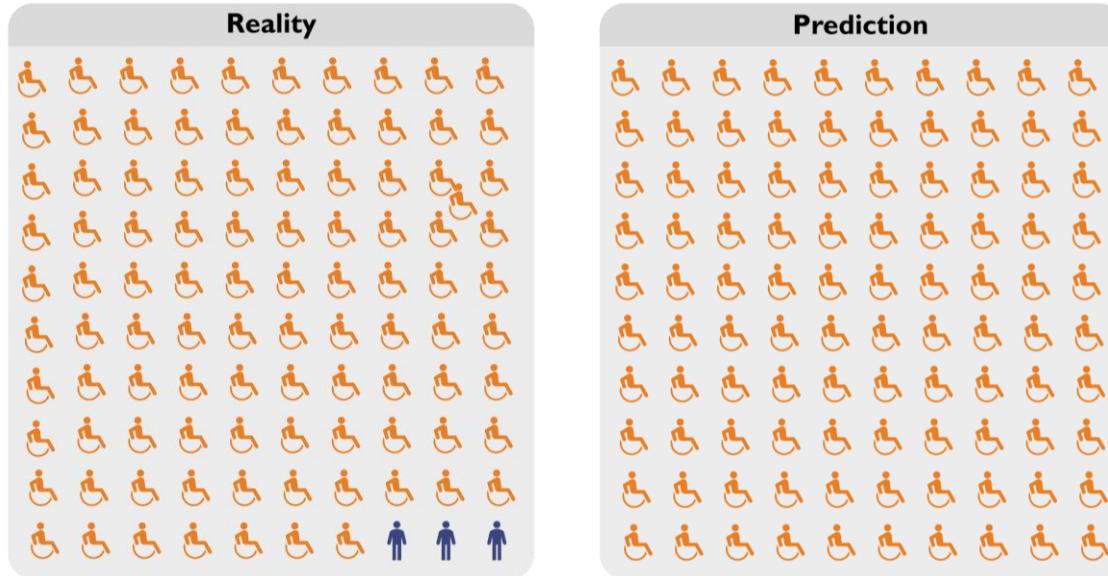
Argmax should not be used indiscriminately as a decision rule. However, it is the optimal decision rule if predicted class scores are calibrated and Accuracy (or EC) is the target metric (see Figs. 7 & 8).

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**

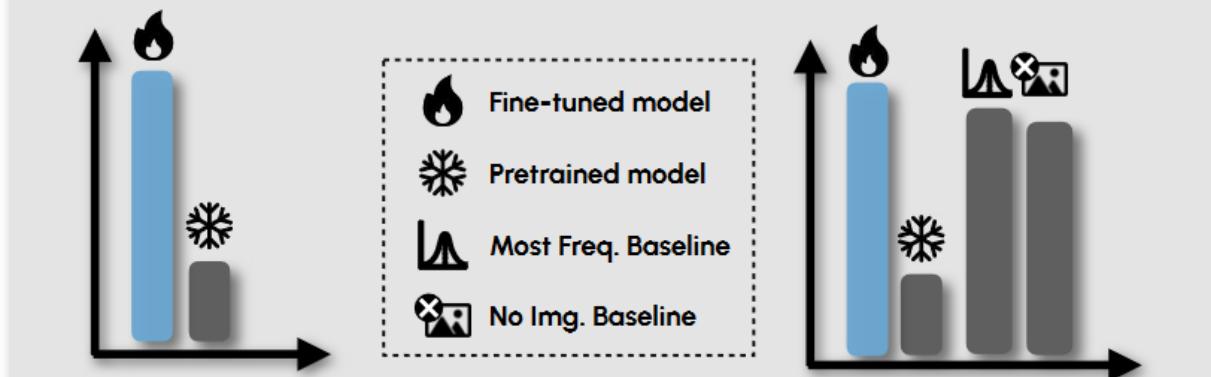
NEW pitfall: New random baselines required



Kim-Celine
Kahl

Selen
Erkan

P3: Model performance lacks interpretability
due to missing sanity baselines



Kahl/Erkan, Maier-Hein, Jäger: SURE-VQA: Systematic understanding of robustness evaluation in medical VQA tasks, [arXiv 2024](#)

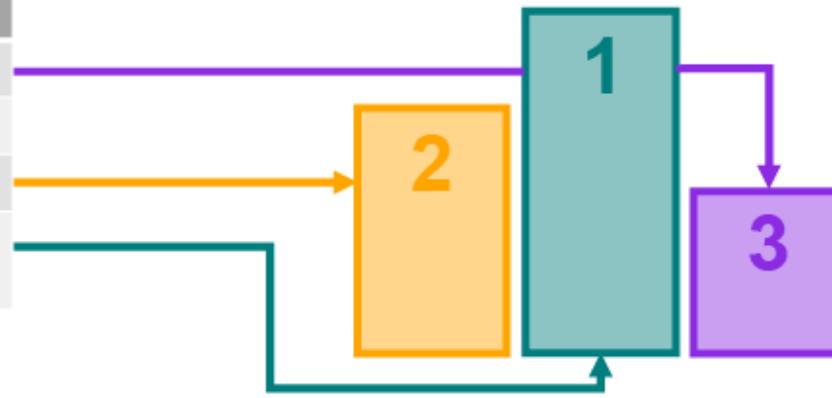


#Rankings

Challenging common practice

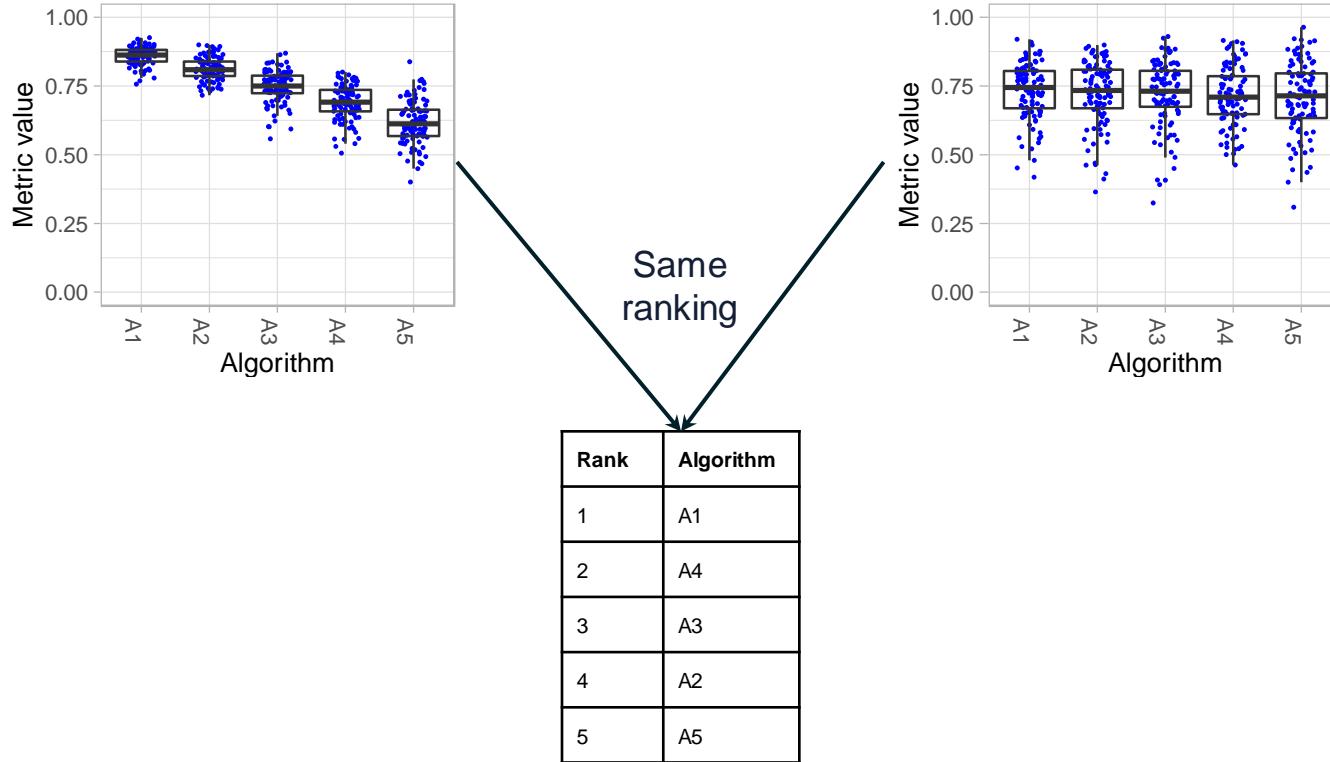
Commonly encountered results tables

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



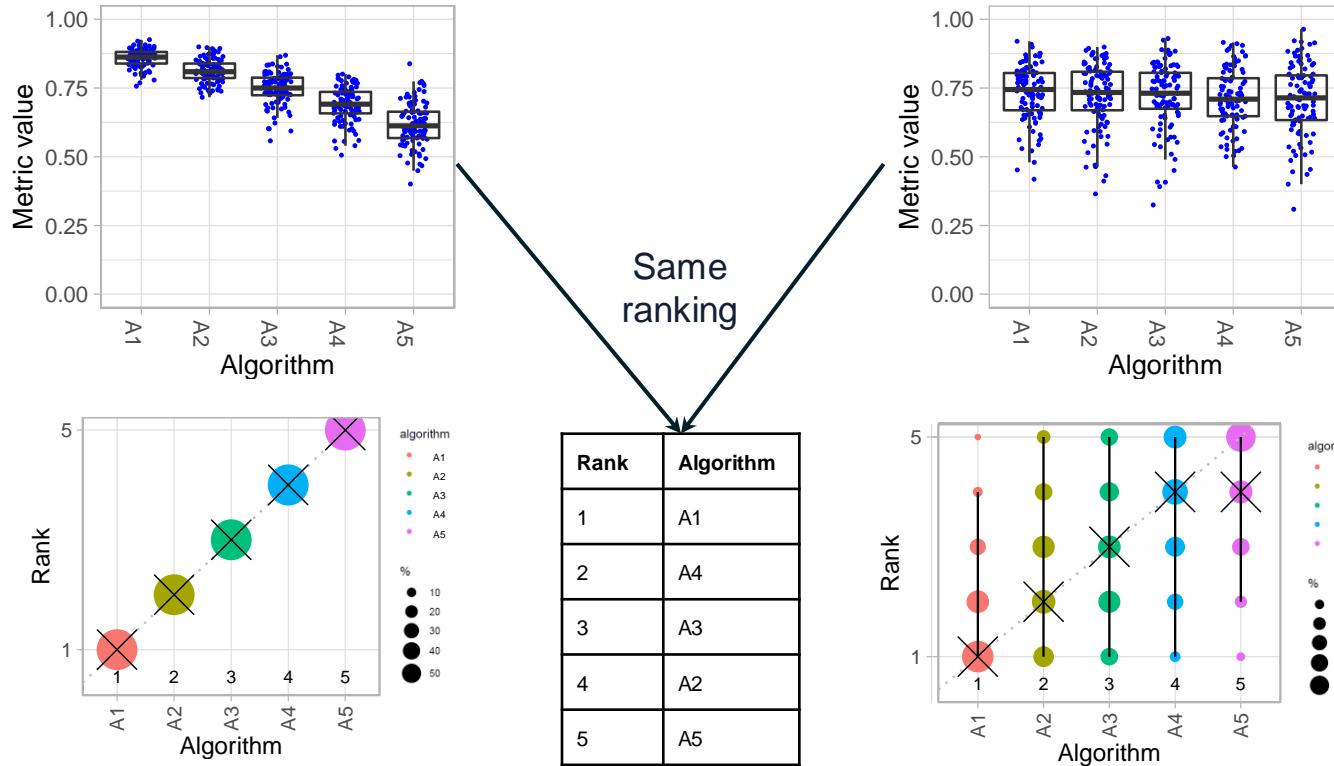
“As shown in Table 1, our method outperforms all previously proposed state-of-the-art methods”

Why result analysis and visualization is critical: Example



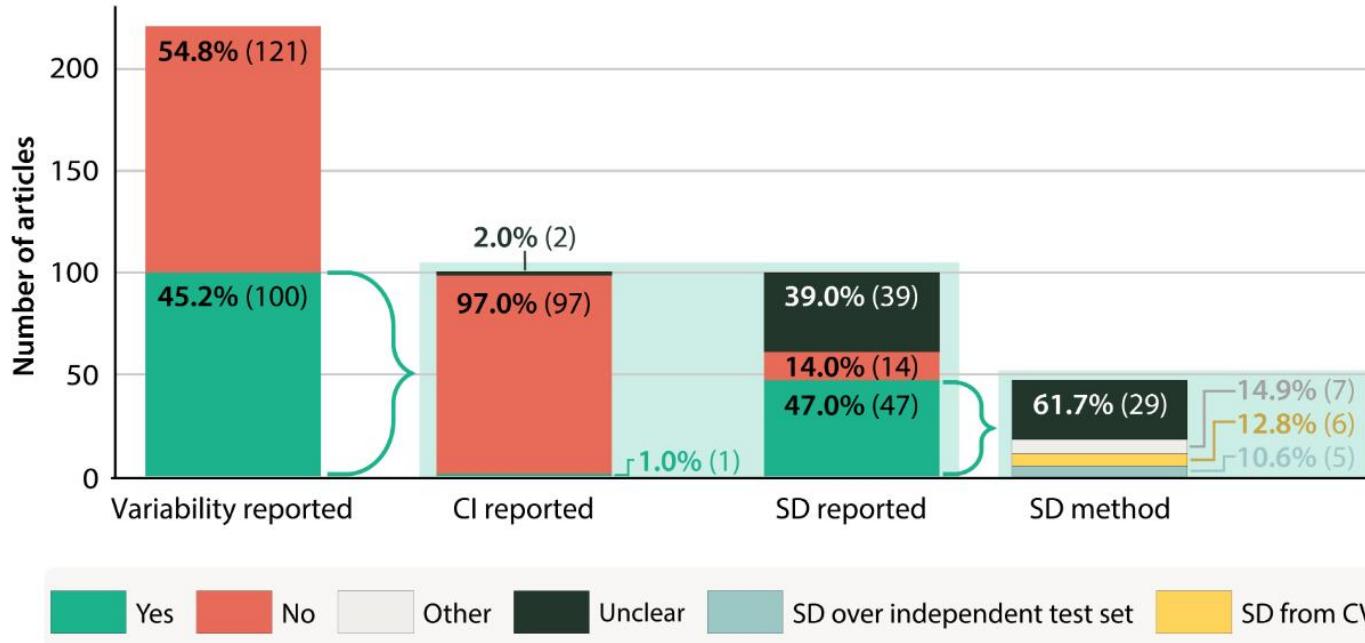
Wiesenfarth, ... Maier-Hein/Kopp-Schneider. Methods and open-source toolkit for analyzing and visualizing challenge results, **Scientific Reports 2021**

Why result analysis and visualization is critical: Example



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

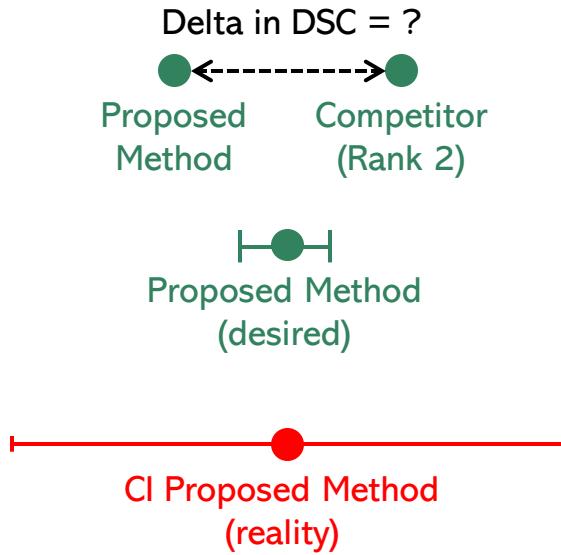


Christodoulou/Reinke et al. Confidence intervals uncovered: Are we ready for real-world medical imaging AI? MICCAI 2024.

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



Eva
Christodoulou



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

Confidence interval width should be smaller ...

Reality: Median CI width: ~0.03



Annika Reinke

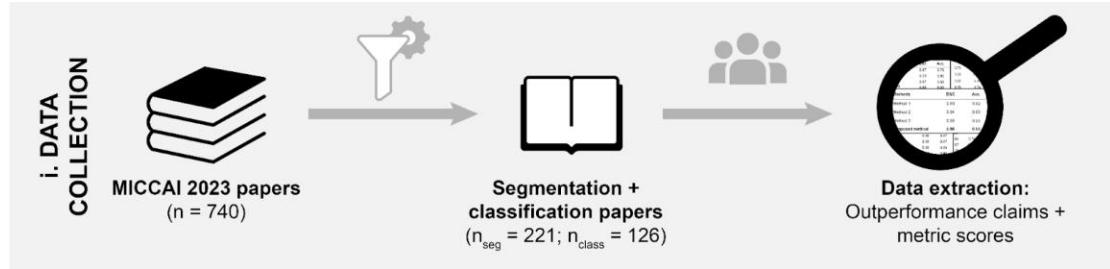


Close collaboration with
Gael Varoquaux
and Olivier Colliot



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

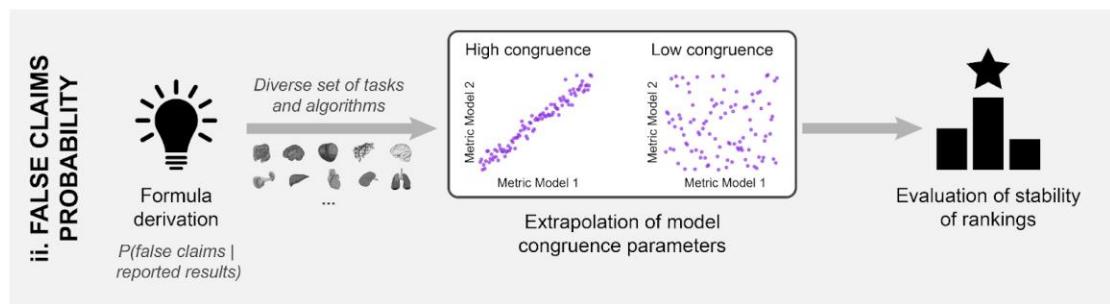
False outperformance claims?



Eva Christodoulou

?

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



Annika Reinke



Close collaboration with Gael Varoquaux and Olivier Colliot

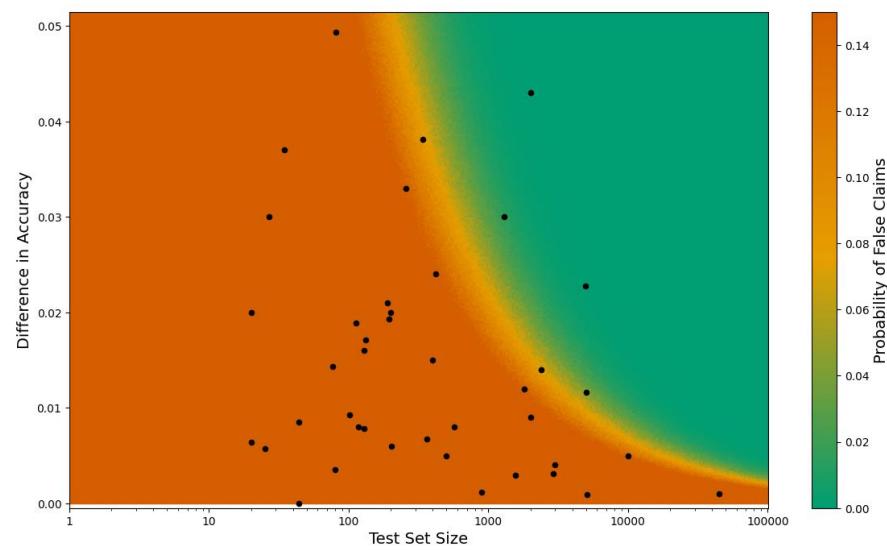
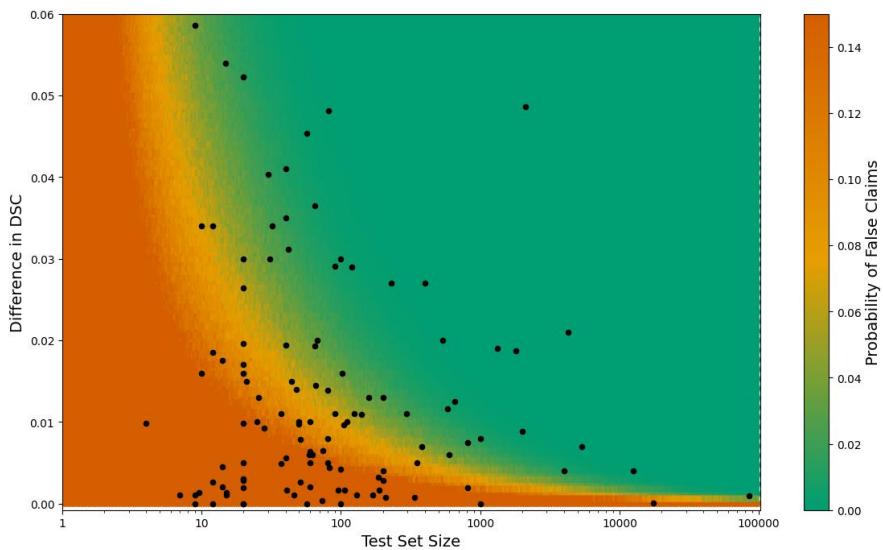


Christodoulou,, Varoquaux/Colliot/Maier-Hein. False Promises in Medical Imaging AI? Assessing Validity of Outperformance Claims, [arXiv 2025](#)

Outperformance claims are not well-substantiated

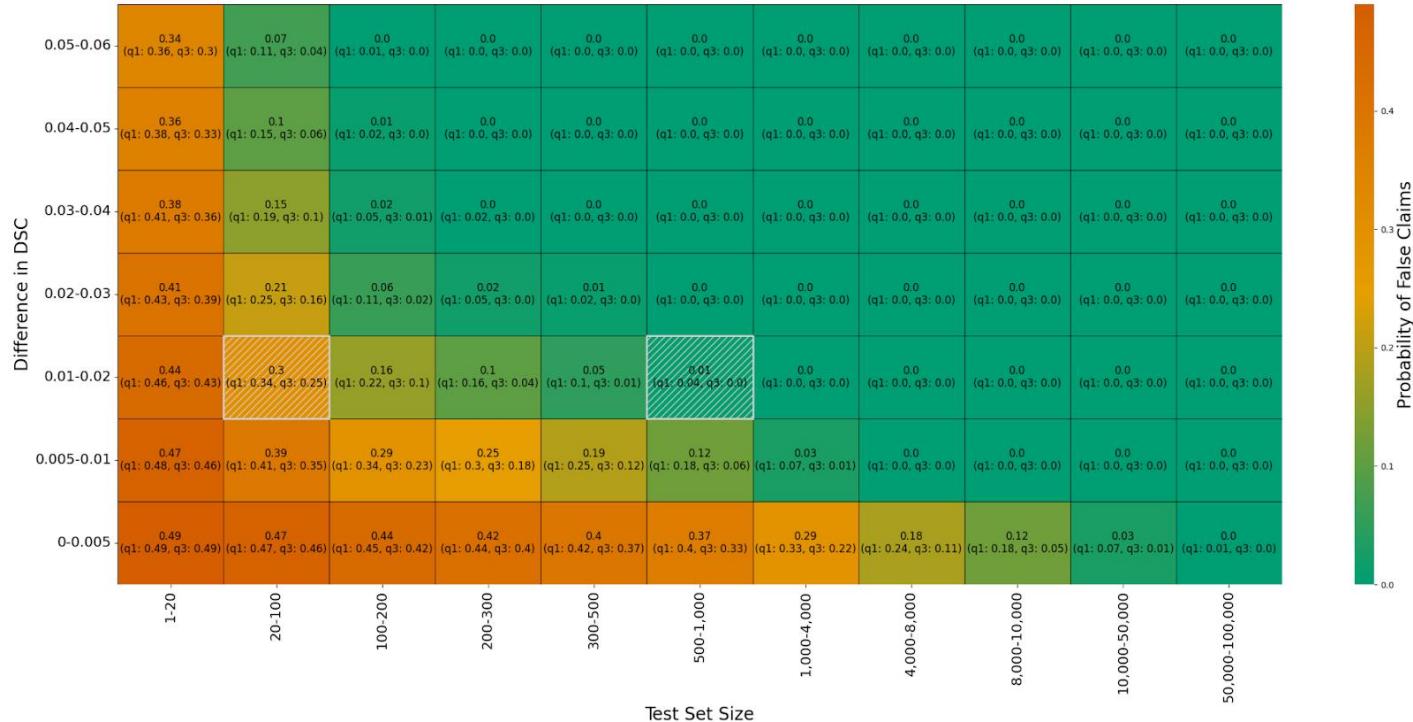
High probability (>5%) of false outperformance claims in

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

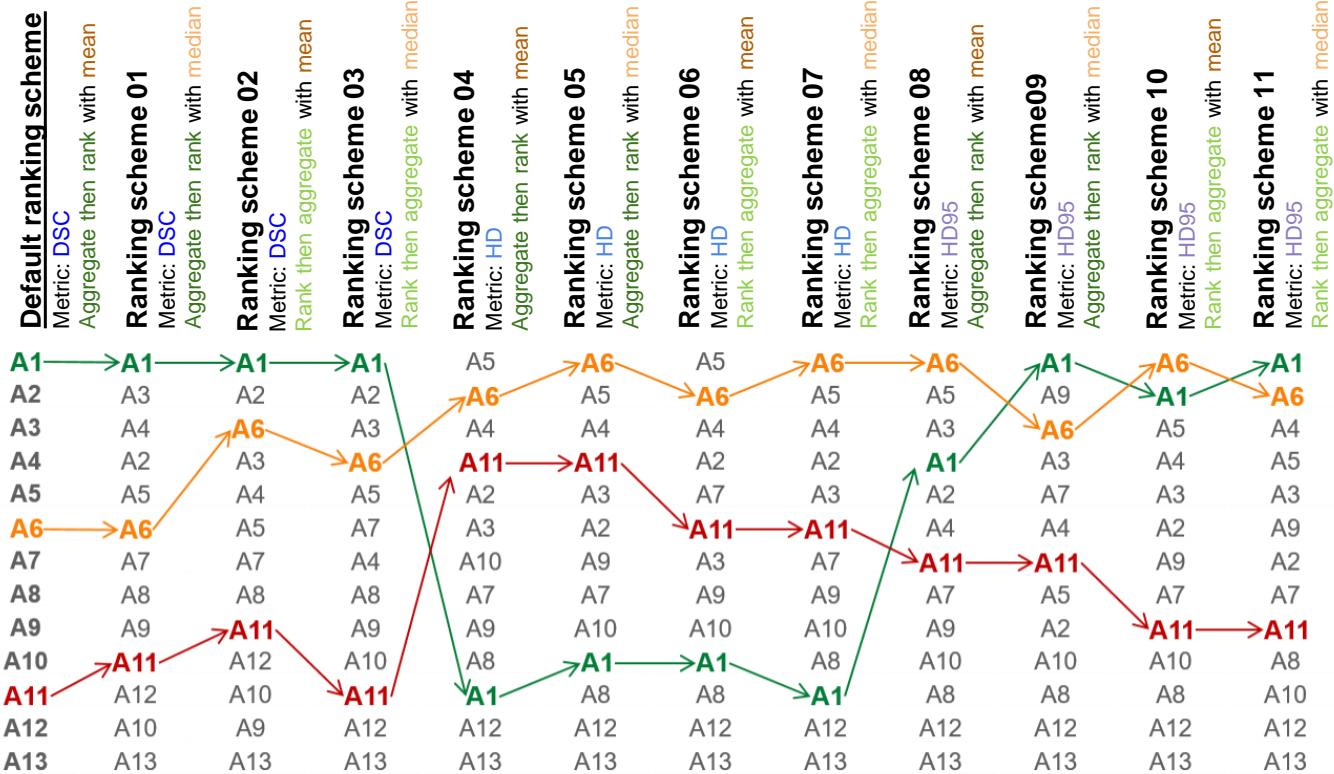


Christodoulou,, Varoquaux/Colliot/Maier-Hein. False Promises in Medical Imaging AI? Assessing Validity of Outperformance Claims, [arXiv 2025](#)

We need larger sample sizes



Pitfalls beyond sample size: Sensitivity to ranking scheme

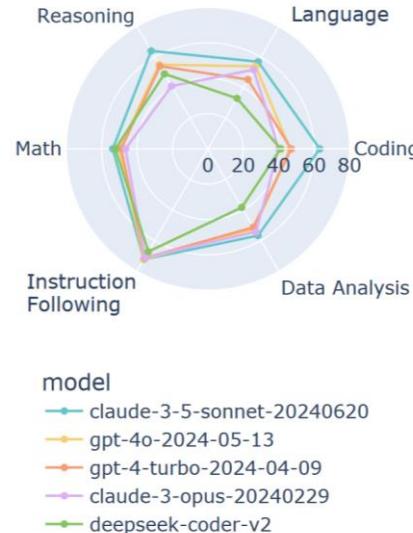


Maier-Hein et al. Why rankings of biomedical image analysis competitions should be interpreted with care. *Nature Commun.* 2018

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 reliable reference
 - Overfitting to arena-preferred examples
 - Reasons of failure?



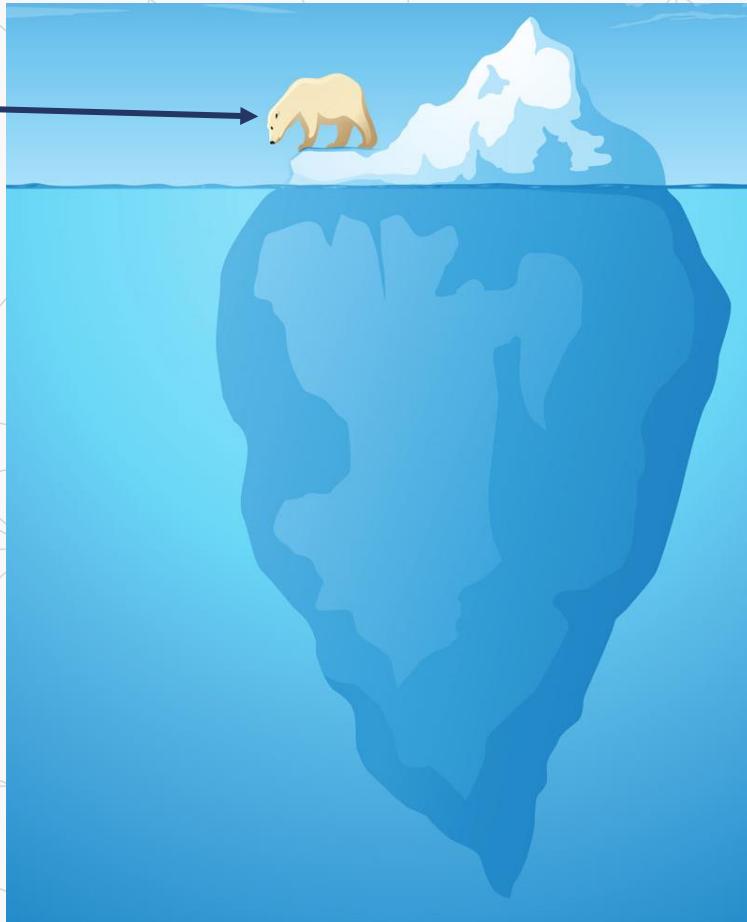
Dataset	Flan-PaLM 540B with non-medical prompt	Flan-PaLM 540B with medical prompt
MedQA 4 options	65.4	67.6
MedMCQA	55.2	57.6
PubMedQA	77.2	75.2



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
- ...

The role of academia in AI?

nature

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

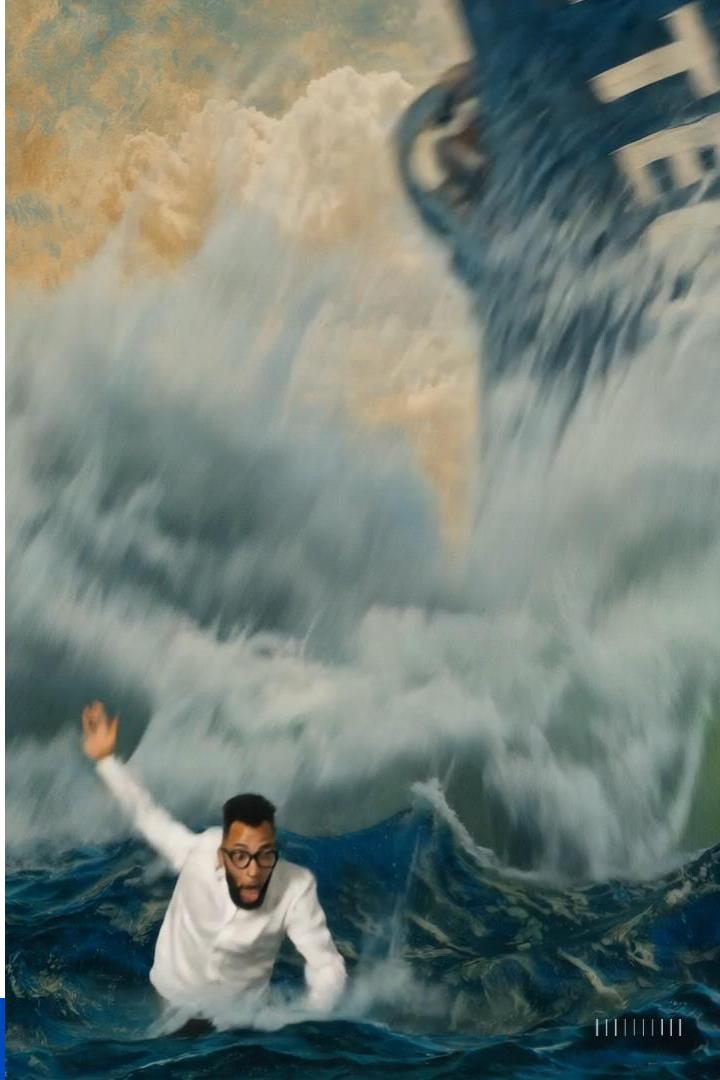
[nature](#) > [nature index](#) > article

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

