(Bench)mark: Pitfalls in AI Validation

Annika Reinke

Div. Intelligent Medical Systems, German Cancer Research Center (DKFZ)

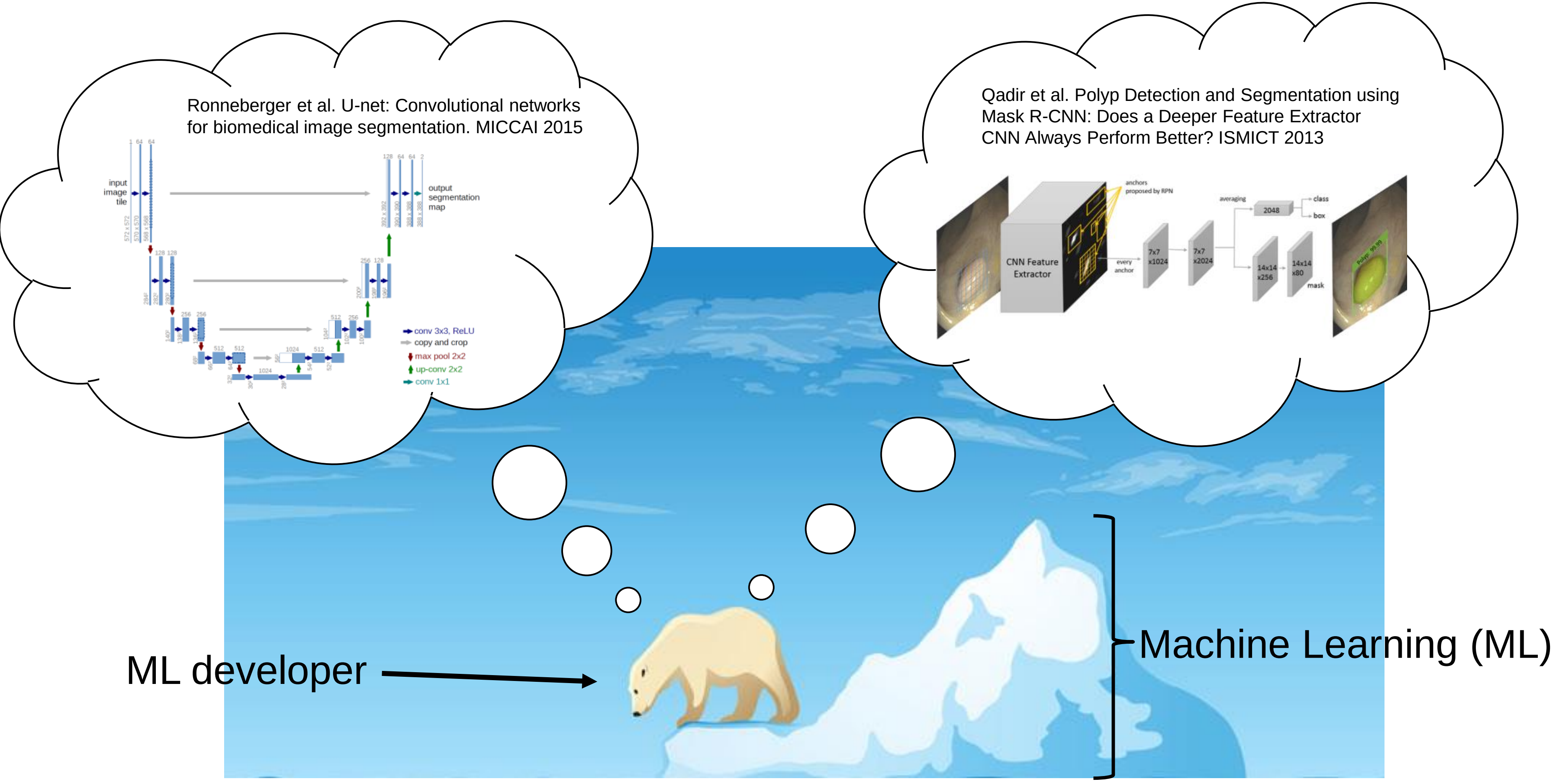
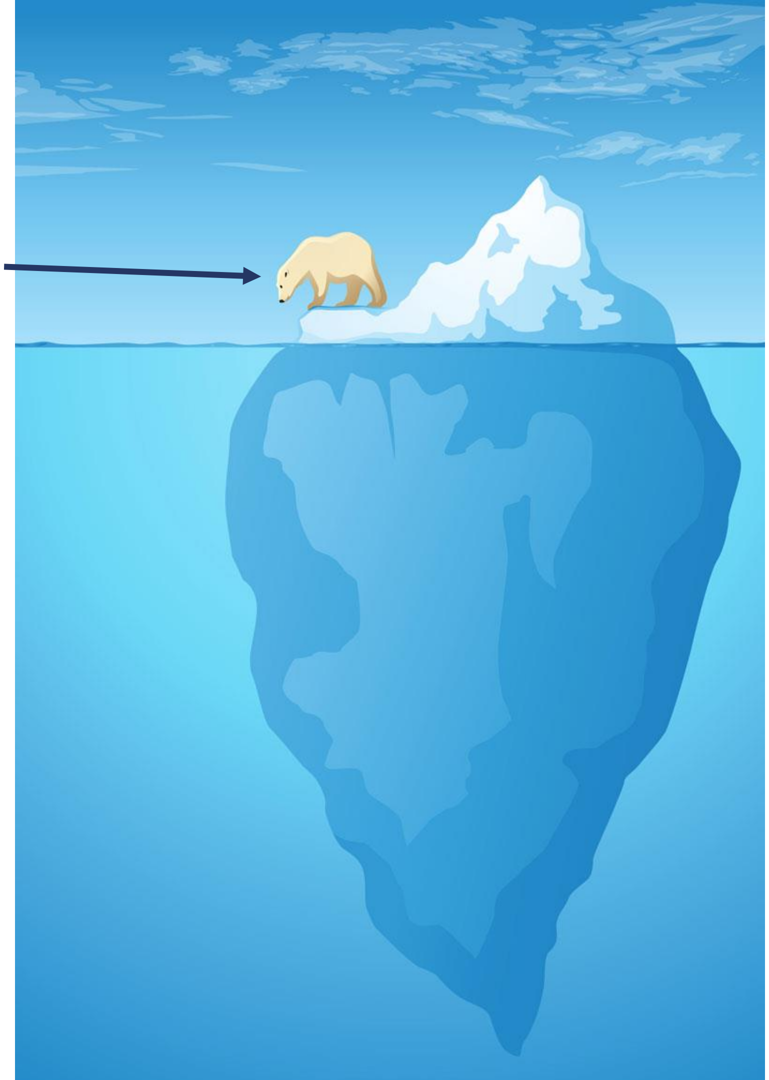


Image source:
Regulatory Affairs Professional Society

ML developer



Machine Learning (ML)

Data set design,

Annotations,

Metrics,

Rankings,

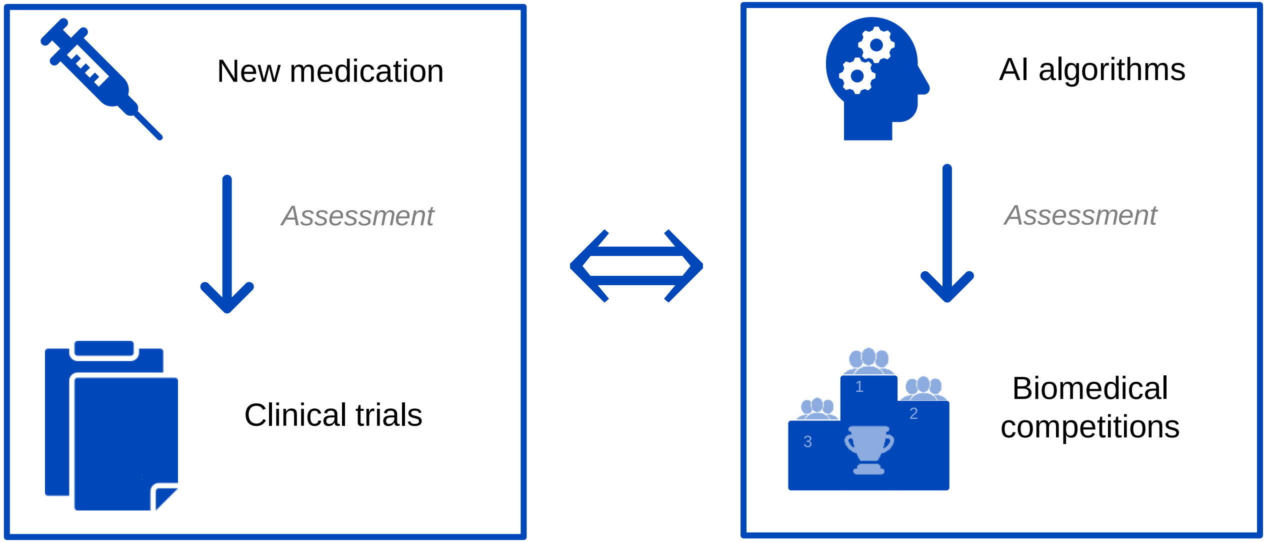
Reporting,

Infrastructure,

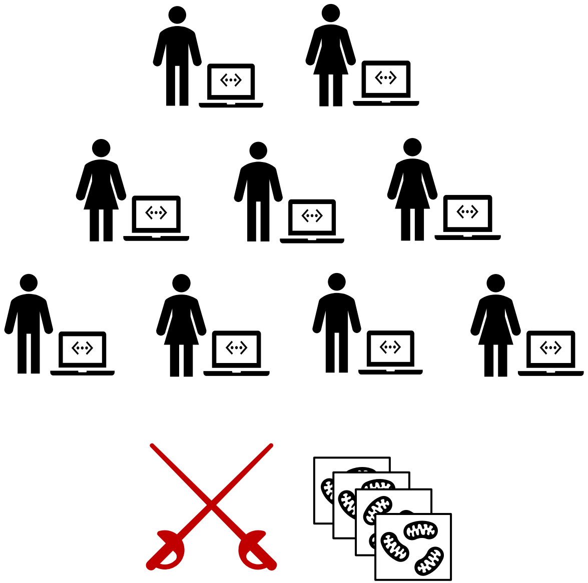
…

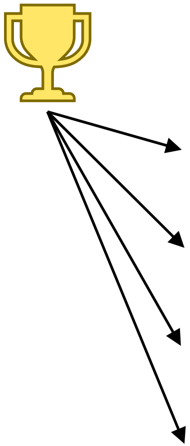
Image source:
Regulatory Affairs Professional Society

Assessment of AI algorithms



Biomedical image analysis competitions





Up to €1 million price money

New state-of-the art method

Fame for researcher

…

- ✓ Challenges have led to common data sets used for validation
- ✓ Various fields of application covered
- ✓ Various modalities covered

Algorithm benchmarking

Table 12
Comparison of existing methods.

Methods (%)	Database	Classifier	Sensitivity (%)	Specificity (%)	Accuracy (%)	AUC (%)
Second technique (best)	DDSM	MLP	96.87	95.94	96.47	95.10
Second technique (average)	DDSM	MLP	96.25	93.78	95.01	94.99
Second technique (best)	MIAS	MLP	92.70	90.54	90.16	95.58
Second technique (average)	MIAS	MLP	87.91	85.40	86.66	88.15
Second technique (best)	DDSM	SCBDL	80.70	79.00	80.00	–
Wang and Yang et al., 2014; Wang, Li, & Gao, 2014	DDSM	SVM	–	–	92.74	96.50
Liu and Tang, 2013	DDSM	SVM	92.00	93.00	93.00	94.39
Saki et al., 2013	MIAS	OWBPE	90.10	88.06	89.28	92.80
Zhang, Tomuro, Furst, & Raicu, 2012	DDSM	SVM	–	–	72.00	–
Tahmasbi et al., 2011	MIAS	MLP	100	94.50	96.43	97.60
Buciu and Gacsadi, 2011	MIAS	PSVM	84.61	80	82.30	78.00
Tahmasbi et al., 2010	MIAS	MLP	90.10	94.44	92.80	98.00
Verma et al., 2010	DDSM	MLP	85.00	92.50	88.75	–
Verma et al., 2010	DDSM	SCBDL	97.50	97.50	97.50	–
Verma et al., 2009	DDSM	SCNN	97.83	90.74	94.28	–
Rojas-Domínguez and Nandi, 2009	DDSM, MIAS	Bayesian, FLD	–	–	81.00	–
Mu et al., 2008	MIAS	S2SP	–	–	–	95.00
Masotti, 2006	DDSM	SVM	90.00	95.50	92.75	97.80

Rouhi, et al. Benign and malignant breast tumors classification based on region growing and CNN segmentation. Expert Systems with Applications 2015.

Model	Top-1 Acc.	Top-5 Acc.	#Params	Ratio-to-EfficientNet	#FLOPs	Ratio-to-EfficientNet
EfficientNet-B0	**77.1%**	**93.3%**	**5.3M**	1x	**0.39B**	1x
ResNet-50 (He et al., 2016)	76.0%	93.0%	26M	4.9x	4.1B	11x
DenseNet-169 (Huang et al., 2017)	76.2%	93.2%	14M	2.6x	3.5B	8.9x
EfficientNet-B1	**79.1%**	**94.4%**	**7.8M**	1x	**0.70B**	1x
ResNet-152 (He et al., 2016)	77.8%	93.8%	60M	7.6x	11B	16x
DenseNet-264 (Huang et al., 2017)	77.9%	93.9%	34M	4.3x	6.0B	8.6x
Inception-v3 (Szegedy et al., 2016)	78.8%	94.4%	24M	3.0x	5.7B	8.1x
Xception (Chollet, 2017)	79.0%	94.5%	23M	3.0x	8.4B	12x
EfficientNet-B2	**80.1%**	**94.9%**	**9.2M**	1x	**1.0B**	1x
Inception-v4 (Szegedy et al., 2017)	80.0%	95.0%	48M	5.2x	13B	13x
Inception-resnet-v2 (Szegedy et al., 2017)	80.1%	95.1%	56M	6.1x	13B	13x
EfficientNet-B3	**81.6%**	**95.7%**	**12M**	1x	**1.8B**	1x
ResNeXt-101 (Xie et al., 2017)	80.9%	95.6%	84M	7.0x	32B	18x
PolyNet (Zhang et al., 2017)	81.3%	95.8%	92M	7.7x	35B	19x
EfficientNet-B4	**82.9%**	**96.4%**	**19M**	1x	**4.2B**	1x
SENet (Hu et al., 2018)	82.7%	96.2%	146M	7.7x	42B	10x
NASNet-A (Zoph et al., 2018)	82.7%	96.2%	89M	4.7x	24B	5.7x
AmoebaNet-A (Real et al., 2019)	82.8%	96.1%	87M	4.6x	23B	5.5x
PNASNet (Liu et al., 2018)	82.9%	96.2%	86M	4.5x	23B	6.0x
EfficientNet-B5	**83.6%**	**96.7%**	**30M**	1x	**9.9B**	1x
AmoebaNet-C (Cubuk et al., 2019)	83.5%	96.5%	155M	5.2x	41B	4.1x
EfficientNet-B6	**84.0%**	**96.8%**	**43M**	1x	**19B**	1x
EfficientNet-B7	**84.3%**	**97.0%**	**66M**	1x	**37B**	1x
GPipe (Huang et al., 2018)	84.3%	97.0%	557M	8.4x	-	-

Tan and Le. EfficientNet: Rethinking model scaling for convolutional neural networks. International conference on machine learning 2019.

	Cats	CelebA	Cars	Chairs	Churches
2D GAN [58]	18	15	**16**	59	19
Plat. GAN [32]	318	321	299	199	242
BlockGAN [64]	47	69	41	41	28
HoloGAN [63]	27	25	17	59	31
GRAF [77]	26	25	39	34	38
Ours	**8**	**6**	**16**	**20**	**17**

Table 1: **Quantitative Comparison.** We report the FID score (↓) at 64^2 pixels for baselines and our method.

	CelebA-HQ	FFHQ	Cars	Churches	Clevr-2
HoloGAN [63]	61	192	34	58	241
w/o 3D Conv	33	70	49	66	273
GRAF [77]	49	59	95	87	106
Ours	**21**	**32**	**26**	**30**	**31**

Table 2: **Quantitative Comparison.** We report the FID score (↓) at 256^2 pixels for the strongest 3D-aware baselines and our method.

Niemeyer and Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. IEEE/CVF 2021.

	UKCF (Binary Targets)		ADNI (Continuous Targets)		MIMIC (Mixed Targets)	
	PRC(I)	PRC(C)	MSE(B)	MSE(C)	PRC	MSE
Base	0.411 ± 0.035*	0.497 ± 0.057*	0.105 ± 0.018*	0.361 ± 0.064	0.142 ± 0.028*	0.153 ± 0.011
REG	0.415 ± 0.030*	0.518 ± 0.052*	0.096 ± 0.014*	0.360 ± 0.066	0.143 ± 0.019*	0.152 ± 0.010
FEA	0.410 ± 0.033*	0.521 ± 0.054*	0.092 ± 0.012*	0.356 ± 0.068	0.144 ± 0.030*	0.152 ± 0.012
TEA	$\mathbf{0.483 \pm 0.045}$	$\mathbf{0.583 \pm 0.072}$	$\mathbf{0.063 \pm 0.010}$	$\mathbf{0.330 \pm 0.066}$	$\mathbf{0.239 \pm 0.039}$	$\mathbf{0.150 \pm 0.012}$
F/TEA	0.457 ± 0.037	0.576 ± 0.071	0.073 ± 0.010*	0.338 ± 0.067	0.166 ± 0.023*	0.154 ± 0.011

Jarrett and van der Schaar. Target-embedding autoencoders for supervised representation learning. arXiv 2020.

Is the winner really the best?

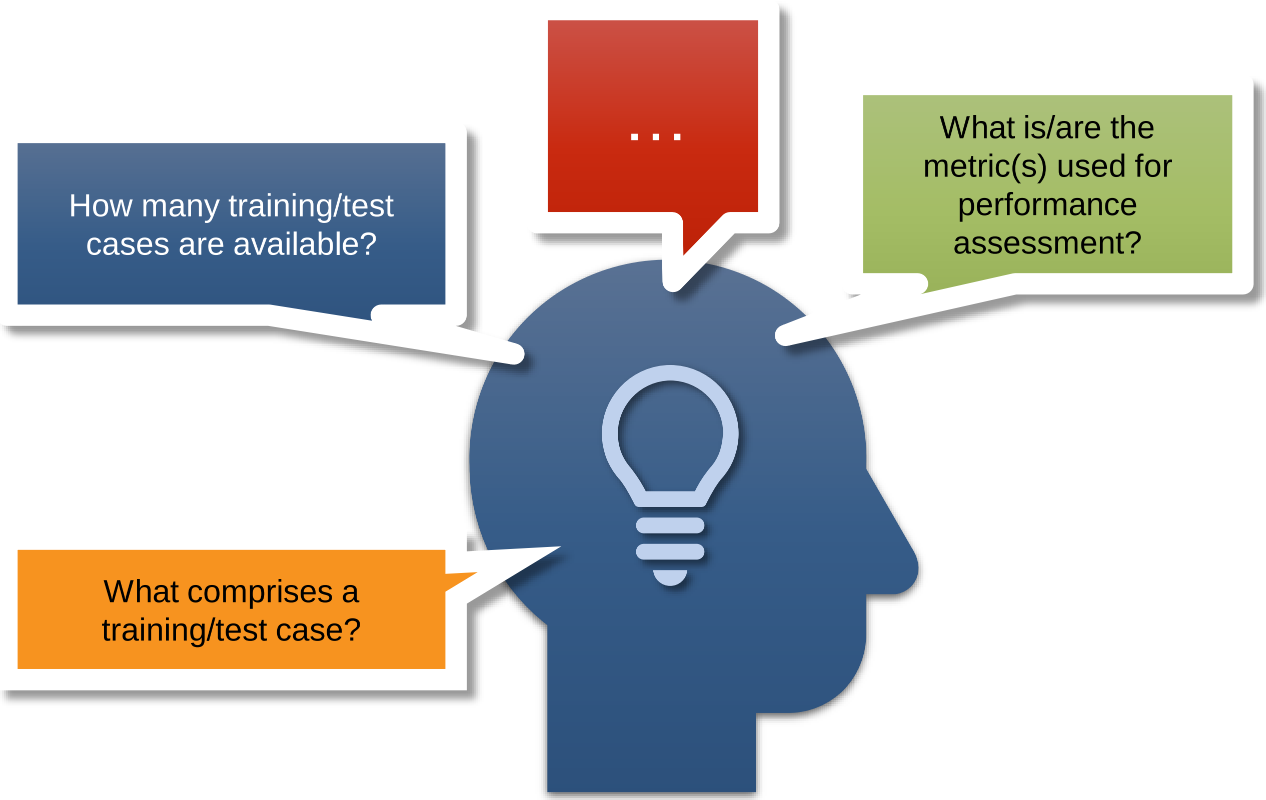
Reporting

“ *The one practice that can universally commended is the* ***transparent and complete reporting of all facets of a study****, allowing a critical reader to evaluate the work and fully understand its strengths and limitations* ”

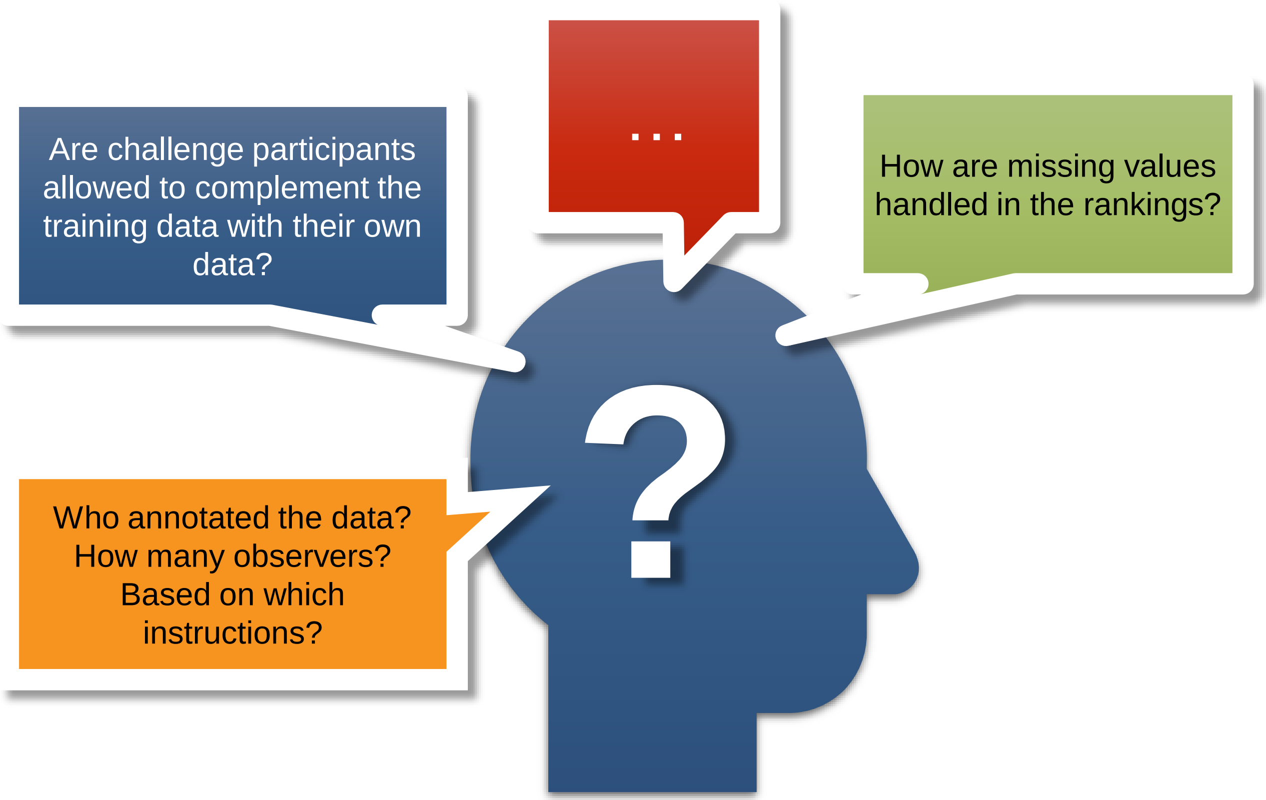
(Nature Neuroscience 2017, https://doi.org/10.1038/nn.4500)

dkfz.

A lot of challenge parameters matter (the obvious)

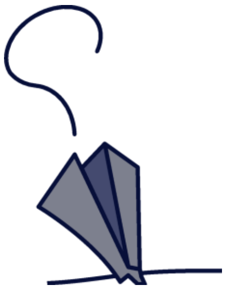


A lot of challenge parameters matter (the “not so obvious”)



Analysis of > 500 competitions

- A median of **64%** of parameters were reported
- Only **6%** of parameters were reported by all challenges
- *Examples:*
  - **85%** of challenges did not give instructions on whether training data provided by challenge organizers may be complemented by other publicly available or private data
  - In **66%** of all tasks, there was no description on how the reference (i.e. gold standard) annotation was performed



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

BIAS Reporting guideline

- BIAS (**B**iomedical **I**mage **A**nalysis **C**hallenge**S**) initiative: bring challenges to next level of quality
- Formed by MICCAI board challenge working group
- Developed **guideline for designing and reporting challenges**
- Registered BIAS with equator network


equator network

Enhancing the QUAlity and Transparency Of health Research

Home | About us | Library | Toolkits | Courses & events | News | Blog

Home > Library > Reporting guideline > BIAS: Transparent reporting of biomedical image analysis challenges

Search for reporting guidelines

Use your browser's Back button to return to your search results

BIAS: Transparent reporting of biomedical image analysis challenges

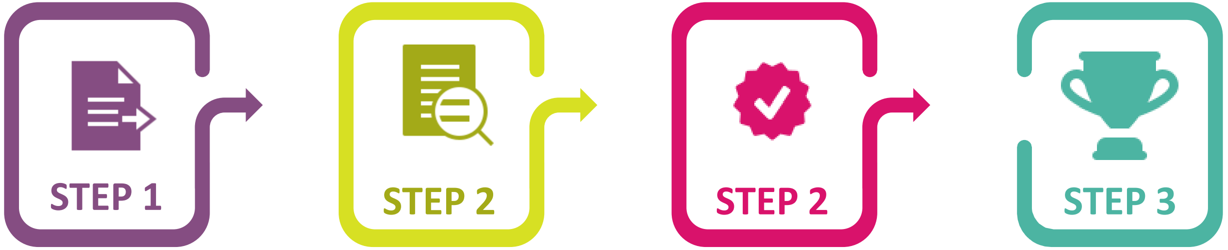
Reporting guideline provided for? (i.e. exactly what the authors state in the paper)	Reporting of a biomedical image analysis challenge.
Full bibliographic reference	Maier-Hein L, Reinke A, Kozubek M, Martel AL, Arbel T, Eisenmann M, Hanbury A, Jannin P, Müller H, Onogur S, Saez-Rodriguez J, van Ginneken B, Kopp-Schneider A, Landman BA. BIAS: Transparent reporting of biomedical image analysis challenges. Med Image Anal. 2020;66:101796.
PubMed ID	32911207
Relevant URLs (full-text if available)	The full-text of the BIAS statement is available at: https://pubmed.ncbi.nlm.nih.gov/32911207/ The BIAS checklist can be accessed at: https://ars.els-cdn.com/content/image/1-s2.0-S1361841520301602-mmc1.pdf
Reporting guideline acronym	BIAS

Maier-Hein, Reinke et al. BIAS: Transparent reporting of biomedical image analysis challenges, **Med Image Anal 2020**

Problem: Quality control after challenge acceptance

Direct quality control

No quality control



Challenge submission

Challenge review

(with modifications of document)

Challenge accepted

Challenge execution

Icons created by the Icons8

Solution: Challenge registration

Direct quality control

Indirect quality control

Challenge submission

Challenge review

(with modifications of document)

Challenge accepted

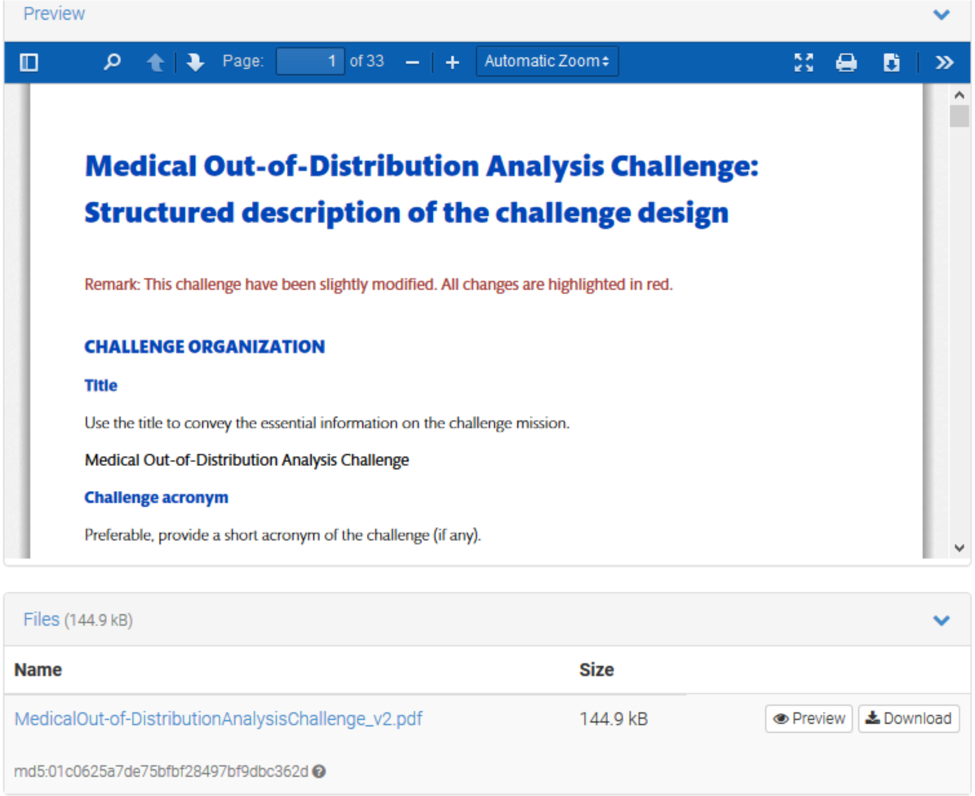
Upload challenge document

Challenge execution

Icons created by the Icons8

Challenge registration

Challenge name	Acronym	DOI
2nd Retinal Fundus Glaucoma Challenge	REFUGE2	10.5281/zenodo.3714946
3D Head and Neck Tumor Segmentation in PET/CT	HECKTOR	10.5281/zenodo.3714956
Anatomical Brain Barriers to Cancer Spread: Segmentation from CT and MR images	ABCs	10.5281/zenodo.3714981
Automated Segmentation of Coronary Arteries	ASOCA	10.5281/zenodo.3714985
Automatic Evaluation of Mycardial Infarction from Delayed-Enhancement Cardiac MRI	EMIDEC	10.5281/zenodo.3714997
Automatic Lung Cancer Detection and Classification in Whole-slide Histopathology	ACDC@LungHP	10.5281/zenodo.3715000
Automatic Structure Segmentation for Radiotherapy Planning Challenge 2020 **(Challenge withdrawn due to COVID-19 pandemic situation)**	StructSeg 2020	10.5281/zenodo.3718884
Cerebral Aneurysm Detection and Analysis	CADA	10.5281/zenodo.3715011
Computational Precision Medicine Challenge on Brain Tumor Classification 2020	CPM-RadPath	10.5281/zenodo.3718893
Diabetic Foot Ulcers Grand Challenge 2020	DFUC 2020	10.5281/zenodo.3715015
Endoscopic Vision Challenge 2020	EndoVis	10.5281/zenodo.3715645
International Skin Imaging Collaboration Challenge: Using Dermoscopic Image Context to Diagnose Melanoma	ISIC 2020	10.5281/zenodo.3715749
Intracranial Aneurysm Detection and Segmentation Challenge	ADAM	10.5281/zenodo.3715847
Large Scale Vertebrae Segmentation Challenge	VerSe'20	10.5281/zenodo.3715865
Learn2Reg - The Challenge	L2R	10.5281/zenodo.3715651
Medical Out-of-Distribution Analysis Challenge	MOOD	10.5281/zenodo.3715869
MICCAI Brain Tumor Segmentation (BraTS) 2020 Benchmark: "Prediction of Survival and Pseudoprogression"	BraTS 2020	10.5281/zenodo.3718903
Multi-Centre, Multi-Vendor & Multi-Disease Cardiac Image Segmentation Challenge	M&Ms	10.5281/zenodo.3715889
Multi-sequence CMR based Mycardial Pathology Segmentation Challenge	MyoPS 2020	10.5281/zenodo.3715931
Quantification of Uncertainties in Biomedical Image Quantification	QUBIQ	10.5281/zenodo.3718911
Rib Fracture Detecion and Classification	RibFrac	10.5281/zenodo.3715933



dkfz.

Rankings

Rankings

	DSC: 0.87	DSC: 0.68	DSC: 0.94	...
	DSC: 0.76	DSC: 0.62	DSC: 0.81	...
	DSC: 0.90	DSC: 0.71	DSC: 0.86	...
	DSC: 0.83	DSC: 0.66	DSC: 0.92	...
...	...	...	...	...

Rankings

Data from MICCAI 2015 segmentation challenges

Challenge rankings are sensitive to a range of challenge design parameters:

- **Metric variant**
- Type of test case **aggregation**
- **Annotator**

	Aggregate by mean	Aggregate by median
RANK 1	**A1**	**A5**
RANK 2	**A2**	**A2**
RANK 3	**A3**	**A3**
RANK 4	**A4**	**A4**
RANK 5	**A5**	**A1**
RANK 6	**A6**	**A8**
RANK 7	**A7**	**A7**
RANK 8	**A8**	**A6**
RANK 9	**A9**	**A9**
RANK 10	**A10**	**A10**
RANK 11	**A11**	**A11**
RANK 12	**A12**	**A12**
RANK 13	**A13**	**A13**

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

dkfz.

Example: Exchange ranking schemes

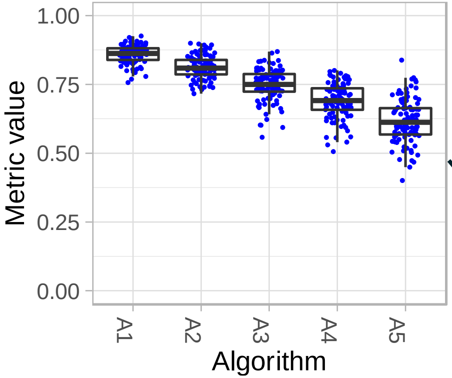
	Default ranking scheme Metric: DSC Aggregate then rank with mean	**Ranking scheme 01** Metric: DSC Aggregate then rank with median	**Ranking scheme 02** Metric: DSC Rank then aggregate with mean	**Ranking scheme 03** Metric: DSC Rank then aggregate with median	**Ranking scheme 04** Metric: HD Aggregate then rank with mean	**Ranking scheme 05** Metric: HD Aggregate then rank with median	**Ranking scheme 06** Metric: HD Rank then aggregate with mean	**Ranking scheme 07** Metric: HD Rank then aggregate with median	**Ranking scheme 08** Metric: HD95 Aggregate then rank with mean	**Ranking scheme09** Metric: HD95 Aggregate then rank with median	**Ranking scheme 10** Metric: HD95 Rank then aggregate with mean	**Ranking scheme 11** Metric: HD95 Rank then aggregate with median
RANK 1	A1											
RANK 2	A2											
RANK 3	A3											
RANK 4	A4											
RANK 5	A5											
RANK 6	A6											
RANK 7	A7											
RANK 8	A8											
RANK 9	A9											
RANK 10	A10											
RANK 11	A11											
RANK 12	A12											
RANK 13	A13											

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

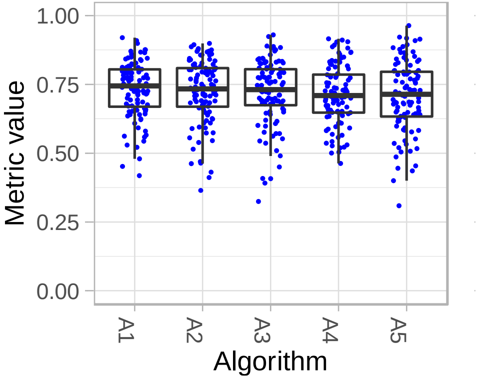
Analysis of Results

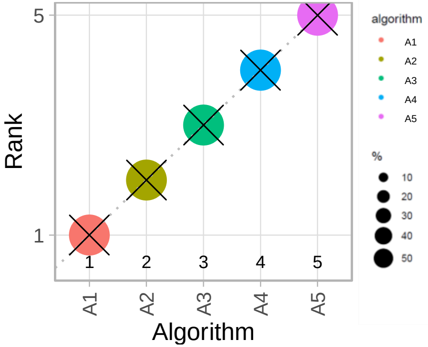
27% of all reports are based solely on ranking lists (without further visualization)

Why result analysis and visualization is critical: Example

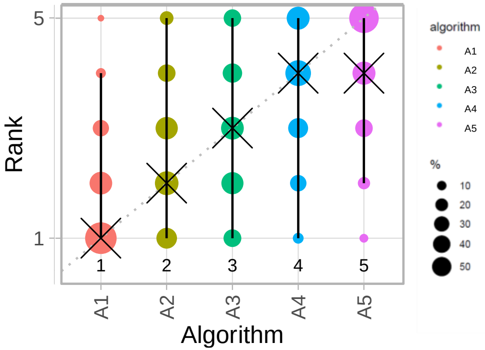


Same ranking





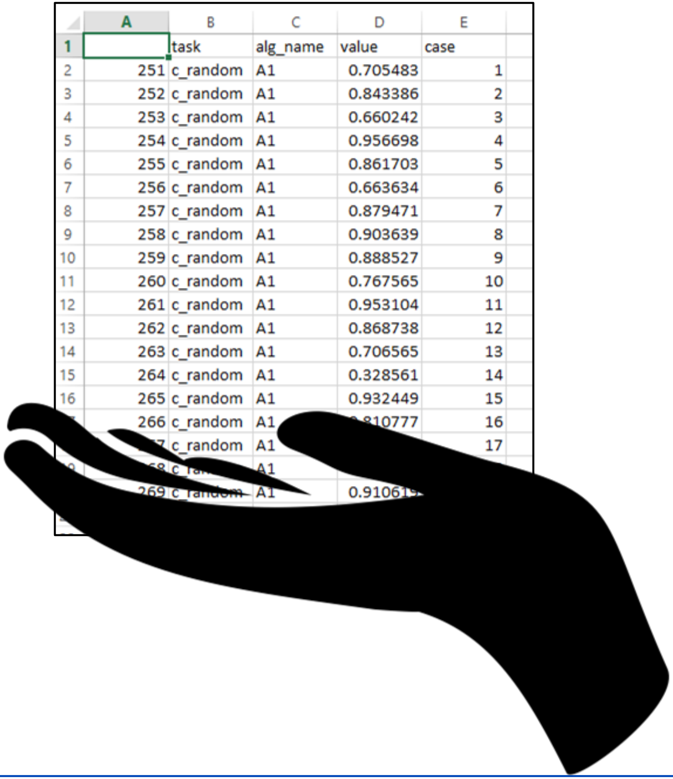
Rank	Algorithm
1	A1
2	A2
3	A3
4	A4
5	A5



Wiesenfarth, Reinke et al. Methods and open-source toolkit for analyzing and visualizing challenge results. **Scientific Reports 2021**

Try it yourself: Metric values in, full PDF report out

Input:
Metric values in csv file



	A	B	C	D	E
1		task	alg_name	value	case
2	251	c_random	A1	0.705483	1
3	252	c_random	A1	0.843386	2
4	253	c_random	A1	0.660242	3
5	254	c_random	A1	0.956698	4
6	255	c_random	A1	0.861703	5
7	256	c_random	A1	0.663634	6
8	257	c_random	A1	0.879471	7
9	258	c_random	A1	0.903639	8
10	259	c_random	A1	0.888527	9
11	260	c_random	A1	0.767565	10
12	261	c_random	A1	0.953104	11
13	262	c_random	A1	0.868738	12
14	263	c_random	A1	0.706565	13
15	264	c_random	A1	0.328561	14
16	265	c_random	A1	0.932449	15
[illegible]	266	c_random	A1	[illegible]810777	16
[illegible]	[illegible]7	c_random	A1	[illegible]	17
[illegible]	[illegible]	[illegible]	A1	[illegible]	[illegible]
[illegible]	269	c_random	A1	0.91061[illegible]	[illegible]

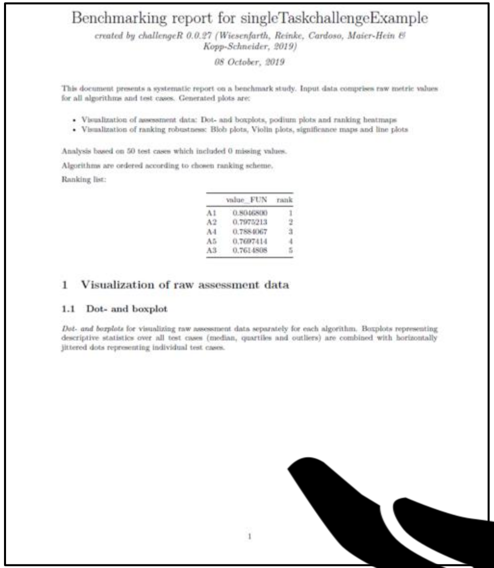
https://github.com/wiesenfa/challengeR

Icons created by the Noun Project

Wiesenfarth et al. Methods and open-source toolkit for analyzing and visualizing challenge results. **Scientific Reports 2021**

Try it yourself: Metric values in, full PDF report out

Output:
Full PDF report



Benchmarking report for singleTaskchallengeExample

created by challengeR 0.0.27 (Wiesenfarth, Reinke, Cardoso, Maier-Hein & Kopp-Schneider, 2019)

08 October, 2019

This document presents a systematic report on a benchmark study. Input data comprises raw metric values for all algorithms and test cases. Generated plots are:

- Visualization of assessment data: Dot- and boxplots, podium plots and ranking heatmaps
- Visualization of ranking robustness: Blob plots, Violin plots, significance maps and line plots

Analysis based on 50 test cases which included 0 missing values.

Algorithms are ordered according to chosen ranking scheme.

Ranking list:

	value_FUN	rank
A1	0.8046800	1
A2	0.7975213	2
A4	0.7884067	3
A5	0.7697414	4
A3	0.7614808	5

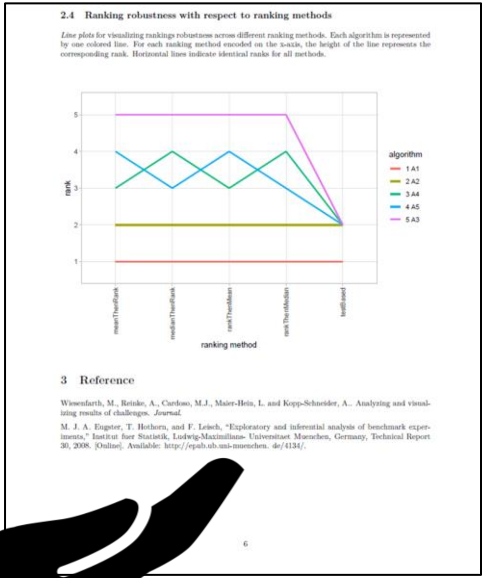
1 Visualization of raw assessment data

1.1 Dot- and boxplot

Dot- and boxplots for visualizing raw assessment data separately for each algorithm. Boxplots representing descriptive statistics over all test cases (median, quartiles and outliers) are combined with horizontally jittered dots representing individual test cases.

1

. . .



2.4 Ranking robustness with respect to ranking methods

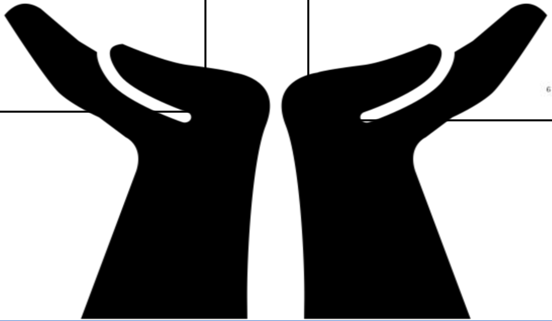
Line plots for visualizing rankings robustness across different ranking methods. Each algorithm is represented by one colored line. For each ranking method encoded on the x-axis, the height of the line represents the corresponding rank. Horizontal lines indicate identical ranks for all methods.

3 Reference

Wiesenfarth, M., Reinke, A., Cardoso, M.J., Maier-Hein, L. and Kopp-Schneider, A.. Analyzing and visualizing results of challenges. *Journal*.

M. J. A. Eugster, T. Hothorn, and F. Leisch, "Exploratory and inferential analysis of benchmark experiments," Institut fuer Statistik, Ludwig-Maximilians- Universitaet Muenchen, Germany, Technical Report 30, 2008. [Online]. Available: http://epub.ub.uni-muenchen. de/4134/.

6



https://github.com/wiesenfa/challengeR

Icons created by the Noun Project

Wiesenfarth et al. Methods and open-source toolkit for analyzing and visualizing challenge results. **Scientific Reports 2021**

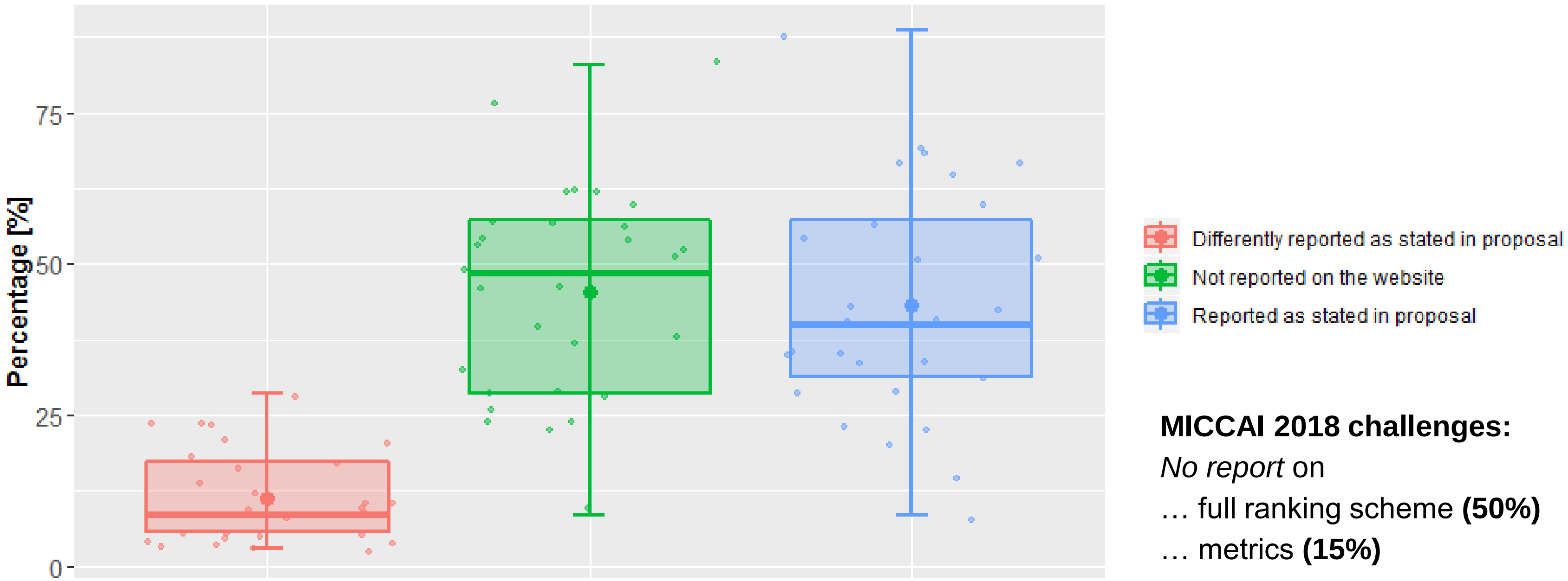
Cheating

You don’t think people cheat?

20%

of the MICCAI 2020 challenge organizers reported cheating!

Example: Weaknesses in challenge design can be exploited



MICCAI 2018 challenges:

No report on

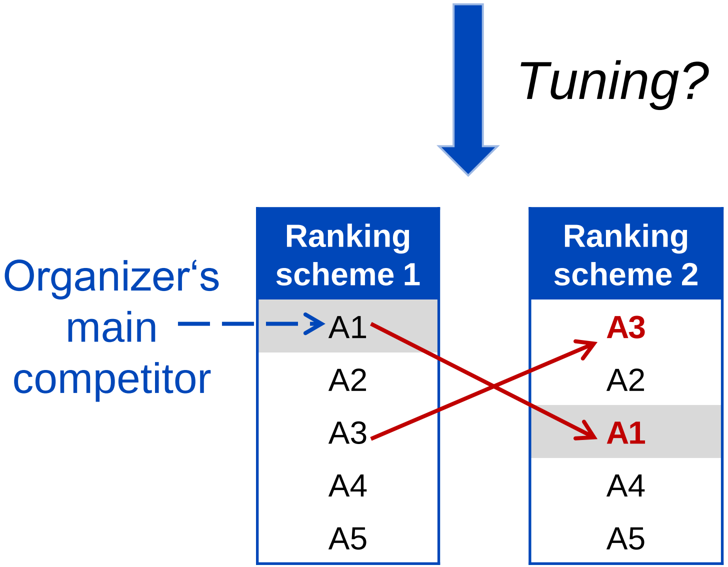
… full ranking scheme **(50%)**

… metrics **(15%)**

Reinke et al. How to Exploit Weaknesses in Biomedical Challenge Design and Organization. **MICCAI 2018**

Example: Weaknesses in challenge design can be exploited

Ranking schemes are often not published before the challenge



Reinke et al. How to Exploit Weaknesses in Biomedical Challenge Design and Organization. **MICCAI 2018**

Example: Missing value handling

82% of tasks provide no information about how missing data is handled

Image	I_1	I_2	I_3	I_4	I_5	I_6
DSC	0.94	**NaN**	0.87	0.90	**NaN**	0.89

Ignore NaNs → **Mean *DSC*:** 0.90

Set NaNs to worst possible value (here: 0) → **Mean *DSC*:** 0.60

Reinke et al. How to Exploit Weaknesses in Biomedical Challenge Design and Organization. **MICCAI 2018**
Reinke et al. Common Limitations of Image Processing Metrics: A Picture Story. **arXiv 2021**

Example: Missing value handling

What happens if algorithms systematically submit only the most plausible results?

- **25%** of non-winning algorithms would have been ranked first
- In **9%** of tasks, every single participating algorithm could have been ranked first

Reinke et al. How to Exploit Weaknesses in Biomedical Challenge Design and Organization. **MICCAI 2018**

Metrics


Volumetric overlap error
Specificity
Hamming loss
Jaccard
Average precision
Dice similarity coefficient
Rand error
False positive rate
Area under curve
Sweet spot coverage
Average mesh distance
F1 score
Jaccard
Recall
Volume
ROC
Adjusted rand index
Spearmans correlation coefficient
Root mean square error
F1 score
Volumetric overlap error
Absolute surface distance
False positive rate
Average perpendicular distance
Mean average precision
Hausdorff distance
Hausdorff distance
Kappa
Average mesh distance
Accuracy
Precision
Root mean square error
Averaged confusion matrix diagonal
Error rate
Average perpendicular distance
Rand error
Error rate
Area under curve
Completeness
Volume
Dice similarity coefficient
Homogeneity
Average recall
Dice similarity coefficient
Accuracy
Kappa
Absolute surface distance
Specificity
Completeness
Area under curve
Hausdorff distance 95
Hausdorff distance 95
Precision
Volumetric overlap error
Precision
F1 score
False negative rate
Mean average precision
Error rate
Hausdorff distance
ROC
False positive rate
Hamming loss
Adjusted rand index
Maximum surface distance
Average precision
Accuracy
Interclass correlation
Average mesh distance
Interclass correlation
Kappa
Average recall
Average symmetric surface distance
Adjusted rand index
Average recall
Interclass correlation
Average symmetric surface distance
Jaccard
Averaged confusion matrix diagonal
Spearmans correlation coefficient
Maximum surface distance
Absolute surface distance
Volume
Euclidean distance
Root mean square error
Specificity
False negative rate
Homogeneity
Rand error
Sweet spot coverage
Euclidean distance
Hausdorff distance 95
Recall
Recall
Sweet spot coverage
Average precision
Maximum surface distance
Euclidean distance

How metrics are currently selected

The two metrics are **widely used** to measure segmentation accuracy.

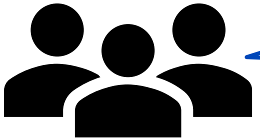
One of the **commonly used** metrics to validate scene segmentation

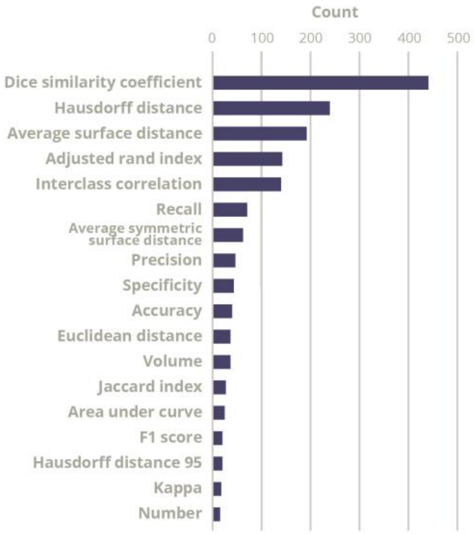
It is **widely used** in the field.

Is **commonly used** for the validation of reconstruction data.

Standard computer vision segmentation metrics.

Is a **common metric** in clinical diagnosis validation.





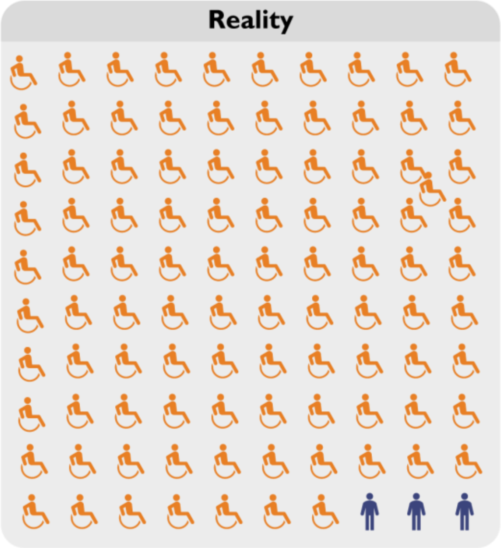
Citations from http://www.miccai.org/special-interest-groups/challenges/miccai-registered-challenges/
Maier-Hein et al. Why rankings of biomedical image analysis competitions should be interpreted with care. **Nature Commun 2018**

Class imbalance

Goal: Classify patients into sick (positive class) and healthy (negative class)

Sick class: 97 patients

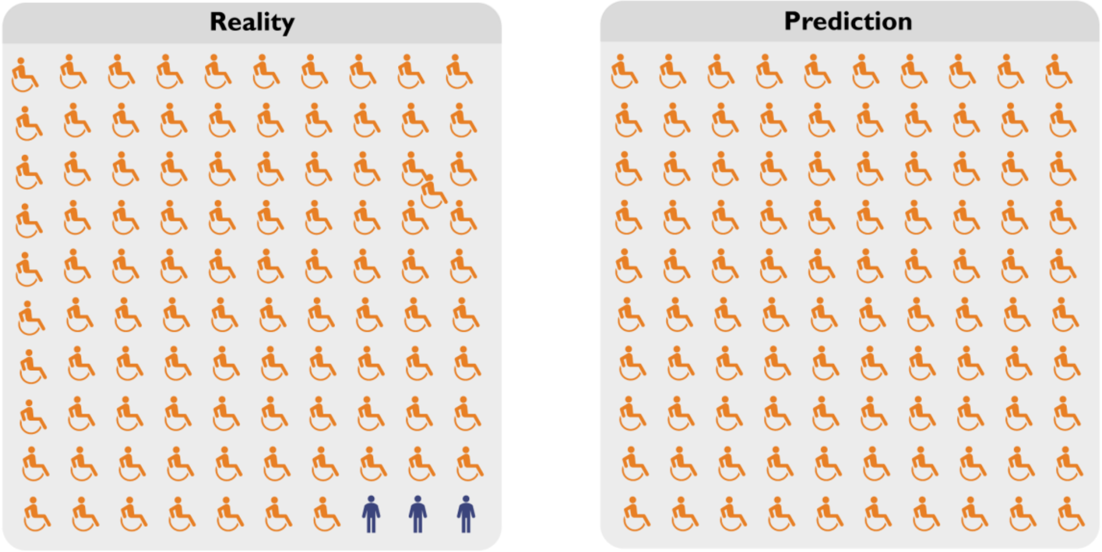
Healthy class: 3 patients



Accuracy = 97%

Class imbalance

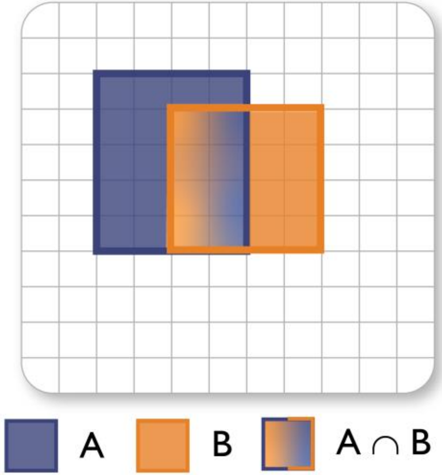
$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{97 + 0}{97 + 0 + 3 + 0} = 0.97$$



Accuracy = 97%

Specificity = 0%

Most common metric: Dice Similarity Coefficient (DSC)

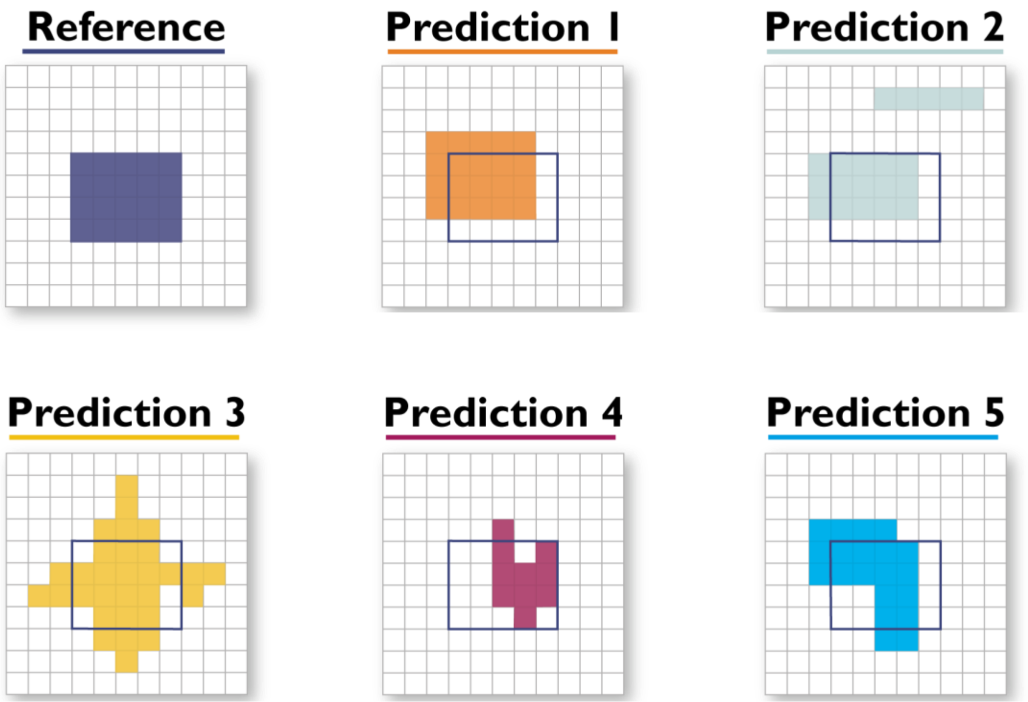


$$DSC(A,B) = \frac{2\,|A \cap B|}{|A| + |B|}$$

$$IoU(A,B) = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} = \frac{|A \cap B|}{|A \cup B|}$$

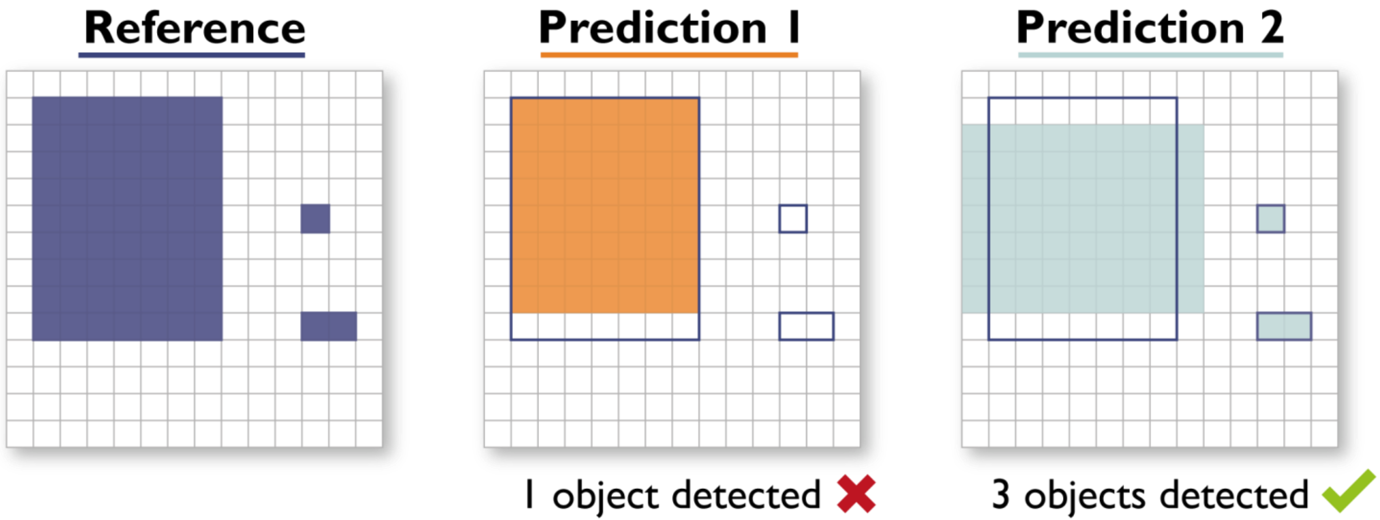
Reinke et al. A discovery dive into the world of evaluation — Do’s, don’ts and other considerations. **Medium Blogpost 2021**

Shape unawareness



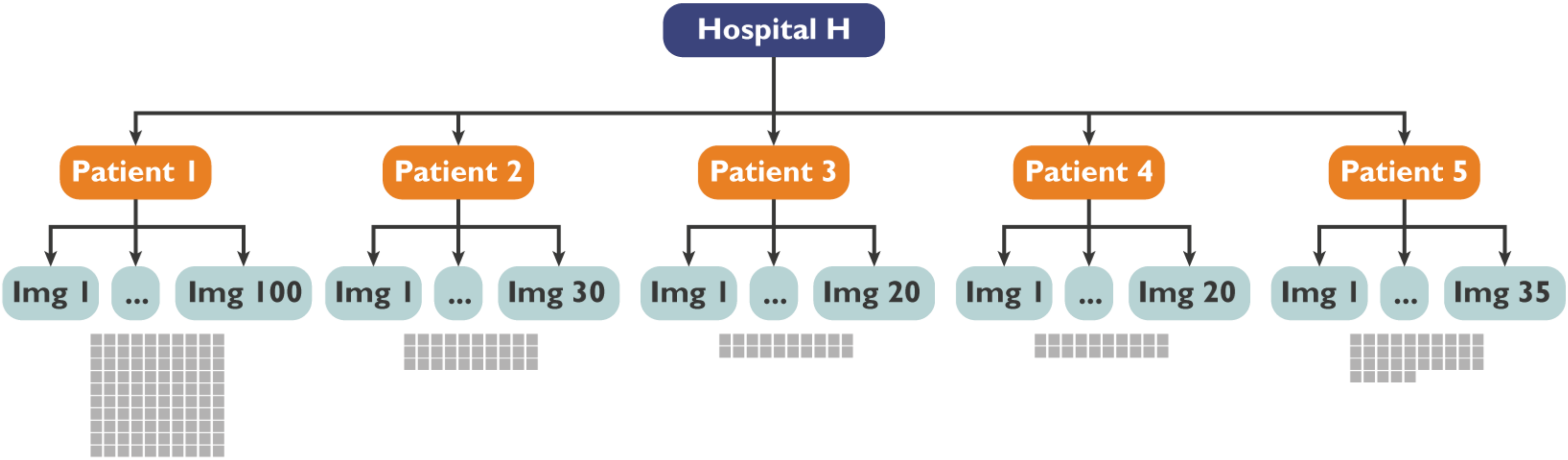
Maier-Hein/Reinke et al. Metrics reloaded: Pitfalls and recommendations for image analysis validation. arXiv 2022
Reinke et al. Common Limitations of Image Processing Metrics: A Picture Story. arXiv 2021

Inappropriate phrasing of the problem:
Object detection vs. segmentation



Maier-Hein/Reinke et al. Metrics reloaded: Pitfalls and recommendations for image analysis validation. arXiv 2022
Reinke et al. Common Limitations of Image Processing Metrics: A Picture Story. arXiv 2021

Metric aggregation



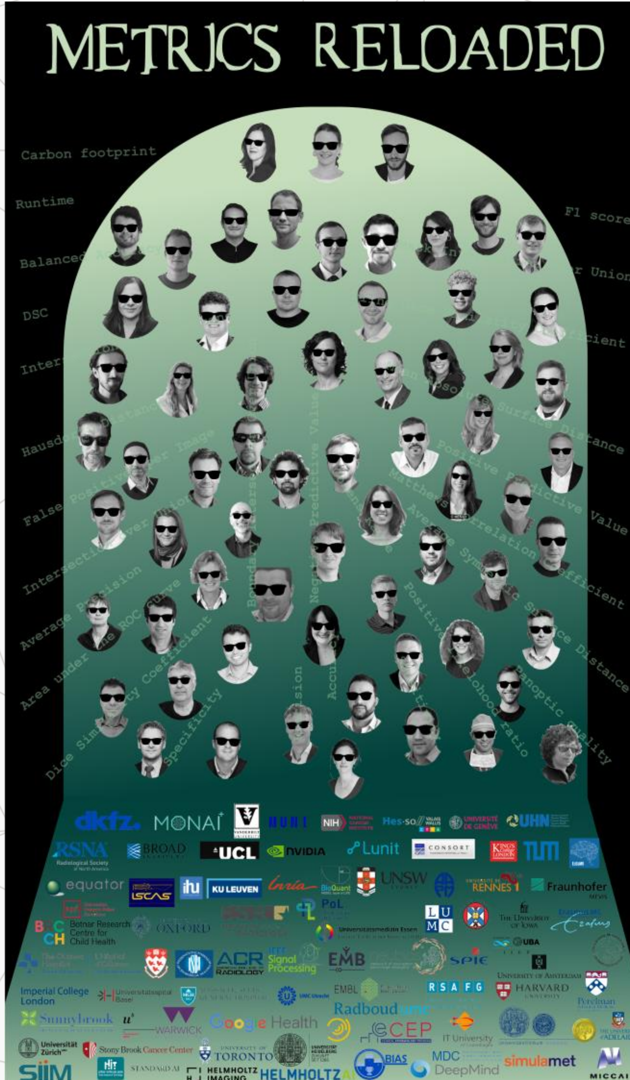
Maier-Hein/Reinke et al. Metrics reloaded: Pitfalls and recommendations for image analysis validation. arXiv 2022
Reinke et al. Common Limitations of Image Processing Metrics: A Picture Story. arXiv 2021

Uncovering problems is good…

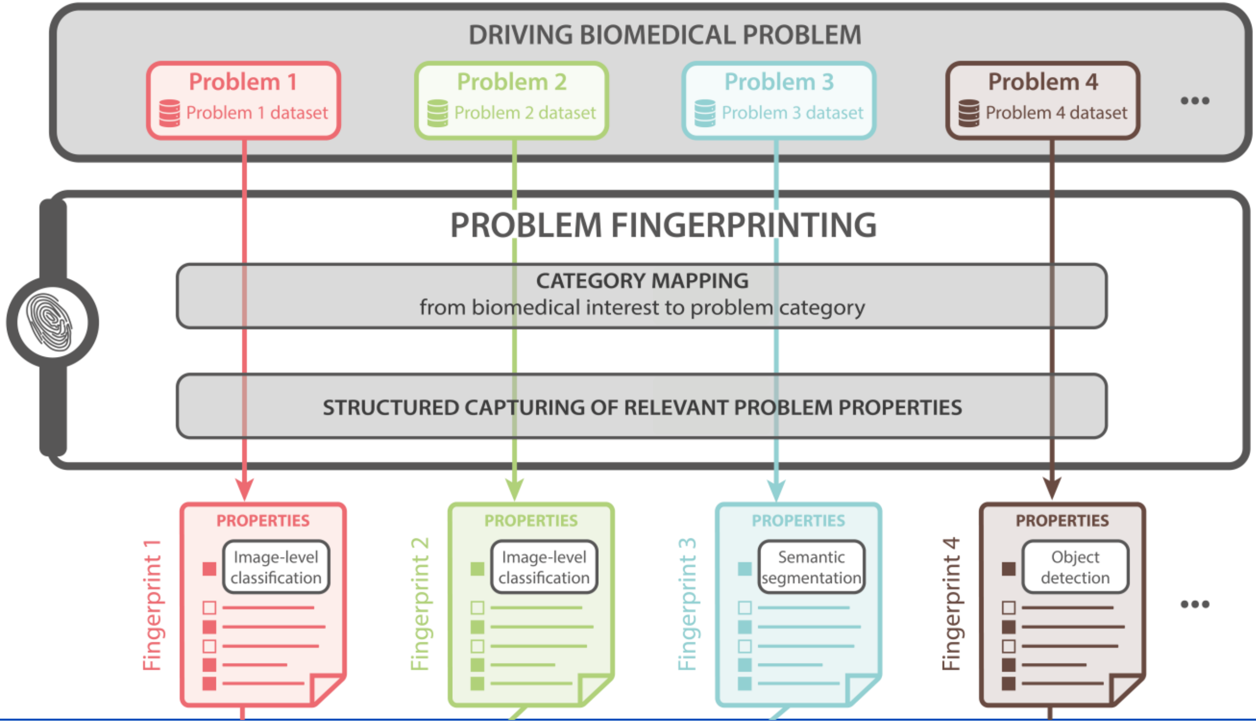
Common Limitations of Image Processing Metrics: A Picture Story

ANNIKA REINKE*, German Cancer Research Center (DKFZ), Germany and Heidelberg University, Germany
MINU D. TIZABI, German Cancer Research Center (DKFZ), Germany
CAROLE H. SUDRE, University College London, UK and King's College London, UK
MATTHIAS EISENMANN, German Cancer Research Center (DKFZ), Germany
TIM RÄDSCH, German Cancer Research Center (DKFZ), Germany and understandAI GmbH, Germany
MICHAEL BAUMGARTNER, German Cancer Research Center (DKFZ), Germany
LAURA ACION, CONICET – Universidad de Buenos Aires, Argentina and University of Iowa, USA
MICHELA ANTONELLI, King's College London, UK and University College London, UK
TAL ARBEL, McGill University, Canada

Solving them is even better!

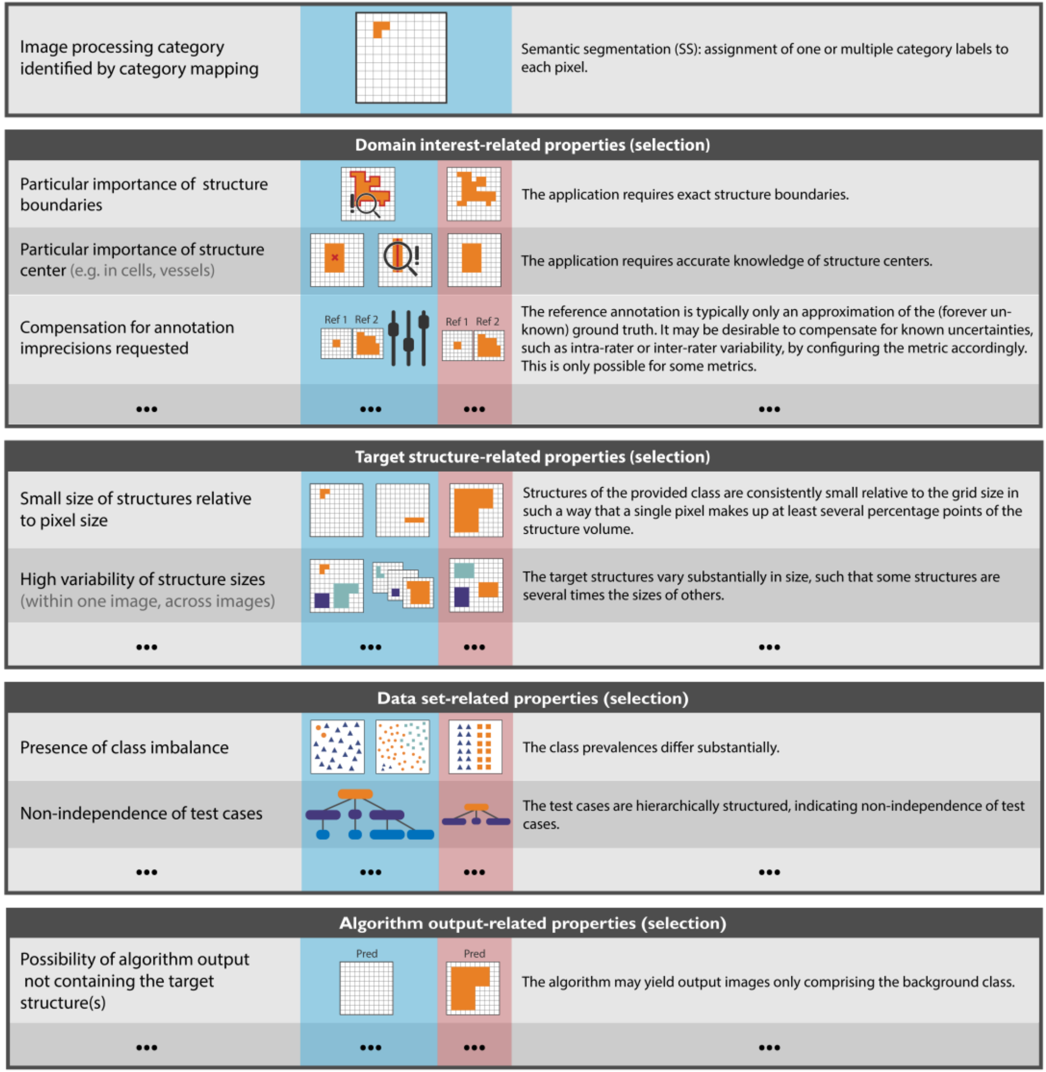


Problem-aware metric recommendation framework

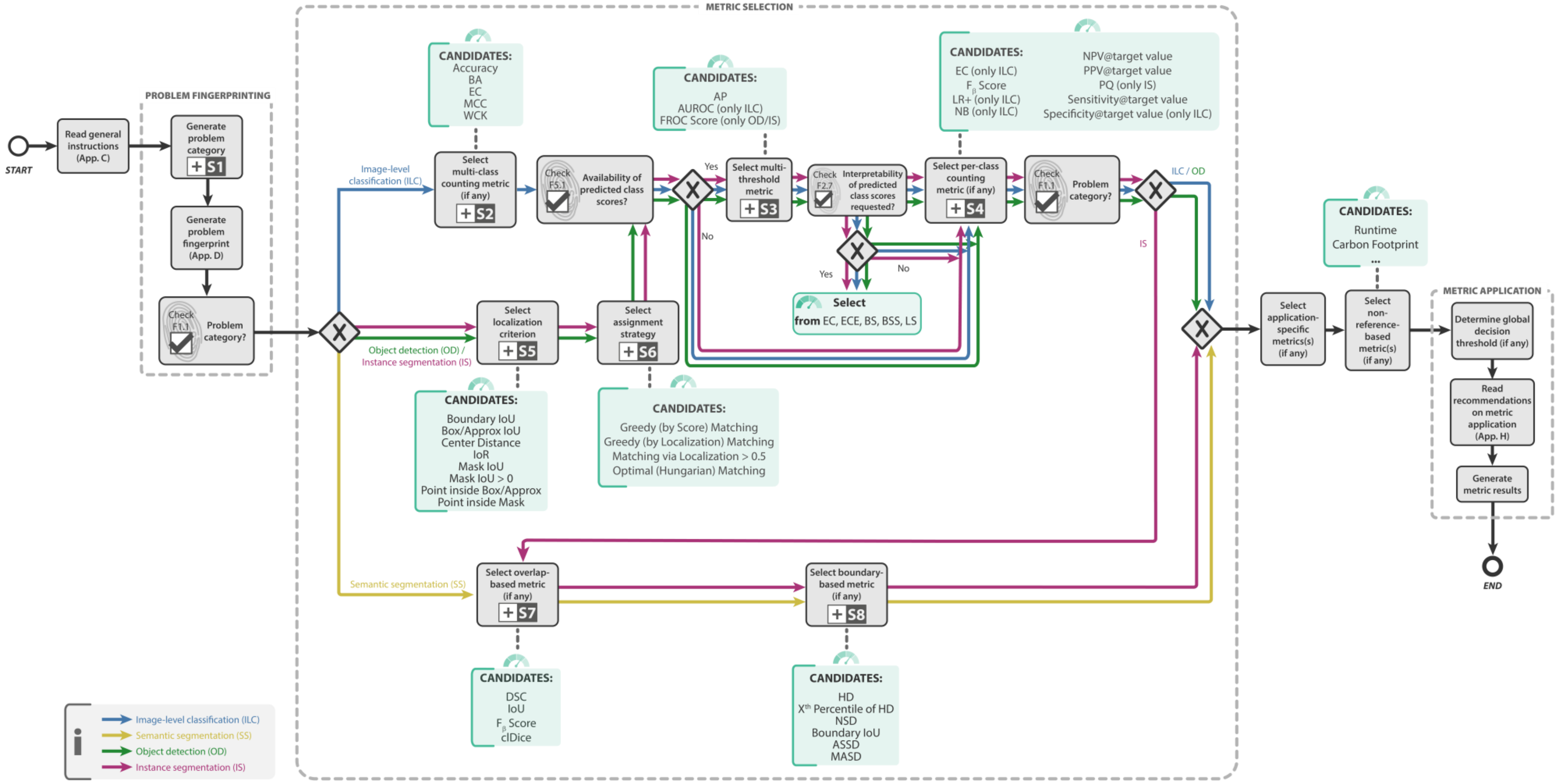


Maier-Hein/Reinke et al. Metrics reloaded: Pitfalls and recommendations for image analysis validation. arXiv 2022

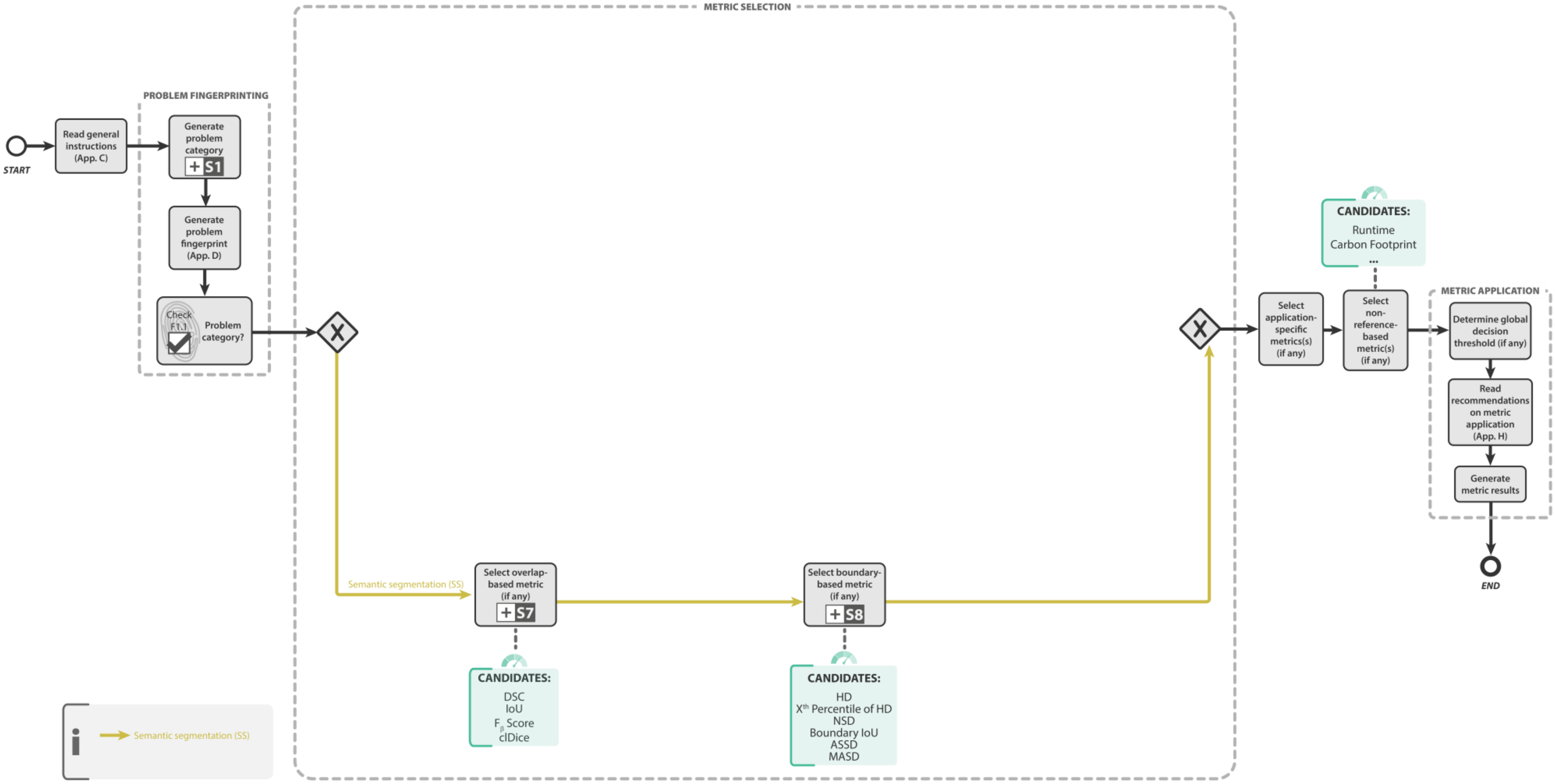
Problem fingerprint



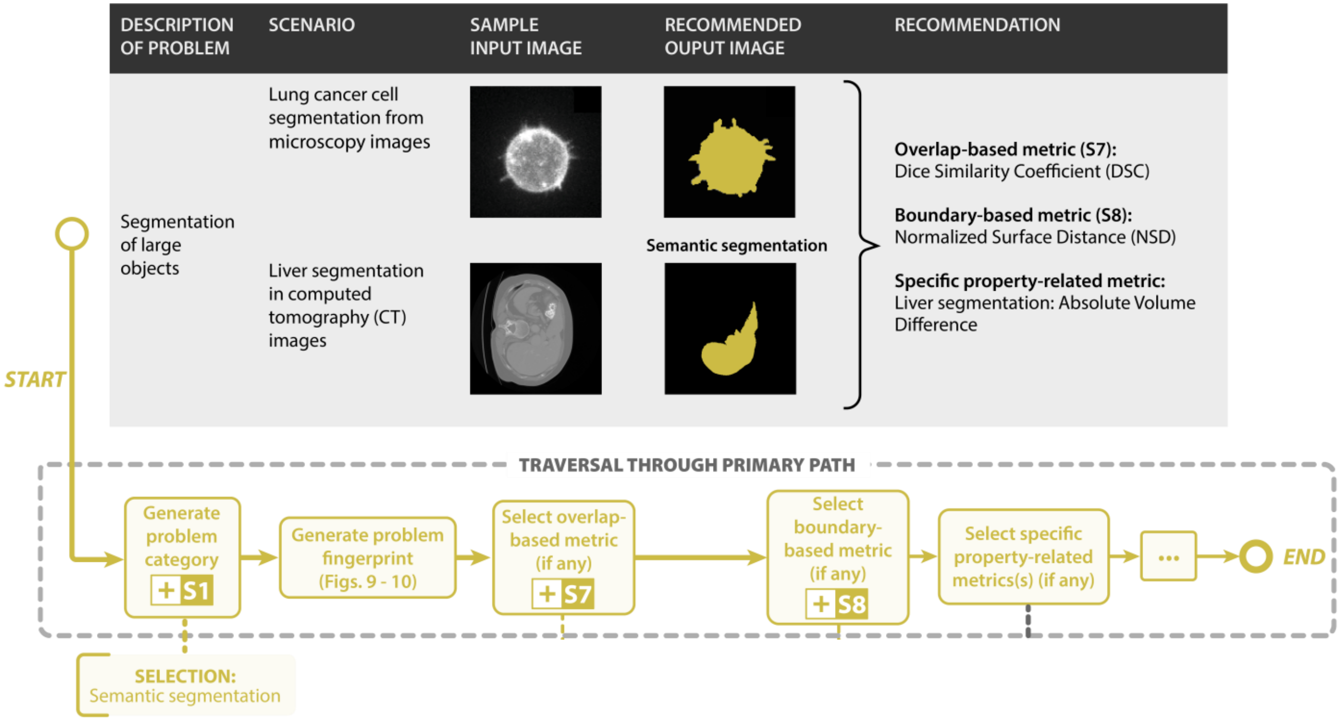
Maier-Hein/Reinke et al. Metrics reloaded: Pitfalls and recommendations for image analysis validation. **arXiv 2022**



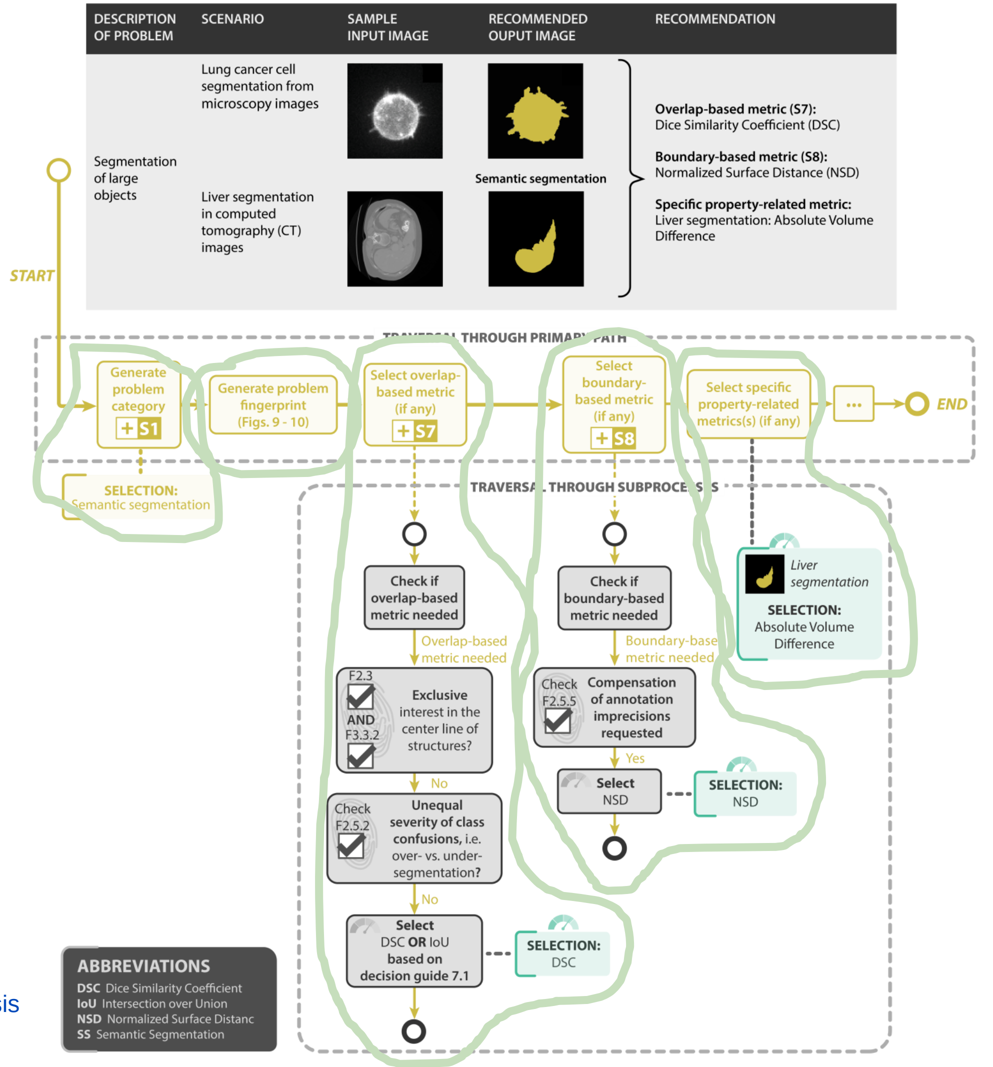
Maier-Hein/Reinke et al. Metrics reloaded: Pitfalls and recommendations for image analysis validation. arXiv 2022



Maier-Hein/Reinke et al. Metrics reloaded: Pitfalls and recommendations for image analysis validation. arXiv 2022

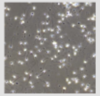
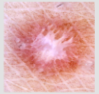
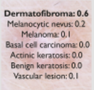
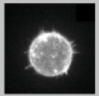
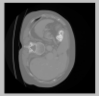
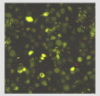
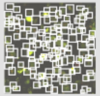
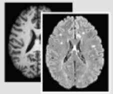
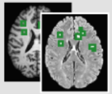
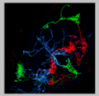
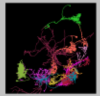


Maier-Hein/Reinke et al.
Metrics reloaded: Pitfalls and recommendations for image analysis validation. arXiv 2022



Maier-Hein/Reinke et al.
Metrics reloaded: Pitfalls and recommendations for image analysis validation. arXiv 2022

Instantiation for common biomedical use cases

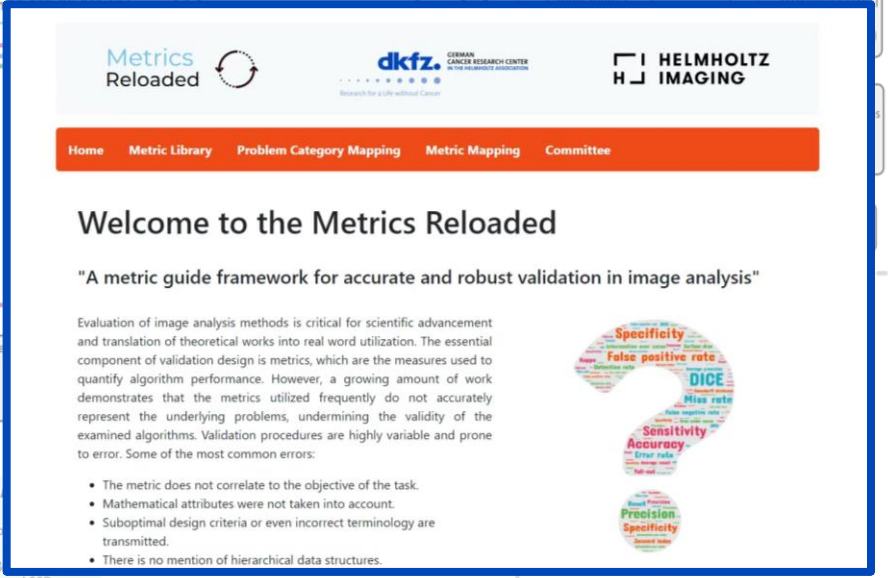
DESCRIPTION OF PROBLEM	SCENARIO	SAMPLE INPUT IMAGE	RECOMMENDED OUPUT IMAGE	RECOMMENDATION
Classification of images	Frame-based sperm motility classification based on microscopy time-lapse video containing human spermatozoa			**Problem category:** Image-level classification **Multi-class counting metric (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			

Maier-Hein/Reinke et al. Metrics reloaded: Pitfalls and recommendations for image analysis validation. **arXiv 2022**

Metrics Reloaded –
A new recommendation framework for biomedical image analysis validation

https://arxiv.org/abs/2206.01653

Working on web-based tool for guiding the user through the process!

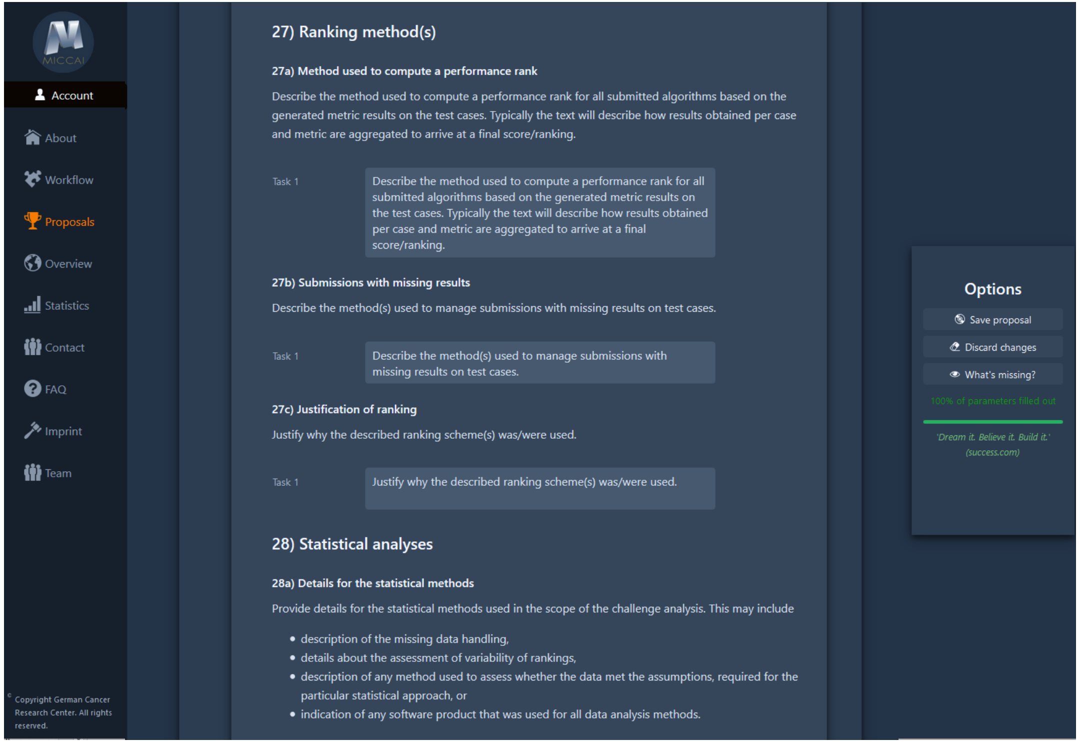


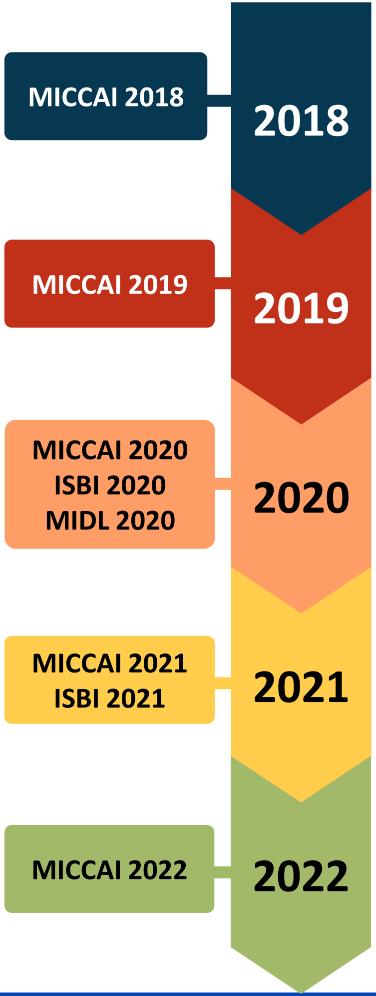
@DKFZ_CAMI_lab
#BiomedicalChallenges
#benchmarking

HELMHOLTZ IMAGING

erc European Research Council

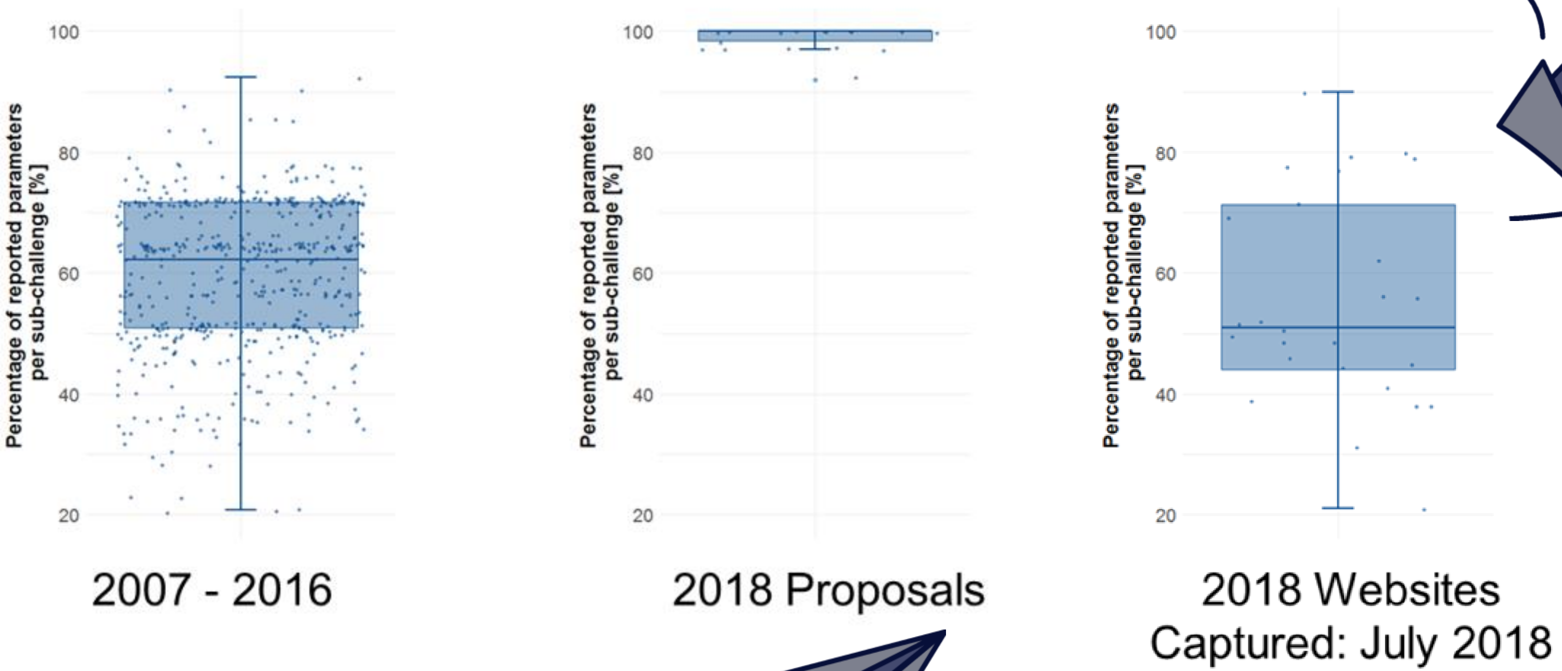
New: Structured challenge submission system





DEVELOPERS:
Annika Reinke, Sinan Onogur, Matthias Eisenmann, Keno März, Sebastian Pirmann
Div. Computer Assisted Medical Interventions (CAMI), German Cancer Research Center, DKFZ)

Problem: Quality control after challenge acceptance



dkfz.