Performance Reporting in Medical Imaging AI:

Current Practices, Strength of Outperformance Claims, and Areas for Improvement

Evangelia Christodoulou

German Cancer Research Center (DKFZ), Heidelberg, Germany

National Center for Tumor Diseases (NCT), NCT Heidelberg, Germany

GERMAN CANCER RESEARCH CENTER IN THE HELMHOLTZ ASSOCIATION

Olivier Colliot

Paris Brain Institute, CNRS, Inria, Inserm, Sorbonne University, France

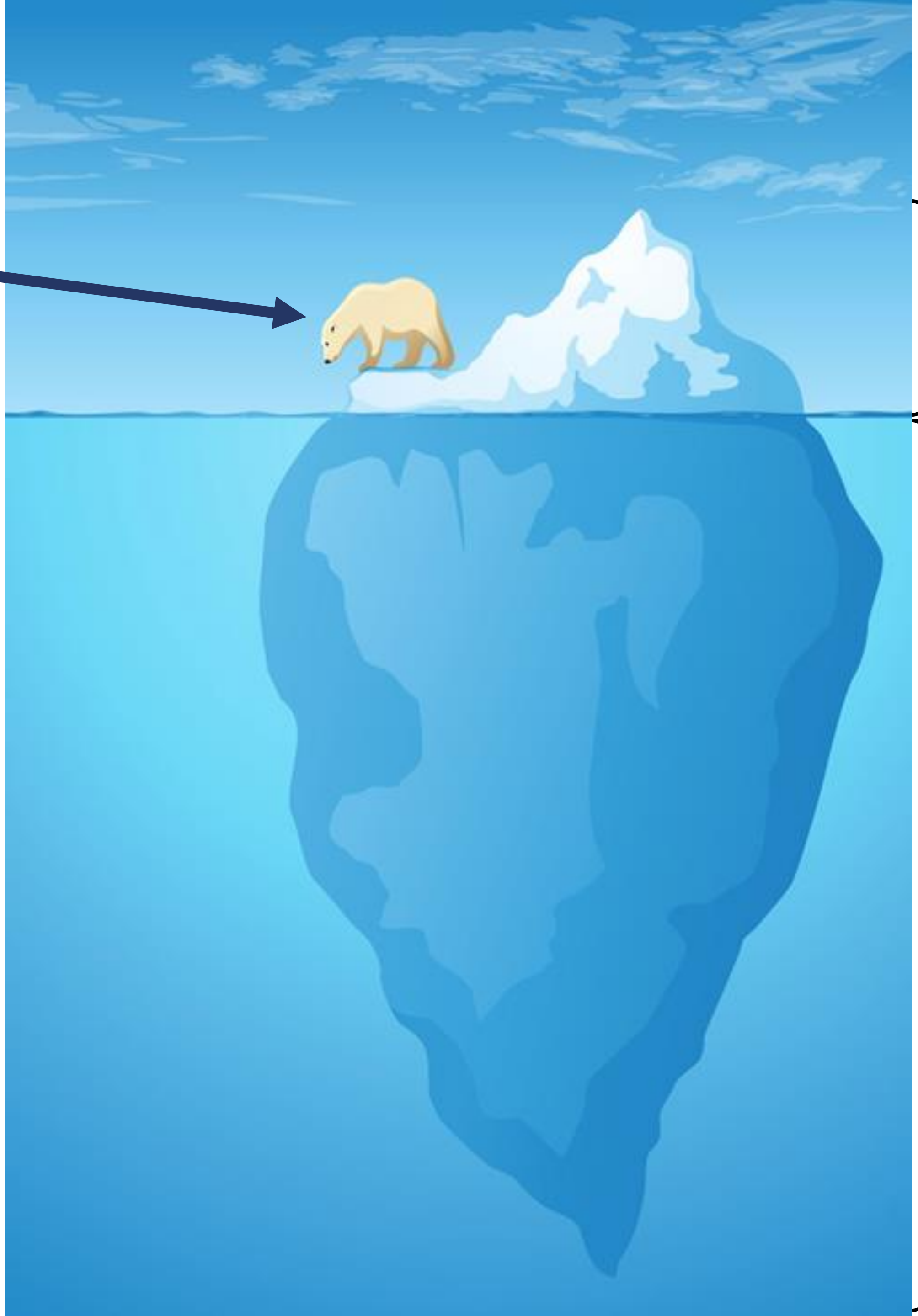
Paris Institute for Artificial Intelligence (PRAIRIE-PSAI), France

MICCAI

SIG for Challenges

Evaluation and Benchmarking WG

AI developer



Machine
Learning (ML)

Dataset design

Annotations

Metrics

Data Splitting

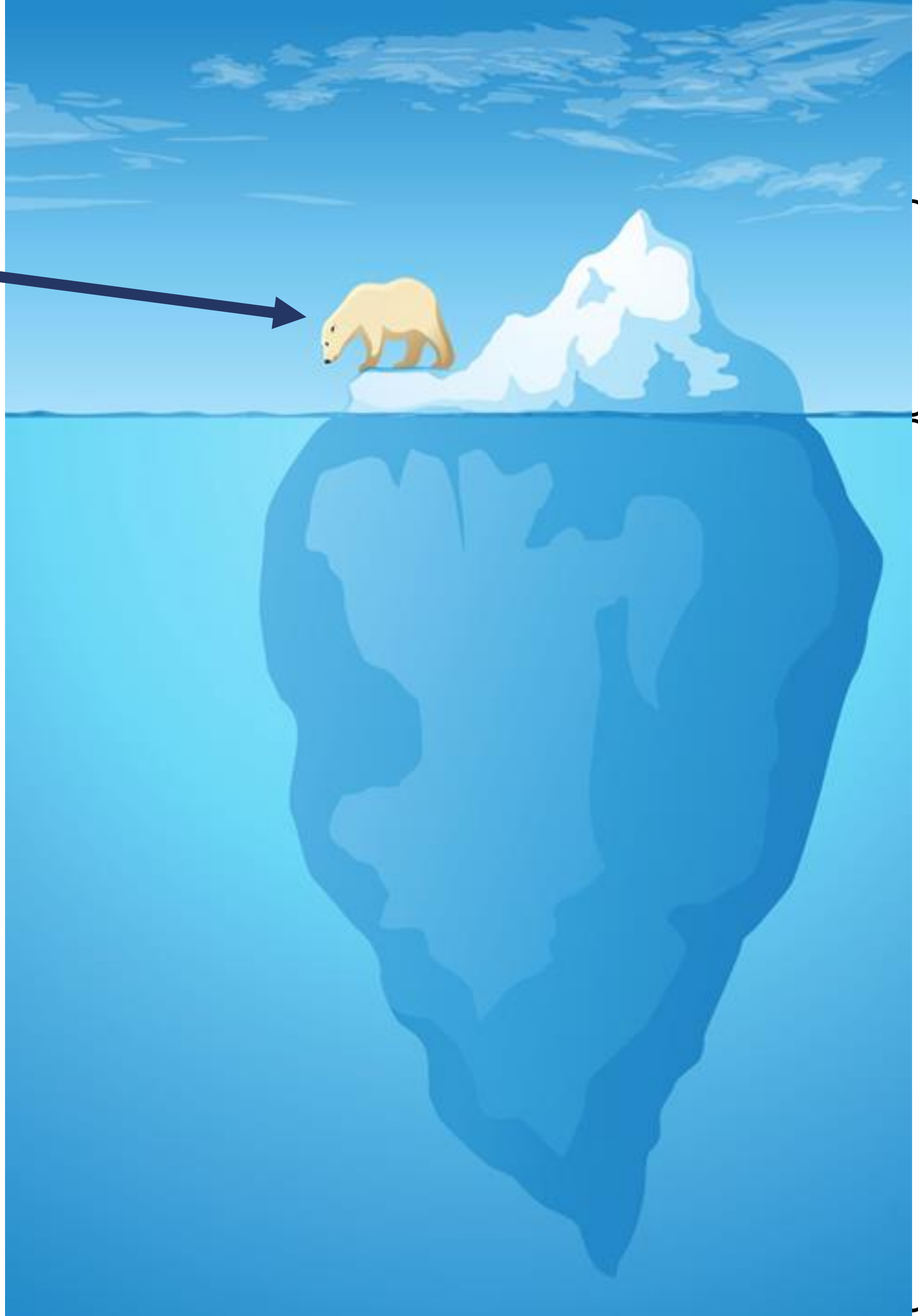
Reporting

Rankings

...

Image source:
Regulatory Affairs Professional Society

AI developer



Machine Learning (ML)

Dataset design

Annotations

Metrics

Data Splitting

Reporting

Rankings

...

Image source:
Regulatory Affairs Professional Society

Previous SIG webinar: metrics reloaded

Central question: which validation metrics?


nature methods

Perspective

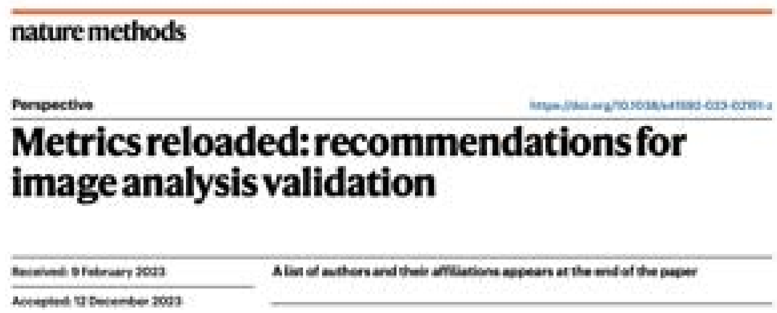
Understanding metric-related pitfalls in image analysis validation

Received: 9 February 2023

Accepted: 12 December 2023

A list of authors and their affiliations appears at the end of the paper

Reinke et al, Nature Methods, 2024
https://www.nature.com/articles/s41592-023-02150-0


nature methods

Perspective

Metrics reloaded: recommendations for image analysis validation

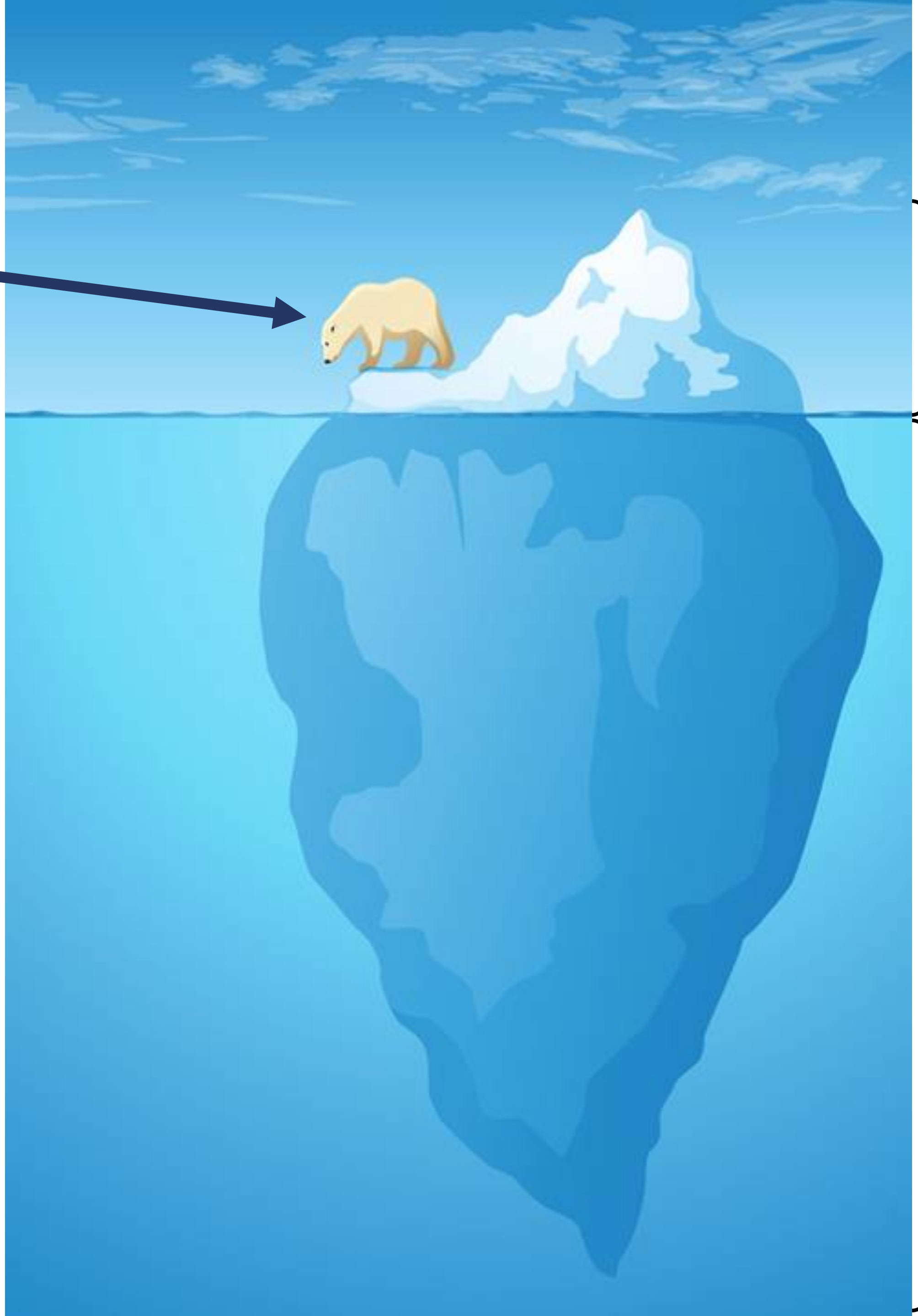
Received: 9 February 2023

Accepted: 12 December 2023

A list of authors and their affiliations appears at the end of the paper

Maier-Hein*, Reinke* et al, Nature Methods, 2024
https://www.nature.com/articles/s41592-023-02151-z

AI developer



Machine Learning (ML)

Dataset design

Annotations

Metrics

Data Splitting

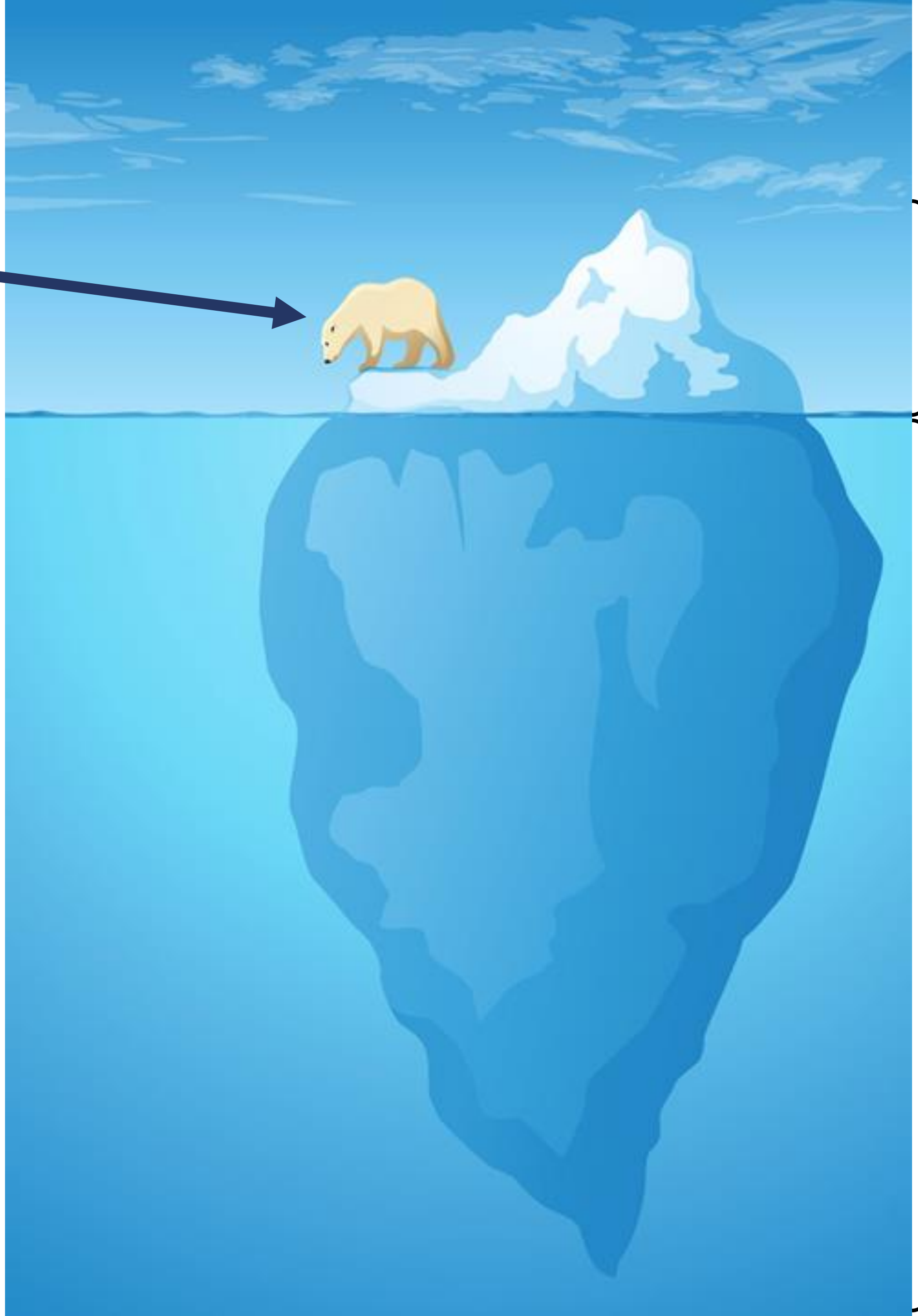
Reporting

Rankings

...

Image source:
Regulatory Affairs Professional Society

AI developer



Machine Learning (ML)

Dataset design

Annotations

Metrics

Data Splitting

Reporting

Rankings

...

Image source:
Regulatory Affairs Professional Society

This SIG webinar

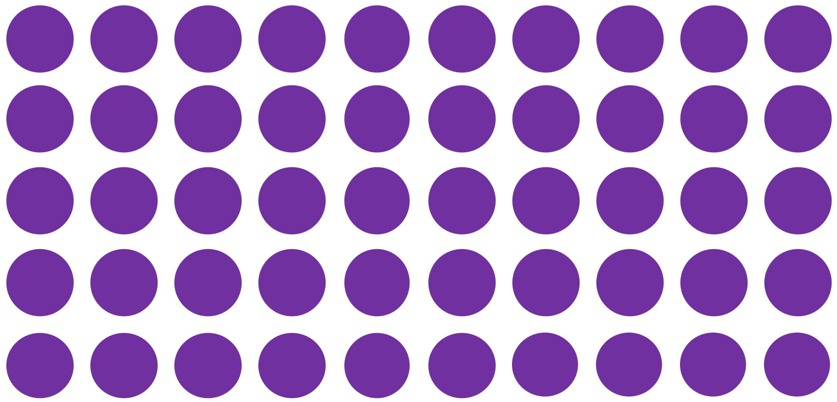
Central question: how variable is model performance?

This SIG webinar

Central question: <u>how variable</u> is model performance?

AI models are evaluated experimentally

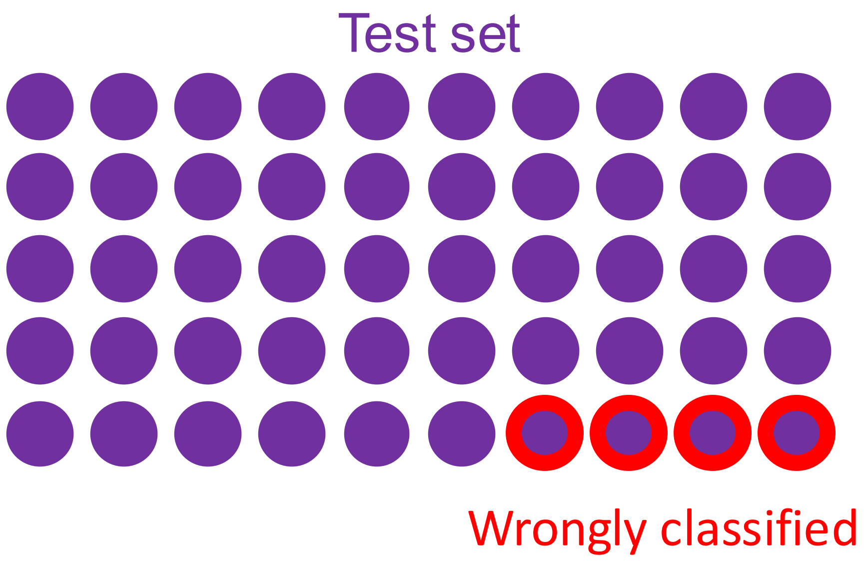
Test set



This SIG webinar

Central question: <u>how variable</u> is model performance?

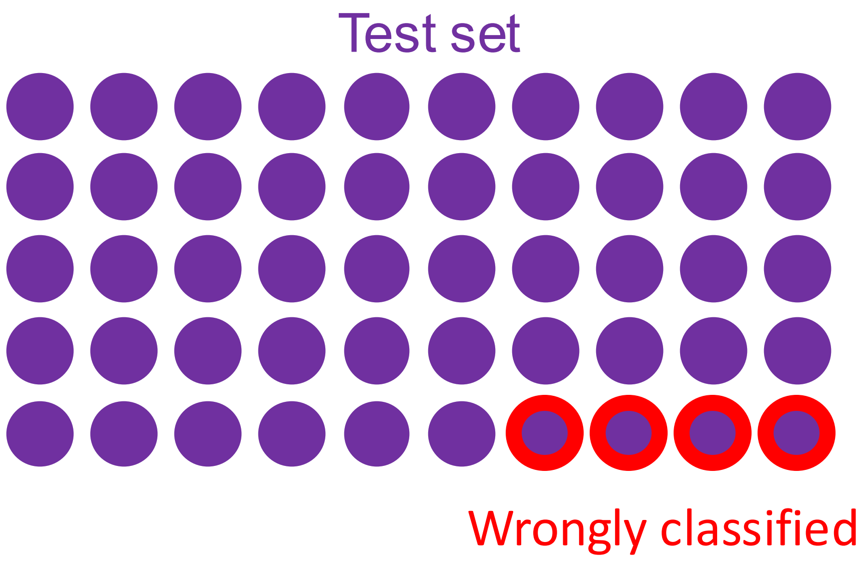
AI models are evaluated experimentally



This SIG webinar

Central question: how variable is model performance?

AI models are evaluated experimentally

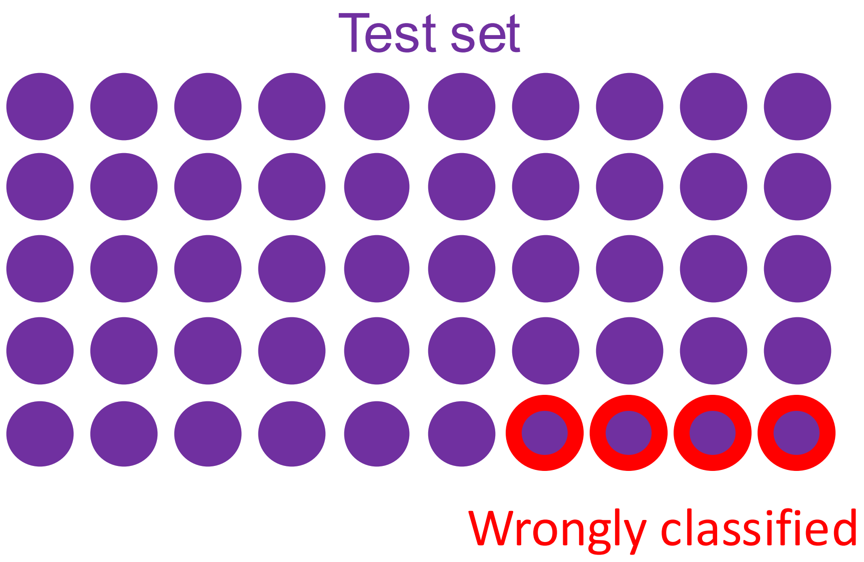


	Accuracy
My model	0.92

This SIG webinar

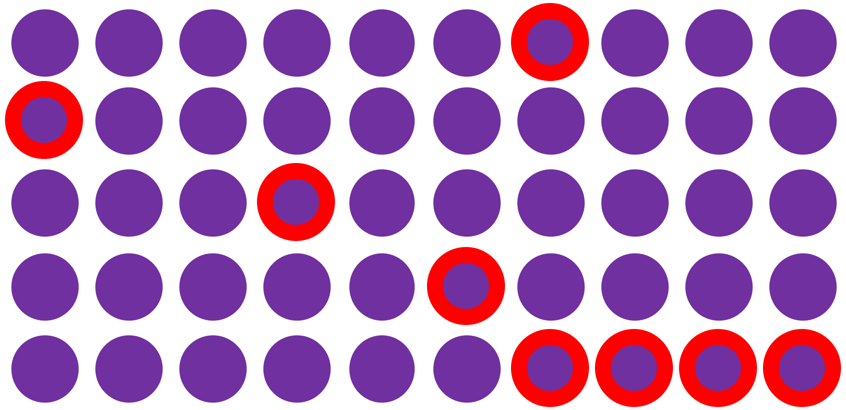
Central question: how variable is model performance?

AI models are evaluated experimentally



	Accuracy
My model	0.92

Estimates are variable

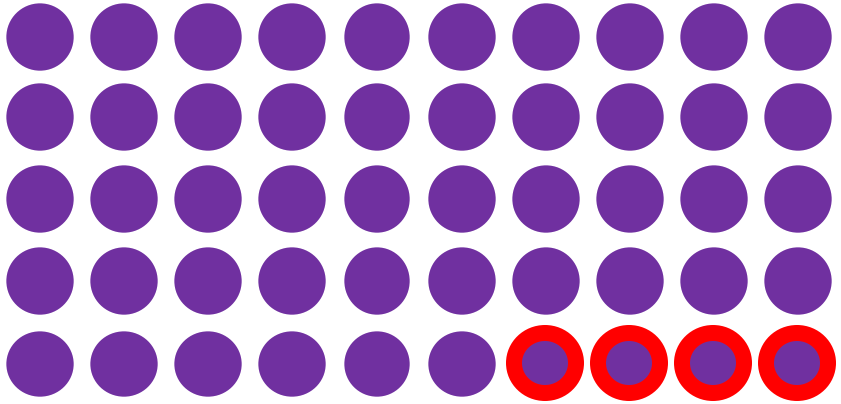


This SIG webinar

Central question: how variable is model performance?

AI models are evaluated experimentally

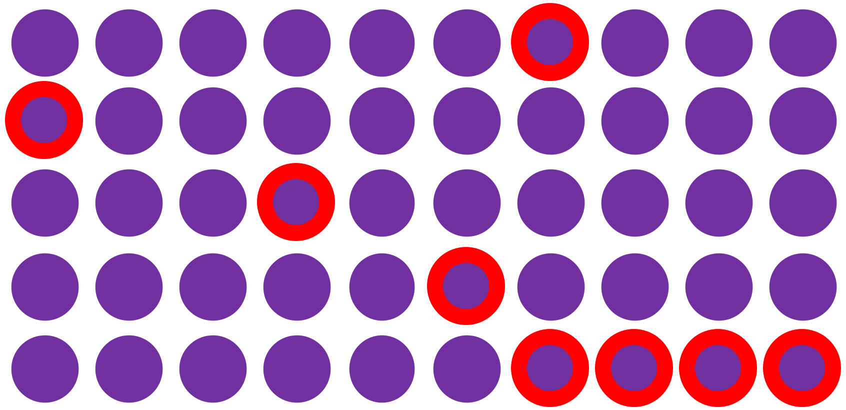
Test set



Wrongly classified

	Accuracy
My model	0.92

Estimates are variable



	Accuracy
My model	0.84

Performance variability is crucial for clinical translation

Commonly encountered results tables

Methods	Accuracy	AUC
Method 1	0.828	0.862
Method 2	0.821	0.857
Method 3	0.847	0.889
Proposed	**0.851**	**0.891**

Performance variability is crucial for clinical translation

Commonly encountered results tables

Methods	Accuracy	AUC
Method 1	0.828	0.862
Method 2	0.821	0.857
Method 3	0.847	0.889
Proposed	**0.851**	**0.891**

[….] All performance estimates should be provided with confidence intervals […]

FDA-2024-D-4488: Artificial Intelligence-Enabled Device Software Functions: Lifecycle Management and Marketing Submission Recommendations

FDA U.S. FOOD & DRUG ADMINISTRATION

Performance variability is crucial for clinical translation

Commonly encountered results tables

Methods	Accuracy	AUC
Method 1	0.828	0.862
Method 2	0.821	0.857
Method 3	0.847	0.889
Proposed	**0.851**	**0.891**

The statistical analysis plays a critical role in the assessment of […] ML performance but may be under-appreciated by many ML developers. […] There are still publications that present point estimates of ML performance without quantification of uncertainties.

Weijie Chen, Daniel Krainak, Berkman Sahiner, Nicholas Petrick, A Regulatory Science Perspective on Performance Assessment of Machine Learning Algorithms in Imaging, 2023

FDA U.S. FOOD & DRUG ADMINISTRATION

1. Current practices

2. Strength of outperformance claims

3. Areas for improvement

Take home messages

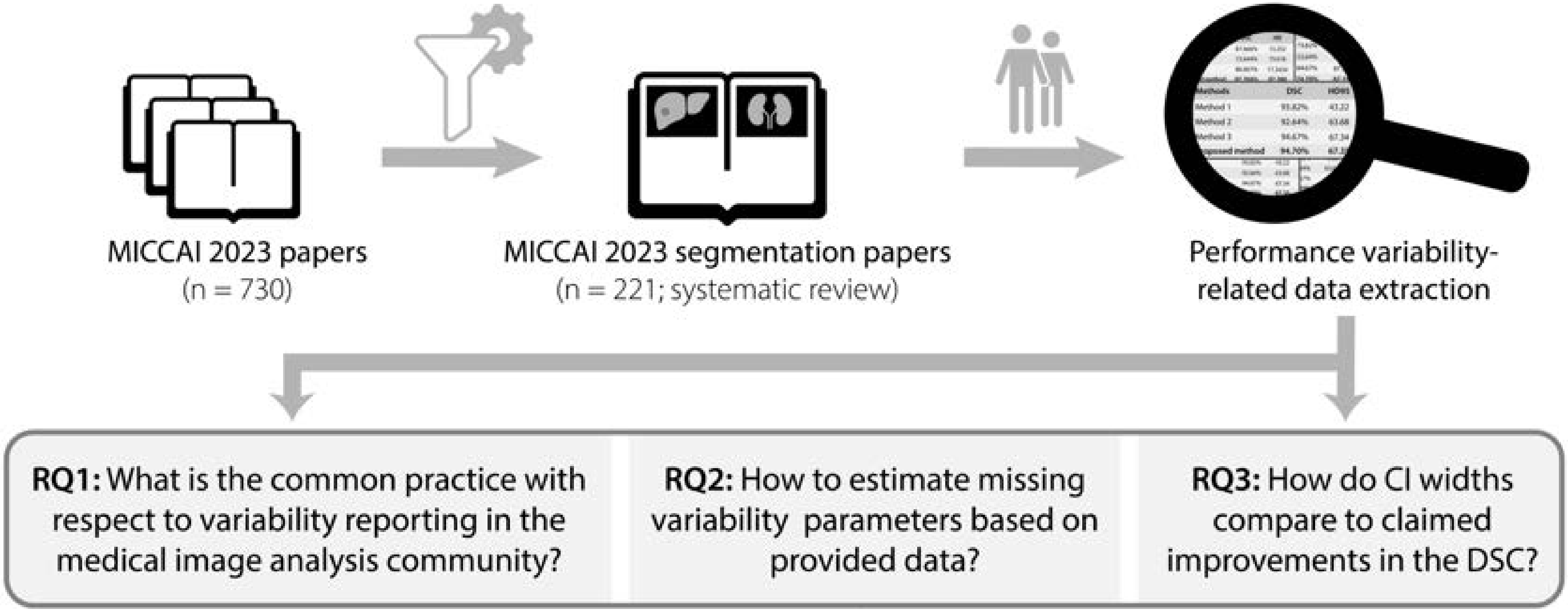
1. Current practices 

2. Strength of outperformance claims

3. Areas for improvement

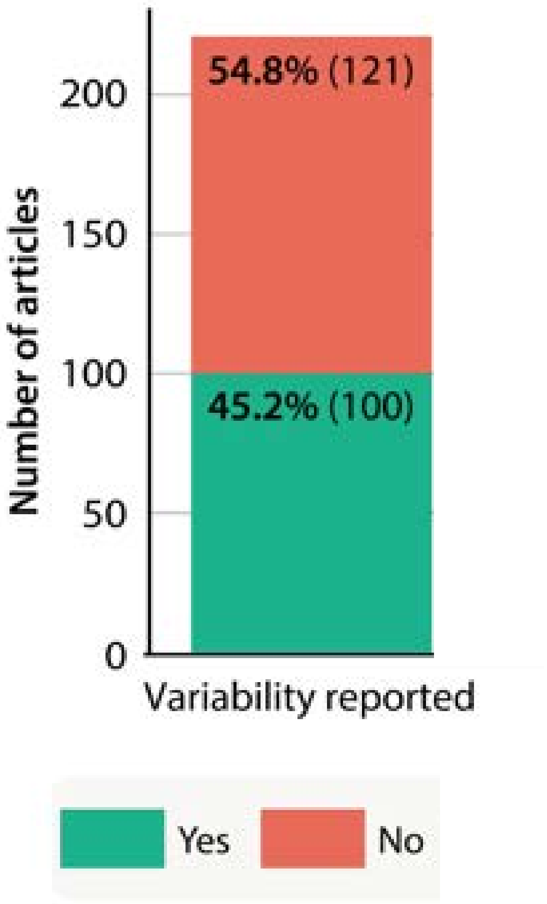
Take home messages

Variability reporting in medical imaging AI



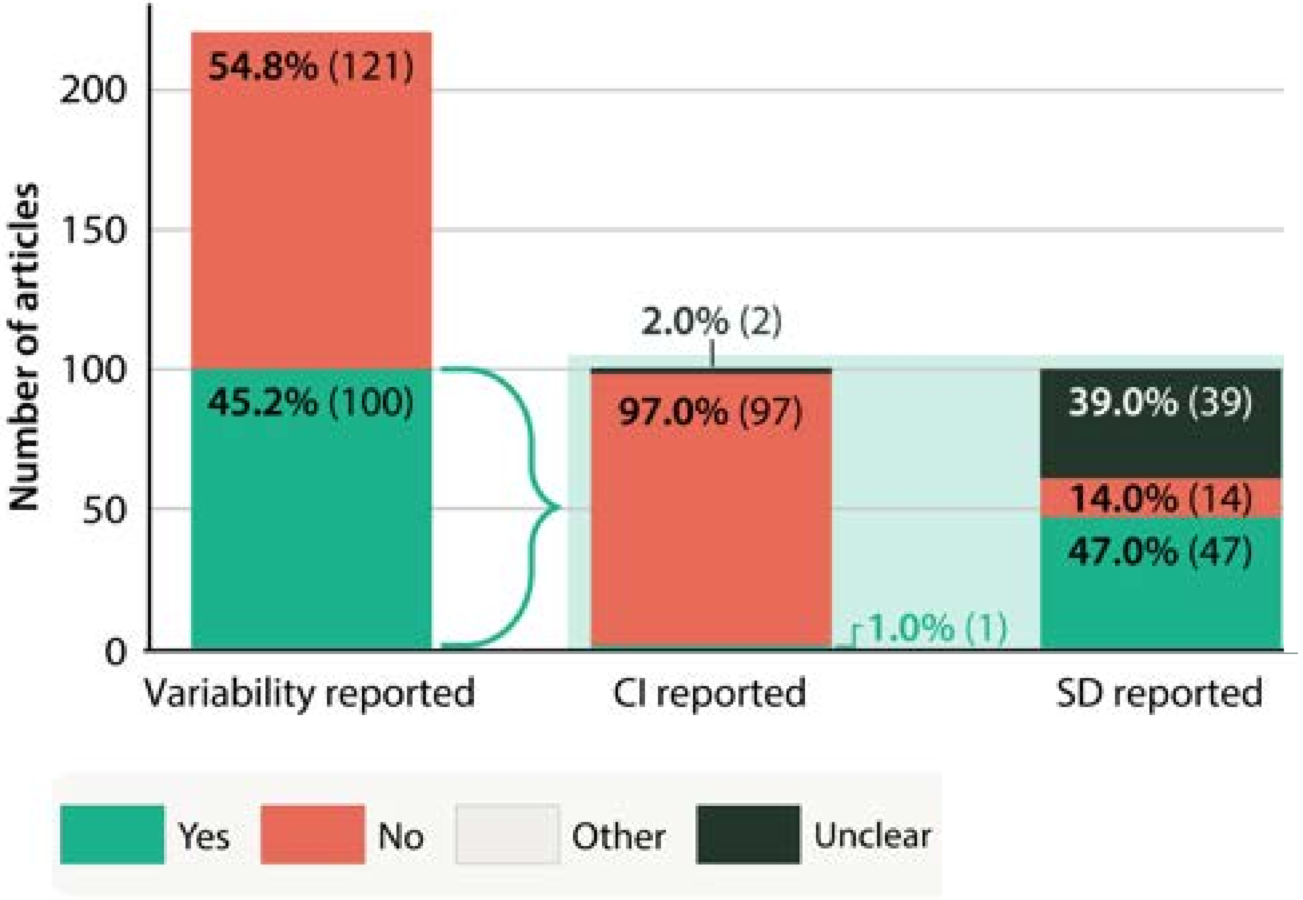
Christodoulou, Evangelia, et al. "Confidence intervals uncovered: Are we ready for real-world medical imaging AI?." International Conference on Medical Image *Computing and Computer-Assisted Intervention*. Cham: Springer Nature Switzerland, 2024.

RQ1: Common reporting practices



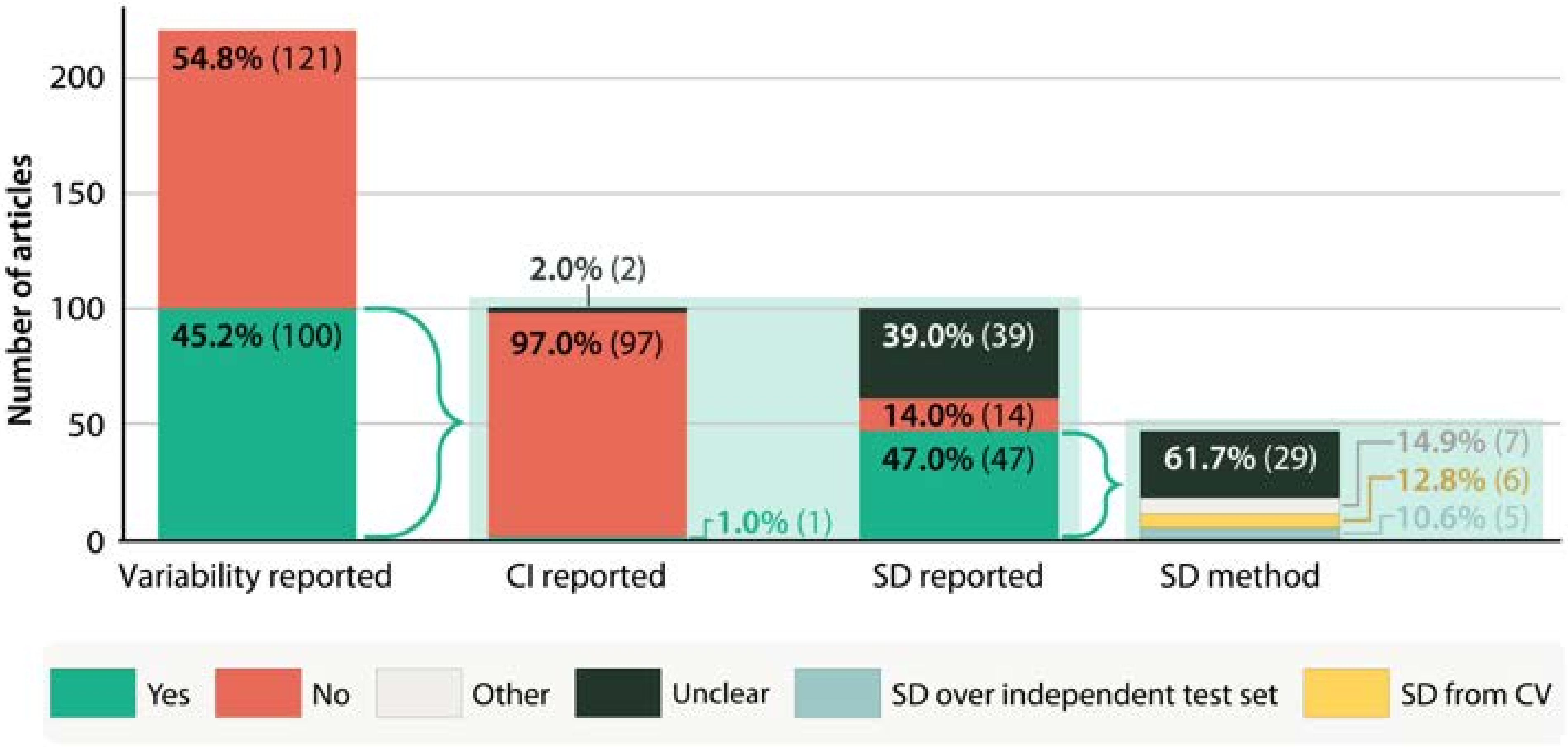
Christodoulou, Evangelia, et al. "Confidence intervals uncovered: Are we ready for real-world medical imaging AI?." International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer Nature Switzerland, 2024.

RQ1: Common reporting practices



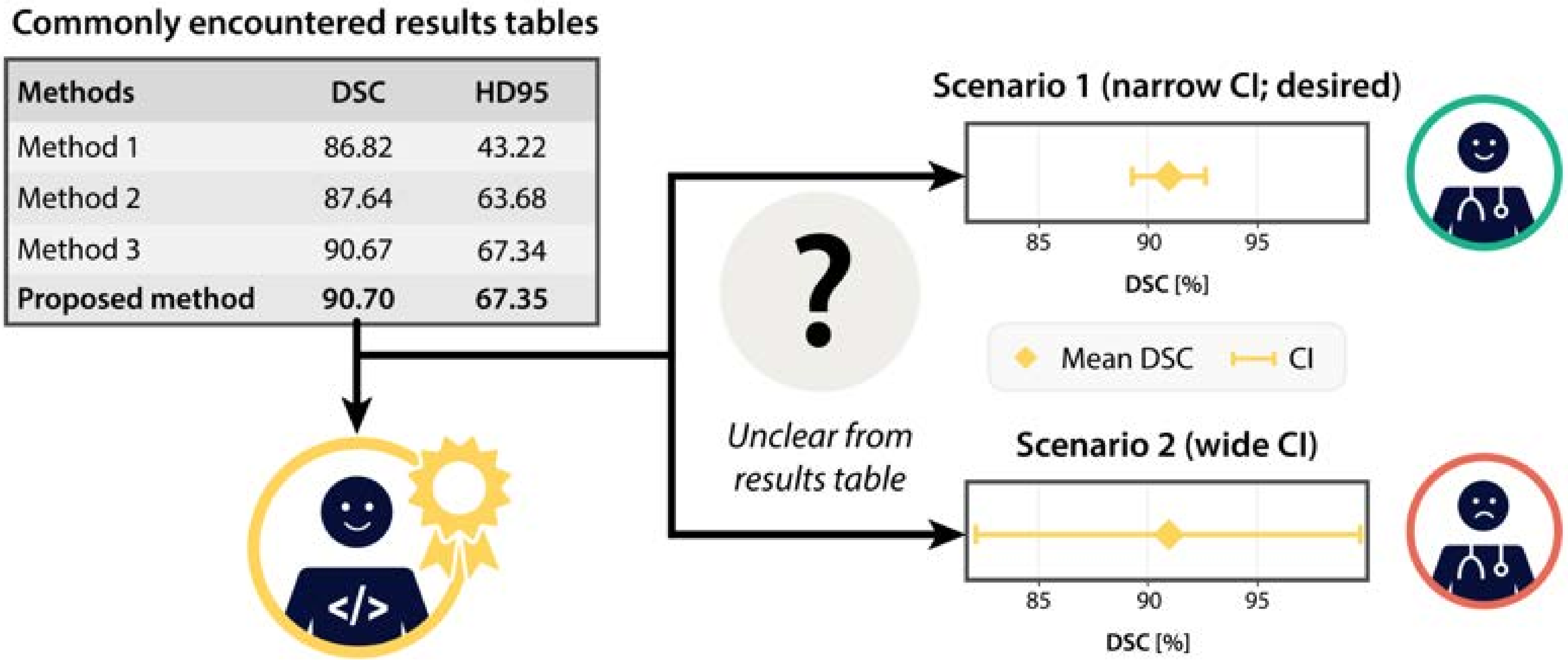
Christodoulou, Evangelia, et al. "Confidence intervals uncovered: Are we ready for real-world medical imaging AI?." International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer Nature Switzerland, 2024.

RQ1: Common reporting practices



Christodoulou, Evangelia, et al. "Confidence intervals uncovered: Are we ready for real-world medical imaging AI?." International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer Nature Switzerland, 2024.

RQ2: Approximation of missing variability parameters



Christodoulou, Evangelia, et al. "Confidence intervals uncovered: Are we ready for real-world medical imaging AI?." International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer Nature Switzerland, 2024.

RQ2: Approximation of missing variability parameters

RQ2: How to estimate missing variability parameters based on provided data?

In other words, can we impute variance from mean?

Christodoulou, Evangelia, et al. "Confidence intervals uncovered: Are we ready for real-world medical imaging AI?." International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer Nature Switzerland, 2024.

RQ2: Approximation of missing variability parameters

RQ2: How to estimate missing variability parameters based on provided data?

In other words, can we impute variance from mean?
There are specific cases with an analytical formula (e.g. accuracy)
In general, there is not

$$E(\hat{p}) = p$$

$$\mathrm{Var}(\hat{p}) = \frac{p(1-p)}{n}$$

Christodoulou, Evangelia, et al. "Confidence intervals uncovered: Are we ready for real-world medical imaging AI?." International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer Nature Switzerland, 2024.

RQ2: Approximation of missing variability parameters


nature communications

Explore content | About the journal | Publish with us

nature > nature communications > articles > article

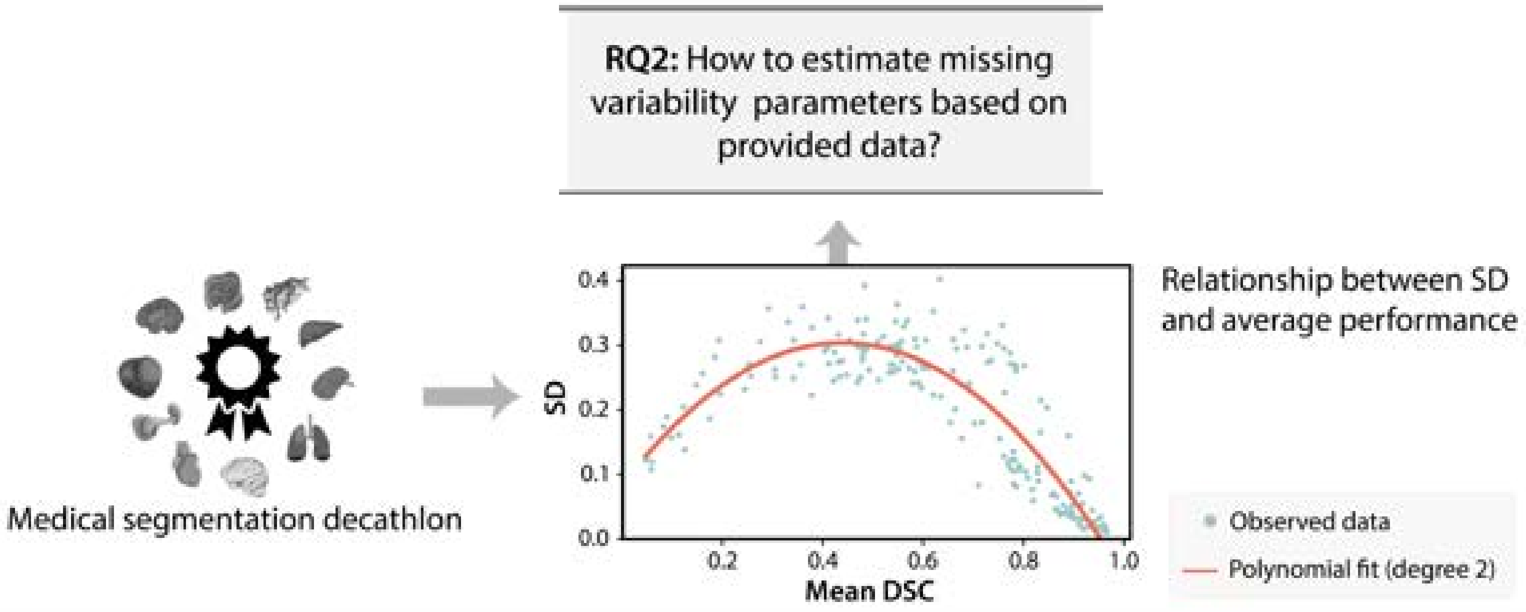
Article | Open access | Published: 15 July 2022

The Medical Segmentation Decathlon

Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A. Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M. Summers, Bram van Ginneken, Michel Bilello, Patrick Bilic, Patrick F. Christ, Richard K. G. Do, Marc J. Gollub, Stephan H. Heckers, Henkjan Huisman, William R. Jarnagin, Maureen K. McHugo, Sandy Napel, Jennifer S. Golia Pernicka, Kawal Rhode, Catalina Tobon-Gomez, … M. Jorge Cardoso + Show authors

Nature Communications 13, Article number: 4128 (2022) | Cite this article

64k Accesses | 420 Citations | 49 Altmetric | Metrics

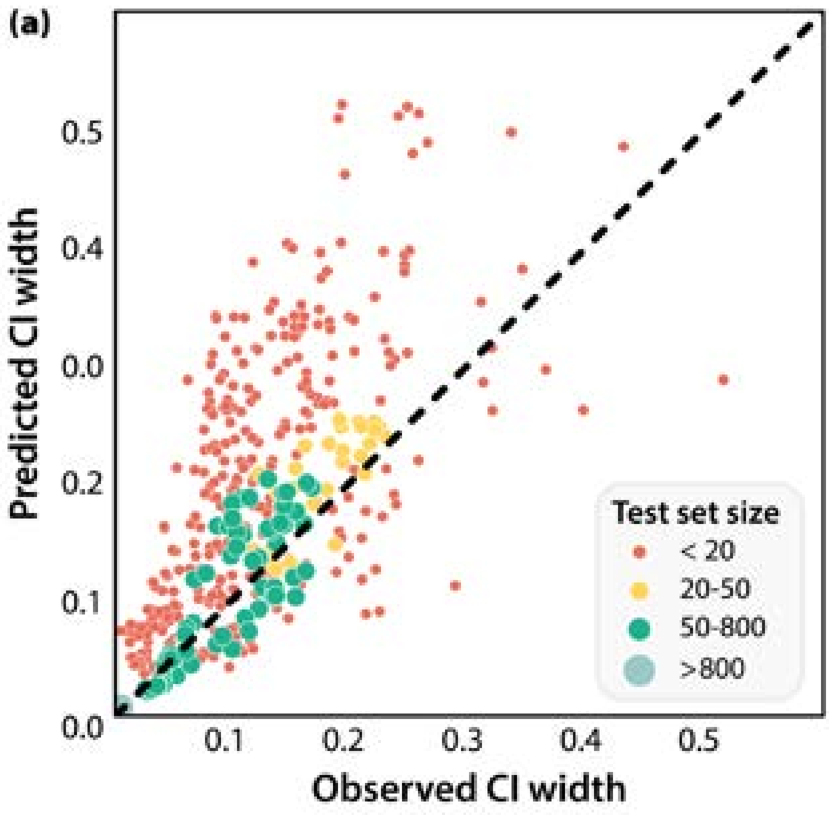


$$\left[DSC_{\mu} - t_{n-1,1-\alpha/2} \cdot \frac{SD}{\sqrt{n}}, DSC_{\mu} + t_{n-1,1-\alpha/2} \cdot \frac{SD}{\sqrt{n}}\right]$$

Christodoulou, Evangelia, et al. "Confidence intervals uncovered: Are we ready for real-world medical imaging AI?." International *Conference on Medical Image Computing and Computer-Assisted Intervention*. Cham: Springer Nature Switzerland, 2024.

RQ2: Approximation of missing variability parameters

Validation of the SD approximation on 56 past segmentation challenges



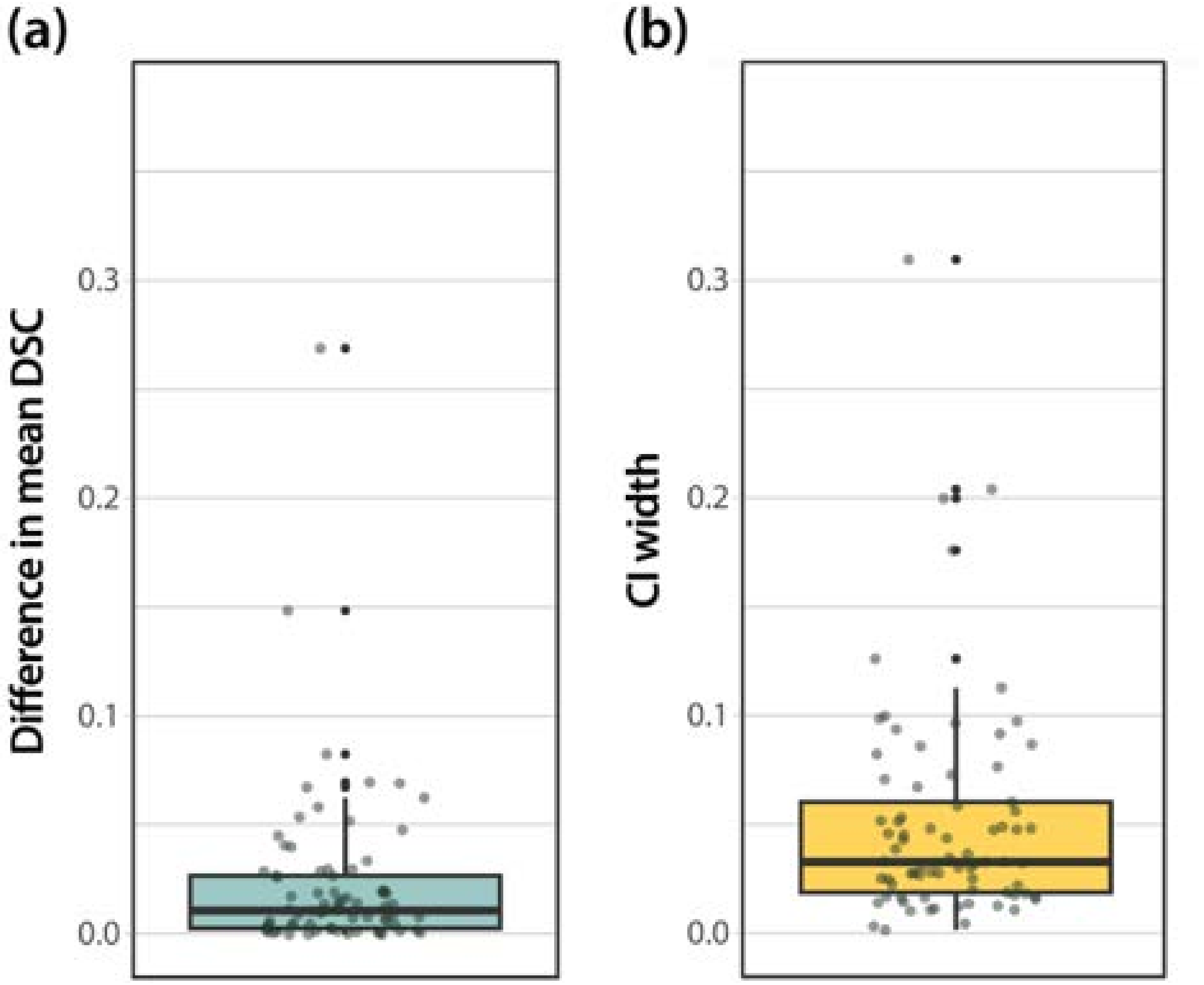
Christodoulou, Evangelia, et al. "Confidence intervals uncovered: Are we ready for real-world medical imaging AI?." International Conference *on Medical Image Computing and Computer-Assisted Intervention*. Cham: Springer Nature Switzerland, 2024.

RQ3: CI widths vs claims for outperformance

⚠️ Median CI width: 3 percent points

⚠️ Median difference between proposed method and second-ranked: 1 percent point

😱 Should we be worried?



Christodoulou, Evangelia, et al. "Confidence intervals uncovered: Are we ready for real-world medical imaging AI?." International Conference *on Medical Image Computing and Computer-Assisted Intervention*. Cham: Springer Nature Switzerland, 2024.

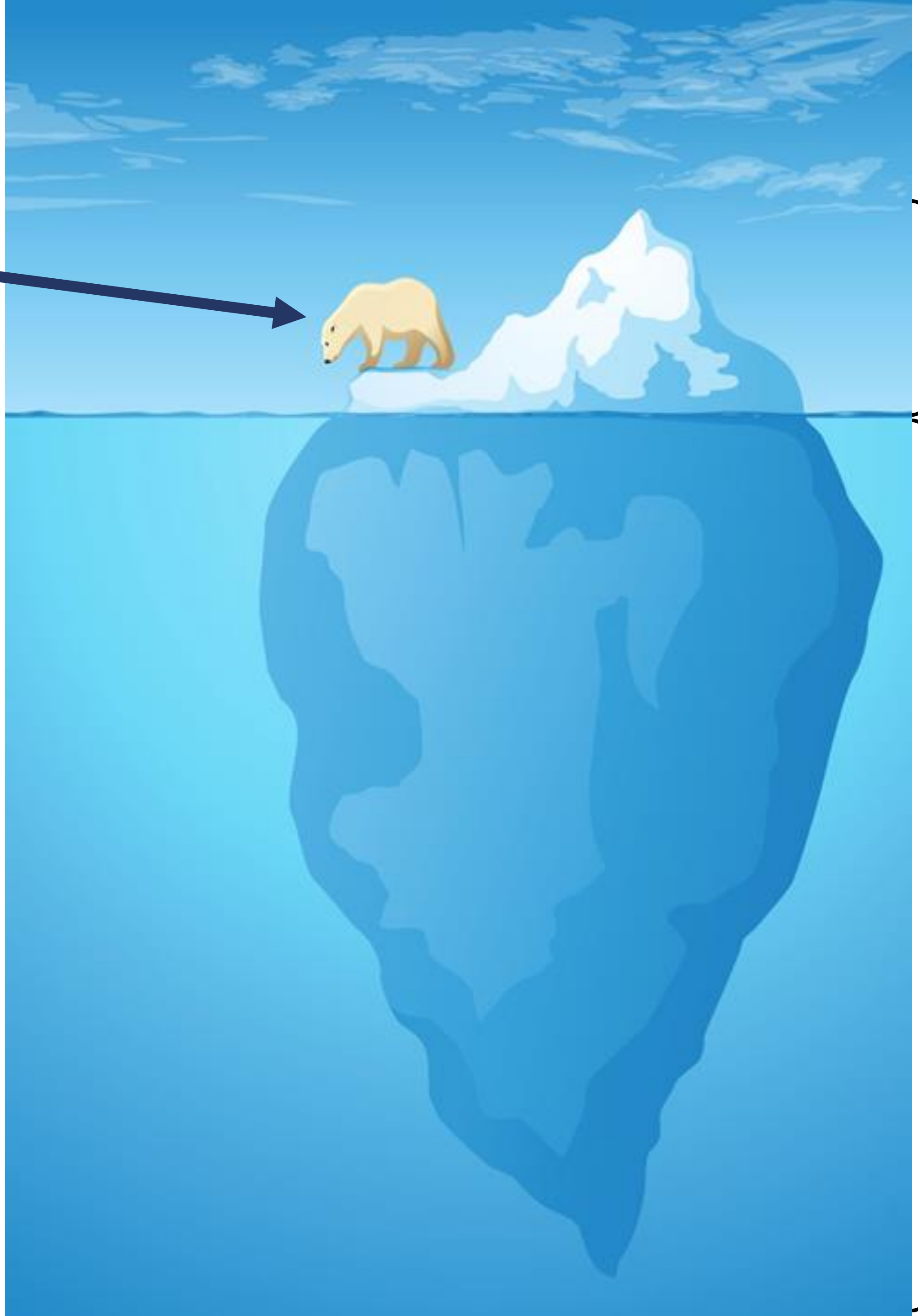
1. Current practices

2. Strength of outperformance claims

3. Areas for improvement

Take home messages

AI developer



Machine Learning (ML)

Dataset design

Annotations

Metrics

Data Splitting

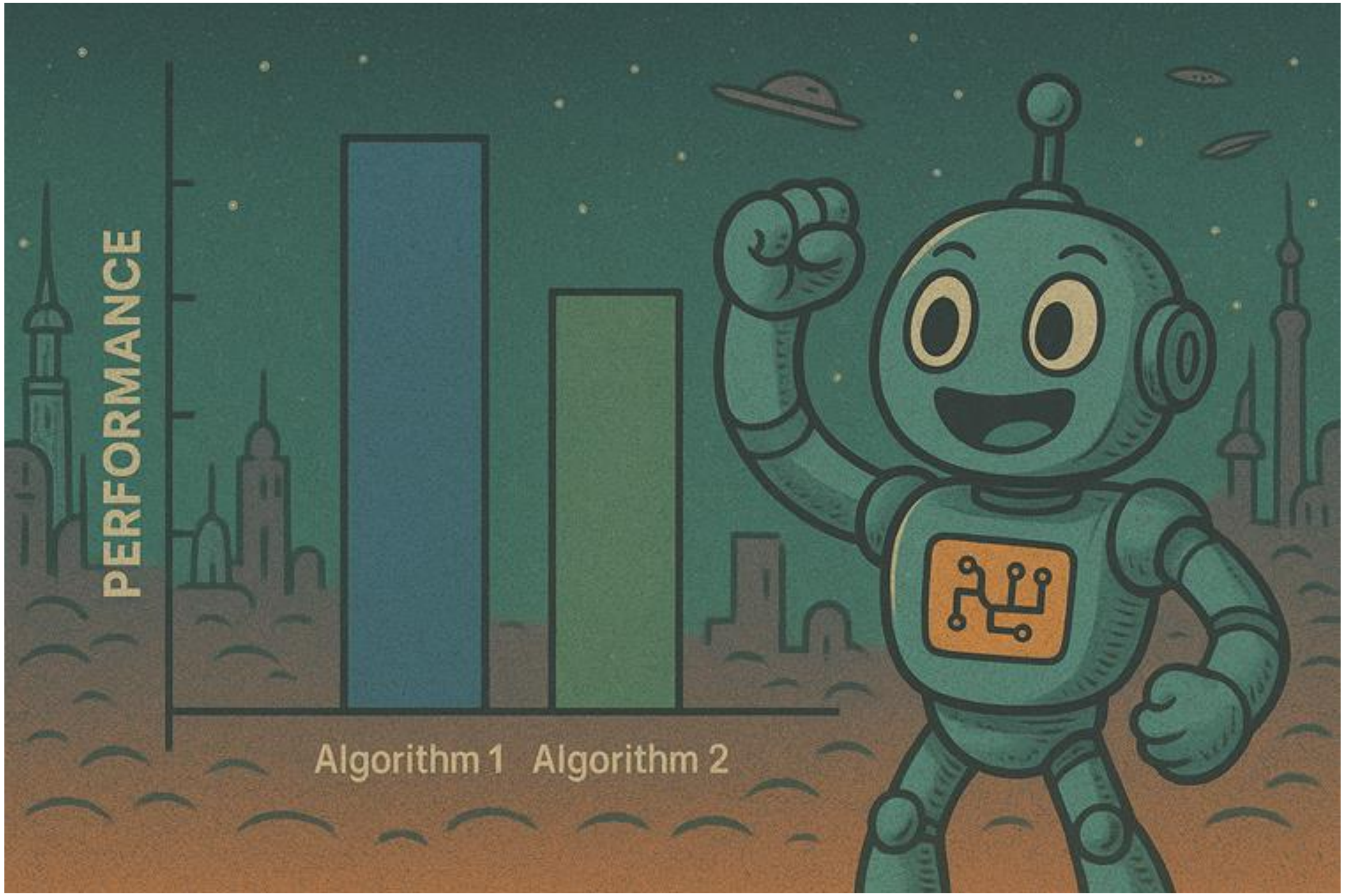
Reporting

Rankings

...

Image source:
Regulatory Affairs Professional Society

How likely is it that the ranks flip?

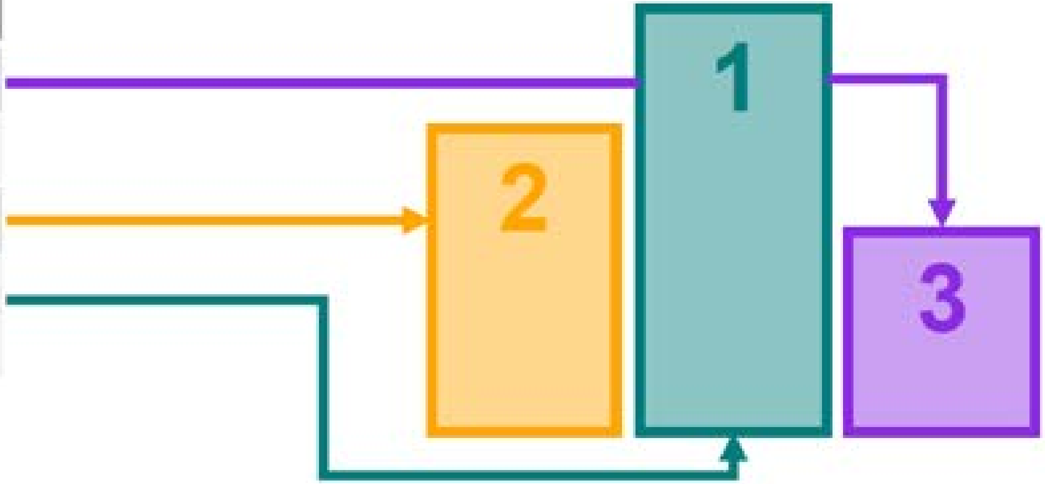


Generated by DALL-E

Outperformed the state-of-the art... (or not?)

Commonly encountered results tables

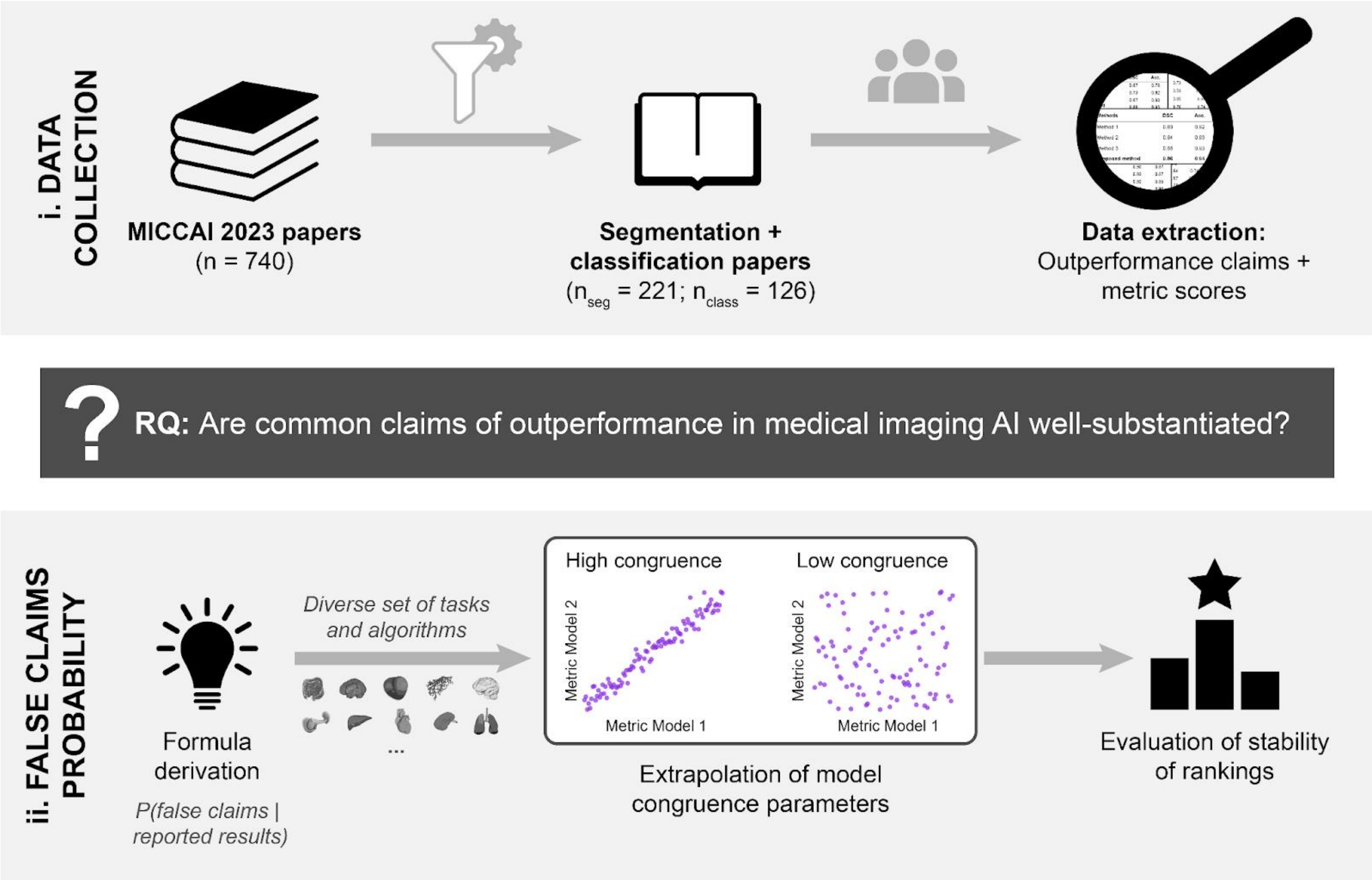
Methods	Accuracy	AUC
Method 1	0.83	0.91
Method 2	0.80	0.89
Method 3	0.83	0.92
Proposed method	**0.84**	**0.92**



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

Christodoulou, Evangelia, et al. "False Promises in Medical Imaging AI? Assessing Validity of Outperformance Claims" Arxiv preprint, 2025. https://arxiv.org/abs/2505.04720

Outperformed the state-of-the art... (or not?)



Christodoulou, Evangelia, et al. "False Promises in Medical Imaging AI? Assessing Validity of Outperformance Claims" Arxiv *preprint,* 2025.
https://arxiv.org/abs/2505.04720

Outperformed the state-of-the art… (or not?)

- **Probability of false claims**
  - Bayesian approach to estimate whether the relative ranking of methods is likely to have occurred by chance
  - Probability that the second-ranked method **(B)** was, in fact, performing equally or better than the first-ranked method **(A)**, given the results reported in the paper

$$P(p_A \leq p_B | \textbf{reported results}) = P(p_A \leq p_B | \widehat{p}_A, \widehat{p}_B)$$

A: first ranked method

B: second ranked method

True performance (random variable)

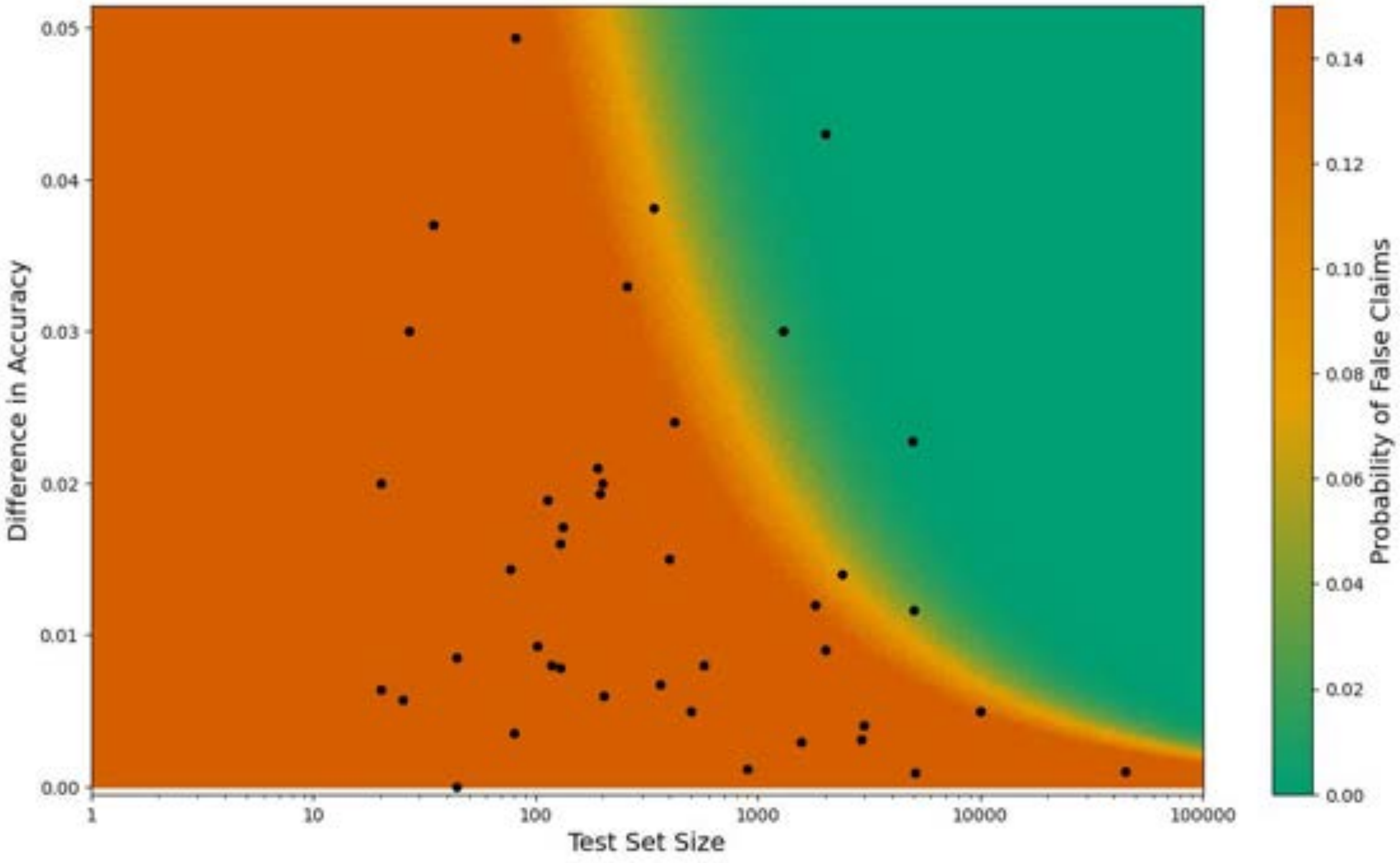
Performance reported in the paper

Christodoulou, Evangelia, et al. "False Promises in Medical Imaging AI? Assessing Validity of Outperformance Claims" Arxiv *preprint,* 2025.
https://arxiv.org/abs/2505.04720

Outperformed the state-of-the art… (or not?)

Classification

Black dots: individual MICCAI 2023 papers



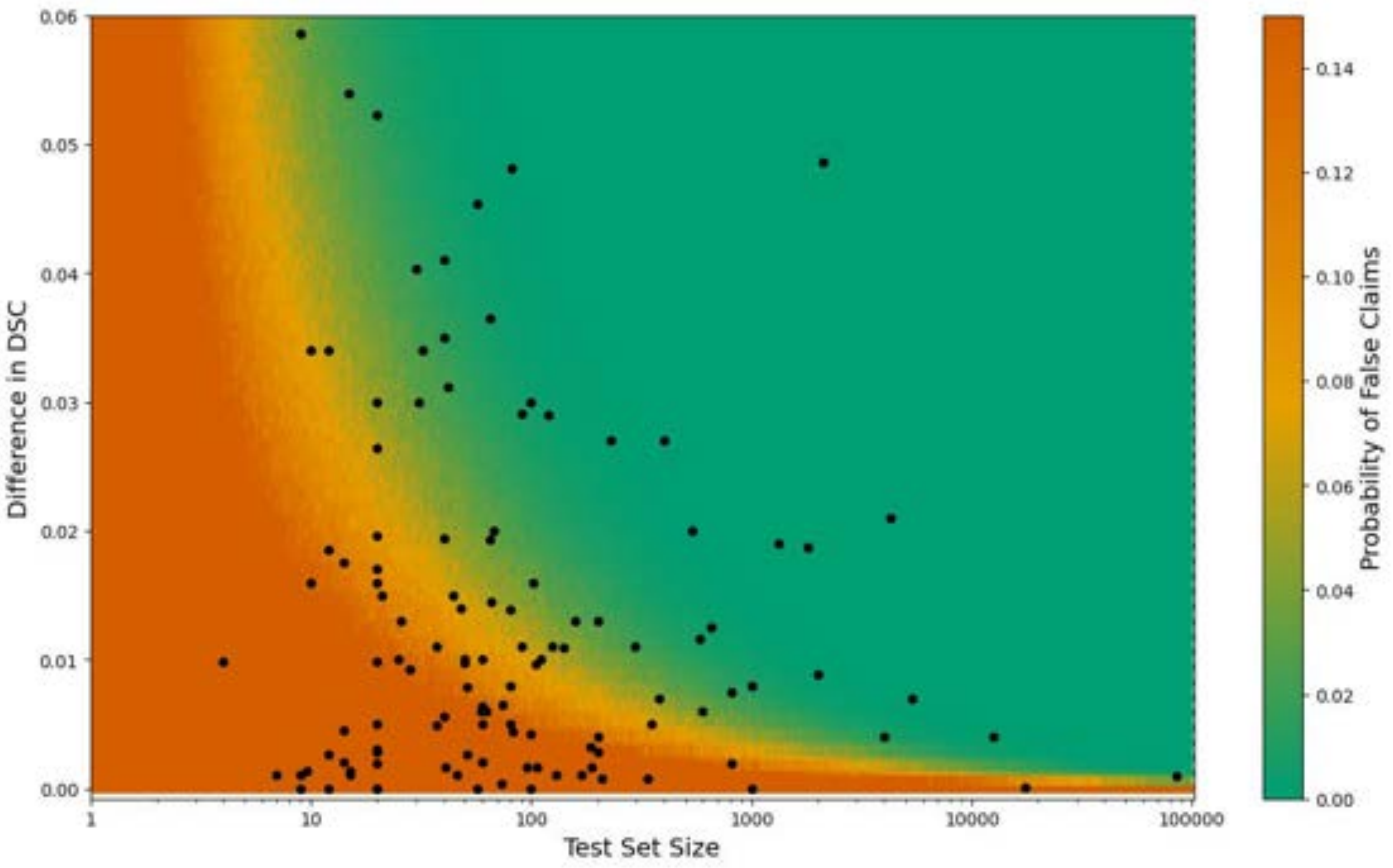
86% of classification papers have a high probability of false claims (>5%)

Christodoulou, Evangelia, et al. "False Promises in Medical Imaging AI? Assessing Validity of Outperformance Claims" Arxiv *preprint,* 2025. https://arxiv.org/abs/2505.04720

Outperformed the state-of-the art… (or not?)

Segmentation

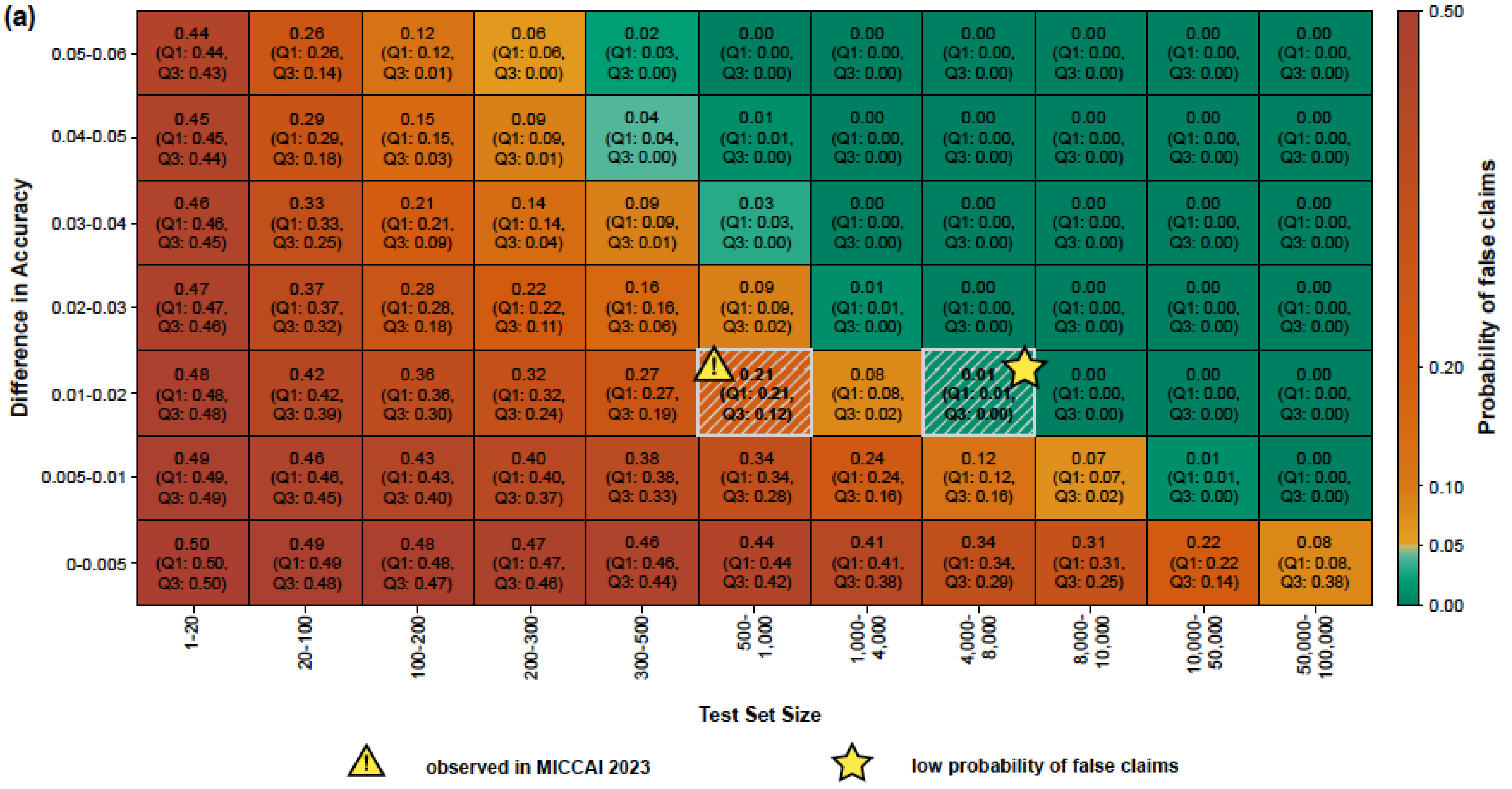
Black dots: individual MICCAI 2023 papers



53% of Segmentation papers have a high probability of false claims (>5%)

Christodoulou, Evangelia, et al. "False Promises in Medical Imaging AI? Assessing Validity of Outperformance Claims" Arxiv *preprint,* 2025.
https://arxiv.org/abs/2505.04720

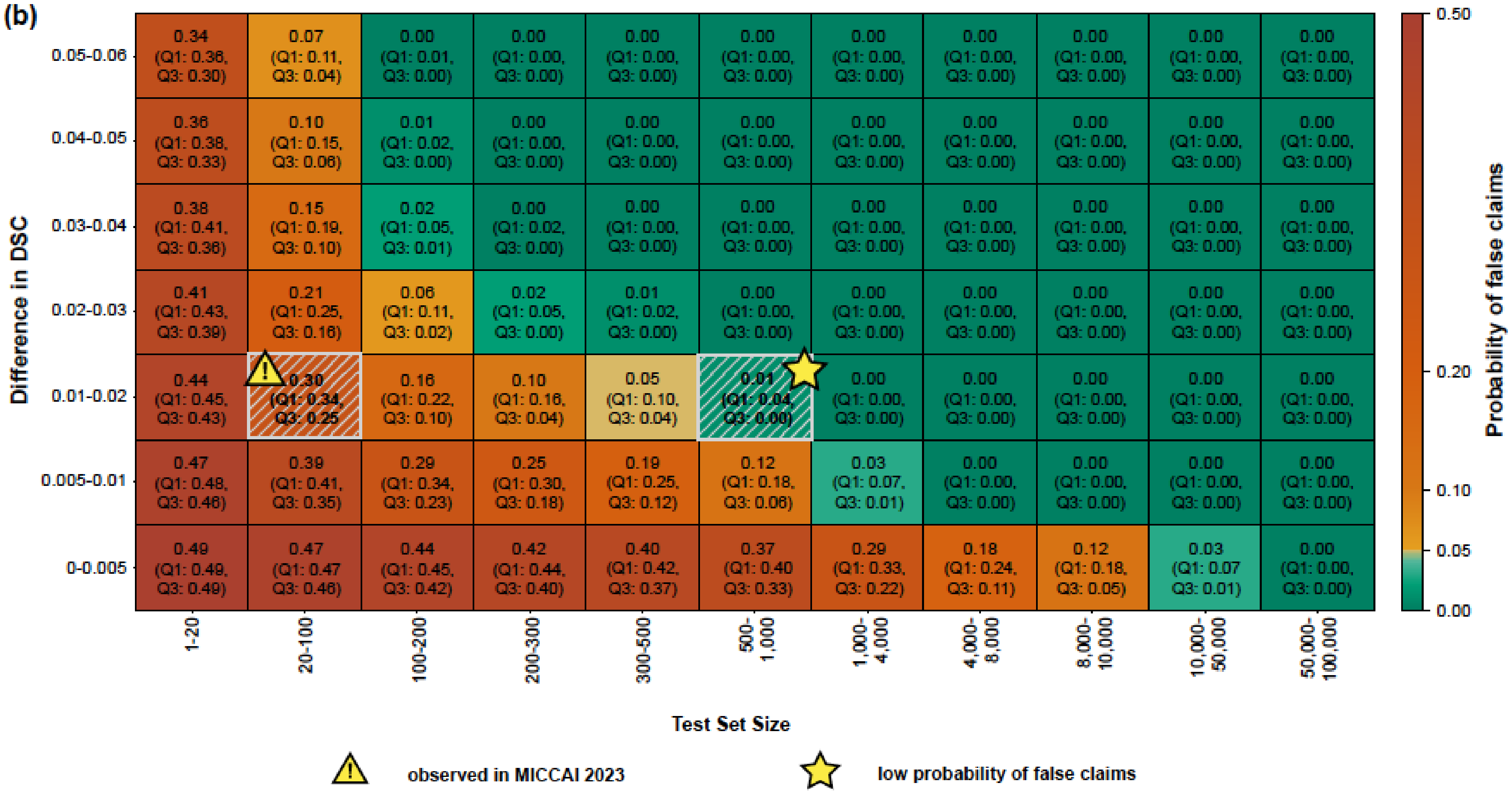
Stronger evidence of outperformance calls for test sets dramatically larger than usual



Classification

Christodoulou, Evangelia, et al. "False Promises in Medical Imaging AI? Assessing Validity of Outperformance Claims" Arxiv *preprint,* 2025. https://arxiv.org/abs/2505.04720

Stronger evidence of outperformance calls for test sets dramatically larger than usual



Segmentation

Christodoulou, Evangelia, et al. "False Promises in Medical Imaging AI? Assessing Validity of Outperformance Claims" Arxiv *preprint,* 2025.
https://arxiv.org/abs/2505.04720

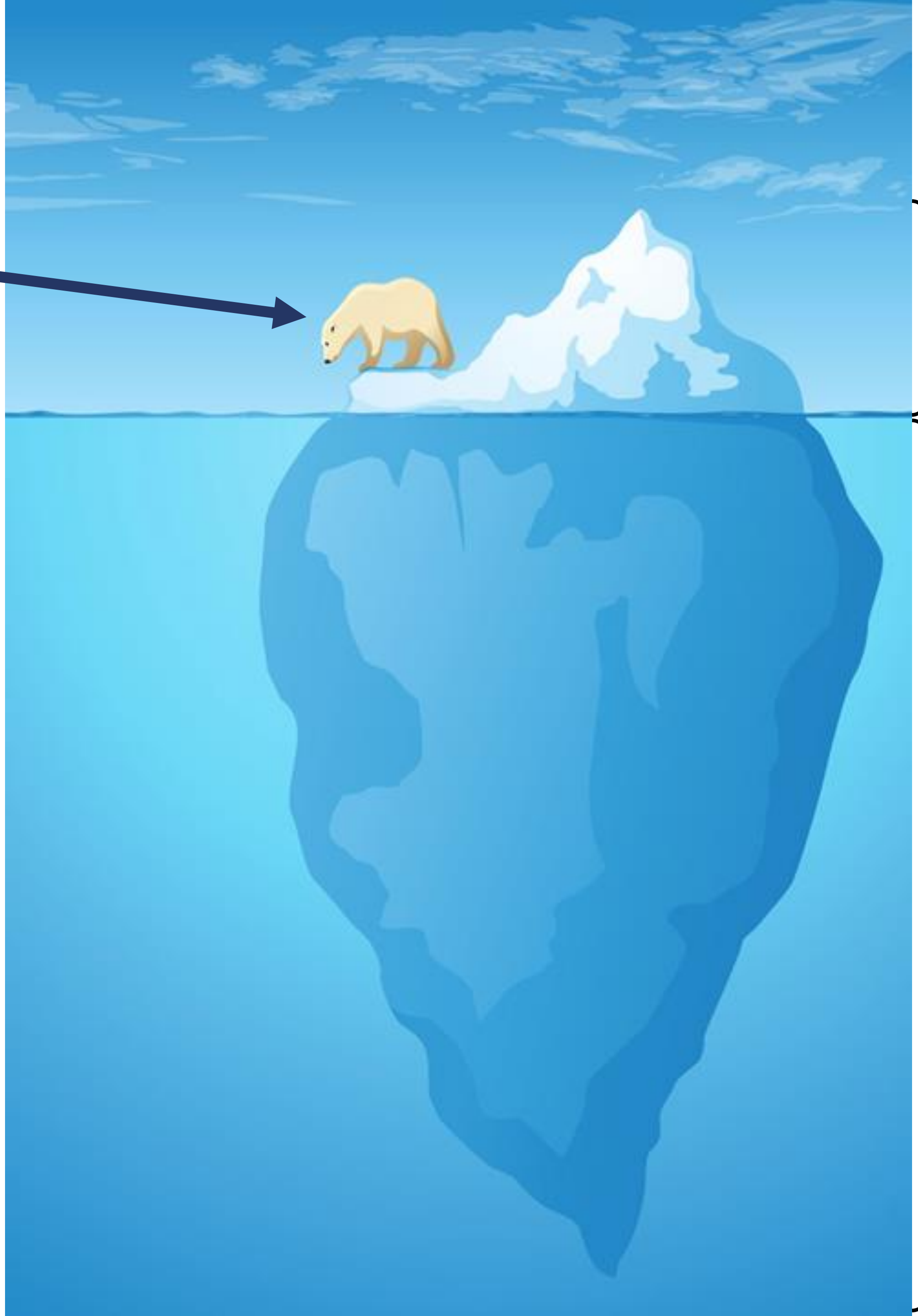
1. Current practices

2. Strength of outperformance claims

3. Areas for improvement

Take home message

AI developer



Machine Learning (ML)

Dataset design

Annotations

Metrics

Data Splitting

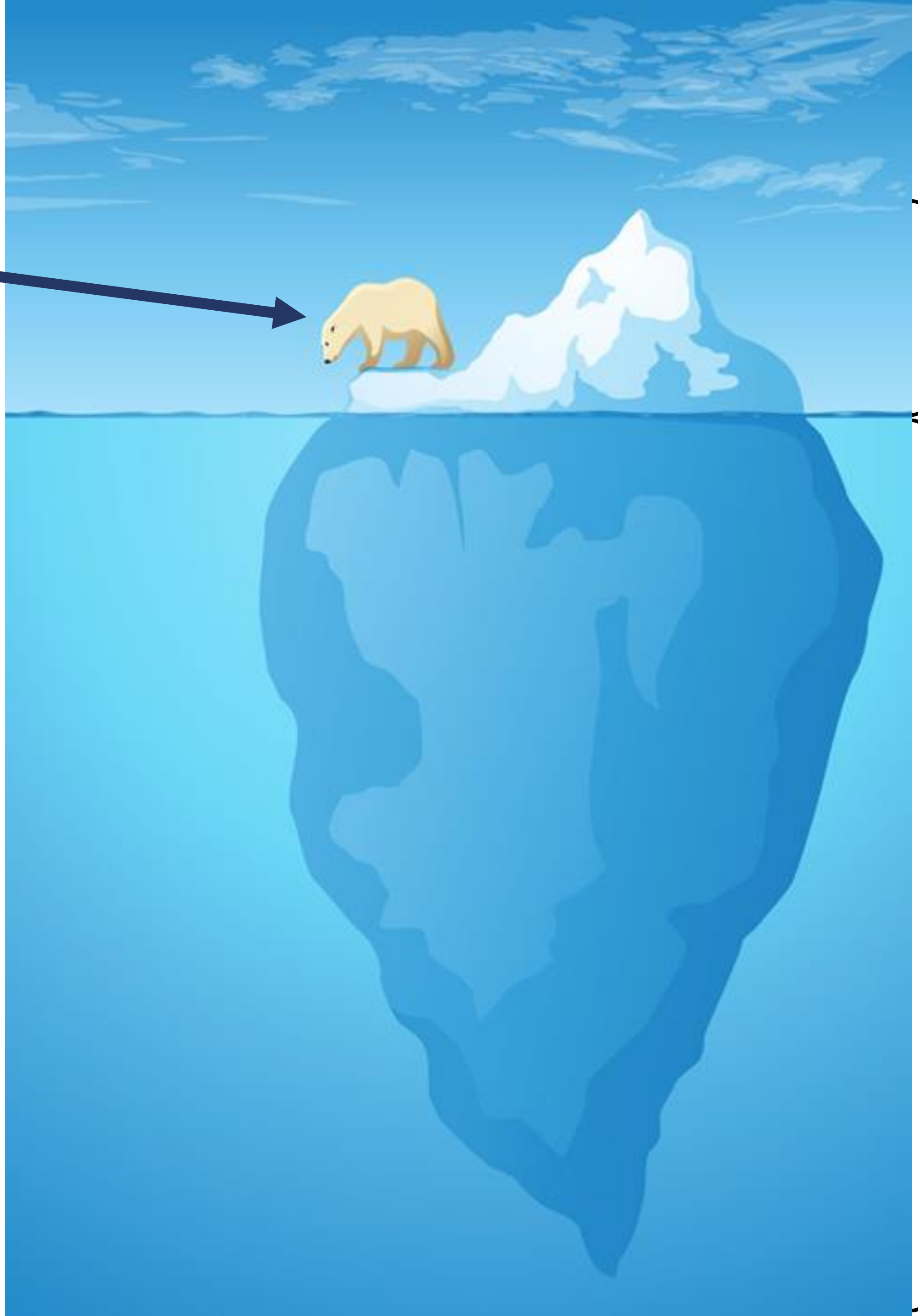
Reporting

Rankings

...

Image source:
Regulatory Affairs Professional Society

AI developer



Machine Learning (ML)

Dataset design

Annotations

Metrics

Data Splitting

Reporting

Rankings

...

Image source:
Regulatory Affairs Professional Society

Computing the mean value: on which dataset?

Mean DSC and HD95

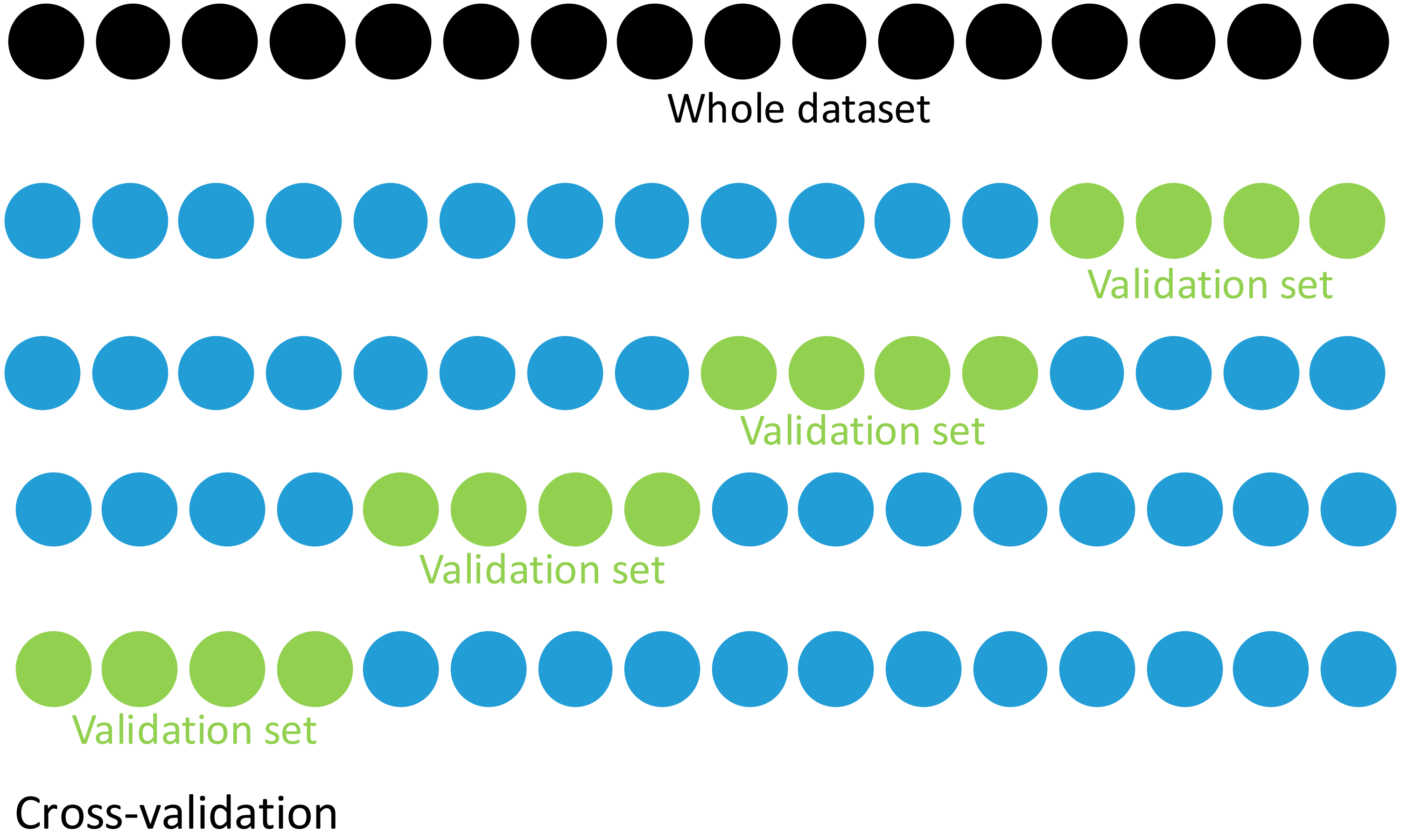
Methods	DSC	HD95
Method 1	0.892	1.23
Method 2	0.895	1.22
Method 3	0.883	1.32
Proposed	**0.897**	**1.21**

Data splitting

Single split

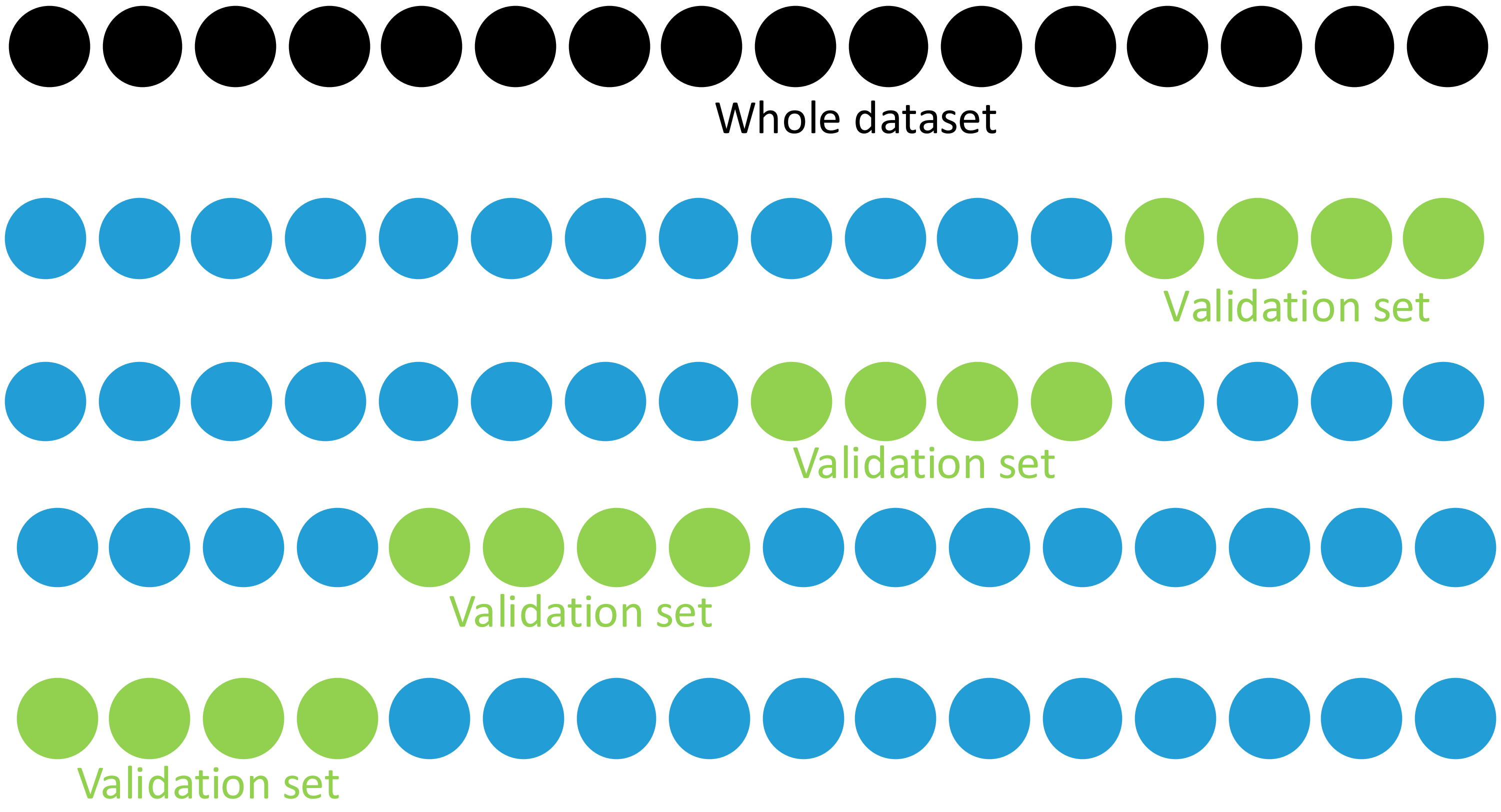
Varoquaux and Colliot, Evaluating machine learning models and their diagnostic value, 2023 https://hal.science/hal-03682454

Data splitting



Varoquaux and Colliot, Evaluating machine learning models and their diagnostic value, 2023 https://hal.science/hal-03682454

Data splitting

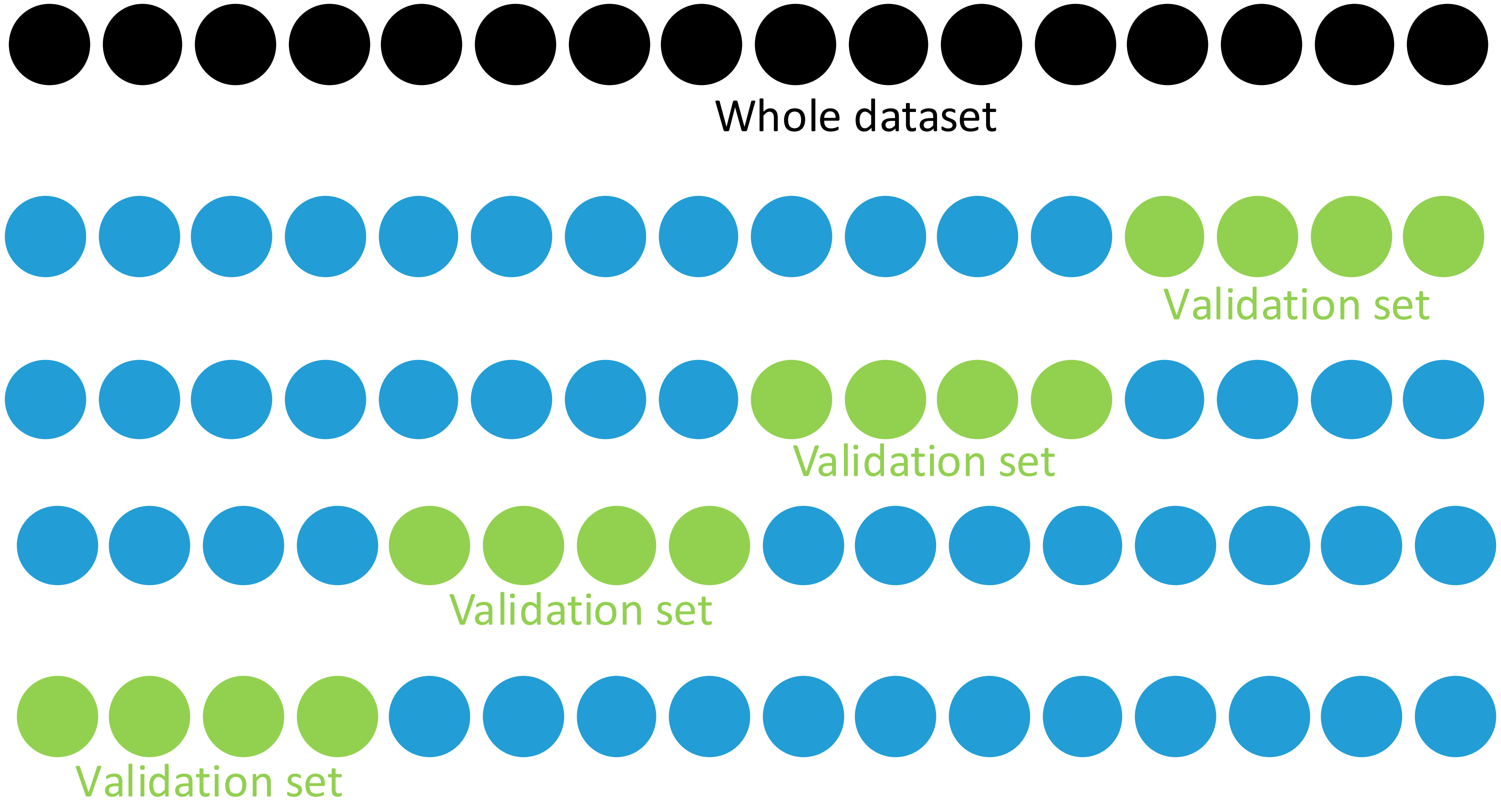


Can be used to report final performance **if no hyperparameter tuning, no architecture modification**

⚠ Not a realistic scenario

Varoquaux and Colliot, Evaluating machine learning models and their diagnostic value, 2023 https://hal.science/hal-03682454

Data splitting

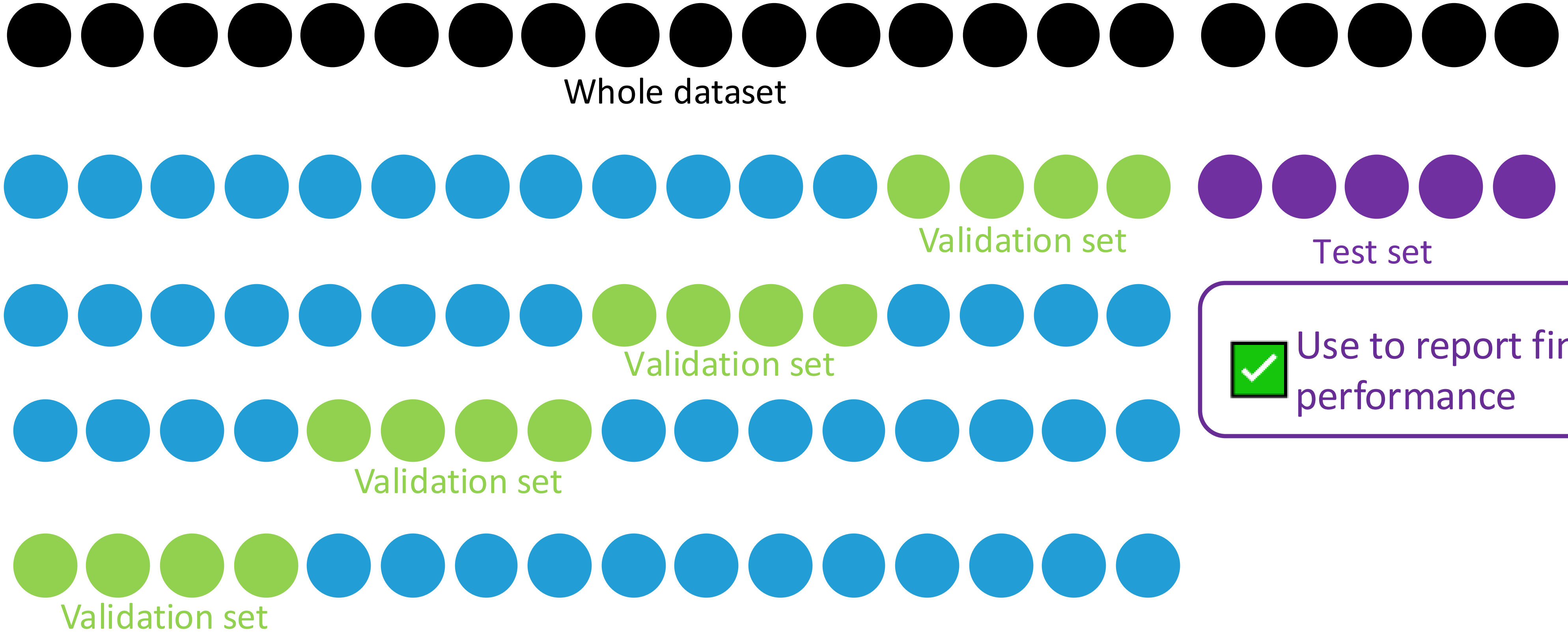


✅ Use to tune hyperparameters, experiment with different architectures…

⛔ Do not use to report final performance (biased)

Varoquaux and Colliot, Evaluating machine learning models and their diagnostic value, 2023 https://hal.science/hal-03682454

Data splitting



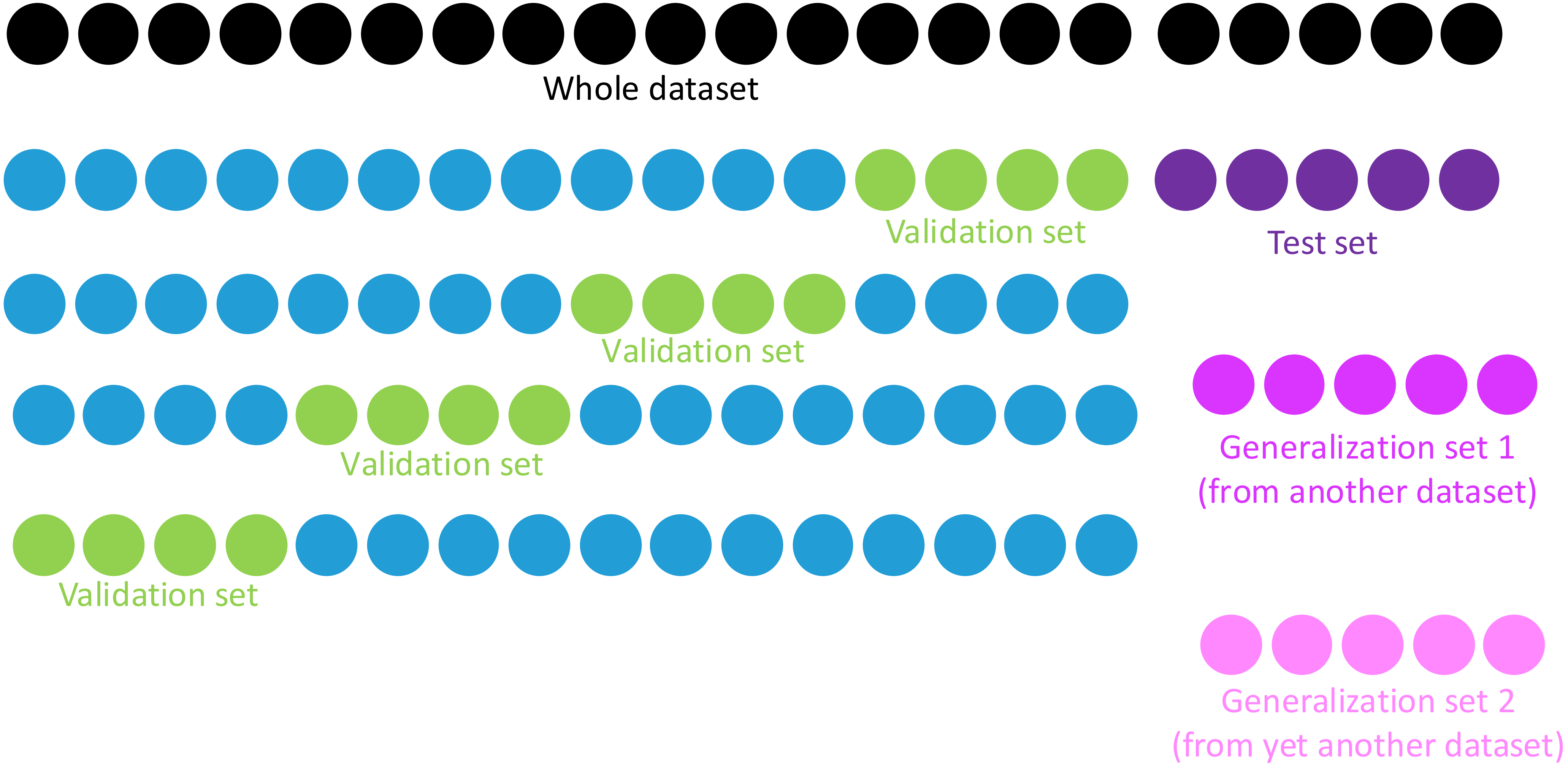
✅ Use to report final performance

✅ Use to tune hyperparameters, experiment with different architectures…

⛔ Do not use to report final performance (biased)

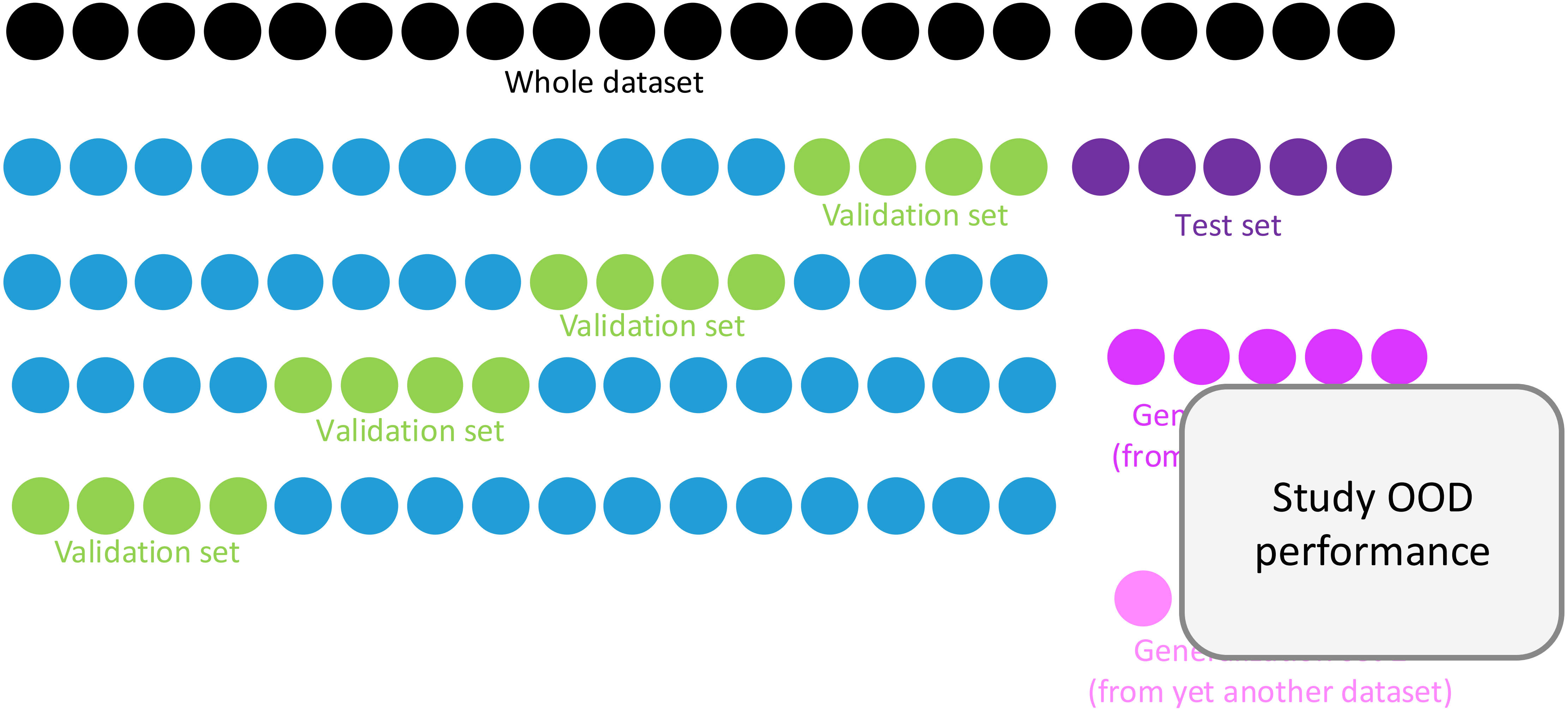
Varoquaux and Colliot, Evaluating machine learning models and their diagnostic value, 2023 https://hal.science/hal-03682454

Data splitting



Varoquaux and Colliot, Evaluating machine learning models and their diagnostic value, 2023 https://hal.science/hal-03682454

Data splitting



Varoquaux and Colliot, Evaluating machine learning models and their diagnostic value, 2023 https://hal.science/hal-03682454

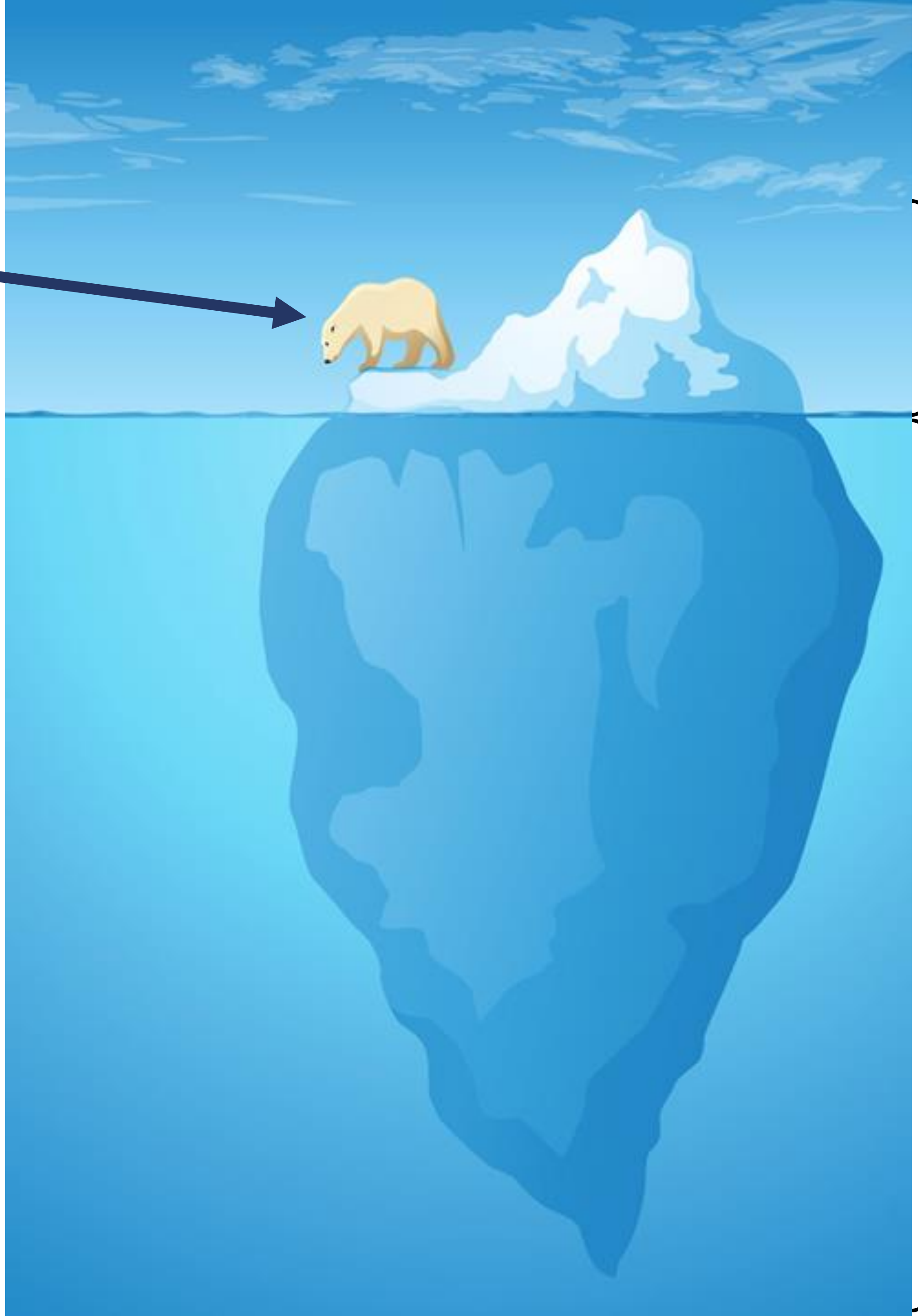
Computing the mean value: on which dataset?

Mean DSC and HD95 on the test set

Methods	DSC	HD95
Method 1	0.892	1.23
Method 2	0.895	1.22
Method 3	0.883	1.32
Proposed	**0.897**	**1.21**

Paper includes text describing precisely the data splitting and which splits were used for what purpose

AI developer



Machine Learning (ML)

Dataset design

Annotations

Metrics

Data Splitting

Reporting

Rankings

...

Image source:
Regulatory Affairs Professional Society

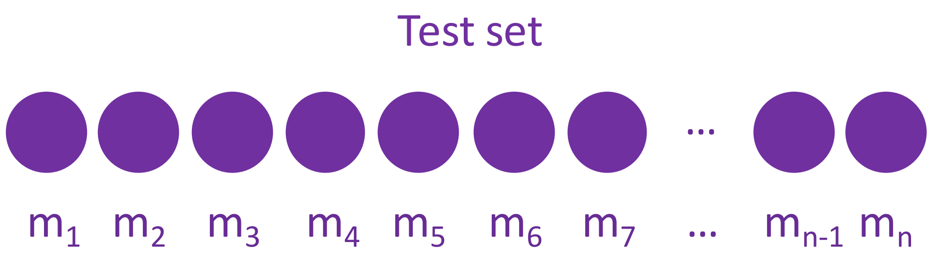
Summary statistics

Mean DSC and HD95 on the test set

Methods	DSC	HD95
Method 1	0.892	1.23
Method 2	0.895	1.22
Method 3	0.883	1.32
Proposed	**0.897**	**1.21**

Individual values of the metric

e.g. DSC on each individual



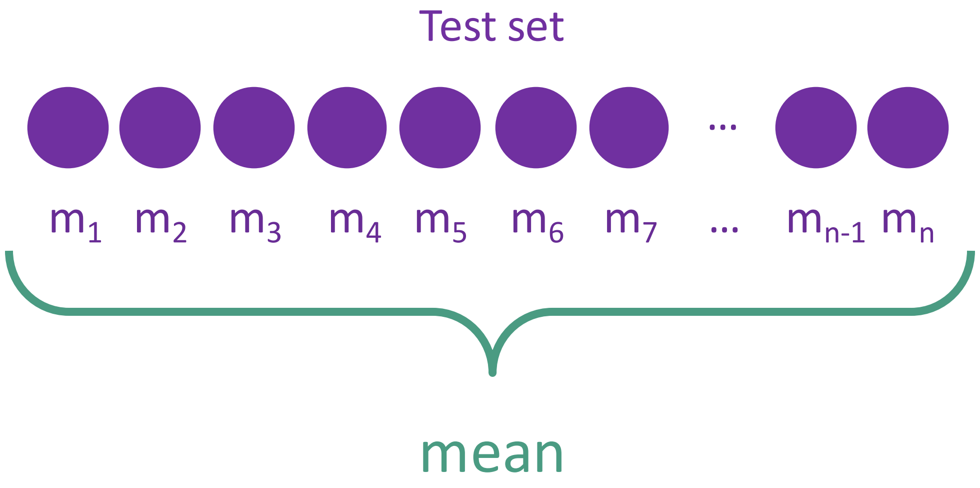
Summary statistics

Mean DSC and HD95 on the test set

Methods	DSC	HD95
Method 1	0.892	1.23
Method 2	0.895	1.22
Method 3	0.883	1.32
Proposed	**0.897**	**1.21**

Individual values of the metric

e.g. DSC on each individual



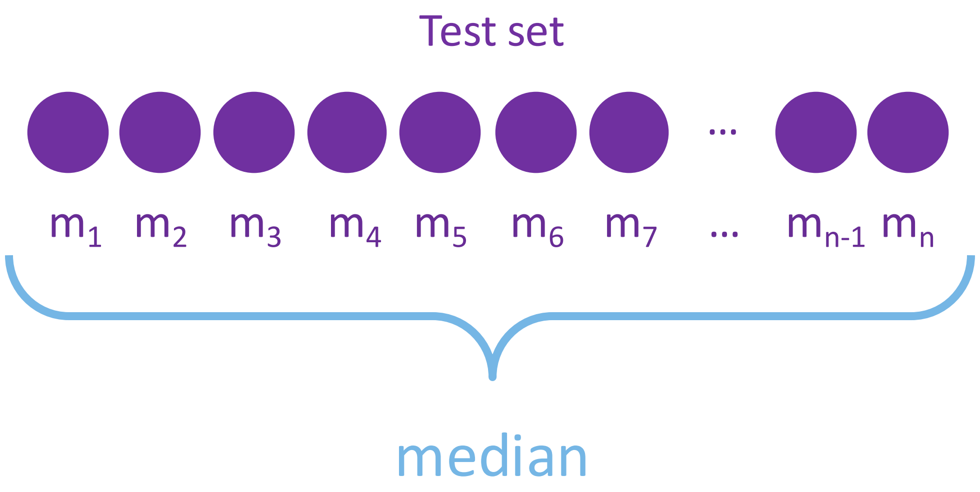
Summary statistics

Median DSC and HD95 on the test set

Methods	DSC	HD95
Method 1	0.892	1.23
Method 2	0.895	1.22
Method 3	0.883	1.32
Proposed	**0.897**	**1.21**

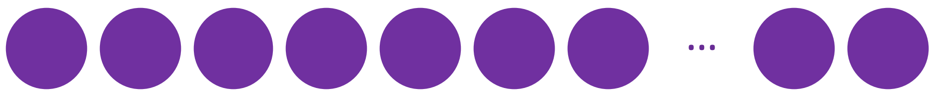
Individual values of the metric

e.g. DSC on each individual



Summary statistics

Test set

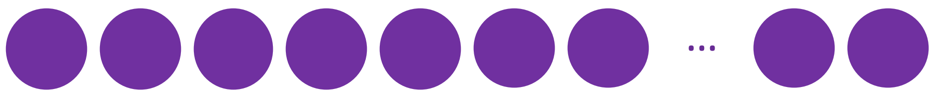


m_1 m_2 m_3 m_4 m_5 m_6 m_7 … m_{n-1} m_n

Summary statistics of central tendency

Mean, median, X% trimmed mean…

Test set



m_1 m_2 m_3 m_4 m_5 m_6 m_7 … m_{n-1} m_n

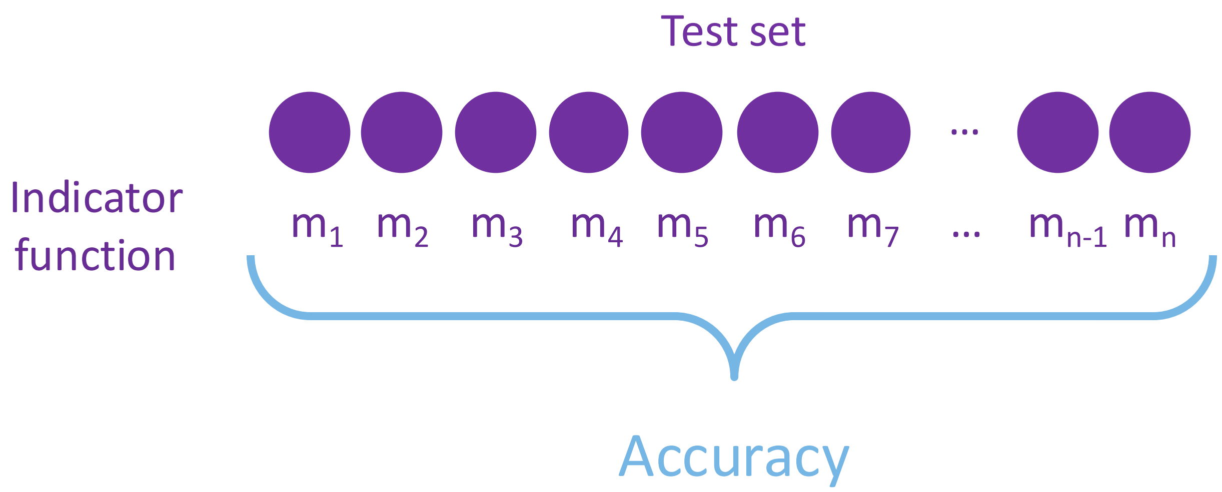
Summary statistics of dispersion

Standard-deviation, inter-quartile range…

Some metrics are only defined on a set

Accuracy and AUC on the test set

Methods	Accuracy	AUC
Method 1	0.828	0.862
Method 2	0.821	0.857
Method 3	0.847	0.889
Proposed	**0.851**	**0.891**



Important implications for variability

What do we mean by SD of accuracy?
SD of its sampling distribution

Reporting variability: which variability?

Methods	DSC
Method 1	0.892 ± 0.017
Method 2	0.895 ± 0.013
Method 3	0.883 ± 0.012
Proposed	**0.897** ± 0.013

± what?

At least 3 possibilities

Standard-deviation (SD) of the metric over the test set

2 Standard-error (SE) of the summary statistic

3 Standard-deviation (SD) over cross-validation (CV)

Reporting variability: which variability?

Methods	DSC
Method 1	0.892 ± 0.017
Method 2	0.895 ± 0.013
Method 3	0.883 ± 0.012
Proposed	**0.897** ± 0.013

± what?

At least 3 possibilities

Standard-deviation (SD) of the metric over the test set

2 Standard-error (SE) of the summary statistic

3 Standard-deviation (SD) over cross-validation (CV)

SD vs SE

Standard-deviation (SD)

SD of your metric across individuals (e.g. over test set)

💡 **Meaning: How variable is your performance across your set**

Its magnitude is independent of n

➡️ Descriptive statistic

Standard error (SE)

SD of the **sampling** distribution of a statistic (e.g. the mean)

💡 **Meaning: How precise is the estimate of the statistic**

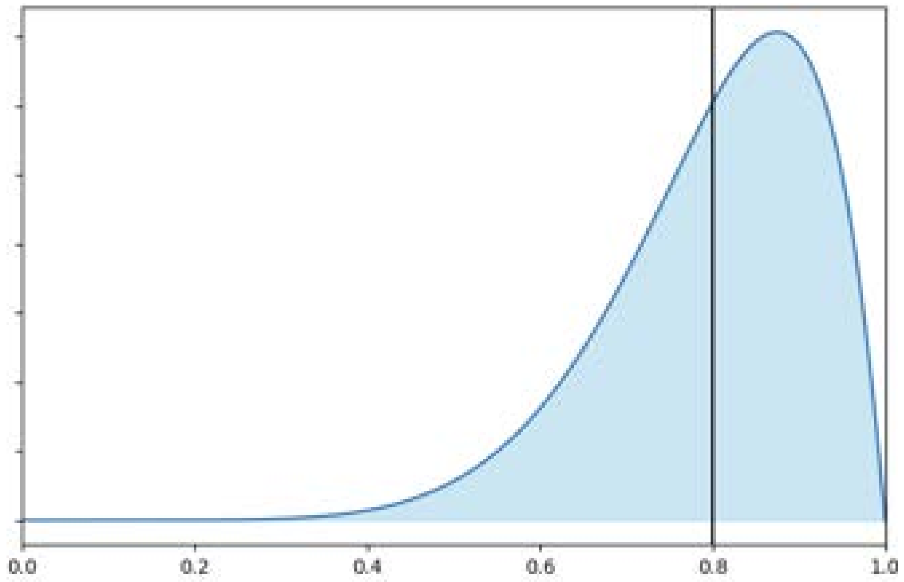
Shrinks with n (with $\sqrt{n}$)

➡️ Inferential statistic

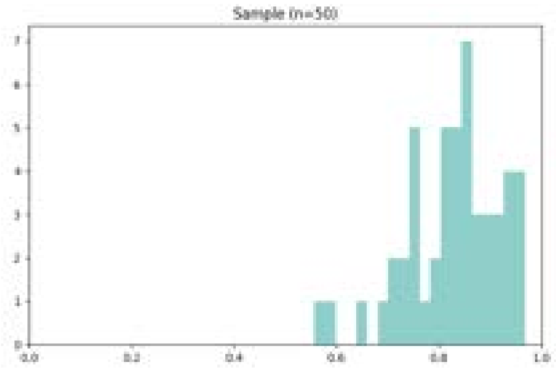
Sampling distribution

Distribution of a statistic (here the mean) across random samples

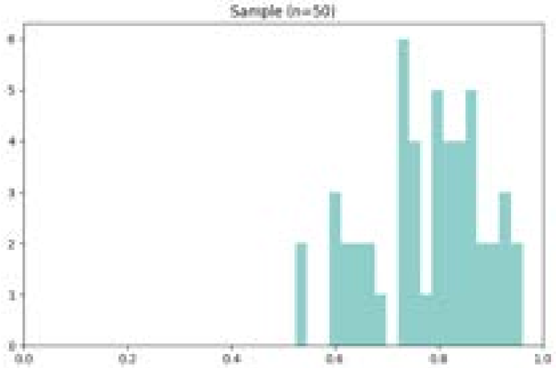
True population



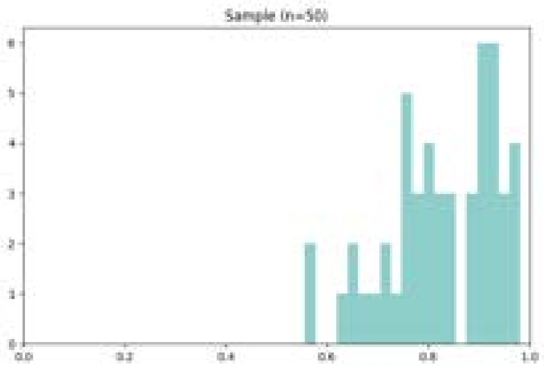
$\mu = 0.8$



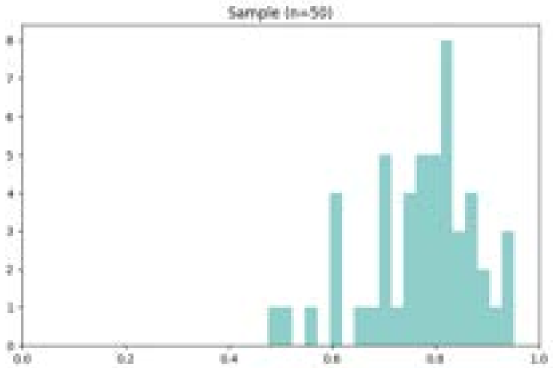
$\hat{\mu} = 0.79$



$\hat{\mu} = 0.78$



$\hat{\mu} = 0.83$

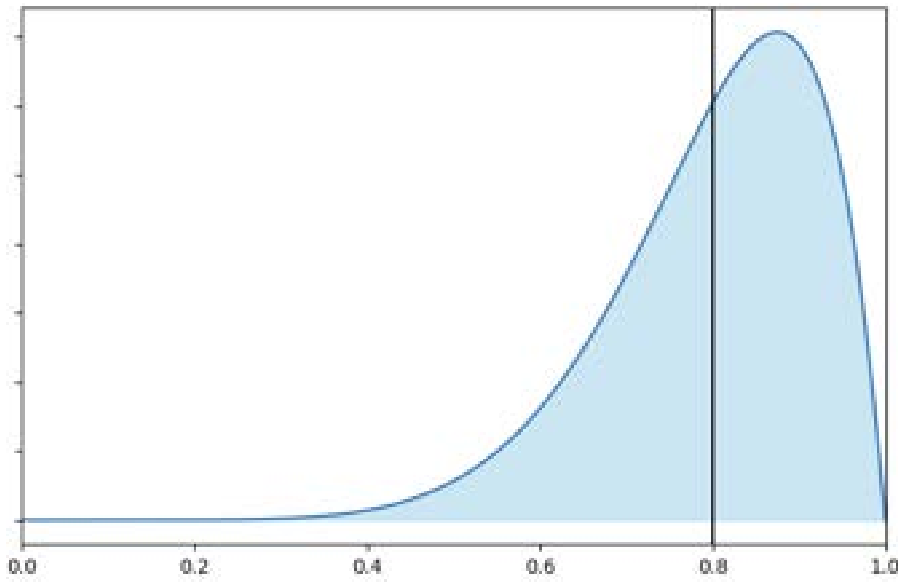


$\hat{\mu} = 0.77$

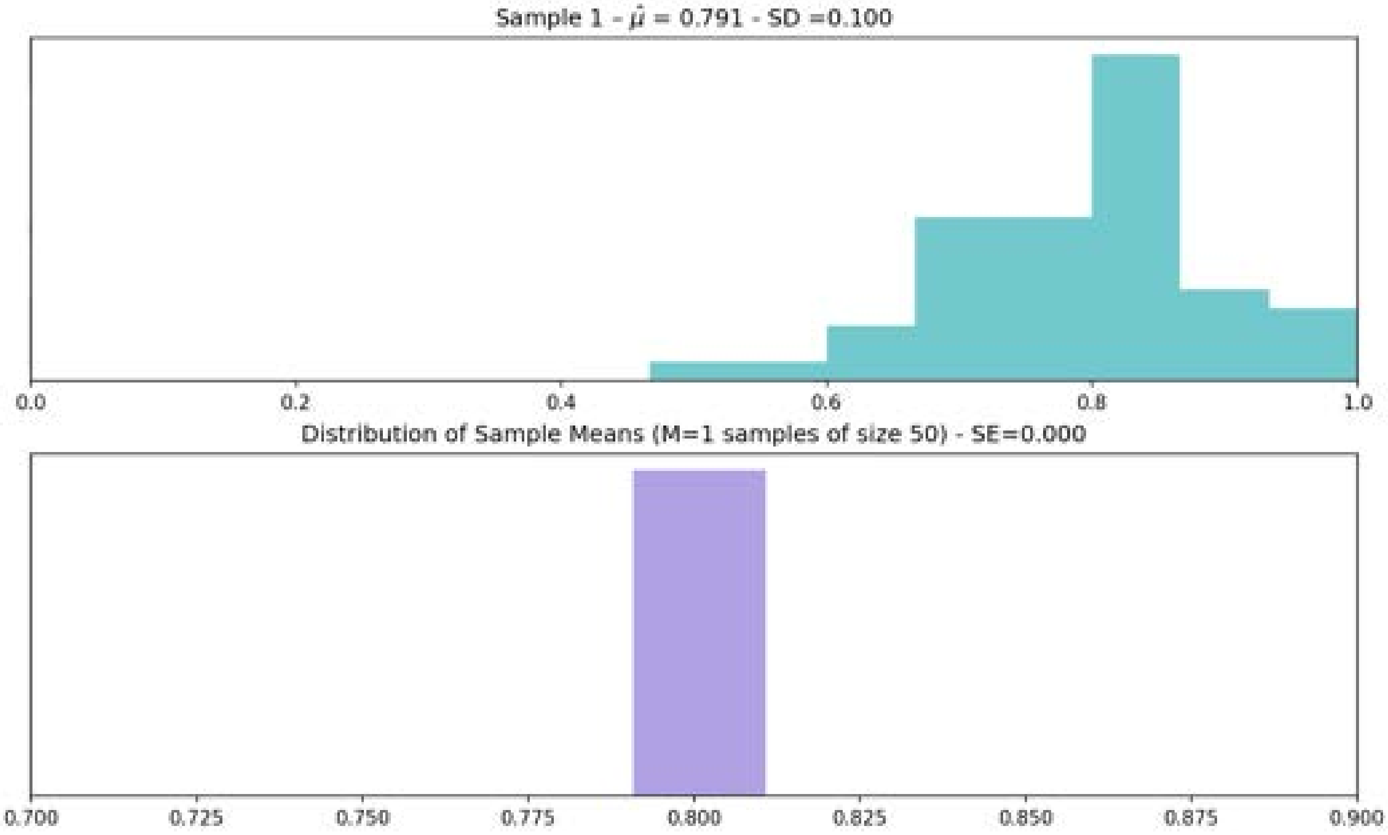
Sampling distribution

Distribution of a statistic (here the mean) across random samples

True population



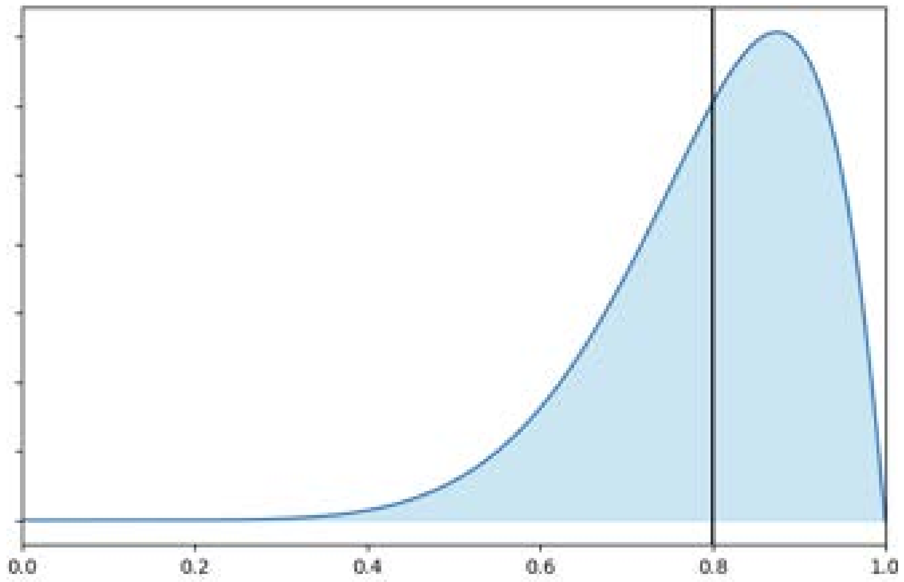
$\mu = 0.8$



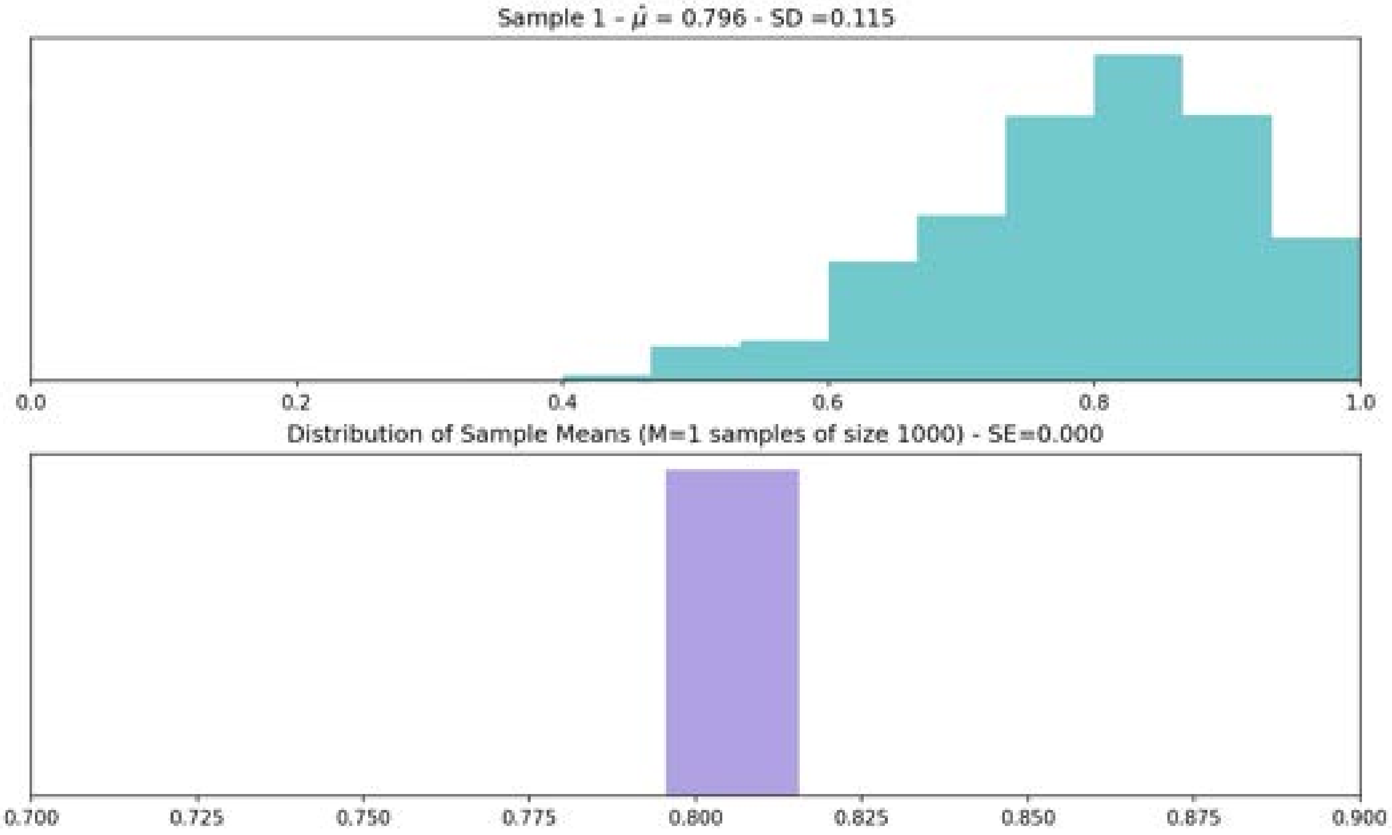
Sampling distribution

Shrinks with $\sqrt{n}$

True population

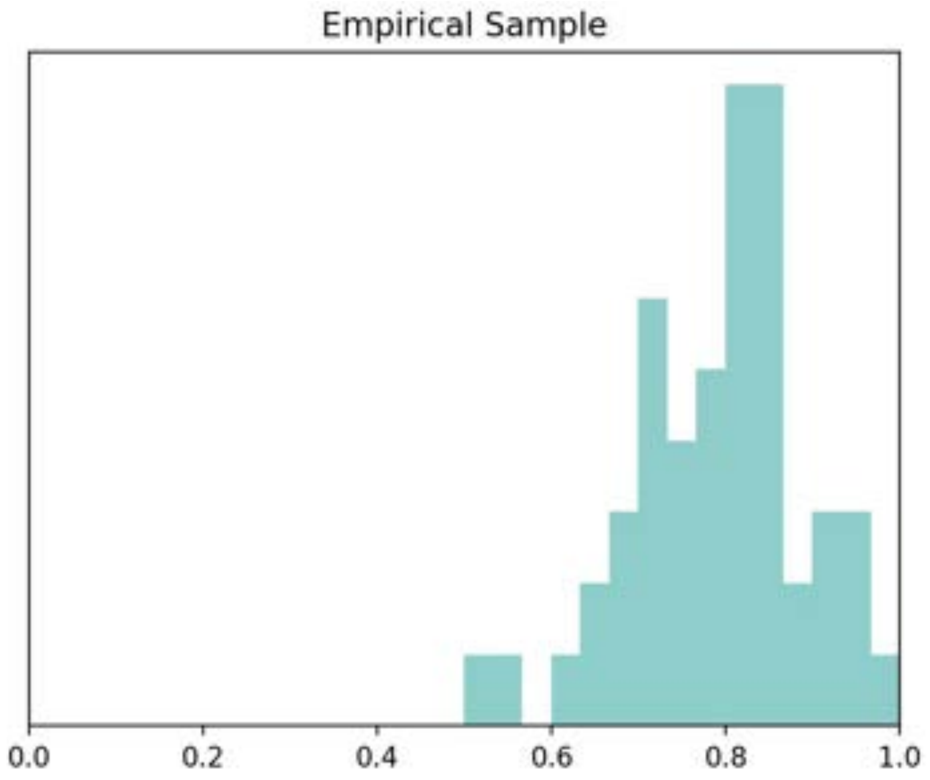


$\mu = 0.8$



Sampling distribution

Distribution of a statistic across random samples



OK but I have only one dataset!

Parametric methods

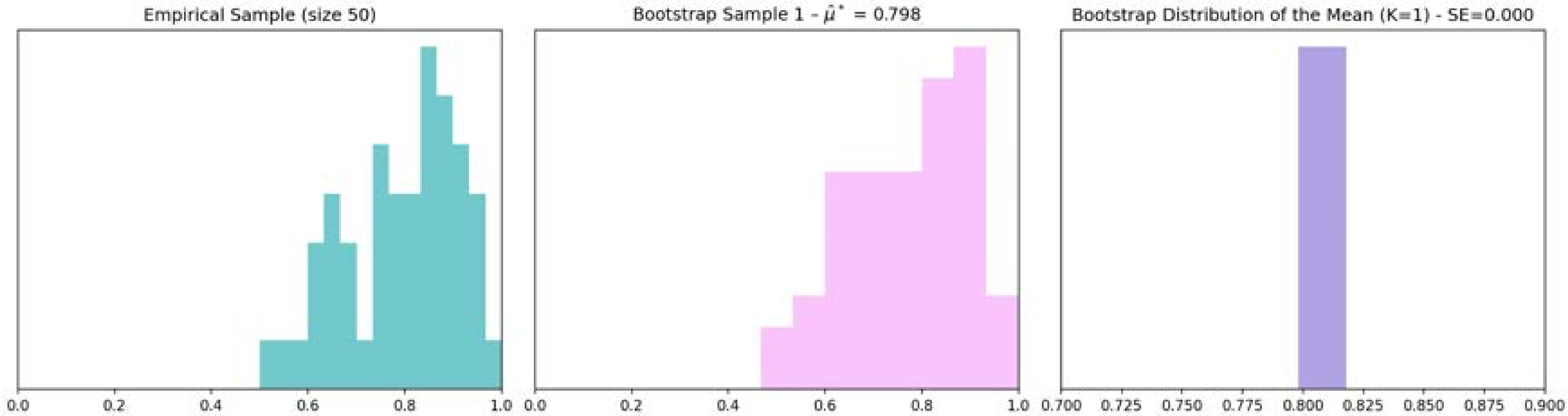
Underlying distribution is known (e.g. Gaussian)

Or asymptotic results

Non-parametric methods

In particular the boostrap

Bootstrap: approximating the sampling distribution



You have a sample of size n

Generate bootstrap samples

- Randomly draw n values with replacement from your sample
- Repeat this process many times (e.g., 9999 times)
- Each time, compute the statistic of interest (e.g., the mean) on the bootstrap sample

These values form the bootstrap distribution

This is an approximation of the sampling distribution of your statistic.

Reporting variability: which variability?

± what?

Methods	Accuracy
Method 1	0.892 ± 0.017
Method 2	0.895 ± 0.013
Method 3	0.883 ± 0.012
Proposed	**0.897** ± 0.013

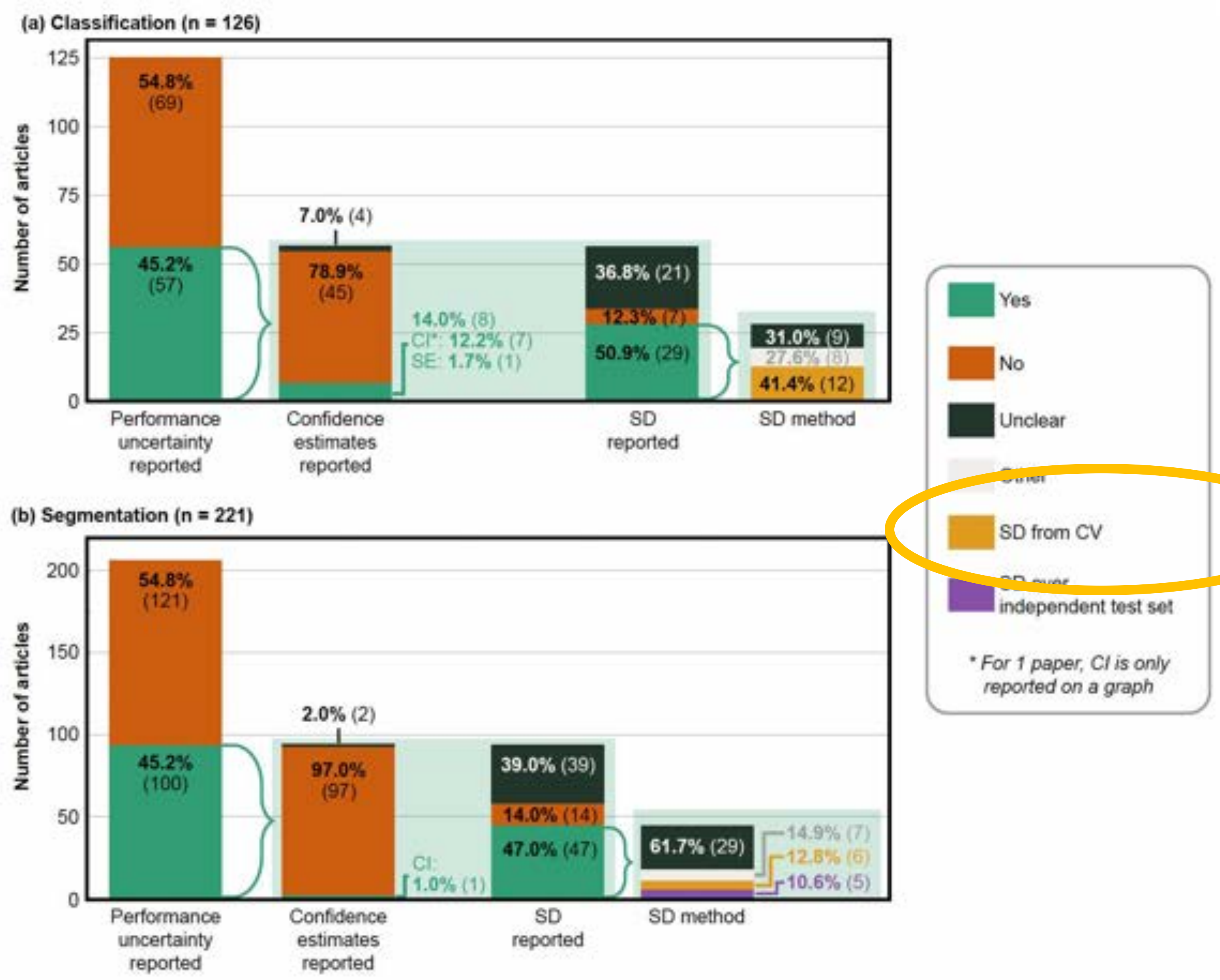
At least 3 possibilities

Standard-deviation (SD) of the metric over the test set

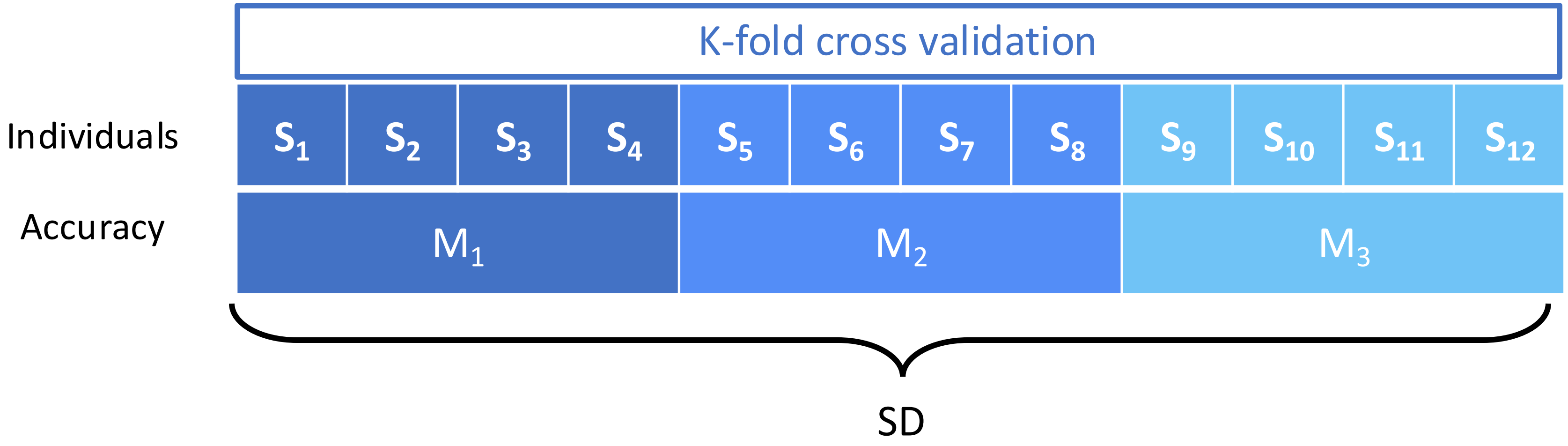
2 Standard-error (SE) of the summary statistic

3 Standard-deviation (SD) over cross-validation (CV)

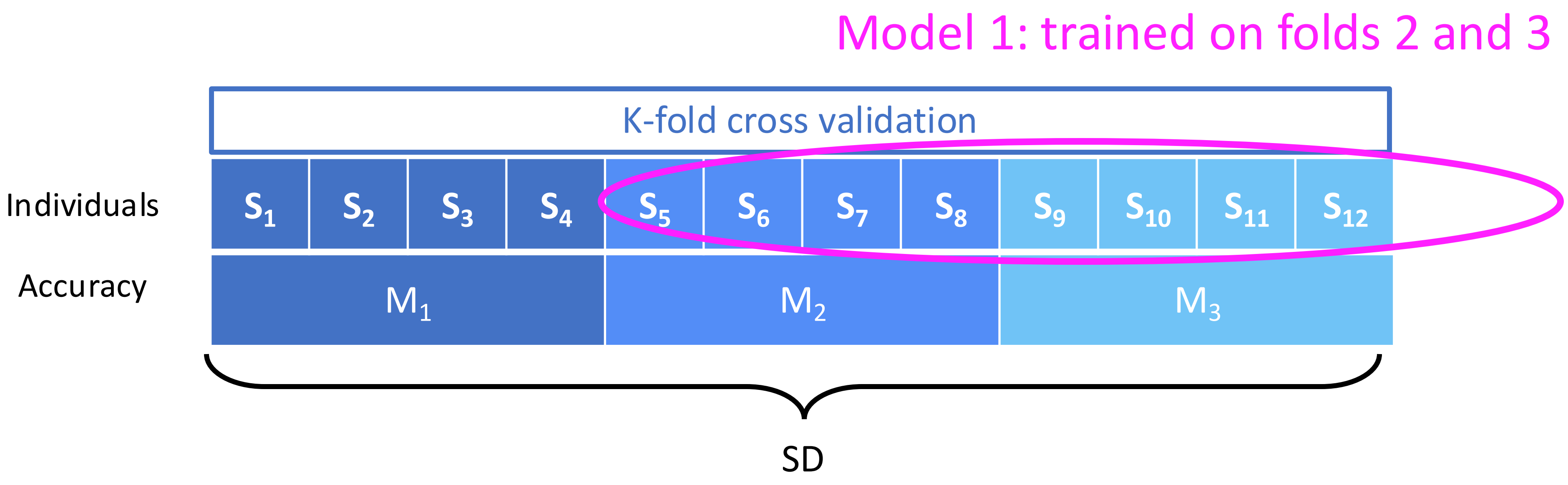
SD from cross-validation



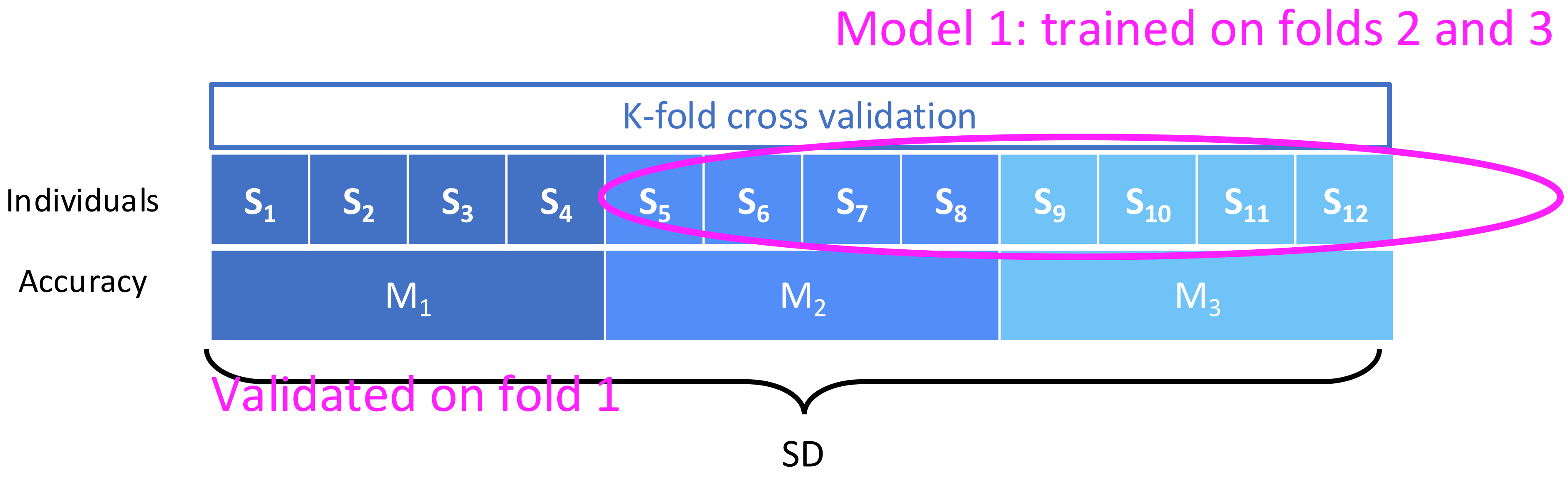
SD from cross-validation: how is it computed?



SD from cross-validation: how is it computed?

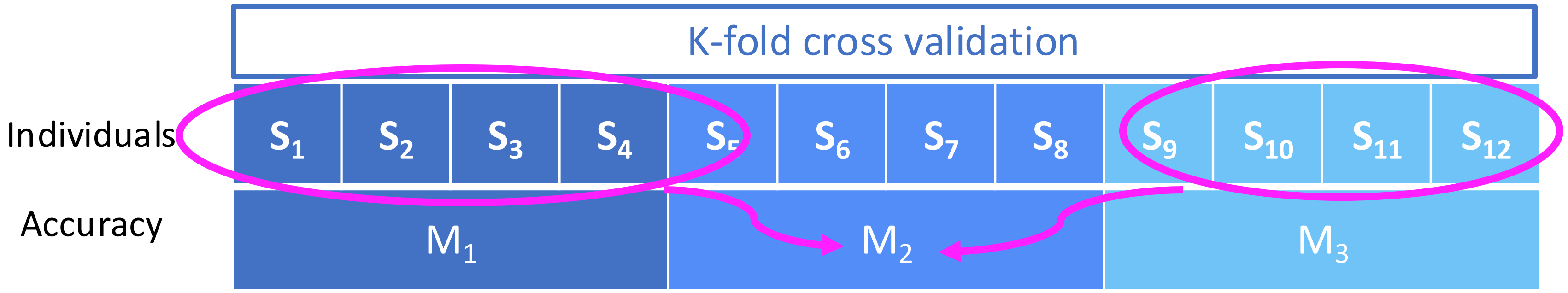


SD from cross-validation: how is it computed?



SD from cross-validation: how is it computed?

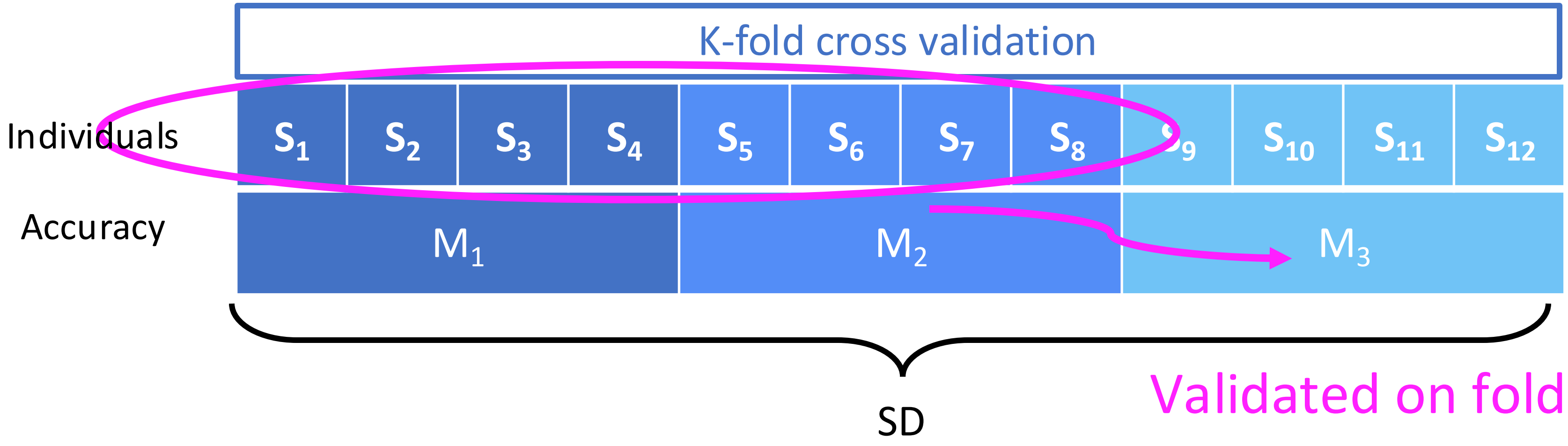
Model 2: trained on folds 1 and 3



Validated on fold 2

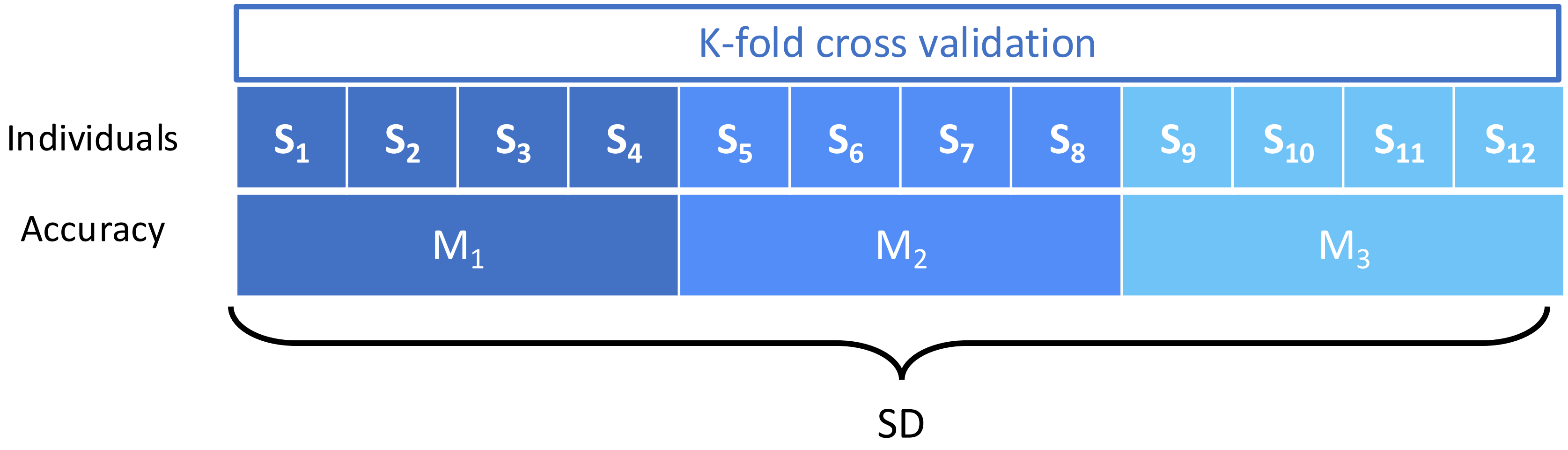
SD from cross-validation: how is it computed?

Model 3: trained on folds 1 and 2



Validated on fold 3

SD from cross-validation: the downside



SD is a biased estimator because of the induced covariance structure

No Unbiased Estimator of the Variance of K-Fold Cross-Validation

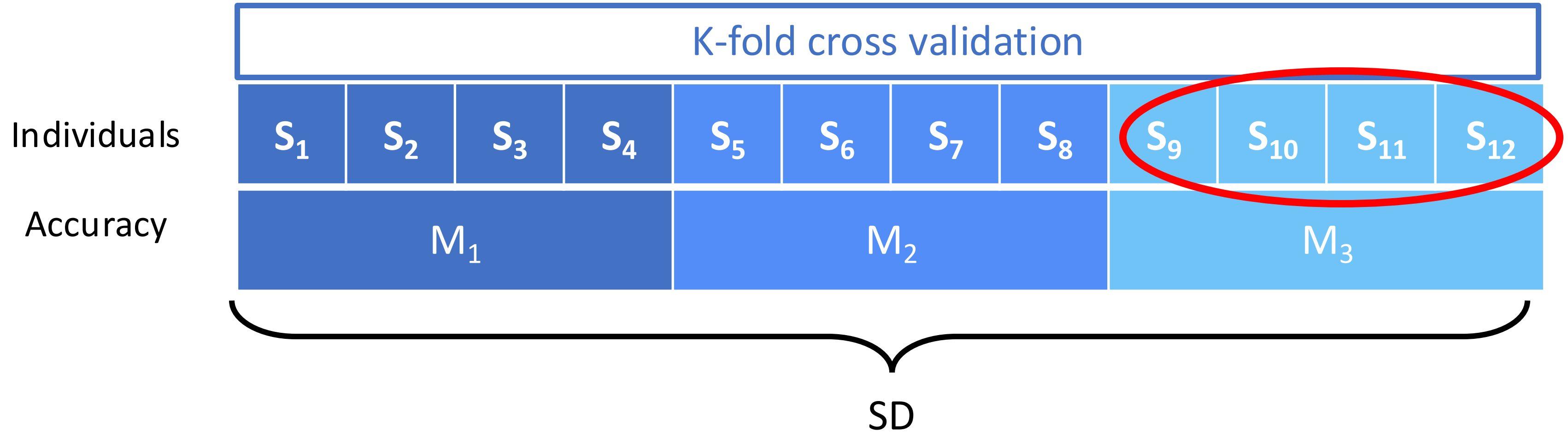
Yoshua Bengio BENGIOY@IRO.UMONTREAL.CA
Dept. IRO, Université de Montréal
C.P. 6128, Montreal, Qc, H3C 3J7, Canada

Yves Grandvalet YVES.GRANDVALET@UTC.FR
Heudiasyc, UMR CNRS 6599
Université de Technologie de Compiègne, France

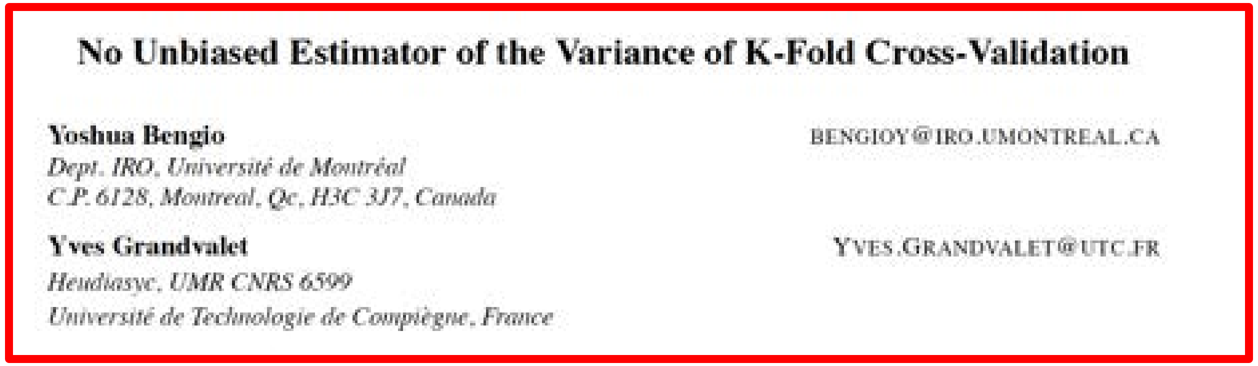
(Bengio and Grandvalet, 2004; Nadeau and Bengio, 2003)

SD from cross-validation: the downside

E.g. Model 1 and Model 2 share fold 3



SD is a biased estimator because of the induced covariance structure



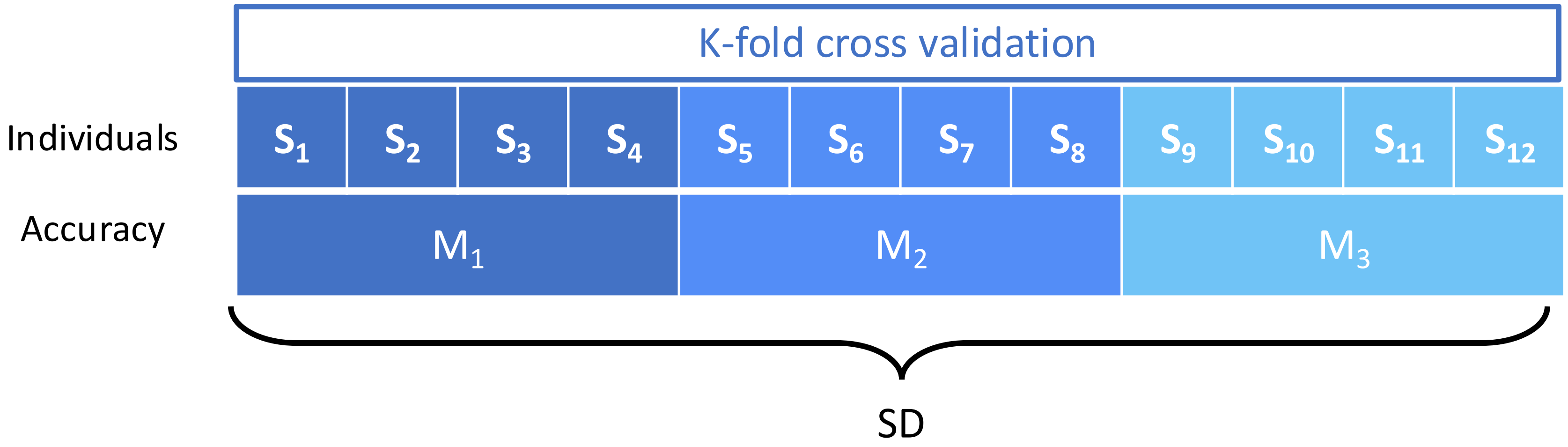
No Unbiased Estimator of the Variance of K-Fold Cross-Validation

Yoshua Bengio BENGIOY@IRO.UMONTREAL.CA
Dept. IRO, Université de Montréal
C.P. 6128, Montreal, Qc, H3C 3J7, Canada

Yves Grandvalet YVES.GRANDVALET@UTC.FR
Heudiasyc, UMR CNRS 6599
Université de Technologie de Compiègne, France

(Bengio and Grandvalet, 2004; Nadeau and Bengio, 2003)

SD from cross-validation: the downside



SD is a biased estimator because of the induced covariance structure

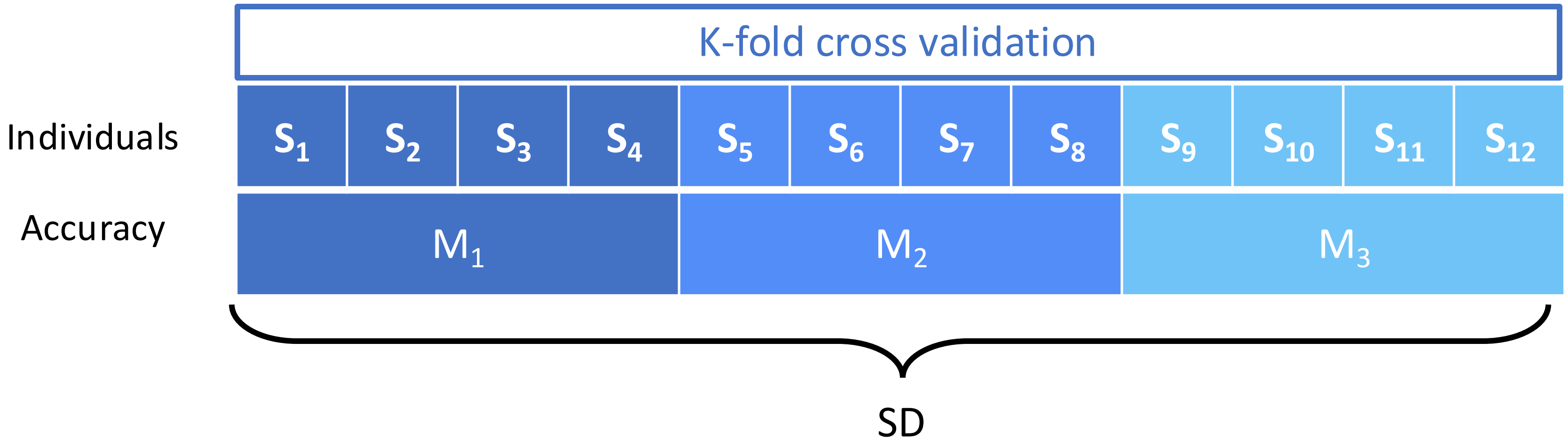
⚠ Important consequences

⛔ No statistical inference (e.g. statistical testing)

‼ Is only an **empirical descriptive statistic**

SD from cross-validation: the benefit

A tool for studying variability of learning procedures

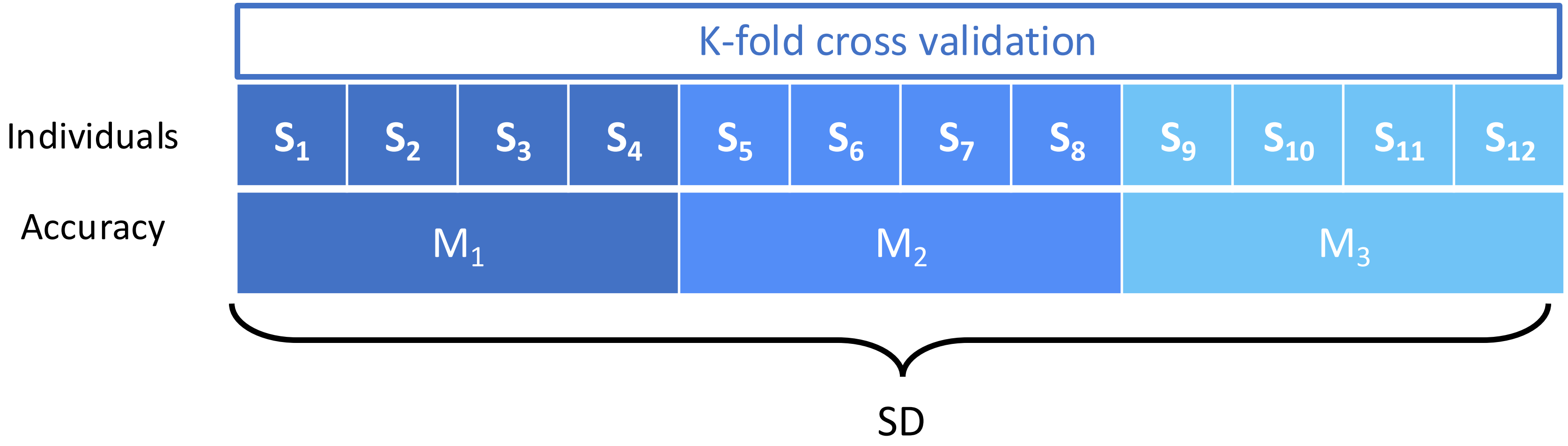


SD from CV provides **empirical** information about variability of a **learning procedure** not of the **trained model**

This is still useful information

SD from cross-validation: the benefit

A tool for studying variability of learning procedures



You can enrich this information:

- letting other factors vary: random seeds, optimized hyperparameters…
- doing more runs/data splits (e.g. repeated shuffle split)

Back to FDA recommendations: confidence intervals

Methods	DSC	HD95
Method 1	79.9 [76.6, 82.2]	8.05 [6.85, 9.37]
Method 2	79.7 [76.4, 82.3]	8.11 [6.93, 9.42]
Method 3	80.1 [76.9, 82.5]	7.91 [6.71, 9.22]
Proposed	**80.2 [77.1, 82.6]**	**7.73 [6.65, 8.91]**

[….] All performance estimates should be provided with confidence intervals […]

FDA-2024-D-4488: Artificial Intelligence-Enabled Device Software Functions: Lifecycle Management and Marketing Submission Recommendations

FDA U.S. FOOD & DRUG ADMINISTRATION

Confidence intervals

➡ Need to be computed from independent test set

Various methods including

1

Parametric methods

✅ Theoretical guarantees **when distributional assumptions met**

❌ Each summary statistic requires special treatment

2

Bootstrap

❌ Less theoretical guarantees

✅ No distributional assumptions

✅ Can be applied to many summary statistics

Confidence intervals

No guidance on CI on medical imaging AI

Confidence intervals

No guidance on CI on medical imaging AI

Unlike other fields

Psychology

Educational and Psychological Measurement

Impact Factor: 2.1 / 5-Year Impact Factor: 4.1 | Journal Homepage | Submis

Restricted access | Research article | First published February 2005

The Effects of Nonnormal Distributions on Confidence Intervals Around the Standardized Mean Difference: Bootstrap and Parametric Confidence Intervals

Ken Kelley View all authors and affiliations

Volume 65, Issue 1 | https://doi.org/10.1177/0013164404264850

Contents | Get access | Cite article | Share options | Information, rights and permissions | Metrics and citations

Confidence intervals

No guidance on CI on medical imaging AI

Unlike other fields

⚡Psychology

⚡Genetics

Confidence intervals

No guidance on CI on medical imaging AI

Unlike other fields

Psychology

Genetics

Economics

Confidence intervals

No guidance on CI on medical imaging AI

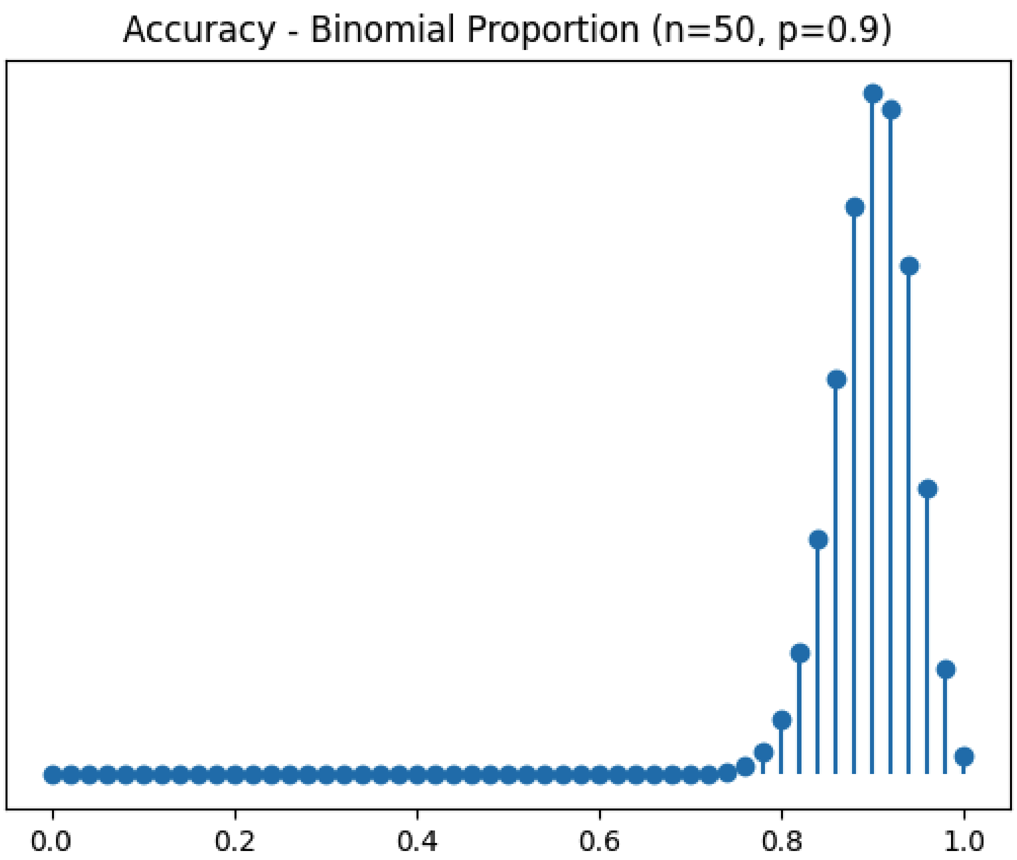
⚠ Even though we have so many metrics

Image source: https://metrics-reloaded.dkfz.de/

Confidence intervals

Metric distribution

✅ In **a few** cases, it is known

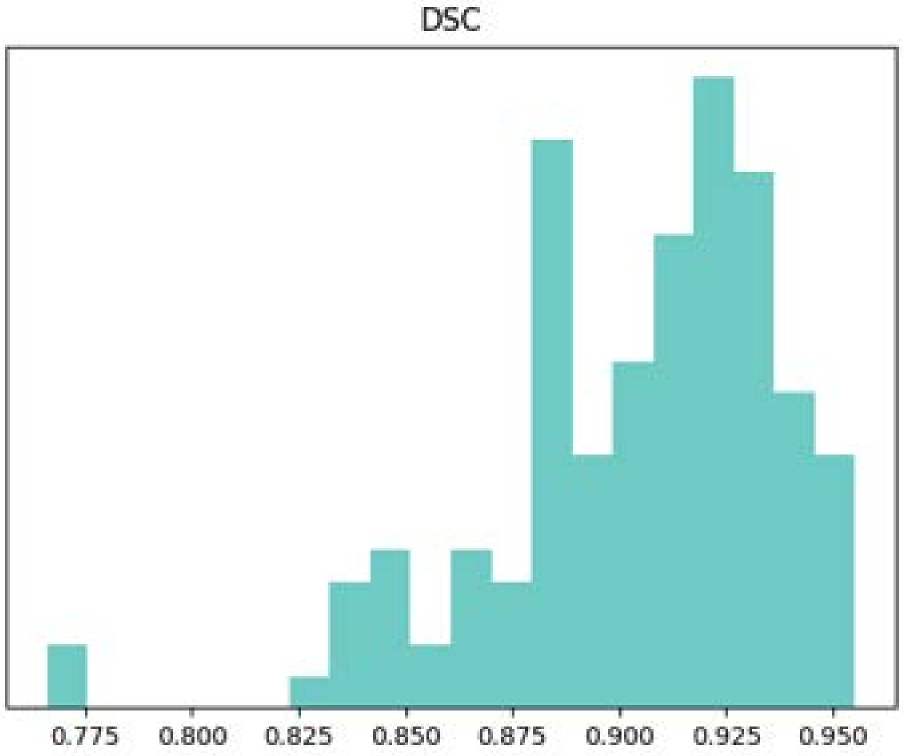


Accuracy follows a binomial proportion

Confidence intervals

Metric distribution

❌ In **most** cases, it is not

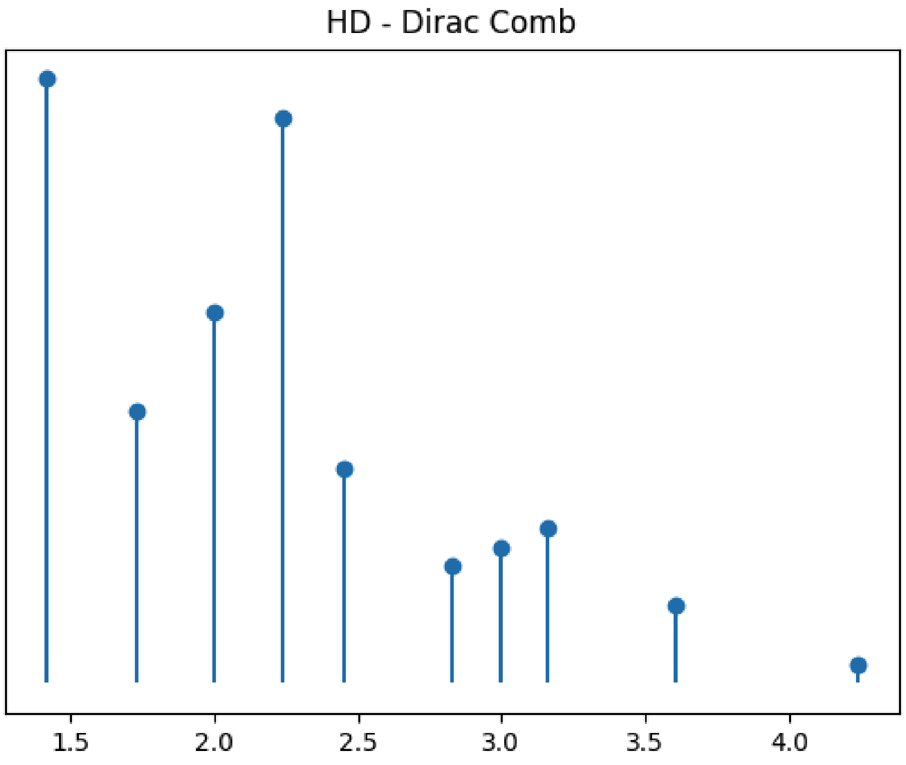


Some are semi-continuous

Confidence intervals

Metric distribution

❌ In **most** cases, it is not



Some are discrete

Confidence intervals

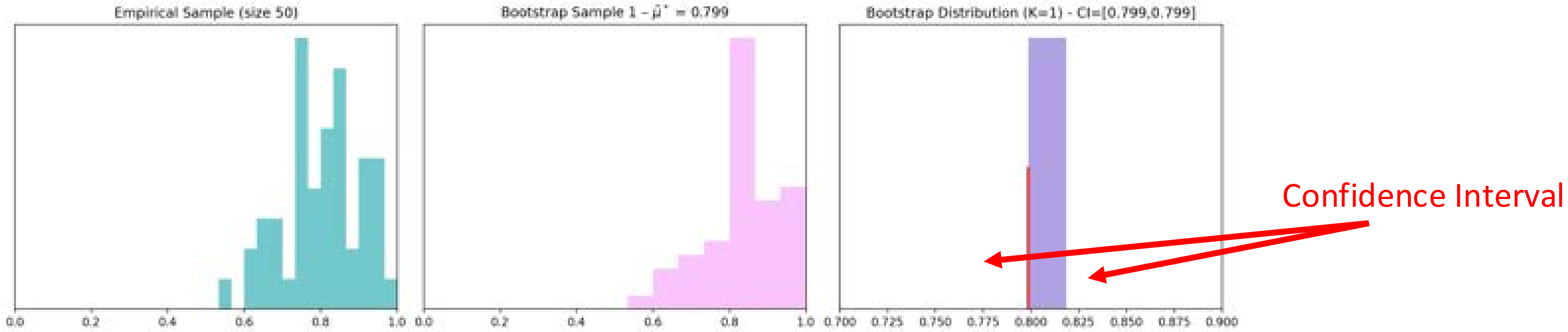
In the absence of specific guidelines for medical imaging AI

➡ Bootstrap on the test set results

- ✅ No distributional assumptions
- ⚠ Test set observations need to be independent

💡 Which bootstrap variant to choose?

- ➡ Percentile bootstrap: robust (safest choice in the absence of more precise guidance)



1. Current practices

2. Strength of outperformance claims

3. Areas for improvement

Take home messages

Take home message (1)

→ Variability reporting is essential for clinical translation

Commonly encountered results tables

Methods	Accuracy	AUC
Method 1	0.828	0.862
Method 2	0.821	0.857
Method 3	0.847	0.889
Proposed	**0.851**	**0.891**

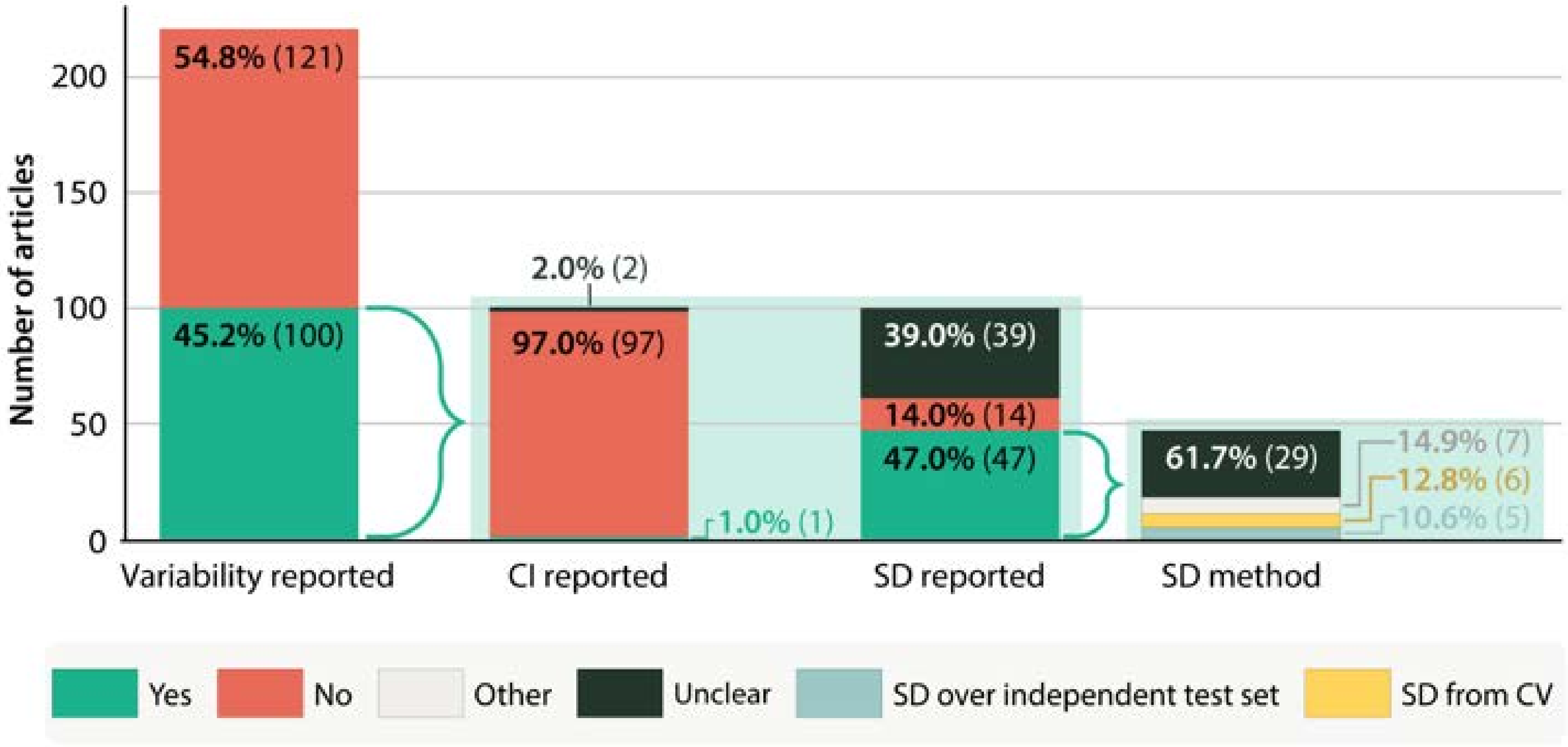
The statistical analysis plays a critical role in the assessment of […] ML performance but may be under-appreciated by many ML developers. […] There are still publications that present point estimates of ML performance without quantification of uncertainties.

Weijie Chen, Daniel Krainak, Berkman Sahiner, Nicholas Petrick, A Regulatory Science Perspective on Performance Assessment of Machine Learning Algorithms in Imaging, 2023

FDA U.S. FOOD & DRUG ADMINISTRATION

Take home message (2)

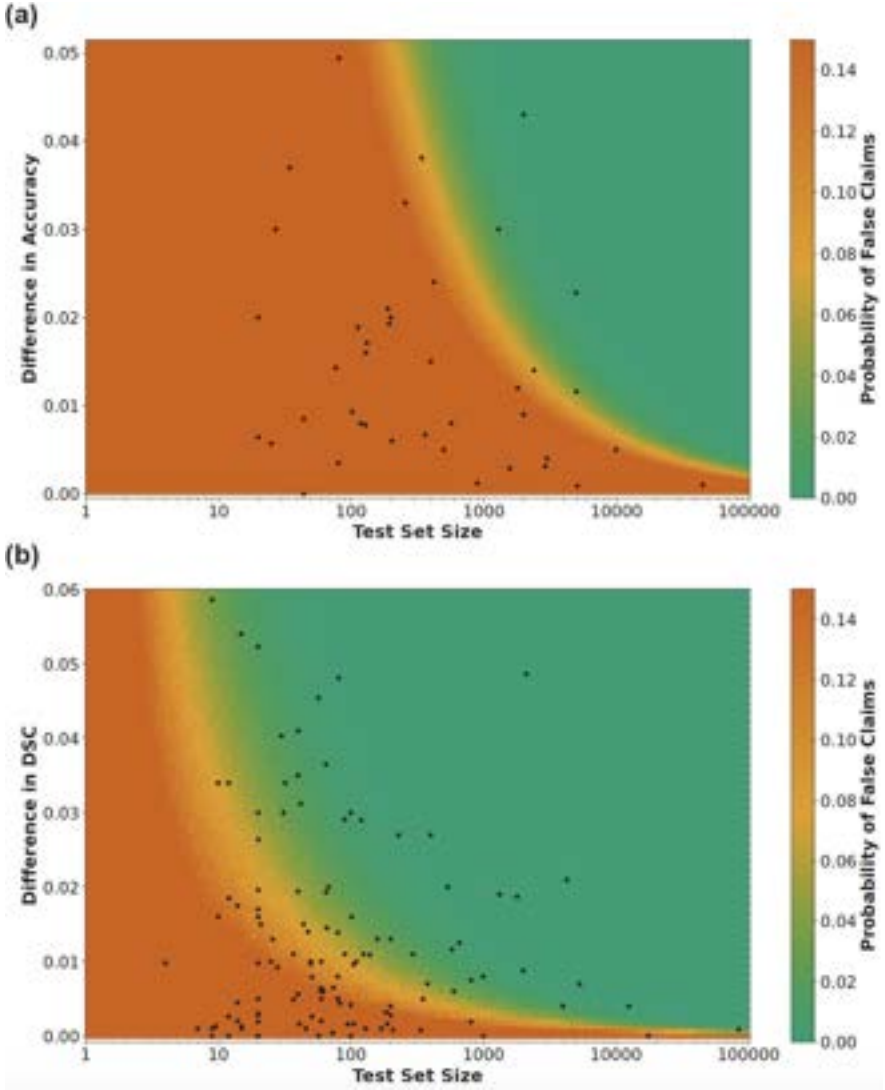
➡ Majority of papers do not report variability



Christodoulou, Evangelia, et al. "Confidence intervals uncovered: Are we ready for real-world medical imaging AI?." International Conference *on Medical Image Computing and Computer-Assisted Intervention*. Cham: Springer Nature Switzerland, 2024.

Take home message (3)

➡ Claims of outperformance are often unsubstantiated



>5% probability of false claims of outperformance

(a) classification: **>86%**

(b) segmentation: **>53%**

MICCAI 2023 papers

Take home message (4)

Generated by DALL-E

Take home message (4)

Generated by DALL-E

✅ Use appropriate data splitting

✅ Report variability on trained models using a test set

✅ Bootstrap on the test set is a reasonable first choice

⚠️ Community needs guidelines for variability reporting

Funding acknowledgements

Funded by the European Union

European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement no. 101002198, NEURAL SPICING)

HELMHOLTZ IMAGING

A platform of the Helmholtz Incubator on Information and Data Science

NCT NATIONAL CENTER FOR TUMOR DISEASES HEIDELBERG

National Center for Tumor Diseases (NCT) Heidelberg's Surgical Oncology Program

Health + Life Science Alliance Heidelberg Mannheim

State of Baden-Württemberg Innovation Campus Health + Life Science Alliance Heidelberg Mannheim

Funded by the European Union

European Union's Horizon Europe Framework Programme: grant number 101136607, project CLARA

FRANCE 2030 · INVESTISSEMENTS D'AVENIR · PR[AI]RIE Paris School

"France 2030" program (reference ANR-23-IACL-0008, project PRAIRIE-PSAI), "Investissements d'avenir" program (reference ANR-19-P3IA-0001, project PRAIRIE 3IA Institute)

INVESTISSEMENTS D'AVENIR · ICM Paris Brain Institute

"Investissements d'avenir" program (reference ANR-10-IAIHU-06, project Agence Nationale de la Recherche-10-IA Institut Hospitalo-Universitaire-6)

A collaborative effort

GERMAN CANCER RESEARCH CENTER IN THE HELMHOLTZ ASSOCIATION

HELMHOLTZ IMAGING

Dr Evangelia Christodoulou
Dr Annika Reinke
Patrick Godau
Piotr Kalinowski
Dr Rola Houhou
Selen Erkan
Leon D. Mayer
Dr Minu D. Tizabi
Prof Dr Annette Kopp-Schneider
Prof Dr Lena Maier-Hein

Paris Brain Institute ICM

Pascaline André
Dr Ninon Burgos
Sofiène Boutaj
Dr Sophie Loizillon
Maëlys Solal
Charles Heitz
Antoine Gilson
Dr Olivier Colliot

SODA team

Dr Gaël Varoquaux

Dr Michella Antonelli
Dr Jorge Cardoso

Dr Carole Sudre

IT UNIVERSITY OF COPENHAGEN

Dr Veronika Cheplygina

Dr Nicola Rieke

MASARYK UNIVERSITY

Prof RNDr. Michal Kozubek

Dr Amber Simpson

MICCAI

SIG for Challenges

Evaluation and Benchmarking WG