Common limitations of performance metrics in biomedical image analysis

A. Reinke[1,2,], M. Eisenmann, M.D. Tizabi, C.H. Sudre, T. Rädsch, M. Antonelli, T. Arbel, S. Bakas, M.J. Cardoso, V. Cheplygina, K. Farahani, B. Glocker, P. Godau, D. Heckmann-Nötzel, F. Isensee, P. Jannin, C.E. Kahn, J. Kleesiek, M. Kozubek, T. Kurc, B.A. Landman, G. Litjens, K. Maier-Hein, A.L. Martel, B. Menze, H. Müller, J. Petersen, M. Reyes, M. Riegler, N. Rieke, B. Stieltjes, R.M. Summers, S.A. Tsaftaris, B. van Ginneken, A. Kopp-Schneider, P. Jäger, L. Maier-Hein*

[1] Div. Computer Assisted Medical Interventions and HIP Helmholtz Imaging Platform, German Cancer Research Center (DKFZ)
[2] Faculty of Mathematics and Computer Science, Heidelberg University
**** Full list of affiliations:** https://arxiv.org/abs/2104.05642*

dkfz. GERMAN CANCER RESEARCH CENTER IN THE HELMHOLTZ ASSOCIATION
Research for a Life without Cancer

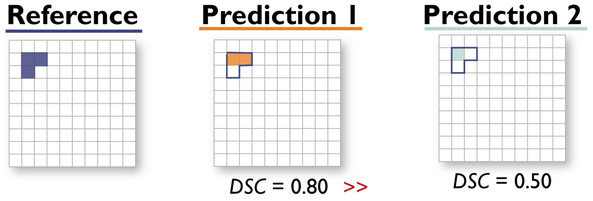
Motivation

While the importance of biomedical image analysis is increasing at an enormous pace, recent meta-research revealed **major flaws with respect to algorithm validation**. Performance metrics are key for objective, transparent and comparative performance assessment, but **little attention has been given to their pitfalls**.

An international consortium has the mission to generate **best practice recommendations for metrics in medical image analysis**. Consensus building is achieved via a **Delphi process**.

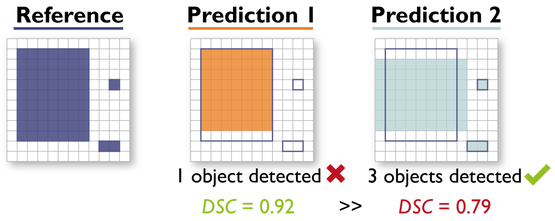
Numerous pitfalls related to metrics

Example of failure:
Effect of small structures in *segmentation tasks*



Single-pixel differences can have huge effects on the metric scores, especially relevant given high inter-rater variability and non-determinism of AI.

Example of failure:
Effect of task/metric mismatch



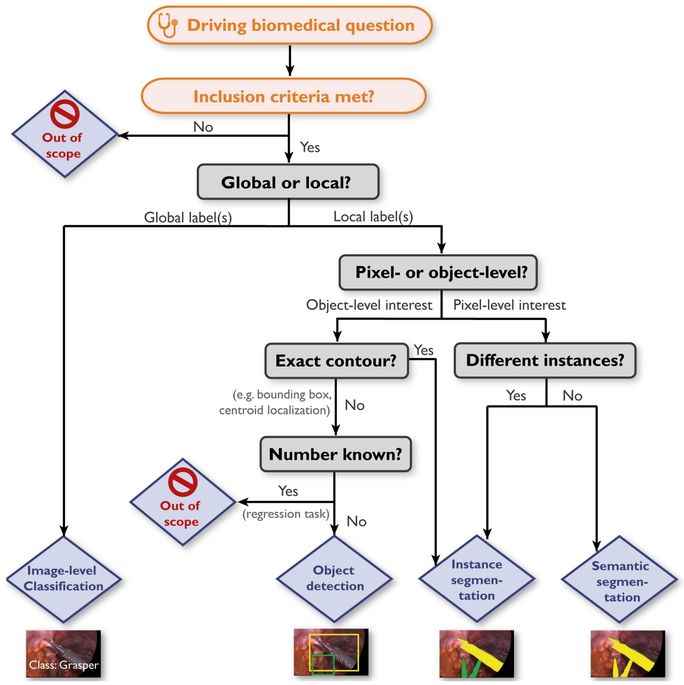
The DSC is strongly biased against single objects, therefore not appropriate to measure the detection of multiple objects.

Objective 1: Choosing the right task

Task mapping

Maps from the driving biomedical question to the correct image processing task.

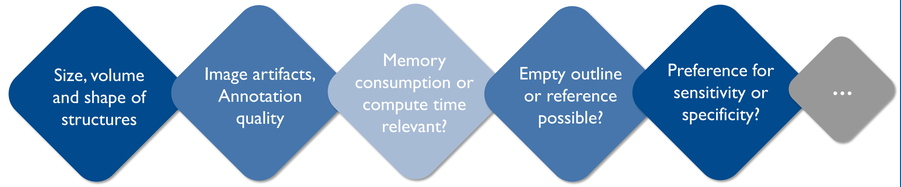
Inclusion criteria for tasks: Interpretable as *classification task* - Image-level classification, object detection (object-level), instance and semantic segmentation (pixel-level)



Objective 2: Choosing the right metric for a given task

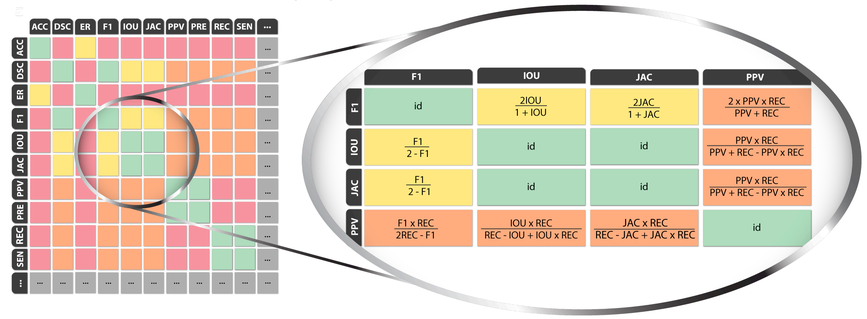
Problem characteristics

The choice of the metrics is dependent on the **driving biomedical question** and its specific **problem characteristics**.



Metric relations

Many metrics are very similar in mathematical terms. Only metrics measuring different properties should be combined.



Synonyms: computing one exactly produces the other value
Closely related: given one, there is little effort needed to compute the other
Loosely related: share some computation step
Not related

Coming up:
Concrete recommendations for most common problems!

Full paper: *Reinke et al. Common Limitations of Image Processing Metrics: A Picture Story. https://arxiv.org/abs/2104.05642*

You would like to contribute? Contact us!

HIP HELMHOLTZ IMAGING PLATFORM

Common limitations of performance metrics in biomedical image analysis

A. Reinke[1,2,], M. Eisenmann, M.D. Tizabi, C.H. Sudre, T. Rädsch, M. Antonelli, T. Arbel, S. Bakas, M.J. Cardoso, V. Cheplygina, K. Farahani, B. Glocker, P. Godau, D. Heckmann-Nötzel, F. Isensee, P. Jannin, C.E. Kahn, J. Kleesiek, M. Kozubek, T. Kurc, B.A. Landman, G. Litjens, K. Maier-Hein, A.L. Martel, B. Menze, H. Müller, J. Petersen, M. Reyes, M. Riegler, N. Rieke, B. Stieltjes, R.M. Summers, S.A. Tsaftaris, B. van Ginneken, A. Kopp-Schneider, P. Jäger, L. Maier-Hein*

[1] Div. Computer Assisted Medical Interventions and HIP Helmholtz Imaging Platform, German Cancer Research Center (DKFZ)
[2] Faculty of Mathematics and Computer Science, Heidelberg University

Full list of affiliations: *https://arxiv.org/abs/2104.05642*

dkfz. GERMAN CANCER RESEARCH CENTER IN THE HELMHOLTZ ASSOCIATION

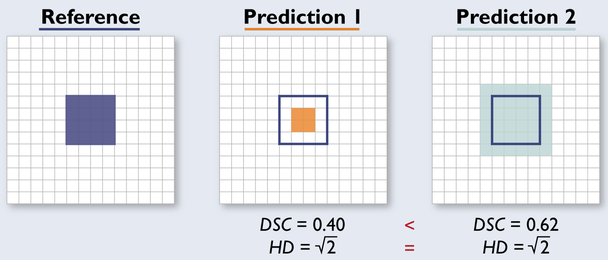
Research for a Life without Cancer

Common limitations of segmentation metrics

Mathematical properties

Awareness of a metric's mathematical properties is crucial when determining its suitability for a given task.

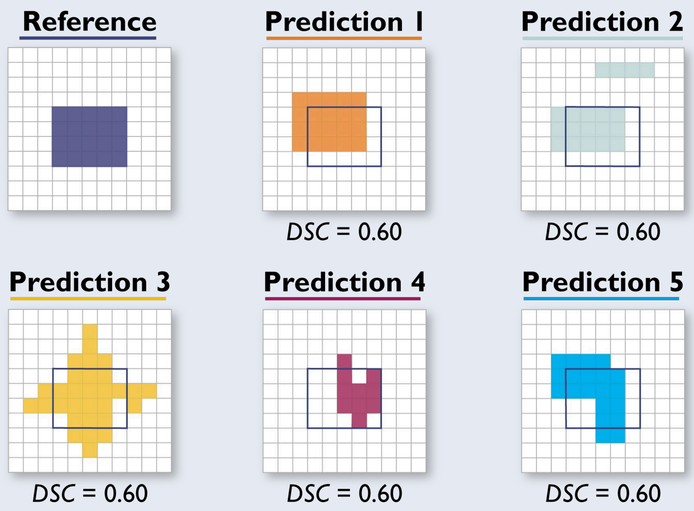
Example of failure: Effect of over- and undersegmentation



Over- and undersegmentation are not treated equally in many overlap-based metrics although the difference may be clinically relevant.

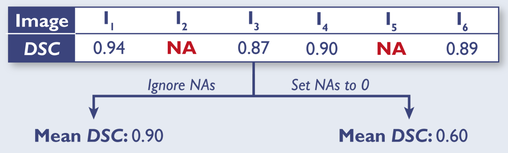
Example of failure: Effect of shape unawareness

Overlap measures are not designed to uncover differences in shape. This is an important problem for many applications, such as radiotherapy.



Metric aggregation

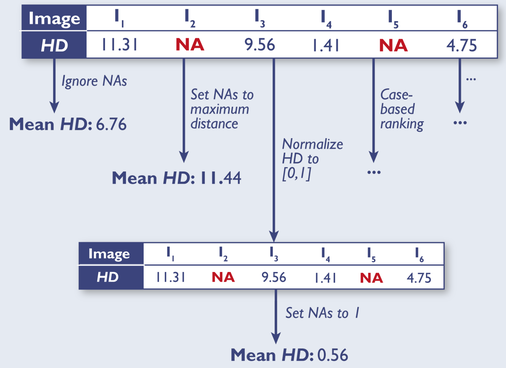
To produce an aggregated metric value over many images, multiple merging strategies may be applied. Special care has to be given to missing values.



Example of failure: *Ignoring missing values leads to a substantially higher DSC compared to setting missing values to the worst possible value (here: 0).*

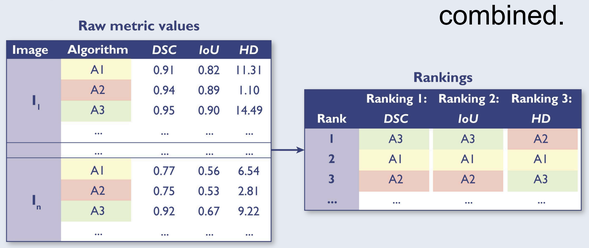
Example of failure: *For distance-based measures without lower/upper bounds, the strategy of how to deal with missing values is not trivial. One may choose the maximum distance of the image or normalize the metric values to [0,1] and use the worst possible value (here: 1).*

Crucially, however, every choice will produce a different aggregated value, thus potentially affecting the ranking.



Metric combination

A single metric typically does not reflect all important aspects that are essential for algorithm validation. Hence, multiple metrics with different properties should be combined.



Example of failure: *Mutually dependent metrics (DSC and IoU) will lead to the same ranking and should be used interchanging, whereas metrics measuring different properties (HD) will lead to a different ranking.*

Full paper: *Reinke et al. Common Limitations of Image Processing Metrics: A Picture Story. https://arxiv.org/abs/2104.05642*

You would like to contribute? Contact us!

HIP HELMHOLTZ IMAGING PLATFORM

Common limitations of performance metrics in biomedical image analysis

A. Reinke[1,2,], M. Eisenmann, M.D. Tizabi, C.H. Sudre, T. Rädsch, M. Antonelli, T. Arbel, S. Bakas, M.J. Cardoso, V. Cheplygina, K. Farahani, B. Glocker, P. Godau, D. Heckmann-Nötzel, F. Isensee, P. Jannin, C.E. Kahn, J. Kleesiek, M. Kozubek, T. Kurc, B.A. Landman, G. Litjens, K. Maier-Hein, A.L. Martel, B. Menze, H. Müller, J. Petersen, M. Reyes, M. Riegler, N. Rieke, B. Stieltjes, R.M. Summers, S.A. Tsaftaris, B. van Ginneken, A. Kopp-Schneider, P. Jäger, L. Maier-Hein*

[1] Div. Computer Assisted Medical Interventions and HIP Helmholtz Imaging Platform, German Cancer Research Center (DKFZ)
[2] Faculty of Mathematics and Computer Science, Heidelberg University

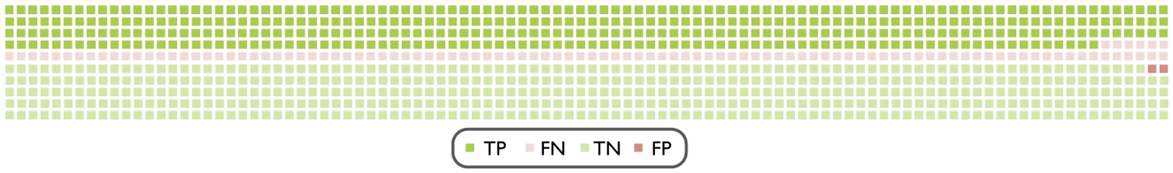
Full list of affiliations: *https://arxiv.org/abs/2104.05642*

dkfz.
GERMAN CANCER RESEARCH CENTER IN THE HELMHOLTZ ASSOCIATION
Research for a Life without Cancer

Common limitations of classification metrics

Example of failure:

Effect of missing prevalence correction



Case-control study, prevalence = 0.5

Uncorrected *PPV* = 0.99
Uncorrected *NPV* = 0.83

General population, prevalence = 0.005

Prevalence-corrected *PPV* = 0.50
Prevalence-corrected *NPV* = 0.99

In case-control studies, the positive predictive value (PPV) and the negative predictive value (NPV) should be seen as conditional probability of a disease being present based on a test result and the prevalence in a general population. Often, the prevalence of the case-control study is used (incorrectly) instead of the general prevalence. Without prevalence-correction, PPV and NPV may lead to completely wrong results.*

Prevalence = 0.5

$$PPV = \frac{TP}{TP + FP}$$

$$NPV = \frac{TN}{TN + FN}$$

Case-control studies with prevalence ≠ 0.5

$$PPV = \frac{Sensitivity \cdot Prevalence}{Sensitivity \cdot Prevalence + (1 - Specificity) \cdot (1 - Prevalence)}$$

$$NPV = \frac{Specificity \cdot (1 - Prevalence)}{(1 - Sensitivity) \cdot Prevalence + Specificity \cdot (1 - Prevalence)}$$

* Thanks to Bernhard Kainz for sharing the figure.

Common limitations of detection metrics

Definition of True Positives

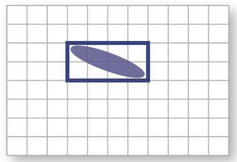
Definition of True Positives (TP) in object detection tasks is typically done by measuring the overlap of bounding boxes with the Intersection over Union (IoU). Depending on a threshold, the predicted object will be interpreted as TP or False Positive (FP). This definition is the pre-requisite for metric computation.

Example of failure:

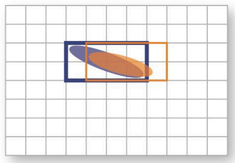
Effect of threshold choice

IoU **> 0.50:** True positive (TP)
IoU **≤ 0.50:** False positive (FP)

Reference

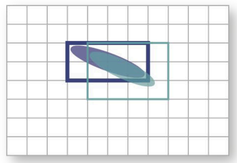


Prediction 1



IoU = 0.60 > 0.50
TP

Prediction 2



IoU = 0.43 < 0.50
FP

The threshold chosen to define TP and FP highly influences the metric values computed from them. Especially for small, diagonal structures, the size of bounding boxes changes quickly, leading to FP although the visual agreement would indicate TP (Prediction 2).

Full paper: *Reinke et al. Common Limitations of Image Processing Metrics: A Picture Story. https://arxiv.org/abs/2104.05642*

You would like to contribute? Contact us!

HIP HELMHOLTZ IMAGING PLATFORM