Common Limitations of Image Processing Metrics: A Picture Story

Annika Reinke[1,2], Matthias Eisenmann[1], Minu D. Tizabi[1], Carole H. Sudre[4,5,6], Tim Rädsch[1,3], Michela Antonelli[4,5], Tal Arbel[7], Spyridon Bakas[8,9,10], M. Jorge Cardoso[4,11], Veronika Cheplygina[12], Keyvan Farahani[13], Ben Glocker[14], Doreen Heckmann-Nötzel[1], Fabian Isensee[15], Pierre Jannin[16], Charles E. Kahn[17], Jens Kleesiek[18], Tahsin Kurc[19], Michal Kozubek[20], Bennett A. Landman[21], Geert Litjens[22,23], Klaus Maier-Hein[15], Bjoern Menze[24], Henning Müller[25,26], Jens Petersen[15], Mauricio Reyes[27], Nicola Rieke[28,29], Bram Stieltjes[30], Ronald M. Summers[31], Sotirios A. Tsaftaris[32], Bram van Ginneken[23,33], Annette Kopp-Schneider[34], Paul Jäger[15], Lena Maier-Hein[1,2,35]

[1]Div. Computer Assisted Medical Interventions, German Cancer Research Center (DKFZ), Germany
[2]Faculty of Mathematics and Computer Science, Heidelberg University, Germany
[3]understandAI GmbH, Germany
[4]School of Biomedical Engineering and Imaging Science, King's College London, UK
[5]Centre for Medical Image Computing, University College London, UK
[6]MRC Unit for Lifelong Health and Ageing at UCL, University College London, UK
[7]Centre for Intelligent Machines, McGill University, MILA, Canada
[8]Center for Biomedical Image Computing & Analytics, University of Pennsylvania, USA
[9]Department of Radiology, Perelman School of Medicine, University of Pennsylvania, USA
[10]Department of Pathology & Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, USA
[11]Department of Medical Physics and Biomedical Engineering, University College London, UK
[12]IT University of Copenhagen, Denmark
[13]Center for Biomedical Informatics and Information Technology, National Cancer Institute, USA
[14]Biomedical Image Analysis Group, Department of Computing, Imperial College London, UK
[15]Div. Medical Image Computing, German Cancer Research Center (DKFZ), Germany
[16]Laboratoire Traitement du Signal et de l'Image – UMR_S 1099, Université de Rennes 1, Inserm, France
[17]Perelman School of Medicine, University of Pennsylvania, USA
[18]Translational Image-guided Oncology (TIO), Institute for AI in Medicine (IKIM), University Medicine Essen, Essen, Germany
[19]Stony Brook Cancer Center, Stony Brook University, USA
[20]Centre for Biomedical Image Analysis, Masaryk University, Czech Republic
[21]Electrical Engineering, Vanderbilt University, USA
[22]Department of Pathology, Radboud University Medical Center, The Netherlands
[23]Radboud University Medical Center, Radboud Institute for Health Sciences, The Netherlands
[24]Department of Quantitative Biomedicine, University of Zurich, Switzerland
[25]University of Applied Sciences Western Switzerland (HES-SO), Switzerland
[26]Medical Faculty, University of Geneva, Switzerland
[27]Healthcare Imaging A.I., Insel Data Science Center, Bern University Hospital, Switzerland
[28]NVIDIA GmbH, Munich, Germany
[29]Technical University of Munich, Munich, Germany
[30]Department of Radiology, University Hospital of Basel, Switzerland
[31]Radiology and Imaging Sciences, Clinical Center, National Institutes of Health, USA
[32]School of Engineering, The University of Edinburgh, Scotland
[33]Fraunhofer MEVIS, Germany
[34]Div. Biostatistics, German Cancer Research Center (DKFZ), Germany
[35]Medical Faculty, Heidelberg University, Germany

Disclaimer

This is a continuous paper on limitations of commonly used metrics in image analysis. The current version discusses segmentation metrics only, while future versions will also include metrics for classification and detection tasks. For missing references, use cases, other comments or questions, please contact `a.reinke@dkfz-heidelberg.de` or `l.maier-hein@dkfz-heidelberg.de`. Substantial contributions to this dynamic document will be acknowledged with a co-authorship.

ABSTRACT
While the importance of automatic image analysis is increasing at an enormous pace, recent meta-research revealed major flaws with respect to algorithm validation. Specifically, performance metrics are key for objective, transparent and comparative performance assessment, but relatively little attention has been given to the practical pitfalls when using specific metrics for a given image analysis task. A common mission of several international initiatives is therefore to provide researchers with guidelines and tools to choose the performance metrics in a problem-aware manner. This dynamically updated document has the purpose to illustrate important limitations of performance metrics commonly applied in the field of image analysis. The current version is based on a Delphi process on metrics conducted by an international consortium of image analysis experts.

KEYWORDS
Good Scientific Practice, Validation, Metrics, Image Processing, Segmentation, Medical Imaging, Dice similarity coefficient

1. Purpose

Metrics are key to assessing the performance of image analysis algorithms in an objective and meaningful manner. So far, however, relatively little attention has been given to the practical pitfalls when using specific metrics for a given image analysis task. An international survey (Maier-Hein et al., 2018), for example, revealed the choice of inappropriate metrics as one of the core problems related to performance assessment in medical image analysis. Similar problems are present in other fields of imaging research (Correia & Pereira, 2006; Honauer, Maier-Hein, & Kondermann, 2015).

Under the umbrella of the Helmholtz Imaging Platform (HIP)[1], three international initiatives have now joined forces to address these issues: the Biomedical Image Analysis Challenges (BIAS) initiative[2], the Medical Image Computing and Computer Assisted Interventions (MICCAI) Society's challenge working group, as well as the benchmarking working group of the MONAI framework[3]. A core mission is to provide researchers with guidelines and tools to choose the performance metrics in a problem-aware manner. This dynamically updated document aims to illustrate important pitfalls and drawbacks of metrics commonly applied in the field of image analysis. The current version is based on a Delphi process (Brown, 1968) on metrics conducted with an international consortium of medical image analysis experts.

[1] https://www.helmholtz-imaging.de/

[2] https://www.dkfz.de/en/cami/research/topics/biasInitiative.html?m=1611915160&

[3] https://monai.io/

2. Segmentation metrics

Image segmentation is one of the most popular image processing tasks. In fact, an international meta-analysis revealed segmentation as the most frequent medical image processing task in international competitions (*challenges*) (Maier-Hein et al., 2018). The chosen metrics in segmentation challenges radically influence the resulting rankings (Maier-Hein et al., 2018; Reinke et al., 2018), and although several papers highlight specific strengths and weaknesses of common metrics (Gooding et al., 2018; Kofler et al., 2021; Konukoglu, Glocker, Ye, Criminisi, & Pohl, 2012; Margolin, Zelnik-Manor, & Tal, 2014; Vaassen et al., 2020), researchers are missing guidelines for choosing the right metric for a given problem (Maier-Hein et al., 2018). To address this community request, this document summarizes common pitfalls related to the most frequently used metrics in medical image segmentation, namely the Dice Similarity Coefficient (DSC) (Dice, 1945), the Hausdorff Distance (HD) (Huttenlocher, Klanderman, & Rucklidge, 1993), and the Intersection over Union (IoU) (Jaccard, 1912) (see Figure 1). To this end, the problems related to segmentation metrics are assigned to four categories, namely (1) awareness of **fundamental mathematical properties** of metrics, necessary to determine the applicability of a metric, (2) **suitability for the underlying image processing task**, (3) **metric aggregation** to combine metric values of single images into one accumulated score and (4) **metric combination** to reflect different aspects in algorithm validation.

Most common segmentation metrics

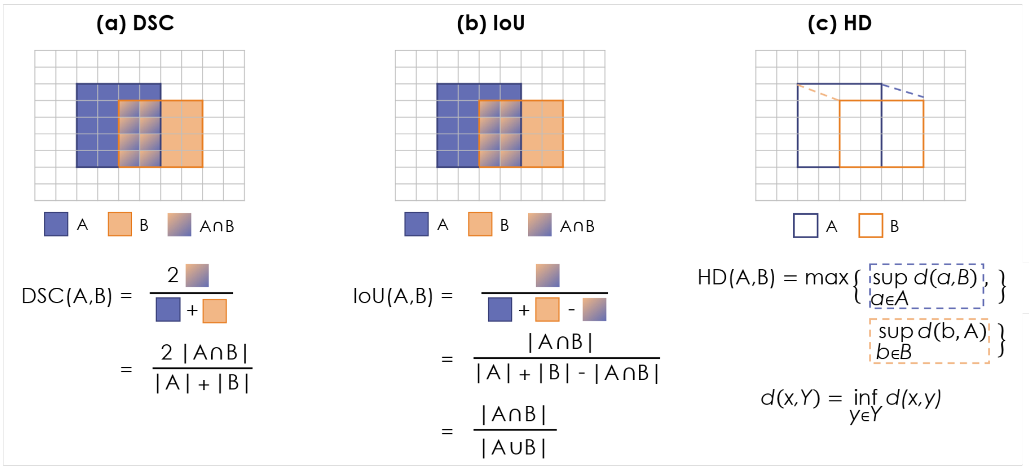


Figure 1 Most commonly used overlap-based (a/b) and contour-based (c) segmentation metrics: (a) the Dice Similarity Coefficient (DSC), (b) the Intersection over Union (IoU) and (c) the Hausdorff Distance (HD), with $|A|$ denoting the cardinality of set A, $A \cap B$ the intersection between sets A and B, $A \cup B$ the union between sets A and B and $d(x, y)$ the distance between points x and y.

2.1. Fundamental mathematical properties

Awareness of the mathematical properties of a metric is crucial when determining its suitability for a given application. In this section, we focus on the *DSC* and the *HD*, but the properties also apply to other related metrics, such as the *IoU* (also called *Jaccard Index* (Jaccard, 1912)).

As illustrated in Figure 1a, the *DSC* was designed to measure the overlap between two given objects and yields a value between 0 (no overlap) and 1 (full overlap). The metric is straightforward to compute and interpret, but comes with several pitfalls highlighted in the following paragraphs:

Small structures Segmentation of small structures, such as brain lesions, cells imaged at low magnification or distant cars, is essential for many image processing applications. In these cases, the *DSC* may not be an appropriate metric, as illustrated in Figure 2. In fact, a single-pixel difference between two predictions can have a large impact on the metric difference. Given that the correct outlines (e.g. of pathologies) are often unknown and taking into account the potentially high inter-observer variability related to generating reference annotations (Joskowicz, Cohen, Caplan, & Sosna, 2019), it is typically not desirable for few pixels to influence the metrics as much.

Problem: Small structures

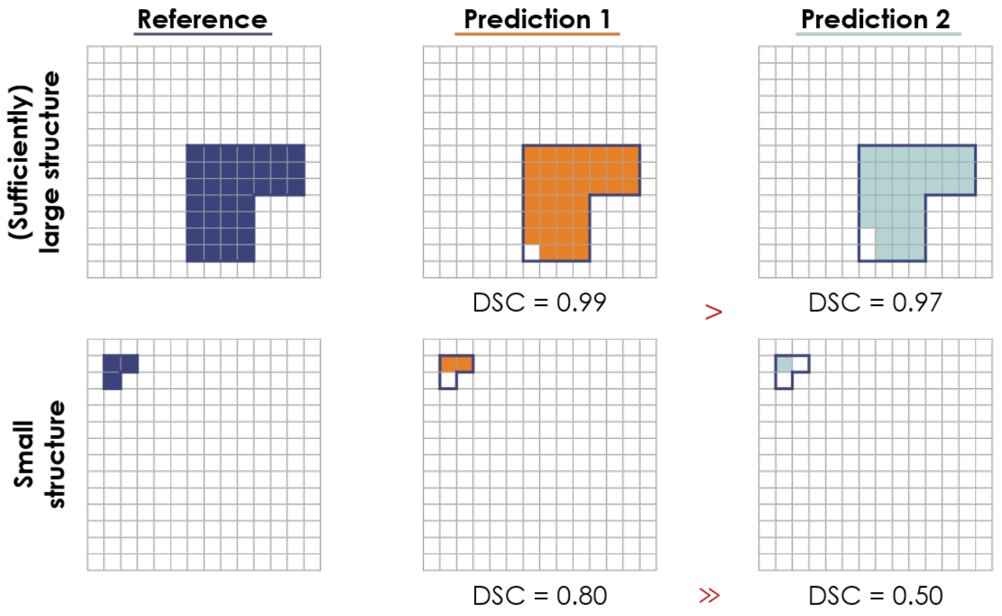


Figure 2 Effect of structure size on the *DSC*. The predictions of two algorithms (*Prediction 1/2*) differ in only a single pixel. In case of the small structure (bottom row), this has a substantial effect on the corresponding metric value.

Noise/errors in the reference annotations Similar problems may arise in the presence of annotation artifacts. Figure 3 demonstrates that a single erroneous pixel in the reference annotation may lead to a substantial decrease in the measured performance, especially in the case of the *HD*.

Problem: Noise and artifacts

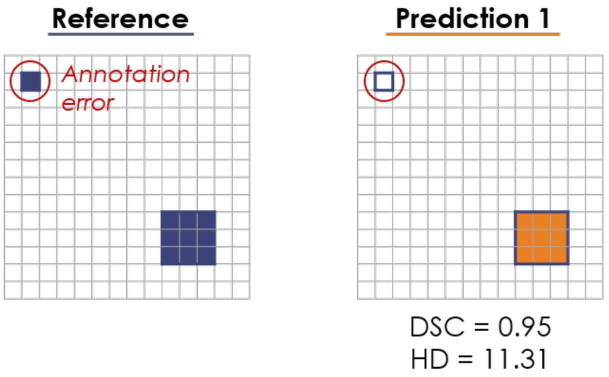


Figure 3 Effect of annotation errors/noise. A single erroneously annotated pixel may lead to a large decrease in performance, especially in the case of the *HD* or in the case of the *DSC* when applied to small structures.

Shape unawareness Metrics measuring the overlap between objects are not designed to uncover differences in shapes. This is an important problem for many applications, such as radiotherapy. Figure 4 illustrates that completely different object shapes may lead to the exact same *DSC* value.

Problem: Shape unawareness

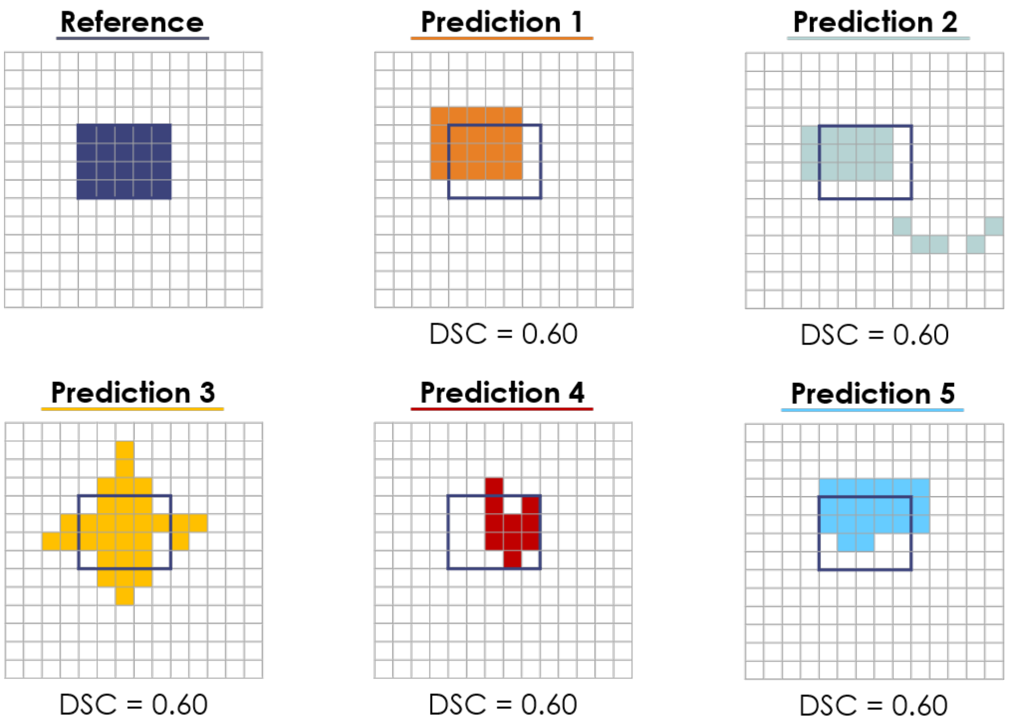


Figure 4 Effect of different shapes. The shapes of the predictions of five algorithms (*Prediction 1-5*) differ substantially, but lead to the exact same *DSC*.

Oversegmentation *vs.* undersegmentation In some applications such as autonomous driving or radiotherapy, it may be highly relevant whether an algorithm tends to over- or under-segment the target structure. The *DSC* metric, however, does not represent over- and under-segmentation equally (Yeghiazaryan & Voiculescu, 2018). As depicted in Figure 5, a difference of a single pixel in the outline yields different *DSC* scores (oversegmentation preferred). Other distance-based performance values such as the *HD* are invariant to these properties.

Problem: Oversegmentation *vs.* undersegmentation

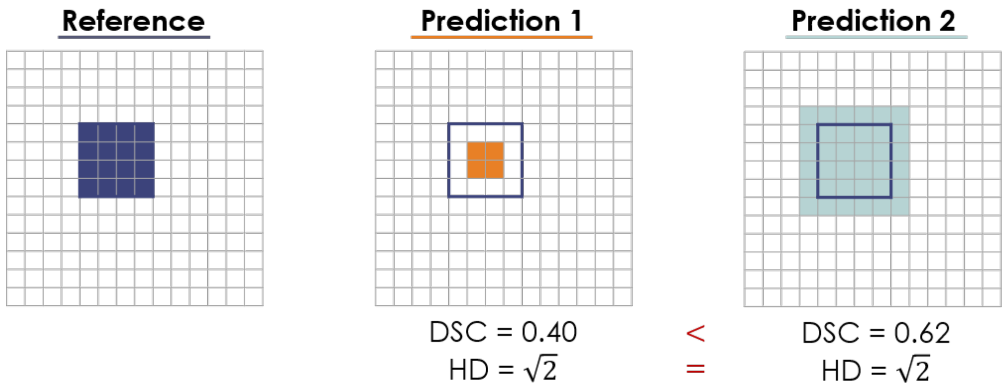


Figure 5 Effect of undersegmentation vs. oversegmentation. The outlines of the predictions of two algorithms (*Prediction 1/2*) differ in only a single pixel (Prediction 1: undersegmentation, Prediction 2: oversegmentation). This has no effect on the *HD*, but yields a substantially different *DSC* score.

2.2. Suitability for underlying image processing task

Performance metrics are typically expected to reflect a domain-specific validation goal (e.g. clinical goal). Previous research, however, suggests, that this is often not the case. A common problem is that segmentation metrics, such as the *DSC*, are applied to *detection and localization* tasks (Jäger, 2020), as illustrated in Figure 6. From a clinical perspective, for example, the algorithm producing *Prediction 2* and covering all three structures of interest (e.g. tumors) would be clinically much more valuable compared to the one producing a highly accurate segmentation for one structure but missing the other two in *Prediction 1*. This is not reflected in the metric values, which are substantially higher for *Prediction 1*. In general, the *DSC* is strongly biased against single objects, therefore not appropriate for a detection task of multiple structures (Yeghiazaryan & Voiculescu, 2018).

Problem: Task/metric mismatch

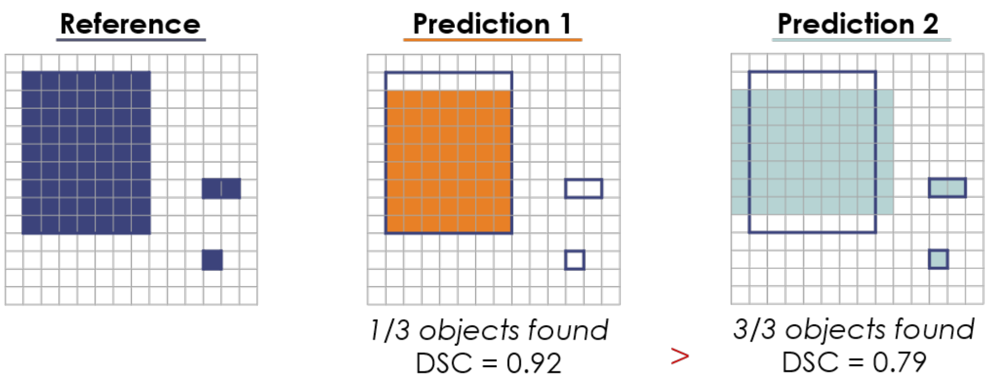


Figure 6 Effect of using a segmentation metric for object detection. In this example, the prediction of one algorithm only detecting one of three structures (*Prediction 1*) leads to a higher *DSC* compared to that of a second algorithm (*Prediction 2*) detecting all structures.

2.3. Metric aggregation

In international competitions (*challenges*), metric values are often aggregated over all test cases to produce a challenge ranking (Maier-Hein et al., 2018). Figures 7 and 8 illustrate why this may be problematic in the presence of missing values.

Problem: Missing values for metrics with fixed upper/lower bounds

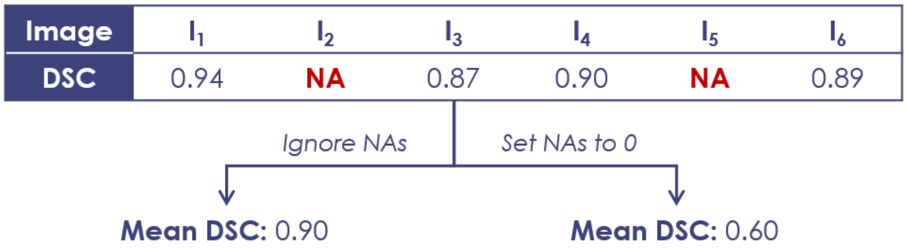


Image	I_1	I_2	I_3	I_4	I_5	I_6
DSC	0.94	NA	0.87	0.90	NA	0.89

Figure 7 Effect of missing values when aggregating metric values. In this example, ignoring missing values leads to a substantially higher *DSC* compared to setting missing values to the worst possible value (here: 0).

In the case of metrics with fixed boundaries, like the *DSC* or the *IoU*, missing values can easily be set to the worst possible value (here: 0). For distance-based measures without lower/upper bounds, the strategy of how to deal with missing values is not trivial. In the case of the *HD*, one may choose the maximum distance of the image and add 1 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 (Figure 8), thus potentially affecting the ranking.

Problem: Missing values for metrics without fixed upper/lower bounds

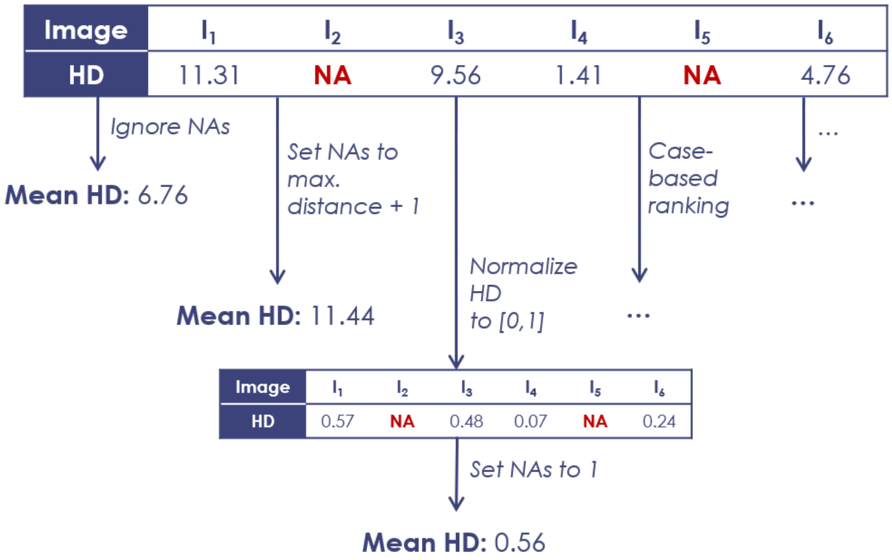


Image	I_1	I_2	I_3	I_4	I_5	I_6
HD	11.31	NA	9.56	1.41	NA	4.76

Image	I_1	I_2	I_3	I_4	I_5	I_6
HD	0.57	NA	0.48	0.07	NA	0.24

Figure 8 Effect of missing values when aggregating metric values for metrics without fixed boundaries (here: HD). In this example, ignoring or treating missing values in different ways leads to substantially different HD values.

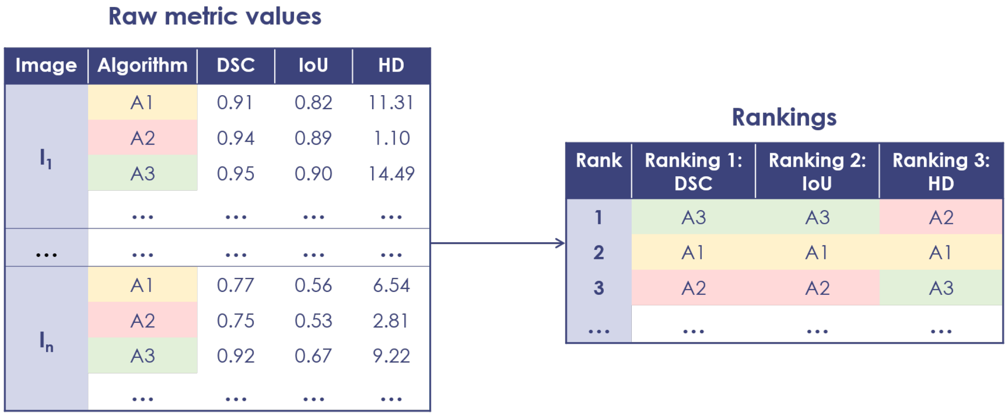
2.4. Metric combination

A single metric typically does not reflect all important aspects that are essential for algorithm validation. Hence, multiple metrics with different properties are often combined. However, the selection of metrics should be well considered as some metrics are mathematically related to each other (Taha & Hanbury, 2015; Taha, Hanbury, & del Toro, 2014). A prominent example is the IoU – the most popular segmentation metric in computer vision – which highly correlates with the DSC – the most popular segmentation metric in medical image analysis. In fact, the IoU and the DSC are mathematically related (Taha & Hanbury, 2015):

$$IoU = \frac{DSC}{2 - DSC}, \qquad (1) \qquad\qquad DSC = \frac{2IoU}{1 + IoU} \qquad (2)$$

Combining metrics that are related will not provide additional information for a ranking. Figure 9 illustrates how the ranking can change when adding a metric that measures different properties.

Problem: Related metrics



Raw metric values

Image	Algorithm	DSC	IoU	HD
I_1	A1	0.91	0.82	11.31
	A2	0.94	0.89	1.10
	A3	0.95	0.90	14.49
	...	...	...	...
...	...	...	...	...
I_n	A1	0.77	0.56	6.54
	A2	0.75	0.53	2.81
	A3	0.92	0.67	9.22
	...	...	...	...

Rankings

Rank	Ranking 1: DSC	Ranking 2: IoU	Ranking 3: HD
1	A3	A3	A2
2	A1	A1	A1
3	A2	A2	A3
...	...	...	...

Figure 9 Effect of combining different metrics for a ranking. Mutually dependent metrics (*DSC* and *IoU*) will lead to the same ranking, whereas metrics measuring different properties (*HD*) will lead to a different ranking.

3. Conclusion

Choosing the right metric for a specific image processing task is a non-trivial task. With this (dynamic) paper, we wish to raise awareness about some of the common flaws of the most frequently used metrics in the field of image processing, encouraging researchers to reconsider common workflows.

Acknowledgements

This work was initiated by the Helmholtz Association of German Research Centers in the scope of the Helmholtz Imaging Platform (HIP). It was further supported in part by the Intramural Research Program of the National Institutes of Health Clinical Center as well as by the National Cancer Institute (NCI) and the National Institute of Neurological Disorders and Stroke (NINDS) of the National Institutes of Health (NIH), under award numbers NCI:U01CA242871 and NINDS:R01NS042645. The content of this publication is solely the responsibility of the authors and does not represent the official views of the NIH.

References

Brown, B. B. (1968). *Delphi process: a methodology used for the elicitation of opinions of experts* (Tech. Rep.). Rand Corp Santa Monica CA.

Correia, P., & Pereira, F. (2006). Video object relevance metrics for overall segmentation quality evaluation. *EURASIP Journal on Advances in Signal Processing*, *2006*, 1–11.

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, *26*(3), 297–302.

Gooding, M. J., Smith, A. J., Tariq, M., Aljabar, P., Peressutti, D., van der Stoep, J., ... others (2018). Comparative evaluation of autocontouring in clinical practice: a practical method using the turing test. *Medical physics*, *45*(11), 5105–5115.

Honauer, K., Maier-Hein, L., & Kondermann, D. (2015). The hci stereo metrics: Geometry-aware performance analysis of stereo algorithms. In *Proceedings of the ieee international conference on computer vision* (pp. 2120–2128).

Huttenlocher, D. P., Klanderman, G. A., & Rucklidge, W. J. (1993). Comparing images using the hausdorff distance. *IEEE Transactions on pattern analysis and machine intelligence*, *15*(9), 850–863.

Jaccard, P. (1912). The distribution of the flora in the alpine zone. 1. *New phytologist*, *11*(2), 37–50.

Jäger, P. F. (2020). Challenges and opportunities of end-to-end learning in medical image classification.

Joskowicz, L., Cohen, D., Caplan, N., & Sosna, J. (2019). Inter-observer variability of manual contour delineation of structures in ct. *European radiology*, *29*(3), 1391–1399.

Kofler, F., Ezhov, I., Isensee, F., Berger, C., Korner, M., Paetzold, J., ... others (2021). Are we using appropriate segmentation metrics? Identifying correlates of human expert perception for CNN training beyond rolling the DICE coefficient. *arXiv preprint arXiv:2103.06205v1*.

Konukoglu, E., Glocker, B., Ye, D. H., Criminisi, A., & Pohl, K. M. (2012). Discriminative segmentation-based evaluation through shape dissimilarity. *IEEE transactions on medical imaging*, *31*(12), 2278–2289.

Maier-Hein, L., Eisenmann, M., Reinke, A., Onogur, S., Stankovic, M., Scholz, P., ... others (2018). Why rankings of biomedical image analysis competitions should be interpreted with care. *Nature communications*, *9*(1), 1–13.

Margolin, R., Zelnik-Manor, L., & Tal, A. (2014). How to evaluate foreground maps? In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 248–255).

Reinke, A., Eisenmann, M., Onogur, S., Stankovic, M., Scholz, P., Full, P. M., ... others (2018). How to exploit weaknesses in biomedical challenge design and organization. In *International conference on medical image computing and computer-assisted intervention* (pp. 388–395).

Taha, A. A., & Hanbury, A. (2015). Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. *BMC medical imaging*, *15*(1), 1–28.

Taha, A. A., Hanbury, A., & del Toro, O. A. J. (2014). A formal method for selecting evaluation metrics for image segmentation. In *2014 ieee international conference on image processing (icip)* (pp. 932–936).

Vaassen, F., Hazelaar, C., Vaniqui, A., Gooding, M., van der Heyden, B., Canters, R., & van Elmpt, W. (2020). Evaluation of measures for assessing time-saving of automatic organ-at-risk segmentation in radiotherapy. *Physics and Imaging in Radiation Oncology*, *13*, 1–6.

Yeghiazaryan, V., & Voiculescu, I. D. (2018). Family of boundary overlap metrics for the evaluation of medical image segmentation. *Journal of Medical Imaging*, *5*(1), 015006.