



HHS Public Access

Author manuscript

Nat Methods. Author manuscript; available in PMC 2024 August 12.

Published in final edited form as:

Nat Methods. 2024 February ; 21(2): 182–194. doi:10.1038/s41592-023-02150-0.

* **Corresponding authors:** Annika Reinke: a.reinke@dkfz-heidelberg.de; Minu D. Tizabi: Lena Maier-Hein, l.maier-hein@dkfz-heidelberg.de; Paul F. Jäger: p.jaeger@dkfz-heidelberg.de; Lena Maier-Hein, l.maier-hein@dkfz-heidelberg.de.

† **Shared first authors:** Annika Reinke and Minu D. Tizabi

‡ **Shared last authors:** Paul F. Jäger and Lena Maier-Hein

AUTHOR CONTRIBUTIONS

A.R. initiated and led the study, was a member of the Delphi core team, wrote and reviewed the manuscript, prepared and evaluated all surveys, organized all workshops, suggested pitfalls, and designed all figures. M.D.T. was a member of the extended Delphi core team and wrote and reviewed the manuscript. P.F.J. initiated and led the study, was a member of the Delphi core team, led the Object Detection (ObD) and Instance Segmentation (InS) expert group, wrote and reviewed the manuscript, prepared and evaluated all surveys, organized all workshops, suggested pitfalls, and participated in surveys. L.M.-H. initiated and led the study, was a member of the Delphi core team, wrote and reviewed the manuscript, prepared and evaluated all surveys, organized all workshops, and suggested pitfalls. M.B. was a member of the extended Delphi core team, was an active member of the ObD and InS expert group, wrote and reviewed the manuscript, and participated in surveys and workshops. M.E. was a member of the extended Delphi core team, reviewed the document, assisted in survey preparation, and participated in surveys and workshops. D.H.-N. was a member of the extended Delphi core team and prepared all surveys. A.E.K. was a member of the extended Delphi core team and participated in surveys. T.R. was a member of the extended Delphi core team, was an active member of the ObD and InS expert group, reviewed the document, assisted in survey preparation, tested all metric examples, suggested pitfalls, and participated in surveys and workshops. C.H.S. was an active member of the ObD and InS expert group, reviewed the manuscript, suggested pitfalls, tested all metric examples, and participated in surveys and workshops. L.A. reviewed the manuscript and participated in surveys and workshops. M.A. was an active member of the Semantic Segmentation (SemS) expert group and participated in surveys and workshops. T.A. was an active member of the ObD and InS expert group, suggested pitfalls, reviewed the manuscript, and participated in surveys and workshops. S.B. co-lead the SemS expert group, reviewed the manuscript, suggested pitfalls, and participated in surveys and workshops. A.B. was an active member of the biomedical and cross-topic expert groups, reviewed the manuscript, and participated in surveys and workshops. F.B. led the calibration expert group, suggested pitfalls, reviewed the manuscript, and participated in surveys. M.J.C. was an active member of the Image-level Classification (ImLC) expert group and participated in surveys and workshops. V.C. was an active member of the ImLC and cross-topic expert groups, reviewed the manuscript, and participated in surveys and 3 workshops. J.C. reviewed the manuscript, suggested pitfalls, and participated in surveys. E.C. led the cross topic expert group, was a member of the extended Delphi core team, wrote and reviewed the manuscript, suggested pitfalls, and participated in surveys. B.A.C. was an active member of the ObD and InS expert group, reviewed the manuscript, suggested pitfalls, and participated in surveys and workshops. K.F. was an active member of the biomedical and cross-topic expert groups and participated in surveys and workshops. L.F. was an active member of the calibration expert group, reviewed the manuscript, suggested pitfalls, and participated in surveys. A.G. was an active member of the calibration expert group, reviewed the manuscript, suggested pitfalls, and participated in surveys. B.v.G. participated in surveys and workshops. B.G. led the cross-topic expert group and was an active member of the SemS expert group, reviewed the manuscript, suggested pitfalls, and participated in surveys and workshops. P.G. led the ImLC expert group, was a member of the extended Delphi core team, wrote and reviewed the manuscript, suggested pitfalls, and participated in surveys and workshops. D.A.H. was an active member of the biomedical and cross-topic expert groups, reviewed the manuscript suggested pitfalls, and participated in surveys and workshops. M.M.H. was an active member of the ImLC expert group, reviewed the manuscript, and participated in surveys and workshops. M.H. co-lead the biomedical expert group, was an active member of the cross-topic expert group, reviewed the manuscript, and participated in surveys and workshops. F.I. led the SemS expert group, reviewed the manuscript, and participated in surveys and workshops. P.J. co-lead the cross-topic expert group, was an active member of the ObD and InS expert group, reviewed the manuscript, and participated in surveys and workshops. C.E.K. was an active member of the biomedical expert group, reviewed the manuscript, and participated in surveys and workshops. D.K. suggested pitfalls and participated in surveys. B.K. suggested pitfalls and participated in surveys. J.K. led the biomedical expert group, reviewed the manuscript, and participated in surveys and workshops. F.K. suggested pitfalls and participated in surveys. Th.K. suggested pitfalls and participated in surveys. A.K.-S. was a member of the extended Delphi core team and was an active member of the cross-topic group. M.K. led the ObD and InS expert group, reviewed the manuscript, suggested pitfalls, and participated in surveys and workshops. An.K. was an active member of the biomedical expert group, reviewed the manuscript, and participated in surveys and workshops. Ta.K. participated in surveys and workshops. B.A.L. was an active member of the SemS expert group and participated in surveys and workshops. G.L. was an active member of the ImLC expert group, reviewed the manuscript, and participated in surveys and workshops. A.M. was an active member of the biomedical and SemS expert groups, suggested pitfalls, and participated in surveys and workshops. K.M.-H. was an active member of the SemS expert group, reviewed the manuscript, and participated in surveys and workshops. A.L.M. participated in surveys and workshops. E.M. was an active member of the ImLC expert group, reviewed the manuscript, and participated in surveys. B.M. participated in surveys and workshops. K.G.M.M. was an active member of the cross-topic expert group, reviewed the manuscript, suggested pitfalls, and participated in surveys and workshops. H.M. was an active member of the ImLC expert group, reviewed the manuscript, and participated in surveys and workshops. B.N. was an active member of the ObD and InS expert group, and participated in surveys. F.N. was an active member of the biomedical expert group and participated in surveys and workshops. J.P. participated in surveys and workshops. S.M.R. reviewed the manuscript, suggested pitfalls, and participated in surveys. Na.R. participated in surveys and workshops. M.R. led the SemS expert group, reviewed the manuscript, suggested pitfalls, and participated in surveys and workshops. M.A.R. led the ImLC expert group, reviewed the manuscript, suggested pitfalls, and participated in surveys and workshops. Ni.R. was an active member of the SemS expert group and participated in surveys and workshops. R.M.S. was an active member of the ObD and InS, the biomedical, and the cross-topic expert groups, reviewed the manuscript, and participated in surveys and workshops. A.A.T. co-lead the SemS expert group, suggested pitfalls, and participated in surveys and workshops. A.T. was an active member of the calibration group, reviewed the manuscript, and participated in surveys. S.A.T. was an active member of the ObD and InS expert group, reviewed the manuscript, and participated in surveys and workshops. B.v.C. was an active member of the cross-topic expert group and participated in surveys. G.V. was an active member of the ImLC and cross-topic expert groups, reviewed the manuscript,

Understanding metric-related pitfalls in image analysis validation

A full list of authors and affiliations appears at the end of the article.

Abstract

Validation metrics are key for tracking scientific progress and bridging the current chasm between artificial intelligence (AI) research and its translation into practice. However, increasing evidence shows that particularly in image analysis, metrics are often chosen inadequately. While taking into account the individual strengths, weaknesses, and limitations of validation metrics is a critical prerequisite to making educated choices, the relevant knowledge is currently scattered and poorly accessible to individual researchers. Based on a multi-stage Delphi process conducted by a multidisciplinary expert consortium as well as extensive community feedback, the present work provides the first reliable and comprehensive common point of access to information on pitfalls related to validation metrics in image analysis. While focused on biomedical image analysis, the addressed pitfalls generalize across application domains and are categorized according to a newly created, domain-agnostic taxonomy. The work serves to enhance global comprehension of a key topic in image analysis validation.

Measuring performance and progress in any given field critically depends on the availability of meaningful outcome metrics. In a field such as athletics, this process is straightforward because the performance measurements (e.g., the time it takes an athlete to run a given distance) exactly reflect the underlying interest (e.g., which athlete runs a given distance the fastest?). In image analysis, the situation is much more complex. Depending on the underlying research question, vastly different aspects of an algorithm's performance might be of interest (Fig. 1) and meaningful in determining its future practical, for example clinical, applicability. If the performance of an image analysis algorithm is not measured according to relevant validation metrics, no reliable statement can be made about the suitability of this algorithm in solving the proposed task, and the algorithm is unlikely to ever reach the stage of real-life application. Moreover, unsuitable algorithms could be wrongly regarded as the best-performing ones, sparking entirely futile resource investment and follow-up research while obscuring true scientific advancements. In determining new state-of-the-art methods and informing future directions, the use of validation metrics actively shapes the evolution of research. In summary, *validation metrics are the key for both measuring and informing scientific progress, as well as bridging the current chasm between image analysis research and its translation into practice.*

and suggested pitfalls. Z.R.Y. suggested pitfalls and participated in surveys. Al.K., J.S.-R., C.I.S., and S.S. served on the expert Delphi panel and participated in workshops and surveys.

CODE AVAILABILITY STATEMENT

We provide reference implementations for all *Metrics Reloaded* metrics within the MONAI open-source framework. They are accessible at <https://github.com/Project-MONAI/MetricsReloaded>.

In image analysis, while for some applications it might, for instance, be sufficient to draw a box around the structure of interest (e.g., detecting individual mitotic cells or regions with apoptotic cell debris) and optionally associate that region with a classification (e.g., a mitotic vs an interphase cell), other applications (e.g., cell tracing for fluorescent signal quantification) could require determining the exact structure boundaries. The suitability of any individual validation metric thus depends crucially on the properties of the driving image analysis problem. As a result, numerous metrics have so far been proposed in the field of image processing. In our previous work, we analyzed all biomedical image analysis competitions conducted within a period of about 15 years [21]. We found a total of 97 different metrics reported in the field of biomedicine alone, each with its own individual strengths, weaknesses, and limitations, and hence varying degrees of suitability for meaningfully measuring algorithm performance on any given research problem. Such a vast range of options makes tracking all related information impossible for any individual researcher and consequently renders the process of metric selection error-prone. Thus, the frequent reliance on flawed, historically grown validation practices in current literature comes as no surprise. To make matters worse, there is currently no comprehensive resource that can provide an overview of the relevant definitions, (mathematical) properties, limitations, and pitfalls pertaining to a metric of interest. *While taking into account the individual properties and limitations of metrics is imperative for choosing adequate validation metrics, the required knowledge is thus largely inaccessible.*

As a result, numerous flaws and pitfalls are prevalent in image analysis validation, with researchers often being unaware of them due to a lack of knowledge of intricate metric properties and limitations. Accordingly, increasing evidence shows that metrics are often selected inadequately in image analysis (e.g., [11, 17, 35]). In the absence of a central information resource, it is common for researchers to resort to popular validation metrics, which, however, can be entirely unsuitable, for instance due to a mismatch of the metric's inherent mathematical properties with the underlying research question and specifications of the data set at hand (see Fig. 1).

The present work addresses this important roadblock in image analysis research with a crowd-sourcing-based approach that involved both a Delphi process undergone by a multidisciplinary expert consortium as well as a social media campaign. It represents the *first comprehensive collection, visualization, and detailed discussion of pitfalls, drawbacks, and limitations regarding validation metrics commonly used in image analysis*. Our work provides researchers with a *reliable, single point of access* to this critical information. Owing to the enormous complexity of the matter, the metric properties and pitfalls are discussed in the specific context of classification problems, i.e., image analysis problems that can be considered classification tasks at either the image, object, or pixel level. Specifically, these encompass the four problem categories of image-level classification, semantic segmentation, object detection, and instance segmentation. Our contribution includes a dedicated profile for each metric (Suppl. Note 3) as well as the creation of a new common taxonomy that categorizes pitfalls in a domain-agnostic manner (Fig. 2). The taxonomy is depicted for individual metrics in provided tables (see Extended Data Tabs. 1–5) and enables researchers to quickly grasp whether using a certain metric comes with pitfalls in a given use case.

While our work grew out of image analysis research and practice in the field of biomedicine, a field of high complexity and particularly high stakes due to its direct impact on human health, we believe the identified pitfalls to be transferable to other application areas of imaging research. It should be noted that this work focuses on identifying, categorizing, and illustrating metric pitfalls, while the sister publication of this work gives specific recommendations on which metrics to apply under which circumstances [22].

Information on metric pitfalls is largely inaccessible

Researchers and algorithm developers seeking to validate image analysis algorithms frequently face the problem of choosing adequate validation metrics while at the same time navigating a range of potential pitfalls. Following common practice is often not the best option, as evidenced by a number of recent publications [11, 17, 21, 35]. Making an educated choice is notably complicated by the absence of any comprehensive databases or reviews covering the topic and thus the lack of a central resource for reliable information on validation metrics.

This lack of accessibility is considered by experts to be a major bottleneck in image analysis validation [21]. To illustrate this point, we searched the literature for available information on commonly used validation metrics. The search was conducted on the platform Google Scholar using search strings that combined different notations of the metric name, including synonyms and acronyms, with search terms indicating problems, such as “pitfall” or “limitation”. The mean and median number of hits for the metrics addressed in the present work were 159,329 and 22,100, respectively, and ranged between 49 for centerline Dice Similarity Coefficient (cIDice) and 962,000 for Sensitivity. Moreover, despite valuable literature on individual relevant aspects (e.g., [5, 6, 13, 17, 32, 33, 35]), we did not find a common point of entry to metric-related pitfalls in image analysis in the form of a review paper or other credible source. We conclude that *the key knowledge required for making educated decisions and avoiding pitfalls related to the use of validation metrics is highly scattered and not accessible by individuals.*

Historically grown practices are not always justified

To obtain an initial insight into current common practice regarding validation metrics, we prospectively captured the designs of challenges organized by the IEEE Society of the International Symposium of Biomedical Imaging (ISBI), the Medical Image Computing and Computer Assisted Interventions (MICCAI) Society and the Medical Imaging with Deep Learning (MIDL) foundation. The organizers of the respective competitions were asked to provide a rationale for the choice of metrics in their competition. An analysis of a total of 138 competitions conducted between 2018 and 2022 revealed that metrics are frequently (in 24% of the competitions) based on common practice in the community. We found, however, that common practices are often not well-justified, and poor practices may even be propagated from one generation to the next.

One remarkable example for this issue is the widespread adoption of an incorrect naming and inconsistent mathematical formulation of a metric proposed for cell instance

segmentation. The term “mean Average Precision (mAP)” usually refers to one of the most common metrics in object detection (object-level classification) [20, 28]. Here, Precision denotes the Positive Predictive Value (PPV), which is “averaged” over varying thresholds on the predicted class scores of an object detection algorithm. The “mean” Average Precision (AP) is then obtained by taking the mean over classes [10, 28]. Despite the popularity of mAP, a widely known challenge on cell instance segmentation¹ introduced a new “Mean Average Precision” in 2018. Although the task matches the task of the original “mean” AP, object detection, all terms in the newly proposed metric (mean, average, and precision) refer to entirely different concepts. For instance, the common definition of Precision from literature $TP/(TP + FP)$ was altered to $TP/(TP + FP + FN)$, where TP, FP, and FN refer to the cardinalities of the confusion matrix (i.e., the true/false positives/negatives). The latter formula actually defines the Intersection over Union (IoU) metric. Despite these problems, the terminology was adopted by subsequent influential works [16, 30, 31, 39], indicating widespread propagation and usage within the community.

A multidisciplinary Delphi process reveals numerous pitfalls in biomedical image analysis validation

With the aim of creating a comprehensive, reliable collection and future point of access to biomedical image analysis metric definitions and limitations, we formed an international multidisciplinary consortium of 62 experts from various biomedical image analysis-related fields that engaged in a multi-stage Delphi process [2] for consensus building. The Delphi process comprised multiple surveys, developed by a coordinating team and filled out by the remaining members of the consortium. Based on the survey results, the list of pitfalls was iteratively refined by collecting pitfall sources, specific feedback and suggestions on pitfalls, and final agreement on which pitfalls to include and how to illustrate them. Further pitfalls were crowdsourced through the publication of a dynamic preprint of this work [28] as well as a social media campaign, both of which asked the scientific community for contributions. This approach allowed us to integrate distributed, cross-domain knowledge on metric-related pitfalls within a single resource. In total, the process revealed 37 distinct sources of pitfalls (see Fig. 2). Notably, these pitfall sources (e.g., class imbalances, uncertainties in the reference, or poor image resolution) can occur irrespective of a specific imaging modality or application. As a result, many pitfalls generalize across different problem categories in image processing (image-level classification, semantic segmentation, object detection, and instance segmentation), as well as imaging modalities and domains. A detailed discussion of all pitfalls can be found in Suppl. Note 2.

A common taxonomy enables domain-agnostic categorization of pitfalls

One of our key objectives was to facilitate information retrieval and provide structure within this vast topic. Specifically, we wanted to enable researchers to identify at a glance which metrics are affected by which types of pitfalls. To this end, we created a comprehensive taxonomy that categorizes the different pitfalls in a semantic fashion. The taxonomy was

¹. <https://www.kaggle.com/competitions/data-science-bowl-2018/overview/evaluation>

created in a domain-agnostic manner to reflect the generalization of pitfalls across different imaging domains and modalities. An overview of the taxonomy is presented in Fig. 2, and the relations between the pitfall categories and individual metrics can be found in Extended Data Tabs. 1–5. We distinguish the following three main categories:

[P1] Pitfalls related to the inadequate choice of the problem category.

A common pitfall lies in the use of metrics for a problem category they are not suited for because they fail to fulfill crucial requirements of that problem category, and hence do not reflect the domain interest (Fig. 1). For instance, popular voxel-based metrics, such as the Dice Similarity Coefficient (DSC) or Sensitivity, are widely used in image analysis problems, although they do not fulfill the critical requirement of detecting all objects in a data set. In a cancer monitoring application they fail to measure instance progress, i.e., the potential increase in number of lesions (Fig. 1), which can have serious consequences for the patient. For some problems, there may even be a lack of matching problem category (Fig. SN 2.2), rendering common metrics inadequate. We present further examples of pitfalls in this category in Suppl. Note 2.1.

[P2] Pitfalls related to poor metric selection.

Pitfalls of this category occur when a validation metric is selected while disregarding specific properties of the given research problem or method used that make this metric unsuitable in the particular context. [P2] can be further divided into the following four subcategories:

[P2.1] Disregard of the domain interest.—Commonly, several requirements arise from the domain interest of the underlying research problem that may clash with particular metric limitations. For example, if there is particular interest in the structure boundaries, it is important to know that overlap-based metrics such as the DSC do not take the correctness of an object's boundaries into account, as shown in Fig. 4(a). Similar issues may arise if the structure volume (Fig. SN 2.4) or center(line) (Fig. SN 2.5) are of particular interest. Other domain interest-related properties may include an unequal severity of class confusions. This may be important in an ordinal grading use case, in which the severity of a disease is categorized by different scores. Predicting a low severity for a patient that actually suffers from a severe disease should be substantially penalized. Common classification metrics do not fulfill this requirement. An example is provided in Fig. 4(b). On pixel level, this property relates to an unequal severity of over- vs. undersegmentation. In applications such as radiotherapy, it may be highly relevant whether an algorithm tends to over- or undersegment the target structure. Common overlap-based metrics, however, do not represent over- and undersegmentation equally [38]. Further pitfalls may occur if confidence awareness (Fig. SN 2.6), comparability across data sets (Fig. SN 2.7), or a cost-benefit analysis (Fig. SN 2.9) are of particular importance, as illustrated in Suppl. Note 2.2.1.

[P2.2] Disregard of the properties of the target structures.—For problems that require capturing local properties (object detection, semantic or instance segmentation), the properties of the target structures to be localized and/or segmented may have important implications for the choice of metrics. Here, we distinguish between *size-related* and *shape-*

and topology-related pitfalls. Common metrics, for example, are sensitive to structure sizes, such that single-pixel differences may hugely impact the metric scores, as shown in Extended Data Fig. 1(a). Shape- and topology-related pitfalls may relate to the fact that common metrics disregard complex shapes (Extended Data Fig. 1(b)) or that bounding boxes do not capture the disconnectedness of structures (Fig. SN 2.14). A high variability of structure sizes (Fig. SN 2.11) and overlapping or touching structures (Fig. SN 2.13) may also influence metric values. We present further examples of [P2.2] pitfalls in Suppl. Note 2.2.2.

[P2.3] Disregard of the properties of the data set.—Various properties of the data set such as class imbalances (Fig. 5(a)), small sample sizes (Fig. 5(b)), or the quality of the reference annotations, may directly affect metric values. Common metrics such as the Balanced Accuracy (BA), for instance, may yield a very high score for a model that predicts many False Positive (FP) samples in an imbalanced setting (see Fig. 5(a)). When only small test data sets are used, common calibration metrics (which are typically biased estimators) either underestimate or overestimate the true calibration error of a model (Fig. 5(b)) [14]. On the other hand, metric values may be impacted by reference annotations (Fig. SN 2.17). Spatial outliers in the reference may have a huge impact on distance-based metrics such as the Hausdorff Distance (HD) (Fig. 5(c)). Additional pitfalls may arise from the occurrence of cases with an empty reference (Extended Data Fig. 2(b)), causing division by zero errors. We present further examples of [P2.3] pitfalls in Suppl. Note 2.2.3.

[P2.4] Disregard of the properties of the algorithm output.—Reference-based metrics compare the algorithm output to a reference annotation to compute a metric score. Thus, the content and format of the prediction are of high importance when considering metric choice. Overlapping predictions in segmentation problems, for instance, may return misleading results. In Extended Data Fig. 2(a), the predictions only overlap to a certain extent, not representing that the reference instances actually overlap substantially. This is not detected by common metrics. Another example are empty predictions that may cause division by zero errors in metric calculations, as illustrated in Extended Data Fig. 2(b), or the lack of predicted class scores (Fig. SN 2.20). We present further examples of [P2.4] pitfalls in Suppl. Note 2.2.3.

[P3] Pitfalls related to poor metric application.

Once selected, the metrics need to be applied to an image or an entire data set. This step is not straightforward and comes with several pitfalls. For instance, when aggregating metric values over multiple images or patients, a common mistake is to ignore the hierarchical data structure, such as data from several hospitals or a varied number of images per patient. We present three examples of [P3] pitfalls in Fig. 6; for more pitfalls in this category, please refer to Suppl. Note 2.3. [P3] can further be divided into five subcategories that are presented in the following paragraphs.

[P3.1] Inadequate metric implementation.—Metric implementation is, unfortunately, not standardized. As shown by [12], different researchers typically employ various different implementations for the same metric, which may yield a substantial variation in the metric

scores. While some metrics are straightforward to implement, others require more advanced techniques and offer different possibilities. In the following, we provide some examples for inadequate metric implementation:

- The method of how identical confidence scores are handled in the computation of the AP metric may lead to substantial differences in the metric scores. Microsoft Common Objects in Context (COCO) [20], for instance, processes each prediction individually, while CityScapes [7] processes all predictions with the same score in one joint step. Fig. 6(a) provides an example with two predictions having the same confidence score, in which the final metric scores differ depending on the chosen handling strategy for identical confidence scores. Similar issues may arise with other curve-based metrics, such as Area under the Receiver Operating Characteristic Curve (AUROC), AP, or Free-Response Receiver Operating Characteristic (FROC) scores (see e.g., [24]).
- Metric implementation may be subject to discretization issues such as the chosen discretization of continuous variables, which may cause differences in the metric scores, as exemplary illustrated in Fig. SN 2.22.
- For metrics assessing structure boundaries, such as the Average Symmetric Surface Distance (ASSD), the exact boundary extraction method is not standardized. Thus, for example, the boundary extraction method implemented by the Liver Tumor Segmentation (LiTS) challenge [1] and that implemented by Google DeepMind² may produce different metric scores for the ASSD. This is especially critical for metrics that are sensitive to small contour changes, such as the HD.
- Suboptimal choices of hyperparameters may also lead to metric scores that do not reflect the domain interest. For example, the choice of a threshold on a localization criterion (see Fig. SN 2.23) or the chosen hyperparameter for the F_β Score will heavily influence the subsequent metric scores [34].

More [P3.1] pitfalls can be found in Suppl. Note 2.3.1.

[P3.2] Inadequate metric aggregation.—A common pitfall with respect to metric application is to simply aggregate metric values over the entire data set and/or all classes. As detailed in Fig. 6(b) and Suppl. Note 2.3.2, important information may get lost in this process, and metric results can be misleading. For example, the popular TorchMetrics framework calculates the DSC metric by default as a global average over all pixels in the data set without considering their image or class of origin³. Such a calculation eliminates the possibility of interpreting the final metric score with respect to individual images and classes. For example, errors in small structures may be suppressed by correctly segmented larger structures in other images (see e.g., Fig. SN 2.26). An adequate aggregation scheme is also crucial for handling hierarchical class structure (Fig. SN 2.27), missing values (Fig. SN

2. <https://github.com/deepmind/surface-distance>

3. <https://torchmetrics.readthedocs.io/en/stable/classification/dice.html?highlight=dice>

2.29), and potential biases (Fig. SN 2.28) of the algorithm. Further [P3.2] pitfalls are shown in Suppl. Note 2.3.2.

[P3.3] Inadequate ranking scheme.—Rankings are often created to compare algorithm performances. In this context, several pitfalls pertain to either metric relationships or ranking uncertainty. For example, to assess different properties of an algorithm, it is advisable to select multiple metrics and determine their values. However, the chosen metrics should assess complementary properties and should not be mathematically related. For example, the DSC and IoU are closely related, so using both in combination would not provide any additional information over using either of them individually (Fig. SN 2.30). Note in this context that unawareness of metric synonyms can equally mislead. Metrics can be known under different names; for instance, Sensitivity and Recall refer to the same mathematical formula. Despite this fact potentially appearing trivial, an analysis of 138 biomedical image analysis challenges [22] found three challenges that unknowingly used two versions of the same metric to calculate their rankings. Moreover, rankings themselves may be unstable (Fig. SN 2.31). [21] and [37] demonstrated that rankings are highly sensitive to altering the metric aggregation operators, the underlying data set, or the general ranking method. Thus, if the robustness of rankings is disregarded, the winning algorithm may be identified by chance rather than true superiority.

[P3.4] Inadequate metric reporting.—A thorough reporting of metric values and aggregates is important both in terms of transparency and interpretability. However, several pitfalls are to be avoided in this regard. Notably, different types of visualization may vary substantially in terms of interpretability, as shown in Figs 6(c). For example, while a box plot provides basic information, it does not depict the distribution of metric values. This may conceal important information, such as specific images on which an algorithm performed poorly. Other pitfalls in this category relate to the non-determinism of algorithms, which introduces a natural variability to the results of a neural network, even with fixed seeds (Fig. SN 2.32). This issue is aggravated by inadequate reporting, for instance, reporting solely the results from the best run instead of proper cross-validation and reporting of the variability across different runs. Generally, shortcomings in reporting, such as providing no standard deviation or confidence intervals in the presented results, are common. Concrete examples of [P3.4] pitfalls can be found in Suppl. Note 2.3.4.

[P3.5] Inadequate interpretation of metric values.—Interpreting metric scores and aggregates is an important step for the analysis of algorithm performances. However, several pitfalls can arise from the interpretation. In rankings, for example, minor differences in metric scores may not be relevant from an application perspective but may still yield better ranks (Fig. SN 2.36). Furthermore, some metrics do not have upper or lower bounds, or the theoretical bounds may not be achievable in practice, rendering interpretation difficult (Fig. SN 2.35). More information on interpretation-based pitfalls can be found in Suppl. Note 2.3.4.

The first illustrated common access point to metric definitions and pitfalls

To underline the importance of a common access point to metric pitfalls, we conducted a search for individual metric-related pitfalls on the platforms Google Scholar and Google, with the purpose of determining how many of the pitfalls identified through our work could be located in existing resources. We were only able to locate a portion of the pitfalls identified by our approach in existing research literature (68%) or online resources such as blog posts (11%; 8% were found in both). Only 27% of the located pitfalls were presented visually.

Our work now provides this key resource in a highly structured and easily understandable form. Suppl. Note 2, contains a dedicated illustration for each of the pitfalls discussed, thus facilitating reader comprehension and making the information accessible to everyone regardless of their level of expertise. A further core contribution of our work are the metric profiles presented in Suppl. Note 2, which, for each metric, summarize the most important information deemed of particular relevance by the *Metrics Reloaded* consortium of the sister work to this publication [22]. The profiles provide the reader with a compact, at-a-glance overview of each metric and an enumeration of the limitations and pitfalls identified in the Delphi process conducted for this work.

DISCUSSION

Flaws in the validation of biomedical image analysis algorithms significantly impede the translation of methods into (clinical) practice and undermine the assessment of scientific progress in the field [19]. They are frequently caused by poor choices due to disregarding the specific properties and limitations of individual validation metrics. The present work represents the first comprehensive collection of pitfalls and limitations to be considered when using validation metrics in image-level classification, semantic segmentation, instance segmentation, and object detection tasks. Our work enables researchers to gain a deep understanding of and familiarity with both the overall topic and individual metrics by providing a common access point to previously largely scattered and inaccessible information — key knowledge they can resort to when conducting validation of image analysis algorithms. This way, our work aims to disrupt the current common practice of choosing metrics based on their popularity rather than their suitability to the underlying research problem. This practice, which, for instance, often manifests itself in the unreflected and inadequate use of the DSC, is concerningly prevalent even among prestigious, high-quality biomedical image analysis competitions (challenges) [8, 11, 15, 17, 18, 21, 23, 35]. The educational aspect of our work is complemented by dedicated ‘metric profiles’ which detail the definitions and properties of all metrics discussed. Notably, our work pioneers the examination of artificial intelligence (AI) validation pitfalls in the biomedical domain, a domain in which they are arguably more critical than in many others as flaws in biomedical algorithm validation can directly affect patient wellbeing and safety.

We posited that shortcomings in current common practice are marked by the low accessibility of information on the pitfalls and limitations of commonly used validation metrics. A literature search conducted from the point of view of a researcher seeking

information on individual metrics confirmed that the number of search results far exceeds any amount that could be overseen within reasonable time and effort, as well as the lack of a common point of entry to reliable metric information. Even when knowing the specific pitfalls and related keywords uncovered by our consortium, only a fraction of those pitfalls could be found in existing literature, indicating the novelty and added value of our work.

For transparency, several constraints regarding our literature search must be noted. First, it must be acknowledged that the remarkably high search result numbers inevitably include duplicates of papers (e.g., the same work in a conference paper and on arXiv) as well as results that are out of scope (e.g., [3], [9]), in the cited examples for instance due to a metric acronym (AUC) simultaneously being an acronym for another entity (a trinucleotide) in a different domain, or the word “sensitivity” being used in its common, non-metric meaning. Moreover, common words used to describe pitfalls such as “problem” or “issue” are by nature present in many publications discussing any kind of research, rendering them unusable for a dedicated search, which could, in turn, account for missing publications that do discuss pitfalls in these terms. Similarly, when searching for specific pitfalls, many of the returned results containing the appropriate keywords did not actually refer to metrics or algorithm validation but to other parts of a model or biomedical problem (e.g., the need for stratification is commonly discussed with regard to the design of clinical studies but not with regard to their validation). Character limits in the Google Scholar search bar further complicate or prevent the use of comprehensive search strings. Finally, it is both possible and probable that our literature search did not retrieve all publications or non-peer-reviewed online resources that mention a particular pitfall, since even extensive search strings might not cover the particular words used for a pitfall description.

None of these observations, however, detracts from our hypothesis. In fact, all of the above observations reinforce our finding that, for any individual researcher, retrieving information on metrics of interest is difficult to impossible. In many cases, finding information on pitfalls only appears feasible if the specific pitfall and its related keywords are exactly known, which, of course, is not the situation most researchers realistically find themselves in. Overall accessibility of such vital information, therefore, currently leaves much to be desired.

Compiling this information through a multi-stage Delphi process allowed us to leverage distributed knowledge from experts across different biomedical imaging domains and thus ensure that the resulting illustrated collection of metric pitfalls and limitations is both comprehensive and of maximum practical relevance. Continued proximity of our work to issues occurring in practical application was achieved through sharing the first results of this process as a dynamic preprint [27] with dedicated calls for feedback, as well as crowdsourcing further suggestions on social media.

Although their severity and practical consequences might differ between applications, we found that the pitfalls generalize across different imaging modalities and application domains. By categorizing them solely according to their underlying sources, we were able to create an overarching taxonomy that goes beyond domain-specific concerns and thus enjoys broad applicability. Given the large number of identified pitfalls, our taxonomy crucially

establishes structure in the topic. Moreover, by relating types of pitfalls to the respective metrics they apply to and illustrating them, it enables researchers to gain a deeper, systemic understanding of the causes of metric failure.

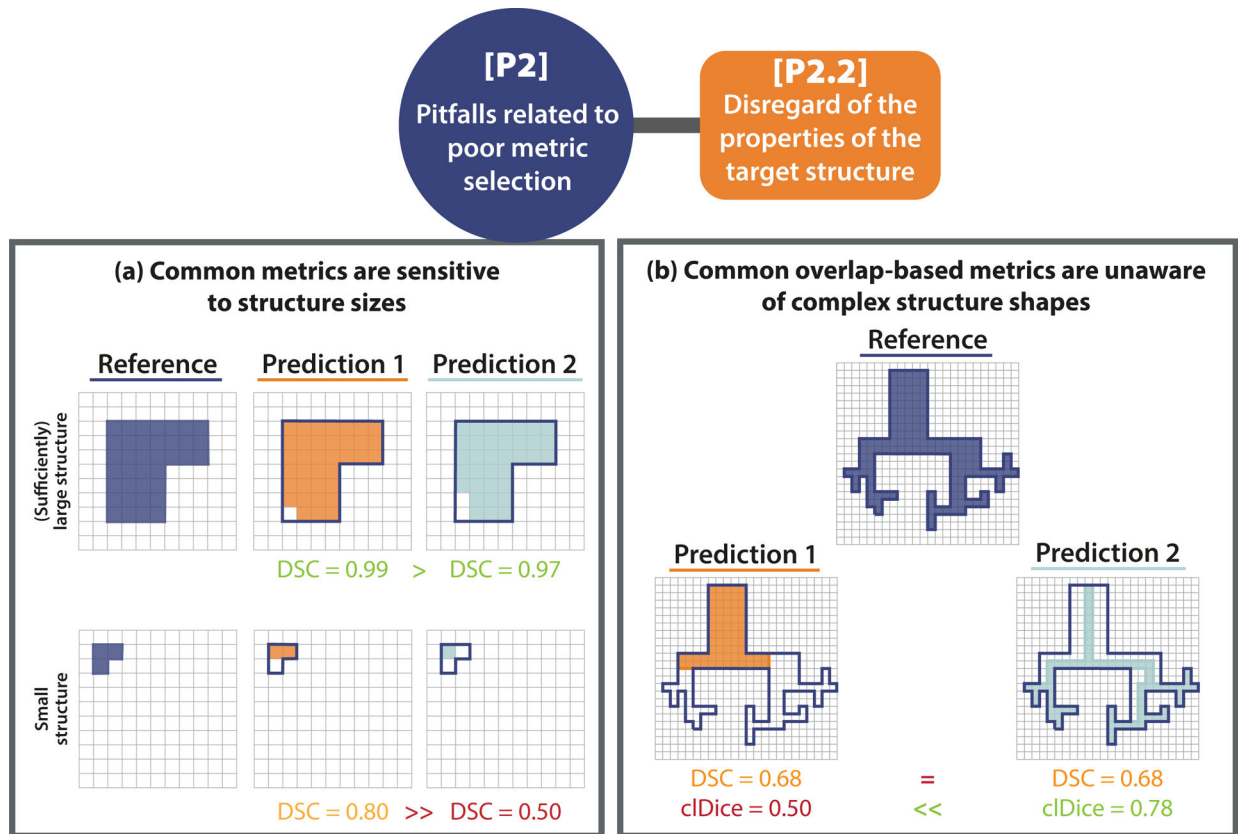
Our complementary *Metrics Reloaded* recommendation framework, which guides researchers towards the selection of appropriate validation metrics for their specific tasks and is introduced in a sister publication to this work [22], shares the same principle of domain independence. Its recommendations are based on the creation of a ‘problem fingerprint’ that abstracts from specific domain knowledge and, informed by the pitfalls discussed here, captures all properties relevant to metric selection for a specific biomedical problem. In this sister publication, we present recommendations to avoid the pitfalls presented in this work. Importantly, the finding that pitfalls generalize and can be categorized in a domain-independent manner opens up avenues for future expansion of our work to other fields of ML-based imaging, such as general computer vision (see below), thus freeing it from its major constraint of exclusively focusing on biomedical problems.

It is worth mentioning that we only examined pitfalls related to the tasks of image-level classification, semantic segmentation, instance segmentation, and object detection, as these can all be considered classification tasks at different levels (image/object/pixel) and hence share similarities in their validation. While including a wider range of biomedical problems not considered classification tasks, such as regression or registration, would have gone beyond the scope of the present work, we envision this expansion in future work. Moreover, our work focused on pitfalls related to reference-based metrics. Including pitfalls pertaining to non-reference-based metrics, such as metrics that assess speed, memory consumption, or carbon footprint, could be a future direction to take. Finally, while we aspired to be as comprehensive as possible in our compilation, we cannot exclude that there are further pitfalls to be taken into account that the consortium and the participating community have so far failed to recognize. Should this be the case, our dynamic *Metrics Reloaded* online platform, which is currently under development and will continuously be updated after release, will allow us to easily and transparently append missed pitfalls. This way, our work can remain a reliable point of access, reflecting the state of the art at any given moment in the future. In this context, we note that we explicitly welcome feedback and further suggestions from the readership of *Nature Methods*.

The expert consortium was primarily compiled in a way to cover the required expertise from various fields but also consisted of researchers of different countries, (academic) ages, roles, and backgrounds (details can be found in the Suppl. Methods). It mainly focused on biomedical applications. The pitfalls presented here are therefore of the highest relevance for biological and clinical use cases. Their clear generalization across different biomedical imaging domains, however, indicates broader generalizability to fields such as general computer vision. Future work could thus see a major expansion of our scope to AI validation well beyond biomedical research. Regardless of this possibility, we strongly believe that by raising awareness of metric-related pitfalls, our work will kick off a necessary scientific debate. Specifically, we see its potential in inducing the scientific communities in other areas of AI research to follow suit and investigate pitfalls and common practices impairing progress in their specific domains.

In conclusion, our work presents the first comprehensive and illustrated access point to information on validation metric properties and their pitfalls. We envision it to not only impact the quality of algorithm validation in biomedical imaging and ultimately catalyze faster translation into practice, but to raise awareness on common issues and call into question flawed AI validation practice far beyond the boundaries of the field.

Extended Data

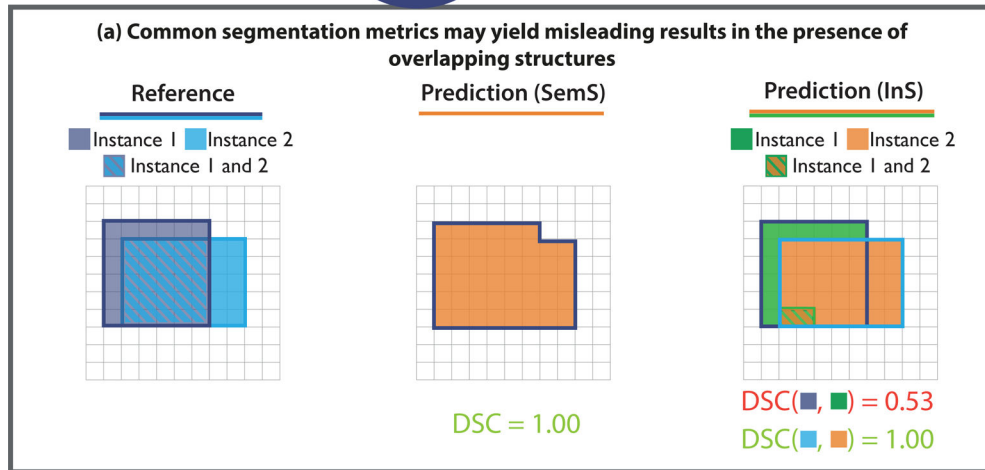
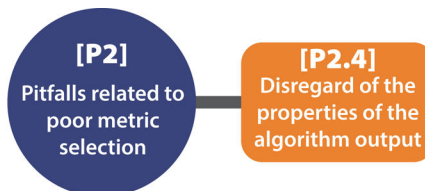


Extended Data Fig. 1. [P2.2] Disregard of the properties of the target structures.

[P2.2] **Disregard of the properties of the target structures. (a) Small structure sizes.**

The predictions of two algorithms (*Prediction 1/2*) differ in only a single pixel. In the case of the small structure (bottom row), this has a substantial effect on the corresponding Dice Similarity Coefficient (DSC) metric value (similar for the Intersection over Union (IoU)).

This pitfall is also relevant for other overlap-based metrics such as the centerline Dice Similarity Coefficient (cIDice), and localization criteria such as Box/Approx/Mask IoU and Intersection over Reference (IoR). **(b) Complex structure shapes.** Common overlap-based metrics (here: DSC) are unaware of complex structure shapes and treat *Predictions 1* and *2* equally. The cIDice uncovers the fact that *Predictions 1* misses the fine-granular branches of the reference and favors *Predictions 2*, which focuses on the center line of the object. This pitfall is also relevant for other overlap-based such as metrics IoU and pixel-level F_{β} Score as well as localization criteria such as Box/Approx/Mask IoU, Center Distance, Mask IoU > 0, Point inside Mask/Box/Approx, and IoR.



(b) Empty reference or prediction leads to invalid scores

Reference								
Prediction								
F_1 Score	NaN (0/0)	1	0	NaN (0/0)	>0	0	1	>0
Sensitivity	NaN (0/0)	1	0	NaN (0/0)	>0	0	1	>0
PPV	NaN (0/0)	1	NaN (0/0)	0	1	0	>0	>0

Extended Data Fig. 2. [P2.4] Disregard of the properties of the algorithm output.
[P2.4] Disregard of the properties of the algorithm output. (a) Possibility of overlapping predictions. If multiple structures of the same type can be seen within the same image (here: reference objects $R1$ and $R2$), it is generally advisable to phrase the problem as instance segmentation (InS; right) rather than semantic segmentation (SemS; left). This way, issues with boundary-based metrics resulting from comparing a given structure boundary to the boundary of the wrong instance in the reference can be avoided. In the provided example, the distance of the red boundary pixel to the reference, as measured by a boundary-based metric in SemS problems, would be zero, because different instances of the same structure cannot be distinguished. This problem is overcome by phrasing the problem as InS. In this case, (only) the boundary of the matched instance (here: $R2$) is considered for distance

computation. **(b) Possibility of empty prediction or reference.** Each column represents a potential scenario for per-image validation of objects, categorized by whether True Positives (TPs), False Negatives (FNs), and False Positives (FPs) are present ($n > 0$) or not ($n = 0$) after matching/assignment. The sketches on the top showcase each scenario when setting “ $n > 0$ ” to “ $n = 1$ ”. For each scenario, Sensitivity, Positive Predictive Value (PPV), and the F_1 Score are calculated. Some scenarios yield undefined values (Not a Number (NaN)).

Extended Data Tab. 1.
Overview of pitfall sources for *image-level classification*
metrics

((a): counting metrics, (b): multi-threshold metrics) related to poor metric selection [P2]. Pitfalls for semantic segmentation, object detection and instance segmentation are provided in Extended Data Tabs. 2–5 respectively. A warning sign indicates a potential pitfall for the metric in the corresponding column, in case the property represented by the respective row holds true. Comprehensive illustrations of pitfalls are available in Suppl. Note 2. A comprehensive list of pitfalls is provided separately for each metrics in the metrics cheat

sheets (Suppl. Note 3). Note that we only list sources of pitfalls relevant to the considered metrics. Other sources of pitfalls are neglected for this table.

(a) **Counting metrics.** Considered metrics: Accuracy (Fig. SN 3.38), Balanced Accuracy (BA) (Fig. SN 3.39), Expected Cost (EC) (Fig. SN 3.42), F_{β} Score (Fig. SN 3.43), Matthews Correlation Coefficient (MCC) (Fig. SN 3.46), Net Benefit (NB) (Fig. SN 3.47), Negative Predictive Value (NPV) (Fig. SN 3.48), Positive Likelihood Ratio (LR+) (Fig. SN 3.50), Positive Predictive Value (PPV) (Fig. SN 3.51), Sensitivity (Sens) (Fig. SN 3.52), Specificity (Spec) (Fig. SN 3.53), Weighted Cohen’s Kappa (WCK) (Fig. SN 3.54).

Source of pitfall	Accuracy	BA	EC	F_{β} Score	LR+	MCC	NB	PPV/ NPV	Sens/ Spec	WCK
Importance of confidence awareness	⚠ *	⚠ *	⚠ *	⚠ *	⚠ *	⚠ *	⚠ *	⚠ *	⚠ *	⚠ *
Importance of comparability across data sets	⚠ (Fig. SN 2.7)		⚠ ** (Fig. SN 2.7)	⚠ (Fig. SN 2.7)		⚠ (Fig. SN 2.7)	⚠ (Fig. SN 2.7)	⚠ (Fig. SN 2.7)		⚠ (Fig. SN 2.7)
Unequal severity of class confusions	⚠ (Fig. 4b)	⚠ (Fig. 4b)		⚠ *** (Fig. 4b)	⚠ (Fig. 4b)	⚠ (Fig. 4b)		⚠ (Fig. 4b)	⚠ (Fig. 4b)	
Importance of cost-benefit analysis	⚠ (Fig. SN 2.9)	⚠ (Fig. SN 2.9)		⚠ *** (Fig. SN 2.9)	⚠ (Fig. SN 2.9)	⚠ (Fig. SN 2.9)		⚠ (Fig. SN 2.9)	⚠ (Fig. SN 2.9)	
High class imbalance	⚠ (Figs. 5a, SN 2.15)	⚠ (Fig. 5a)	⚠ * (Fig. 5a)		⚠ (Fig. 5a)		⚠ (Figs. 5a, SN 2.15)	NPV: ⚠ (Figs. 5a, SN 2.15)	⚠ (Sens: Fig. 5a; Spec: Figs. 5a, SN 2.15)	⚠ (Figs. 5a, SN 2.15)
Small test set size	⚠ (Fig. SN 2.16)	⚠ (Fig. SN 2.16)	⚠ (Fig. SN 2.16)	⚠ (Fig. SN 2.16)	⚠ (Fig. SN 2.16)	⚠ (Fig. SN 2.16)	⚠ (Fig. SN 2.16)	⚠ (Fig. SN 2.16)	⚠ (Fig. SN 2.16)	⚠ (Fig. SN 2.16)

* Discrimination metrics do not assess whether the predicted class scores reflect the confidence of the classifier. This is typically achieved with additional calibration metrics, which come with their own pitfalls (see Figs. SN 2.6 and SN 2.22, Extended Data Fig.1b and the metric profiles in Suppl. Note 3.2).

** The weights in EC can be adjusted to avoid this pitfall.

*** The hyperparameter β can be used as a penalty for class confusions in the binary case. This property is not applicable to multi-class problems.

(b) **Multi-threshold metrics.** Considered metrics: Area under the Receiver Operating Characteristic Curve (AUROC) (Fig. SN 3.55) and Average Precision (AP) (Fig. SN 3.56).

Source of pitfall	AP	AUROC
Importance of confidence awareness	⚠ *	⚠ *
Importance of comparability across data sets	⚠ (Fig. SN 2.7)	

High class imbalance		⚠ (Fig. 5a)
Small test set size	⚠ (Fig. SN 2.16)	⚠ (Fig. SN 2.16)
Lack of predicted class scores	⚠ (Fig. SN 2.20)	⚠ (Fig. SN 2.20)

* Discrimination metrics do not assess whether the predicted class scores reflect the confidence of the classifier. This is typically achieved with additional calibration metrics, which come with their own pitfalls (see Figs. SN 2.6 and SN 2.22, Extended Data Fig.1b and the metric profiles in Suppl. Note 3.2).

Extended Data Tab. 2.
Overview of pitfall sources for *semantic segmentation metrics*

((a): overlap-based metrics, (b): boundary-based metrics) related to poor metric selection [P2]. A warning sign indicates a potential pitfall for the metric in the corresponding column, in case the property represented by the respective row holds true. Comprehensive illustrations of pitfalls are available in Suppl. Note 2. A comprehensive list of pitfalls is provided separately for each metrics in the metrics cheat sheets (Suppl. Note 3). Note that



















we only list sources of pitfalls relevant to the considered metrics. Other sources of pitfalls are neglected for this table.

(a) **Overlap-based metrics.** Considered metrics: centerline Dice Similarity Coefficient (clDice) (Fig. SN 3.40), Dice Similarity Coefficient (DSC) (Fig. SN 3.41), F_{β} Score (Fig. SN 3.43), Intersection over Union (IoU) (Fig. SN 3.45).

Source of potential pitfall	clDice	DSC/IoU	F_{β} Score
Importance of structure boundaries	⚠ (Fig. 4a)	⚠ (Fig. 4a)	⚠ (Fig. 4a)
Importance of structure center(line)		⚠ (Fig. SN 2.5, Extended Data Fig. 1b)	⚠ (Fig. SN 2.5, Extended Data Fig. 1b)
Unequal severity of class confusions	⚠ (Fig. SN 2.8)	⚠ (Fig. SN 2.8)	
Small structure sizes	⚠ (Fig. SN 2.10, Extended Data Fig. 1a)	⚠ (Fig. SN 2.10, Extended Data Fig. 1a)	⚠ (Fig. SN 2.10, Extended Data Fig. 1a)
High variability of structure sizes	⚠ (Fig. SN 2.11)	⚠ (Fig. SN 2.11)	⚠ (Fig. SN 2.11)
Complex structure shapes		⚠ (Fig. SN 2.12)	⚠ (Fig. SN 2.12)
Occurrence of overlapping or touching structures	⚠ (Fig. SN 2.13)	⚠ (Fig. SN 2.13)	⚠ (Fig. SN 2.13)
Imperfect reference standard		⚠ (Fig. SN 2.17)	⚠ (Fig. SN 2.17)
Occurrence of cases with an empty reference	⚠ (Fig. SN 2.18)	⚠ (Fig. SN 2.18)	⚠ (Fig. SN 2.18)
Possibility of empty prediction	⚠ (Fig. SN 2.18)	⚠ (Fig. SN 2.18)	⚠ (Fig. SN 2.18)
Possibility of overlapping predictions	⚠ (Fig. SN 2.19, Extended Data Fig. 2a)	⚠ (Fig. SN 2.19, Extended Data Fig. 2a)	⚠ (Fig. SN 2.19, Extended Data Fig. 2a)

(b) **Boundary-based metrics.** Considered metrics: Average Symmetric Surface Distance (ASSD) (Fig. SN 3.58), Boundary Intersection over Union (Boundary IoU) (Fig. SN 3.59), Hausdorff Distance (HD) (Fig. SN 3.60), Hausdorff Distance 95th Percentile (HD95) (Fig. SN 3.63), Mean Average Surface Distance (MASD) (Fig. SN 3.61), Normalized Surface Distance (NSD) (Fig. SN 3.62).

Source of potential pitfall	ASSD	Boundary IoU	HD	HD95	MASD	NSD
Importance of structure volume	⚠ (Fig. SN 2.4)	⚠ (Fig. SN 2.4)	⚠ (Fig. SN 2.4)	⚠ (Fig. SN 2.4)	⚠ (Fig. SN 2.4)	⚠ (Fig. SN 2.4)
Importance of structure center(line)	⚠ (Fig. SN 2.5, Extended Data Fig. 1b)	⚠ (Fig. SN 2.5, Extended Data Fig. 1b)	⚠ (Fig. SN 2.5, Extended Data Fig. 1b)	⚠ (Fig. SN 2.5, Extended Data Fig. 1b)	⚠ (Fig. SN 2.5, Extended Data Fig. 1b)	⚠ (Fig. SN 2.5, Extended Data Fig. 1b)
Occurrence of overlapping or touching structures	⚠ (Fig. SN 2.13)	⚠ (Fig. SN 2.13)	⚠ (Fig. SN 2.13)	⚠ (Fig. SN 2.13)	⚠ (Fig. SN 2.13)	⚠ (Fig. SN 2.13)
Imperfect reference	⚠ (Figs. 5c,	⚠ (Figs. 5c,	⚠ (Figs. 5c,	⚠ (Figs.	⚠ (Figs. 5c,	

standard	SN 2.17)	SN 2.17)	SN 2.17)	5c*, SN 2.17)	SN 2.17)	
Occurrence of cases with an empty reference	 (Fig. SN 2.18)	 (Fig. SN 2.18)	 (Fig. SN 2.18)	 (Fig. SN 2.18)	 (Fig. SN 2.18)	 (Fig. SN 2.18)
Possibility of empty prediction	 (Fig. SN 2.18)	 (Fig. SN 2.18)	 (Fig. SN 2.18)	 (Fig. SN 2.18)	 (Fig. SN 2.18)	 (Fig. SN 2.18)
Possibility of overlapping predictions	 (Fig. SN 2.19, Extended Data Fig. 2a)	 (Fig. SN 2.19, Extended Data Fig. 2a)	 (Fig. SN 2.19, Extended Data Fig. 2a)	 (Fig. SN 2.19, Extended Data Fig. 2a)	 (Fig. SN 2.19, Extended Data Fig. 2a)	 (Fig. SN 2.19, Extended Data Fig. 2a)
























*Can be mitigated by the choice of the percentile

Extended Data Tab. 3.
Overview of sources of pitfalls for *object detection*
metrics

((a): detection metrics, (b): localization criteria) related to poor metric selection [P2]. A warning sign indicates a potential pitfall for the metric in the corresponding column, in case the property represented by the respective row holds true. Comprehensive illustrations of pitfalls are available in Suppl. Note 2. A comprehensive list of pitfalls is provided separately for each metrics in the metrics cheat sheets (Suppl. Note 3). Note that we only list sources of




















pitfalls relevant to the considered metrics. Other sources of pitfalls are neglected for this table.




(a) **Detection metrics.** Considered counting metrics: F_β Score (Fig. SN 3.43), Positive Predictive Value (PPV) (Fig. SN 3.51), Sensitivity (Sens) (Fig. SN 3.52). Considered multi-threshold metrics: Average Precision (AP) (Fig. SN 3.56) and Free-Response Receiver Operating Characteristic (FROC) (Fig. SN 3.57).

Source of potential pitfall	F_β Score	PPV	Sens	AP	FROC Score
Unequal severity of class confusions	 (Fig. 4b)	 (Fig. 4b)	 (Fig. 4b)	 (Fig. 4b)	 (Fig. 4b)
High class imbalance			 (Fig. 5a)		
Small test set size	 (Fig. SN 2.16)	 (Fig. SN 2.16)	 (Fig. SN 2.16)	 (Fig. SN 2.16)	 (Fig. SN 2.16)
Occurrence of cases with an empty reference	 (Fig. SN 2.18, Extended Data Fig. 2b)	 (Fig. SN 2.18, Extended Data Fig. 2b)	 (Fig. SN 2.18, Extended Data Fig. 2b)	 (Fig. SN 2.18, Extended Data Fig. 2b)	 (Fig. SN 2.18, Extended Data Fig. 2b)
Possibility of empty prediction	 (Fig. SN 2.18, Extended Data Fig. 2b)	 (Fig. SN 2.18, Extended Data Fig. 2b)	 (Fig. SN 2.18, Extended Data Fig. 2b)	 (Fig. SN 2.18, Extended Data Fig. 2b)	 (Fig. SN 2.18, Extended Data Fig. 2b)
Lack of predicted class scores				 (Fig. SN 2.20)	 (Fig. SN 2.20)

* The hyperparameter β can be used as a penalty for class confusions in the binary case. This property is not applicable to multi-class problems.

(b) **Localization criteria.** Considered localization criteria: Box/Approx IoU (Fig. SN 3.74), Center Distance (Fig. SN 3.72), Mask IoU > 0 (Fig. SN 3.75), and Point inside Mask/ Box/ Approx (Fig. SN 3.76).

Source of potential pitfall	Box/Approx IoU	Center Distance	Mask IoU > 0	Point inside Mask/ Box/ Approx
Importance of structure boundaries	 (Fig. 4a)	 (Fig. 4a)	 (Fig. 4a)	 (Fig. 4a)
Importance of structure volume		 (Fig. SN 2.4)	 (Fig. SN 2.4)	 (Fig. SN 2.4)
Importance of structure center(line)	 (Fig. SN 2.5, Extended Data Fig. 1b)		 (Fig. SN 2.5, Extended Data Fig. 1b)	 (Fig. SN 2.5, Extended Data Fig. 1b)
Unequal severity of class confusions	 (Fig. SN 2.8)	 (Fig. SN 2.8)*	 (Fig. SN 2.8)	 (Fig. SN 2.8)*
Small structure sizes	 (Fig. SN 2.10, Extended Data Fig. 1a)			
Complex structure shapes	 (Figs. SN 2.12, SN 2.14)	 (Fig. SN 2.12)	 (Fig. SN 2.12)	 (Fig. SN 2.12)

Occurrence of disconnected structures	 (Fig. SN 2.14)			Point inside Box:  (Fig. SN 2.14)
Imperfect reference standard	 (Fig. 5c)			














* Criterion implies point prediction, thus overlap assessment is not applicable.

Extended Data Tab. 4.
Overview of sources of pitfalls for *instance segmentation metrics* (Part 1)

((a): detection metrics, (b): localization criteria) related to poor metric selection [P2]. A warning sign indicates a potential pitfall for the metric in the corresponding column, in case the property represented by the respective row holds true. Comprehensive illustrations of pitfalls are available in Suppl. Note 2. A comprehensive list of pitfalls is provided separately for each metrics in the metrics cheat sheets (Suppl. Note 3). Note that we only list sources of

















pitfalls relevant to the considered metrics. Other sources of pitfalls are neglected for this table.

(a) Detection metrics. Considered counting metrics: F_{β} Score, Positive Predictive Value (PPV), Panoptic Quality (PQ), Sensitivity (Sens). Considered multi-threshold metrics: Average Precision (AP) (Fig. SN 3.56) and Free-Response Receiver Operating Characteristic (FROC) (Fig. SN 3.57).

Source of potential pitfall	F_{β} Score	PPV	PQ	Sens	AP	FROC Score
Unequal severity of class confusions	 (Fig. 4b)*	 (Fig. 4b)	 (Fig. 4b)	 (Fig. 4b)		
High class imbalance				 (Fig. 5a)		
Small test set size	 (Fig. SN 2.16)	 (Fig. SN 2.16)	 (Fig. SN 2.16)	 (Fig. SN 2.16)	 (Fig. SN 2.16)	 (Fig. SN 2.16)
Lack of predicted class scores					 (Fig. SN 2.20)	 (Fig. SN 2.20)

* The hyperparameter β can be used as a penalty for class confusions in the binary case. This property is not applicable to multi-class problems.

(b) Localization criteria. Considered localization criteria: Boundary Intersection over Union (IoU) (Fig. SN 3.59), Intersection over Reference (IoR) (Fig. SN 3.73), Mask IoU (Fig. SN 3.74).















Source of potential pitfall	Boundary IoU	IoR	Mask IoU
Importance of structure boundaries		 (Fig. 4a)	 (Fig. 4a)
Importance of structure volume	 (Fig. SN 2.4)		
Importance of structure center(line)	 (Fig. SN 2.5, Extended Data Fig. 1b)	 (Fig. SN 2.5, Extended Data Fig. 1b)	 (Fig. SN 2.5, Extended Data Fig. 1b)
Unequal severity of class confusions	 (Fig. SN 2.8)	 (Fig. SN 2.8)	 (Fig. SN 2.8)
Small structure sizes		 (Fig. SN 2.10, Extended Data Fig. 1a)	 (Fig. SN 2.10, Extended Data Fig. 1a)
Complex structure shapes		 (Fig. SN 2.12)	 (Fig. SN 2.12)
Imperfect reference standard	 (Fig. SN 2.17)	 (Fig. SN 2.17)	 (Fig. SN 2.17)

Extended Data Tab. 5.
Overview of sources of pitfalls for *instance segmentation metrics* (Part 2)


















((a) per instance segmentation overlap-based metrics, (b) per instance segmentation boundary-based metrics) related to poor metric selection [P2]. A warning sign indicates a potential pitfall for the metric in the corresponding column, in case the property represented by the respective row holds true. Comprehensive illustrations of pitfalls are available in

Suppl. Note 2. Note that we only list sources of pitfalls relevant to the considered metrics. Other sources of pitfalls are neglected for this table.

(a) Per instance segmentation overlap-based metrics. Considered metrics: Considered metrics: centerline Dice Similarity Coefficient (cDice) (Fig. SN 3.40), Dice Similarity Coefficient (DSC) (Fig. SN 3.41), F_{β} Score (Fig. SN 3.43), Intersection over Union (IoU) (Fig. SN 3.45).

Source of potential pitfall	cDice	DSC/IoU	F_{β} Score
Importance of structure boundaries	 (Fig. 4a)	 (Fig. 4a)	 (Fig. 4a)
Importance of structure center(line)		 (Fig. SN 2.5, Extended Data Fig. 1b)	 (Fig. SN 2.5, Extended Data Fig. 1b)
Unequal severity of class confusions	 (Fig. SN 2.8)	 (Fig. SN 2.8)	
Small structure sizes	 (Fig. SN 2.10, Extended Data Fig. 1a)	 (Fig. SN 2.10, Extended Data Fig. 1a)	 (Fig. SN 2.10, Extended Data Fig. 1a)
Complex structure shapes		 (Fig. SN 2.12)	 (Fig. SN 2.12)
Imperfect reference standard		 (Fig. SN 2.17)	 (Fig. SN 2.17)

(b) Per instance segmentation boundary-based metrics. Considered metrics: Average Symmetric Surface Distance (ASSD) (Fig. SN 3.58), Boundary Intersection over Union (IoU) (Fig. SN 3.59), Hausdorff Distance (HD) (Fig. SN 3.60), Hausdorff Distance 95th Percentile (HD95) (Fig. SN 3.61), Mean Average Surface Distance (MASD) (Fig. SN 3.61) and Normalized Surface Distance (NSD) (Fig. SN 3.62).

Source of potential pitfall	ASSD	Boundary IoU	HD	HD95	MASD	NSD
Importance of structure volume	 (Fig. SN 2.4)	 (Fig. SN 2.4)	 (Fig. SN 2.4)	 (Fig. SN 2.4)	 (Fig. SN 2.4)	 (Fig. SN 2.4)
Importance of structure center(line)	 (Fig. SN 2.5, Extended Data Fig. 1b)	 (Fig. SN 2.5, Extended Data Fig. 1b)	 (Fig. SN 2.5, Extended Data Fig. 1b)	 (Fig. SN 2.5, Extended Data Fig. 1b)	 (Fig. SN 2.5, Extended Data Fig. 1b)	 (Fig. SN 2.5, Extended Data Fig. 1b)
Imperfect reference standard	 (Figs. 5c, SN 2.17)	 (Figs. 5c, SN 2.17)	 (Figs. 5c, SN 2.17)	 (Figs. 5c, SN 2.17)	 (Figs. 5c, SN 2.17)	

* Can be mitigated by the choice of the percentile

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Authors

ANNIKA REINKE^{*,†,1,2,3}, MINU D. TIZABI^{*,†,1,4}, MICHAEL BAUMGARTNER⁵, MATTHIAS EISENMANN¹, DOREEN HECKMANN-NÖTZEL^{1,4}, A. EMRE KAVUR^{1,5,6}, TIM RÄDSCH^{1,2}, CAROLE H. SUDRE^{7,8}, LAURA ACION⁹, MICHELA ANTONELLI^{8,10}, TAL ARBEL¹¹, SPYRIDON BAKAS^{12,13}, ARIEL BENIS^{14,15}, FLORIAN BUETTNER^{16,17,18,19}, M. JORGE CARDOSO⁸, VERONIKA CHEPLYGINA²⁰, JIANXU CHEN²¹, EVANGELIA CHRISTODOULOU¹, BETH A. CIMINI²², KEYVAN FARAHANI²³, LUCIANA FERRER²⁴, ADRIAN GALDRAN^{25,26}, BRAM VAN GINNEKEN^{27,28}, BEN GLOCKER²⁹, PATRICK GODAU^{1,3,4}, DANIEL A. HASHIMOTO^{30,31}, MICHAEL M. HOFFMAN^{32,33,34,35}, MEREL HUISMAN³⁶, FABIAN ISENSEE^{5,6}, PIERRE JANNIN^{37,38}, CHARLES E. KAHN³⁹, DAGMAR KAINMUELLER^{40,41}, BERNHARD KAINZ^{42,43}, ALEXANDROS KARARGYRIS⁴⁴, JENS KLEESIEK⁴⁵, FLORIAN KOFLER⁴⁶, THIJS KOOI⁴⁷, ANNETTE KOPPSCHNEIDER⁴⁸, MICHAL KOZUBEK⁴⁹, ANNA KRESHUK⁵⁰, TAHSIN KURC⁵¹, BENNETT A. LANDMAN⁵², GEERT LITJENS⁵³, AMIN MADANI⁵⁴, KLAUS MAIERHEIN^{5,55}, ANNE L. MARTEL^{33,56}, ERIK MEIJERING⁵⁷, BJOERN MENZE⁵⁸, KAREL G.M. MOONS⁵⁹, HENNING MÜLLER^{60,61}, BRENNAN NICHYPORUK⁶², FELIX NICKEL⁶³, JENS PETERSEN⁵, SUSANNE M. RAFELSKI⁶⁴, NASIR RAJPOOT⁶⁵, MAURICIO REYES^{66,67}, MICHAEL A. RIEGLER^{68,69}, NICOLA RIEKE⁷⁰, JULIO SAEZ-RODRIGUEZ^{71,72}, CLARA I. SÁNCHEZ⁷³, SHRAVYA SHETTY⁷⁴, RONALD M. SUMMERS⁷⁵, ABDEL A. TAHA⁷⁶, ALEKSEI TIULPIN^{77,78}, SOTIRIOS A. TSAFTARIS⁷⁹, BEN VAN CALSTER^{80,81}, GAËL VAROQUAUX⁸², ZIV R. YANIV⁸³, PAUL F. JÄGER^{*,†,2,84}, LENA MAIER-HEIN^{*,†,1,2,3,4,72}

Affiliations

- ¹German Cancer Research Center (DKFZ) Heidelberg, Division of Intelligent Medical Systems, Im Neuenheimer Feld 223, 69120 Heidelberg, Heidelberg, Germany
- ²German Cancer Research Center (DKFZ) Heidelberg, HI Helmholtz Imaging, Im Neuenheimer Feld 223, 69120 Heidelberg, Heidelberg, Germany
- ³Faculty of Mathematics and Computer Science, Heidelberg University, Im Neuenheimer Feld 205, 69120 Heidelberg, Germany
- ⁴National Center for Tumor Diseases (NCT), NCT Heidelberg, a partnership between DKFZ and University Medical Center Heidelberg, Im Neuenheimer Feld 460, 69120 Heidelberg, Germany
- ⁵German Cancer Research Center (DKFZ) Heidelberg, Division of Medical Image Computing, Im Neuenheimer Feld 223, 69120 Heidelberg, Heidelberg, Germany
- ⁶German Cancer Research Center (DKFZ) Heidelberg, HI Applied Computer Vision Lab, Im Neuenheimer Feld 223, 69120 Heidelberg, Heidelberg, Germany
- ⁷MRC Unit for Lifelong Health and Ageing at UCL and Centre for Medical Image Computing, Department of Computer Science, University College London, Gower St, London WC1E 6BT, UK

- ⁸School of Biomedical Engineering and Imaging Science, King's College London, Westminster Bridge Road, London SE1 7EH, UK
- ⁹Instituto de Cálculo, CONICET – Universidad de Buenos Aires, Av. Int. Güiraldes 2160, C1428 Buenos Aires, Argentina
- ¹⁰Centre for Medical Image Computing, University College London, Gower St, London WC1E 6BT, UK
- ¹¹Centre for Intelligent Machines and MILA (Quebec Artificial Intelligence Institute), McGill University, 3480 Rue University, Montréal, QC H3A 2A7, Canada
- ¹²Division of Computational Pathology, Dept of Pathology & Laboratory Medicine, Indiana University School of Medicine, IU Health Information and Translational Sciences Building, 410 W 10th St, Rm.3119, Indianapolis, IN 46202, USA
- ¹³Center for Biomedical Image Computing and Analytics (CBICA), University of Pennsylvania, Richards Medical Research Laboratories FL7, 3700 Hamilton Walk, Philadelphia, PA 19104, USA
- ¹⁴Department of Digital Medical Technologies, Holon Institute of Technology, YGolomb St. 52, 5810201 Holon, Israel
- ¹⁵European Federation for Medical Informatics, Ch de Maillefer 37, CH-1052 Le Mont-sur-Lausanne, Switzerland
- ¹⁶German Cancer Consortium (DKTK), partner site Frankfurt/Mainz, a partnership between DKFZ and UCT Frankfurt-Marburg, Theodor-Stern-Kai 7, 60590 Frankfurt am Main, Germany
- ¹⁷German Cancer Research Center (DKFZ) Heidelberg, Im Neuenheimer Feld 280, 69120 Heidelberg, Germany
- ¹⁸Goethe University Frankfurt, Department of Medicine, 60590 Frankfurt am Main, Germany, Goethe University Frankfurt, Department of Informatics, 60629 Frankfurt am Main, Germany,
- ¹⁹Frankfurt Cancer Institute, Paul-Ehrlich-Straße 42-44, 60596 Frankfurt am Main, Germany
- ²⁰Department of Computer Science, IT University of Copenhagen, Rued Langgaards Vej 7, 2300 Copenhagen, Denmark
- ²¹Leibniz-Institut für Analytische Wissenschaften – ISAS – e.V., Bunsen-Kirchhoff-Straße 11, 44139 Dortmund, Germany
- ²²Imaging Platform, Broad Institute of MIT and Harvard, 415 Main St., Cambridge, MA 02142, USA
- ²³Center for Biomedical Informatics and Information Technology, National Cancer Institute, 37 Convent Dr, Bethesda, MD 20814, USA
- ²⁴Instituto de Investigación en Ciencias de la Computación (ICC), CONICET-UBA, Pabellón 0+inf, Ciudad Universitaria, Ciudad Autónoma de Buenos Aires, Argentina

25. Universitat Pompeu Fabra, Plaça de la Mercè, 10-12, 08002 Barcelona, Spain
26. University of Adelaide, Adelaide SA 5005, Australia
27. Fraunhofer MEVIS, Max-Von-Laue-Straße 2, 28359 Bremen, Germany
28. Radboud Institute for Health Sciences, Radboud University Medical Center, Geert Grooteplein Zuid 10, 6525 GA Nijmegen, The Netherlands
29. Department of Computing, Imperial College London, South Kensington Campus, London SW7 2AZ, UK
30. Department of Surgery, Perelman School of Medicine, 3400 Civic Center Boulevard, Philadelphia, PA 19104, USA
31. General Robotics Automation Sensing and Perception Laboratory, School of Engineering and Applied Science, University of Pennsylvania, 3330 Walnut Street, Philadelphia, PA 19104-6228, USA
32. Princess Margaret Cancer Centre, University Health Network, Princess Margaret Cancer Research Tower 11-311, 101 College St, Toronto, ON M5G 1L7, Canada
33. Department of Medical Biophysics, University of Toronto, Princess Margaret Cancer Research Tower 11-311, 101 College St, Toronto, ON M5G 1L7, Canada
34. Department of Computer Science, University of Toronto, 40 St. George St Room 4283, Toronto, ON M5 2E4, Canada
35. Vector Institute for Artificial Intelligence, 661 University Ave., Suite 710, Toronto, ON M5G 1M1, Canada
36. Department of Radiology and Nuclear Medicine, Radboud University Medical Center, Geert Grooteplein Zuid 10, 6525 GA Nijmegen, The Netherlands
37. Laboratoire Traitement du Signal et de l'Image – UMR_S 1099, Université de Rennes 1, 263 Avenue du Général Leclerc, 35042 Rennes, France
38. INSERM, 101 rue de Tolbiac, 75654 Paris Cedex 13, France
39. Department of Radiology and Institute for Biomedical Informatics, University of Pennsylvania, 3400 Spruce St., Philadelphia, PA 19104-4238, USA
40. Max-Delbrück Center for Molecular Medicine in the Helmholtz Association (MDC), Biomedical Image Analysis and HI Helmholtz Imaging, Robert-Rössle-Straße 10, 13125 Berlin, Germany
41. University of Potsdam, Digital Engineering Faculty, Prof.-Dr.-Helmert-Str. 2-3, 14482 Potsdam, Germany
42. Department of Computing, Faculty of Engineering, Imperial College London, 180 Queen's Gate, South Kensington, London SW7 2RH, UK
43. Department AIBE, Friedrich-Alexander-Universität (FAU), Werner-von-Siemens-Straße 61, 91052 Erlangen-Nürnberg, Germany
44. IHU Strasbourg, 1 Pl. de l'Hôpital, 67000 Strasbourg, France

45. Translational Image-guided Oncology (TIO), Institute for AI in Medicine (IKIM), University Medicine Essen, Girardetstraße 2, 45131 Essen, Germany
46. Helmholtz AI, Ingolstädter Landstraße 1, 85764 Oberschleißheim, Germany
47. Lunit, Gangnam-gu, Gangnam-daero, 374 4 - 9, Seoul, South Korea
48. German Cancer Research Center (DKFZ) Heidelberg, Division of Biostatistics, Germany, Im Neuenheimer Feld 580, Heidelberg, Germany
49. Centre for Biomedical Image Analysis and Faculty of Informatics, Masaryk University, Botanická 68a, 602 00 Brno, Czech Republic
50. Cell Biology and Biophysics Unit, European Molecular Biology Laboratory (EMBL), Meyerhofstraße 1, 69117 Heidelberg, Germany
51. Department of Biomedical Informatics, Stony Brook University, Health Science Center, Stony Brook, NY 11794-8322, USA
52. Electrical Engineering, Vanderbilt University, 2301 Vanderbilt Pl, Nashville, TN 37235, USA
53. Department of Pathology, Radboud University Medical Center, Geert Grooteplein Zuid 10, 6525 GA Nijmegen, The Netherlands
54. Department of Surgery, University Health Network, 3400 Civic Center Boulevard, Philadelphia, PA 19104, Canada
55. Pattern Analysis and Learning Group, Department of Radiation Oncology, Heidelberg University Hospital, Im Neuenheimer Feld 672, 69120 Heidelberg, Germany
56. Physical Sciences, Sunnybrook Research Institute, 2075 Bayview Ave, Toronto, ON M4N 3M5, Canada
57. School of Computer Science and Engineering, University of New South Wales, Engineering Rd, UNSW Sydney, Kensington NSW 2052, Australia
58. Department of Quantitative Biomedicine, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland
59. Julius Center for Health Sciences and Primary Care, UMC Utrecht, Utrecht University, Universiteitsweg 100, 3584 CG Utrecht, The Netherlands
60. Information Systems Institute, University of Applied Sciences Western Switzerland (HES-SO), Rue de l'Industrie 23, 1950 Sierre, Switzerland
61. Medical Faculty, University of Geneva, Rue Michel-Servet 1, 1206 Geneva, Switzerland
62. MILA (Quebec Artificial Intelligence Institute), 6666 Rue Saint-Urbain, Montréal, QC H2S 3H1, Canada
63. Department of General, Visceral and Thoracic Surgery, University Medical Center Hamburg-Eppendorf, Martinistraße 52, 20246 Hamburg, Germany

- ⁶⁴.Allen Institute for Cell Science, 615 Westlake Ave North, Seattle, WA 98109, USA
- ⁶⁵.Tissue Image Analytics Laboratory, Department of Computer Science, University of Warwick, Coventry CV4 7AL, UK
- ⁶⁶.ARTORG Center for Biomedical Engineering Research, University of Bern, Bern, Switzerland
- ⁶⁷.Department of Radiation Oncology, University Hospital Bern, University of Bern, Bern, Switzerland
- ⁶⁸.Simula Metropolitan Center for Digital Engineering, Pilestredet 52, 0167 Oslo, Norway
- ⁶⁹.UiT The Arctic University of Norway, Hansine Hansens veg 18, 9019 Tromsø, Norway
- ⁷⁰.NVIDIA GmbH, Einsteinstraße 172, 81677 München, Germany
- ⁷¹.Institute for Computational Biomedicine, Heidelberg University, Im Neuenheimer Feld 130.3, 69120 Heidelberg, Germany
- ⁷².Faculty of Medicine, Heidelberg University Hospital, 69120 Heidelberg, Germany
- ⁷³.Informatics Institute, Faculty of Science, University of Amsterdam, P.O. Box 94323, 1090 GH Amsterdam, The Netherlands
- ⁷⁴.Google Health, Google, 935 E Meadow Dr, Palo Alto, CA 94303, USA
- ⁷⁵.National Institutes of Health Clinical Center, 10 Center Dr, Bethesda, MD 20892, USA
- ⁷⁶.Institute of Information Systems Engineering, TU Wien, Favoritenstraße 9-11/194, 1040 Vienna, Austria
- ⁷⁷.Research Unit of Health Sciences and Technology, Faculty of Medicine, University of Oulu, Aapistie 5A, Oulu, Finland
- ⁷⁸.Neurocenter Oulu, Oulu University Hospital, Kajaanintie 50, Oulu, Finland
- ⁷⁹.School of Engineering, The University of Edinburgh, Edinburgh EH9 3JL, Scotland
- ⁸⁰.Department of Development and Regeneration and EPI-centre, KU Leuven, Herestraat 49 box 805, 3000 Leuven, Belgium
- ⁸¹.Department of Biomedical Data Sciences, Leiden University Medical Center, Albinusdreef 2, 2333 ZA Leiden, The Netherlands
- ⁸².Parietal project team, INRIA Saclay-Île de France, 1 Rue Honoré d'Estienne d'Orves, 91120 Palaiseau, France
- ⁸³.National Institute of Allergy and Infectious Diseases, National Institutes of Health, 5601 Fishers Ln, Bethesda, MD 20892, USA

⁸⁴.German Cancer Research Center (DKFZ) Heidelberg, Interactive Machine Learning Group, Im Neuenheimer Feld 223, 69120 Heidelberg, Heidelberg, Germany

ACKNOWLEDGEMENTS

This work was initiated by the Helmholtz Association of German Research Centers in the scope of the Helmholtz Imaging Incubator (HI), the MICCAI Special Interest Group for biomedical image analysis challenges, and the benchmarking working group of the MONAI initiative. It has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. [101002198], NEURAL SPICING) and the Surgical Oncology Program of the National Center for Tumor Diseases (NCT) Heidelberg. 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. T.A. acknowledges the Canada Institute for Advanced Research (CIFAR) AI Chairs program, the Natural Sciences and Engineering Research Council of Canada. F.B. was co-funded by the European Union (ERC, TAIPO, 101088594). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. M.J.C. acknowledges funding from Wellcome/EPSRC Centre for Medical Engineering (WT203148/Z/16/Z), the Wellcome Trust (WT213038/Z/18/Z), and the InnovateUK funded London AI Centre for Value-Based Healthcare. J.C. is supported by the Federal Ministry of Education and Research (BMBF) under the funding reference 161L0272. V.C. acknowledges funding from NovoNordisk Foundation (NNF21OC0068816) and Independent Research Council Denmark (1134-00017B). B.A.C. was supported by NIH grant P41 GM135019 and grant 2020-225720 from the Chan Zuckerberg Initiative DAF, an advised fund of the Silicon Valley Community Foundation. G.S.C. was supported by Cancer Research UK (programme grant: C49297/A27294). M.M.H. is supported by the Natural Sciences and Engineering Research Council of Canada (RGPIN- 2022-05134). A.Ka. is supported by French State Funds managed by the "Agence Nationale de la Recherche (ANR)" - "Investissements d'Avenir" (Investments for the Future), Grant ANR-10- IAHU-02 (IHU Strasbourg). M.K. was funded by the Ministry of Education, Youth and Sports of the Czech Republic (Project LM2018129). Ta.K. was supported in part by 4UH3-CA225021- 03, 1U24CA180924-01A1, 3U24CA215109-02, and 1UG3-CA225-021-01 grants from the National Institutes of Health. G.L. receives research funding from the Dutch Research Council, the Dutch Cancer Association, HealthHolland, the European Research Council, the European Union, and the Innovative Medicine Initiative. S.M.R. wishes to acknowledge the Allen Institute for Cell Science founder Paul G. Allen for his vision, encouragement and support. M.R. is supported by Innosuisse grant number 31274.1 and Swiss National Science Foundation Grant Number 205320_212939. C.H.S. is supported by an Alzheimer's Society Junior Fellowship (AS-JF-17-011). R.M.S. is supported by the Intramural Research Program of the NIH Clinical Center. A.T. acknowledges support from Academy of Finland (Profi6 336449 funding program), University of Oulu strategic funding, Finnish Foundation for Cardiovascular Research, Wellbeing Services County of North Ostrobothnia (VTR project K62716), and Terttu foundation. S.A.T. acknowledges the support of Canon Medical and the Royal Academy of Engineering and the Research Chairs and Senior Research Fellowships scheme (grant RCSRF1819\8\25).

We would like to thank Peter Bankhead, Gary S. Collins, Robert Haase, Fred Hamprecht, Alan Karthikesalingam, Hannes Kennigott, Peter Mattson, David Moher, Bram Stieltjes, and Manuel Wiesenfarth for the fruitful discussions on this work.

We would like to thank Sandy Engelhardt, Sven Koehler, M. Alican Noyan, Gorkem Polat, Hassan Rivaz, Julian Schroeter, Anindo Saha, Lalith Sharan, Peter Hirsch, and Matheus Viana for suggesting additional illustrations that can be found in [27].

COMPETING INTERESTS

The authors declare the following competing interests: F.B. is an employee of Siemens AG (Munich, Germany). B.v.G. is a shareholder of Thirona (Nijmegen, NL). B.G. is an employee of HeartFlow Inc (California, USA) and Kheiron Medical Technologies Ltd (London, UK). M.M.H. received an Nvidia GPU Grant. Th. K. is an employee of Lunit (Seoul, South Korea). G.L. is on the advisory board of Canon Healthcare IT (Minnetonka, USA) and is a shareholder of Aiosyn BV (Nijmegen, NL). Na.R. is the founder and CSO of Histofy (New York, USA). Ni.R. is an employee of Nvidia GmbH (Munich, Germany). J.S.-R. reports funding from GSK (Heidelberg, Germany), Pfizer (New York, USA) and Sanofi (Paris, France) and fees from Travers Therapeutics (California, USA), Stadapharm (Bad Vilbel, Germany), Astex Therapeutics (Cambridge, UK), Pfizer (New York, USA), and Grunenthal (Aachen, Germany). R.M.S. receives patent royalties from iCAD (New Hampshire, USA), ScanMed (Nebraska, USA), Philips (Amsterdam, NL), Translation Holdings (Alabama, USA) and PingAn (Shenzhen, China); his lab received

research support from PingAn through a Cooperative Research and Development Agreement. S.A.T. receives financial support from Canon Medical Research Europe (Edinburgh, Scotland). The remaining authors declare no competing interests.

DATA AVAILABILITY STATEMENT

No data was used in this study.

REFERENCES

- [1]. Bilic Patrick, Christ Patrick, Li Hongwei Bran, Vorontsov Eugene, Ben-Cohen Avi, Kaissis Georgios, Szeskin Adi, Jacobs Colin, Mamani Gabriel Efrain Humpire, Chartrand Gabriel, et al. The liver tumor segmentation benchmark (lits). *Medical Image Analysis*, 84:102680, 2023. [PubMed: 36481607]
- [2]. Brown Bernice B. Delphi process: a methodology used for the elicitation of opinions of experts. Technical report, Rand Corp Santa Monica CA, 1968.
- [3]. Carbonell Alberto, De la Pena Marcos, Flores Ricardo, and Gago Selma. Effects of the trinucleotide preceding the self-cleavage site on eggplant latent viroid hammerheads: differences in co-and post-transcriptional self-cleavage may explain the lack of trinucleotide auc in most natural hammerheads. *Nucleic acids research*, 34(19):5613–5622, 2006. [PubMed: 17028097]
- [4]. Chen Jianxu, Ding Liya, Viana Matheus P, Lee HyeonWoo, Sluezwski M Filip, Morris Benjamin, Hendershott Melissa C, Yang Ruian, Mueller Irina A, and Rafelski Susanne M. The allen cell and structure segmenter: a new open source toolkit for segmenting 3d intracellular structures in fluorescence microscopy images. *BioRxiv*, page 491035, 2020.
- [5]. Chicco Davide and Jurman Giuseppe. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21(1):1–13, 2020.
- [6]. Chicco Davide, Tötsch Niklas, and Jurman Giuseppe. The matthews correlation coefficient (mcc) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData mining*, 14(1):1–22, 2021. The manuscript addresses the challenge of evaluating binary classifications. It compares MCC to other metrics, explaining their mathematical relationships and providing use cases where MCC offers more informative results. [PubMed: 33430939]
- [7]. Cordts Marius, Omran Mohamed, Ramos Sebastian, Scharwächter Timo, Enzweiler Markus, Benenson Rodrigo, Franke Uwe, Roth Stefan, and Schiele Bernt. The cityscapes dataset. In *CVPR Workshop on The Future of Datasets in Vision*, 2015.
- [8]. Correia Paulo and Pereira Fernando. Video object relevance metrics for overall segmentation quality evaluation. *EURASIP Journal on Advances in Signal Processing*, 2006:1–11, 2006.
- [9]. Sabatino Antonio Di and Corazza Gino Roberto. Nonceliac gluten sensitivity: sense or sensibility?, 2012.
- [10]. Everingham Mark, Luc Van Gool, Williams Christopher KI, Winn John, and Zisserman Andrew. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.
- [11]. Gooding Mark J, Smith Annamarie J, Tariq Maira, Aljabar Paul, Peressutti Devis, van der Stoep Judith, Reymen Bart, Emans Daisy, Hattu Djoya, van Loon Judith, et al. Comparative evaluation of autocontouring in clinical practice: a practical method using the turing test. *Medical physics*, 45(11):5105–5115, 2018. [PubMed: 30229951]
- [12]. Gooding Mark J, Boukerroui Djamel, Osorio Eliana Vasquez, Monshouwer René, and Brunenberg Ellen. Multicenter comparison of measures for quantitative evaluation of contouring in radiotherapy. *Physics and Imaging in Radiation Oncology*, 24:152–158, 2022. [PubMed: 36424980]
- [13]. Grandini Margherita, Bagli Enrico, and Visani Giorgio. Metrics for multi-class classification: an overview. *arXiv preprint arXiv:2008.05756*, 2020.

- [14]. Gruber Sebastian and Buettner Florian. Trustworthy deep learning via proper calibration errors: A unifying approach for quantifying the reliability of predictive uncertainty. arXiv preprint arXiv:2203.07835, 2022.
- [15]. Honauer Katrin, Maier-Hein Lena, and Kondermann Daniel. The hci stereo metrics: Geometry-aware performance analysis of stereo algorithms. In Proceedings of the IEEE International Conference on Computer Vision, pages 2120–2128, 2015.
- [16]. Kaggle. Satorius Cell Instance Segmentation 2021. <https://www.kaggle.com/c/sartorius-cell-instance-segmentation>, 2021. [Online; accessed 25-April-2022].
- [17]. Kofler Florian, Ezhov Ivan, Isensee Fabian, Berger Christoph, Korner Maximilian, Paetzold Johannes, Li Hongwei, Shit Suprosanna, McKinley Richard, Bakas Spyridon, et al. 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, 2021.
- [18]. Konukoglu Ender, Glocker Ben, Ye Dong Hye, Criminisi Antonio, and Pohl Kilian M. Discriminative segmentation-based evaluation through shape dissimilarity. IEEE transactions on medical imaging, 31(12):2278–2289, 2012. [PubMed: 22955890]
- [19]. Lennerz Jochen K, Green Ursula, Williamson Drew FK, and Mahmood Faisal. A unifying force for the realization of medical ai. npj Digital Medicine, 5(1):1–3, 2022. [PubMed: 35013539]
- [20]. Lin Tsung-Yi, Maire Michael, Belongie Serge, Hays James, Perona Pietro, Ramanan Deva, Dollár Piotr, and Zitnick C Lawrence. Microsoft coco: Common objects in context. In European conference on computer vision, pages 740–755. Springer, 2014.
- [21]. Maier-Hein Lena, Eisenmann Matthias, Reinke Annika, Onogur Sinan, Stankovic Marko, Scholz Patrick, Arbel Tal, Bogunovic Hrvoje, Bradley Andrew P, Carass Aaron, et al. Why rankings of biomedical image analysis competitions should be interpreted with care. Nature communications, 9(1):1–13, 2018. With this comprehensive analysis of biomedical image analysis competitions (challenges), the authors initiated a shift in how such challenges are designed, performed, and reported in the biomedical domain. Its concepts and guidelines have been adopted by reputed organizations such as MICCAI.
- [22]. Maier-Hein Lena, Reinke Annika, Christodoulou Evangelia, Glocker Ben, Godau Patrick, Isensee Fabian, Kleesiek Jens, Kozubek Michal, Reyes Mauricio, Riegler Michael A, et al. Metrics reloaded: Pitfalls and recommendations for image analysis validation. arXiv preprint arXiv:2206.01653, 2022.
- [23]. Margolin Ran, Zelnik-Manor Lihi, and Tal Ayellet. How to evaluate foreground maps? In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 248–255, 2014.
- [24]. Muschelli John. Roc and auc with a binary predictor: a potentially misleading metric. Journal of classification, 37(3): 696–708, 2020. [PubMed: 33250548]
- [25]. Nasa Prashant, Jain Ravi, and Juneja Deven. Delphi methodology in healthcare research: how to decide its appropriateness. World Journal of Methodology, 11(4):116, 2021. [PubMed: 34322364]
- [26]. Ounkomol Chawin, Seshamani Sharmishta, Maleckar Mary M, Collman Forrest, and Johnson Gregory R. Label-free prediction of three-dimensional fluorescence images from transmitted-light microscopy. Nature methods, 15(11): 917–920, 2018. [PubMed: 30224672]
- [27]. Reinke Annika, Eisenmann Matthias, Tizabi Minu D, Sudre Carole H, Rädtsch Tim, Antonelli Michela, Arbel Tal, Bakas Spyridon, Cardoso M Jorge, Cheplygina Veronika, Farahani Keyvan, Glocker Ben, Heckmann-Nötzel Doreen, Isensee Fabian, Jannin Pierre, Kahn Charles, Kleesiek Jens, Kurc Tahsin, Kozubek Michal, Landman Bennett A, Litjens Geert, Maier-Hein Klaus, Martel Anne L, Müller Henning, Petersen Jens, Reyes Mauricio, Rieke Nicola, Stieltjes Bram, Summers Ronald M, Tsaftaris Sotirios A, van Ginneken Bram, Kopp-Schneider Annette, Jäger Paul, and Maier-Hein Lena. Common limitations of image processing metrics: A picture story. arXiv preprint arXiv:2104.05642, 2021.
- [28]. Reinke Annika, Eisenmann Matthias, Tizabi Minu D, Sudre Carole H, Rädtsch Tim, Antonelli Michela, Arbel Tal, Bakas Spyridon, Cardoso M Jorge, Cheplygina Veronika, et al. Common limitations of image processing metrics: A picture story. arXiv preprint arXiv:2104.05642, 2021.
- [29]. Roberts Brock, Haupt Amanda, Tucker Andrew, Grancharova Tanya, Arakaki Joy, Fuqua Margaret A, Nelson Angelique, Hookway Caroline, Ludmann Susan A, Mueller Irina A, et al.

- Systematic gene tagging using crispr/cas9 in human stem cells to illuminate cell organization. *Molecular biology of the cell*, 28(21):2854–2874, 2017. [PubMed: 28814507]
- [30]. Schmidt Uwe, Weigert Martin, Broaddus Coleman, and Myers Gene. Cell detection with star-convex polygons. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 265–273. Springer, 2018.
- [31]. Stringer Carsen, Wang Tim, Michaelos Michalis, and Pachitariu Marius. Cellpose: a generalist algorithm for cellular segmentation. *Nature methods*, 18(1):100–106, 2021. [PubMed: 33318659]
- [32]. Taha Abdel Aziz and Hanbury Allan. Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. *BMC medical imaging*, 15(1):1–28, 2015. The paper discusses the importance of effective metrics for evaluating the accuracy of 3D medical image segmentation algorithms. The authors analyze existing metrics, propose a selection methodology, and develop a tool to aid researchers in choosing appropriate evaluation metrics based on the specific characteristics of the segmentation task. [PubMed: 25645550]
- [33]. Taha Abdel Aziz, Hanbury Allan, and Jimenez del Toro Oscar A. A formal method for selecting evaluation metrics for image segmentation. In *2014 IEEE international conference on image processing (ICIP)*, pages 932–936. IEEE, 2014.
- [34]. Tran Thuy Nuong, Adler Tim, Yamahi Amine, Christodoulou Evangelia, Godau Patrick, Reinke Annika, Tizabi Minu Dietlinde, Sauer Peter, Persicke Tillmann, Albert Jörg Gerhard, et al. Sources of performance variability in deep learning- based polyp detection. *arXiv preprint arXiv:2211.09708*, 2022.
- [35]. Vaassen Femke, Hazelaar Colien, Vaniqui Ana, Gooding Mark, Brent van der Heyden, Richard Canters, and Wouter van Elmpt. Evaluation of measures for assessing time-saving of automatic organ-at-risk segmentation in radiotherapy. *Physics and Imaging in Radiation Oncology*, 13:1–6, 2020. [PubMed: 33458300]
- [36]. Viana Matheus P, Chen Jianxu, Knijnenburg Theo A, Vasani Ritvik, Yan Calysta, Arakaki Joy E, Bailey Matte, Berry Ben, Borensztein Antoine, Brown Eva M, et al. Integrated intracellular organization and its variations in human ips cells. *Nature*, pages 1–10, 2023.
- [37]. Wiesenfarth Manuel, Reinke Annika, Landman Bennett A, Eisenmann Matthias, Saiz Laura Aguilera, Cardoso M Jorge, Maier-Hein Lena, and Kopp-Schneider Annette. Methods and open-source toolkit for analyzing and visualizing challenge results. *Scientific Reports*, 11(1):1–15, 2021. [PubMed: 33414495]
- [38]. Yeghiazaryan Varduhi and Voiculescu Irina D. Family of boundary overlap metrics for the evaluation of medical image segmentation. *Journal of Medical Imaging*, 5(1):015006, 2018. [PubMed: 29487883]
- [39]. Hirling Dominik, Tasnadi Ervin, Caicedo Juan, Caroprese Maria V, Sjögren Rickard, Aubreville Marc, Koos Krisztian, and Horvath Peter. Segmentation metric misinterpretations in bioimage analysis. *Nature methods*, pages 1–4, 2023. [PubMed: 36635552]

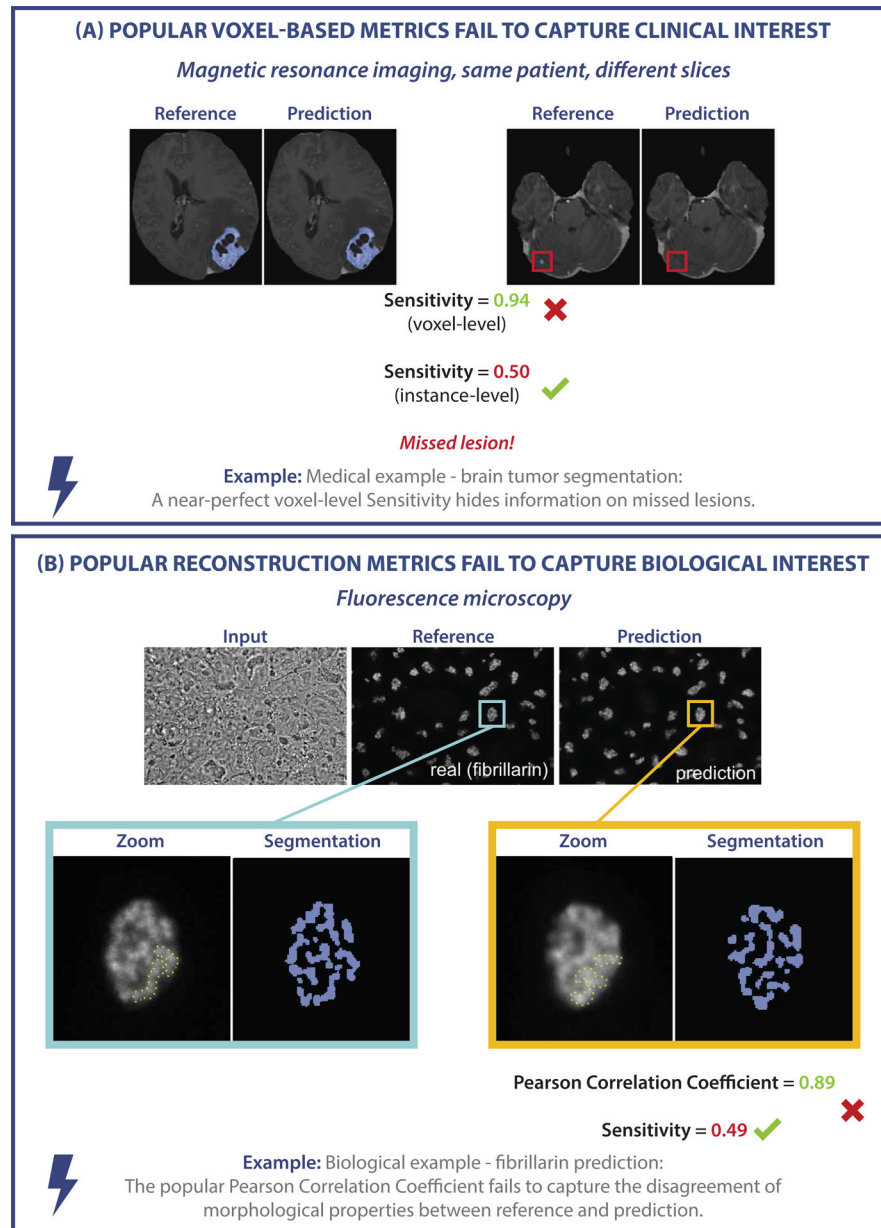


Figure 1: Examples of metric-related pitfalls in image analysis validation. (A) Medical image analysis example: Voxel-based metrics are not appropriate for detection problems. Measuring the voxel-level performance of a prediction yields a near-perfect Sensitivity. However, the Sensitivity at the instance level reveals that lesions are actually missed by the algorithm. (B) Biological image analysis example: The task of predicting fibrillar in the dense fibrillary component of the nucleolus should be phrased as a segmentation task, for which segmentation metrics reveal the low quality of the prediction. Phrasing the task as image reconstruction instead and validating it using metrics such as the Pearson Correlation Coefficient yields misleadingly high metric scores [4, 26, 29, 36, 36].

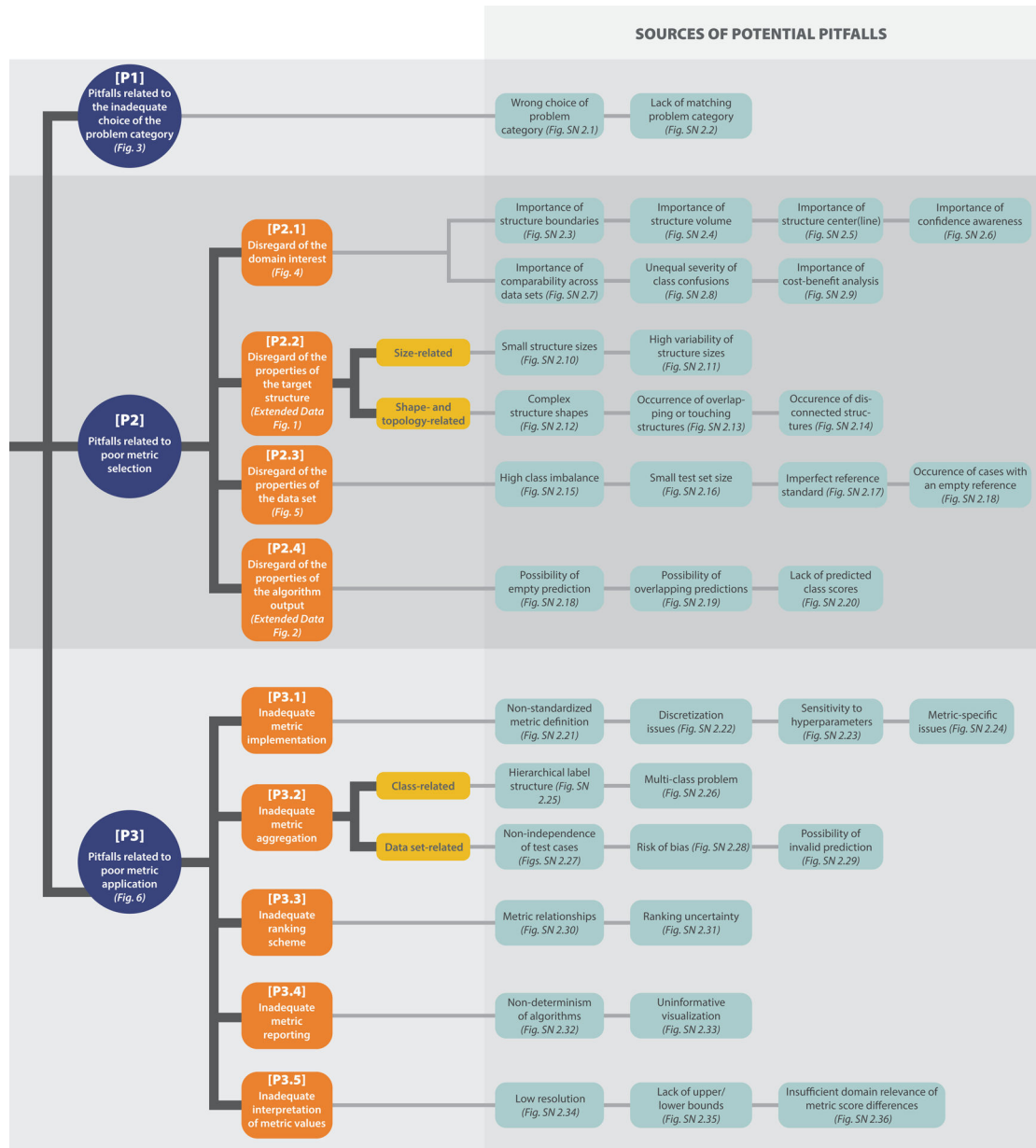
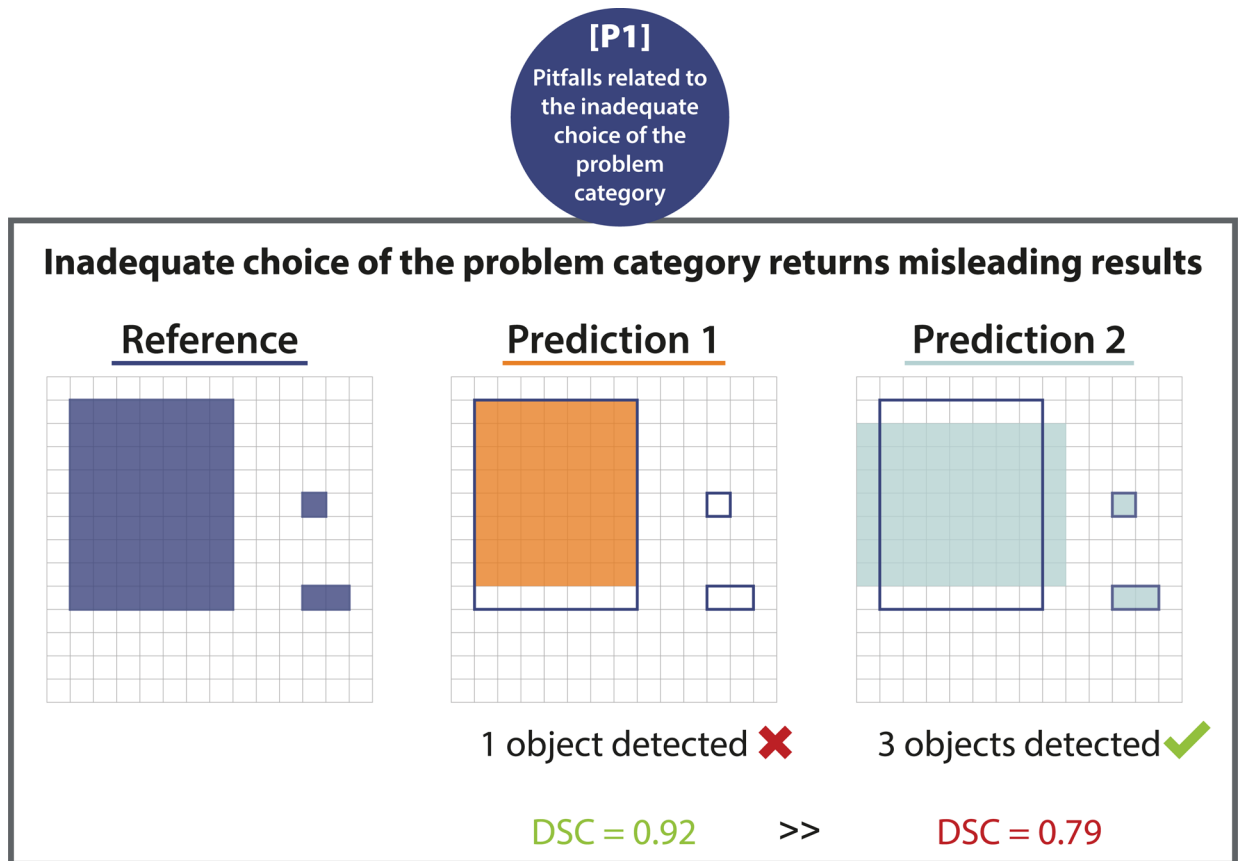


Figure 2: Overview of the taxonomy for metric-related pitfalls. Pitfalls can be grouped into three main categories: [P1] Pitfalls related to the inadequate choice of the problem category, [P2] pitfalls related to poor metric selection, and [P3] pitfalls related to poor metric application. [P2] and [P3] are further split into subcategories. For all categories, pitfall sources are presented (green), with references to corresponding illustrations of representative examples. Note that the order in which the pitfall sources are presented does not correlate with importance.

**Figure 3:**

[P1] Pitfalls related to the inadequate choice of the problem category. **Wrong choice of problem category.** Effect of using segmentation metrics for object detection problems. The pixel-level Dice Similarity Coefficient (DSC) of a prediction recognizing every structure (*Prediction 2*) is lower than that of a prediction that only recognizes one of the three structures (*Prediction 1*).

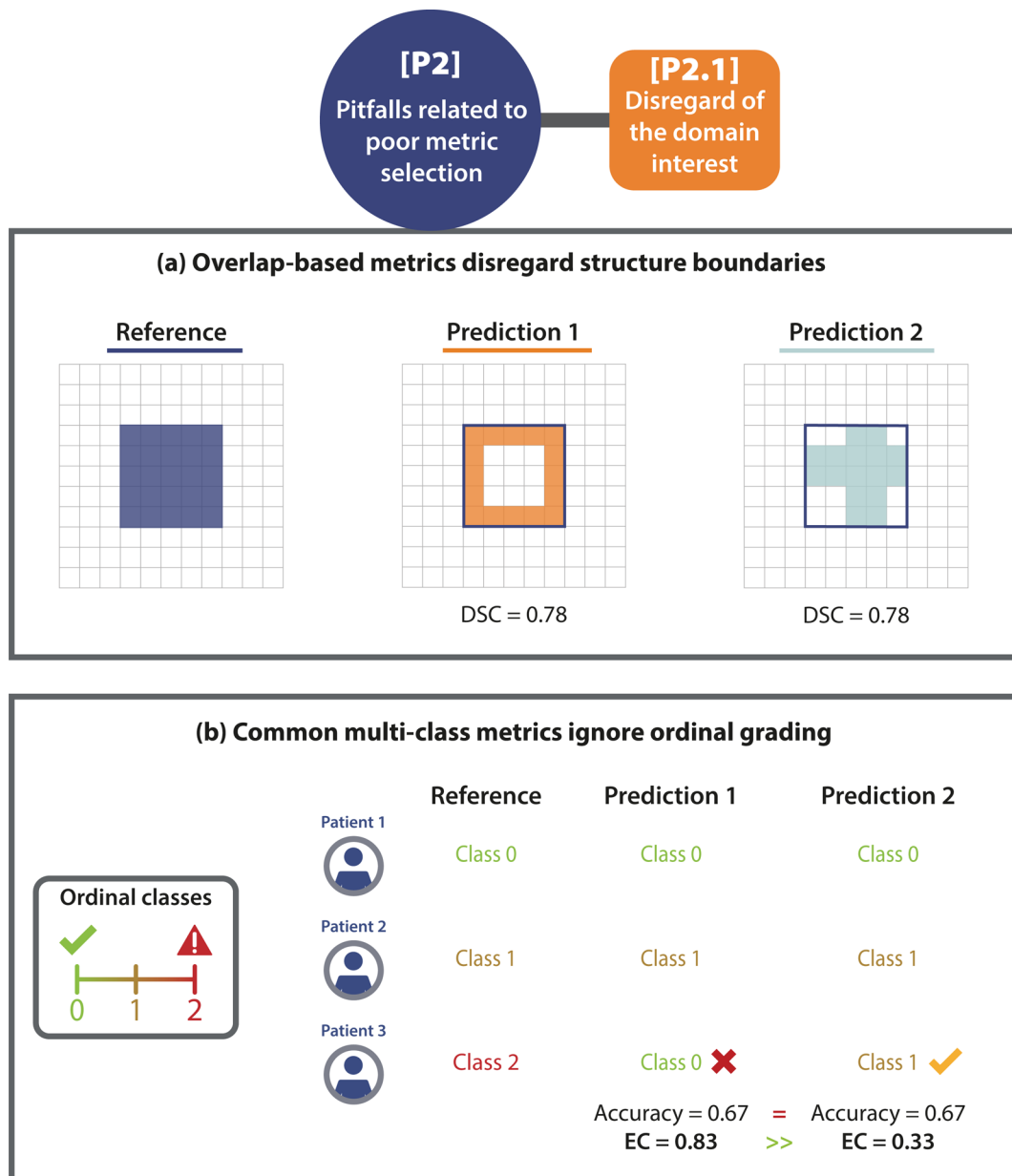


Figure 4: [P2.1] Disregard of the domain interest.

(a) Importance of structure boundaries. The predictions of two algorithms (*Prediction 1/2*) capture the boundary of the given structure substantially differently, but lead to the exact same Dice Similarity Coefficient (DSC), due to its boundary un-awareness. This pitfall is also relevant for other overlap-based metrics such as centerline Dice Similarity Coefficient (clDice), pixel-level F_β Score, and Intersection over Union (IoU), as well as localization criteria such as Box/Approx/Mask IoU, Center Distance, Mask IoU > 0, Point inside Mask/Box/Approx, and Intersection over Reference (IoR). **(b) Unequal severity of class confusions.** When predicting the severity of a disease for three patients in an ordinal classification problem, *Prediction 1* assumes a much lower severity for *Patient 3* than actually observed. This critical issue is overlooked by common metrics (here: Accuracy),

which measure no difference to *Prediction 2*, which assesses the severity much better. Metrics with pre-defined weights (here: Expected Cost (EC)) correctly penalize *Prediction 1* much more than *Prediction 2*. This pitfall is also relevant for other counting metrics, such as Balanced Accuracy (BA), F_β Score, Positive Likelihood Ratio (LR+), Matthews Correlation Coefficient (MCC), Net Benefit (NB), Negative Predictive Value (NPV), Positive Predictive Value (PPV), Sensitivity, and Specificity.

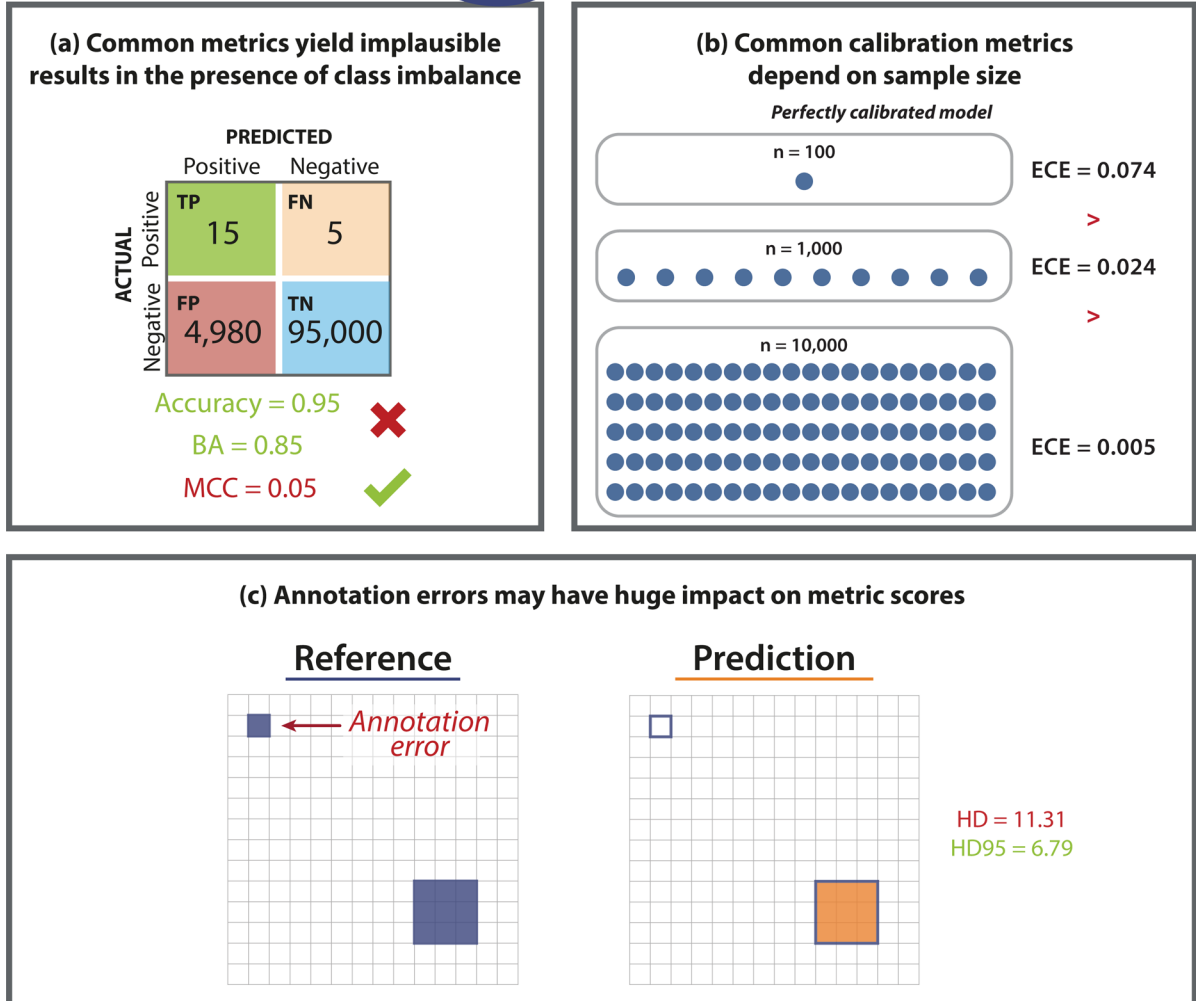
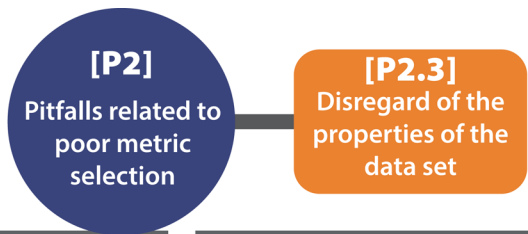


Figure 5: [P2.3] Disregard of the properties of the data set.

(a) High class imbalance. In the case of underrepresented classes, common metrics may yield misleading values. In the given example, Accuracy and Balanced Accuracy (BA) have a high score despite the high amount of False Positive (FP) samples. The class imbalance is only uncovered by metrics considering predictive values (here: Matthews Correlation Coefficient (MCC)). This pitfall is also relevant for other counting and multi-threshold metrics such as Area under the Receiver Operating Characteristic Curve (AUROC), Expected Cost (EC) (depending on the chosen costs), Positive Likelihood Ratio (LR+), Net Benefit (NB), Sensitivity, Specificity, and Weighted Cohen’s Kappa (WCK). **(b) Small test set size.** The values of the Expected Calibration Error (ECE) depend on the sample size. Even for a simulated perfectly calibrated model, the ECE will be substantially greater than

zero for small sample sizes [14]. (e) **Imperfect reference standard.** A single erroneously annotated pixel may lead to a large decrease in performance, especially in the case of the Hausdorff Distance (HD) when applied to small structures. The Hausdorff Distance 95th Percentile (HD95), on the other hand, was designed to deal with spatial outliers. This pitfall is also relevant for localization criteria such as Box/Approx Intersection over Union (IoU) and Point inside Box/Approx. Further abbreviations: True Positive (TP), False Negative (FN), True Negative (TN).

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

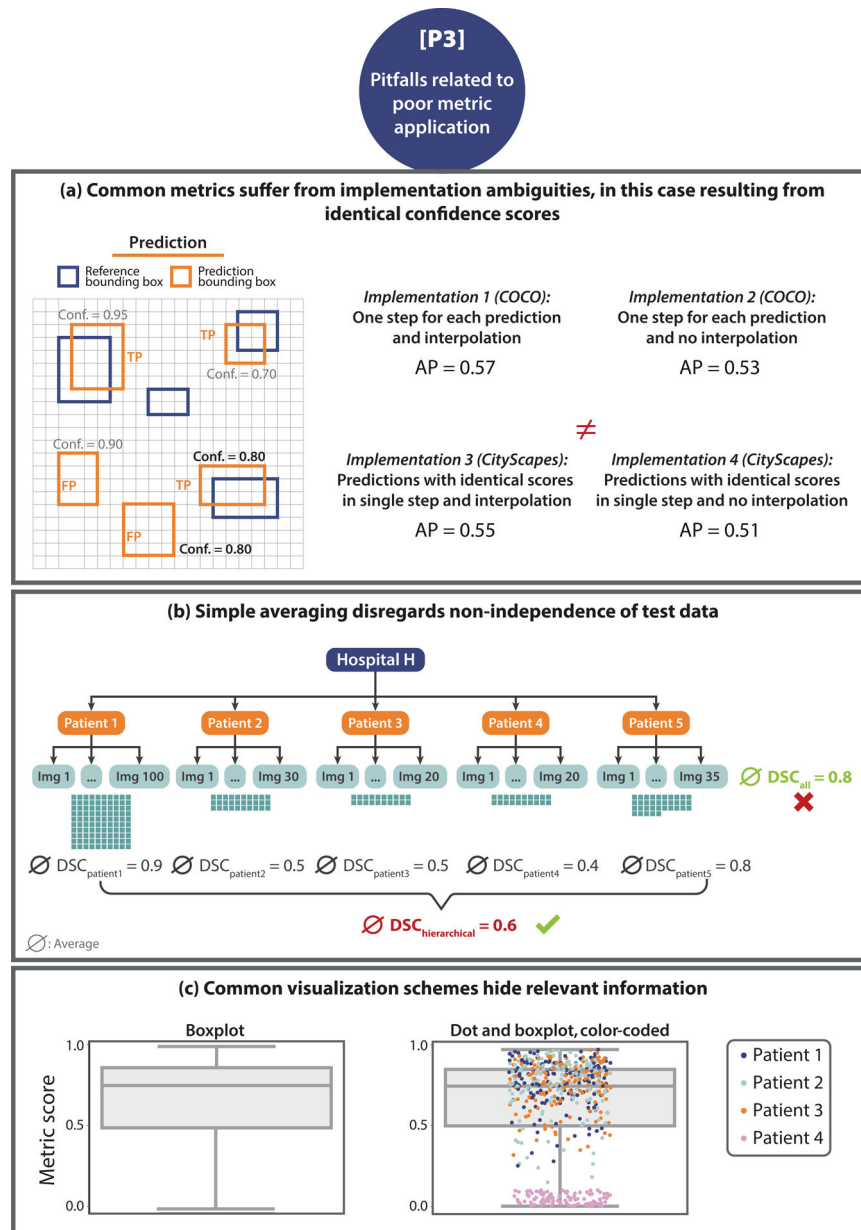


Figure 6: [P3] Pitfalls related to poor metric application.

(a) Non-standardized metric implementation. In the case of the Average Precision (AP) metric and the construction of the Precision- Recall (PR)-curve, the strategy of how identical scores (here: confidence score of 0.80 is present twice) are treated has a substantial impact on the metric scores. Microsoft Common Objects in Context (COCO) [20] and CityScapes [7] are used as examples. **(b) Non-independence of test cases.** The number of images taken from *Patient 1* is much higher compared to that acquired from *Patients 2–5*. Averaging over all Dice Similarity Coefficient (DSC) values, denoted by \emptyset , results in a high averaged score. Aggregating metric values per patient reveals much higher scores for *Patient 1* compared to the others, which would have been hidden by simple aggregation. **(c) Uninformative visualization.** A single box plot (left) does not give sufficient information about the raw

metric value distribution. Adding the raw metric values as jittered dots on top (right) adds important information (here: on clusters). In the case of non-independent validation data, color/shape-coding helps reveal data clusters.