

Understanding metric-related pitfalls in image analysis validation

Received: 9 February 2023

Accepted: 12 December 2023

Published online: 12 February 2024

 Check for updates

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

Validation metrics are key for tracking scientific progress and bridging the current chasm between artificial intelligence research and its translation into practice. However, increasing evidence shows that, particularly in image analysis, metrics are often chosen inadequately. Although 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 multistage Delphi process conducted by a multidisciplinary expert consortium as well as extensive community feedback, the present work provides a reliable and comprehensive common point of access to information on pitfalls related to validation metrics in image analysis. Although 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 domains such as athletics, this process is straightforward because the performance measurements (for example, the time it takes an athlete to run a given distance) exactly reflect the underlying interest (for example, 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 applicability, particularly in clinical settings. If the performance of an image analysis algorithm is not measured using relevant validation metrics, no reliable statement can be made about how suitable the algorithm is for solving the proposed task, and the algorithm is unlikely to ever be used in the real world. 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 key for both measuring and informing scientific progress, as well as bridging the current chasm between image analysis research and its translation into practice.

In image analysis, for some applications, it might be sufficient to draw a box around the structure of interest (for example, detecting individual mitotic cells or regions with apoptotic cell debris) and, optionally, to associate that region with a classification (for example, distinguishing a mitotic versus an interphase cell); however, other applications (for example, cell tracing for fluorescent signal quantification) could require the exact structure boundaries to be determined. The suitability of any validation metric thus hinges on the properties of the driving problem. As a result, several metrics have so far been proposed in the image-processing field. In our previous work, we analyzed all biomedical-image-analysis competitions conducted within a period of about 15 years¹. We found 97 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 a given research problem. The vast range of options makes tracking all related information impossible for any individual, and consequently renders the process of metric selection prone to error. Thus, it is not surprising that researchers often rely on flawed validation practices that have been traditionally used in the literature. To make matters worse, there is currently no comprehensive resource providing an overview of the relevant definitions, (mathematical) properties,

✉ e-mail: a.reinke@dkfz-heidelberg.de; m.tizabi@dkfz-heidelberg.de; p.jaeager@dkfz-heidelberg.de; l.maier-hein@dkfz-heidelberg.de

a Popular voxel-based metrics fail to capture clinical interest

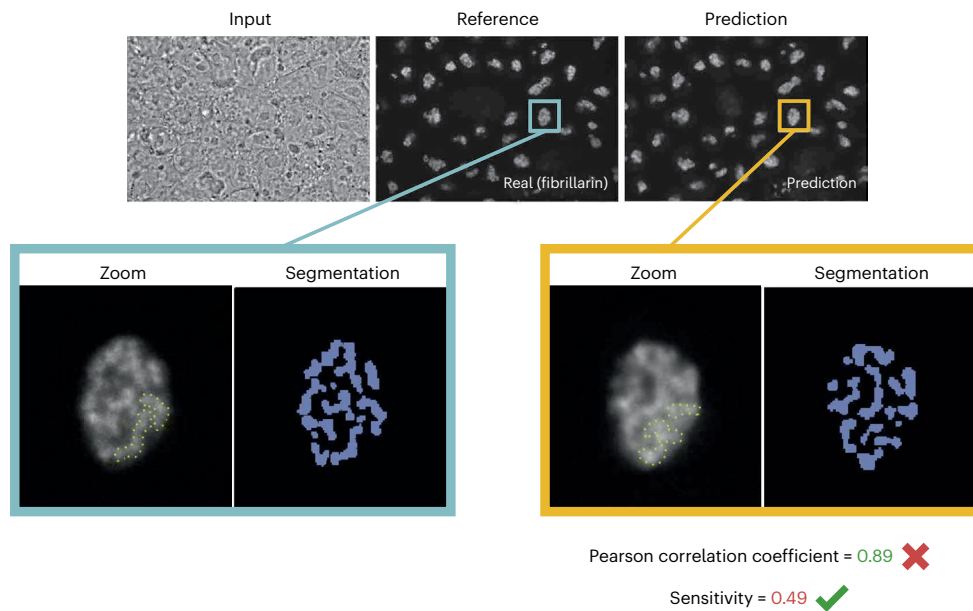
Magnetic resonance imaging, same patient, different slices



Medical example: brain-tumor segmentation
A near-perfect voxel-level sensitivity hides information on missed lesions

b Popular reconstruction metrics fail to capture biological interest

Fluorescence microscopy



Biological example: fibrillar prediction
The popular Pearson correlation coefficient fails to capture the disagreement of morphological properties between reference and prediction

Fig. 1 | Examples of metric-related pitfalls in image analysis validation.

a, An example of medical image analysis. 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. Green metric values correspond to a good metric value, whereas red values correspond to a poor value. Green check marks indicate desirable behavior of metrics; red crosses

indicate undesirable behavior. **b**, An example of biological image analysis. 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^{12,35–38}.

limitations and pitfalls for a metric of interest. Although taking into account the individual properties and limitations of metrics is imperative for choosing suitable validation metrics, the required knowledge is largely inaccessible.

As a result, numerous flaws and pitfalls are prevalent in image analysis validation, and researchers are often unaware of them owing to a lack of knowledge of intricate metric properties and limitations. Accordingly, increasing evidence shows that metrics are often selected inadequately in image analysis (for example, refs. 2–4). In the absence of a central information resource, it is common for researchers to resort to popular validation metrics; however, these can be entirely unsuitable, for instance owing to a mismatch of the metric's inherent mathematical properties with the underlying research question and specifications of the dataset at hand (see Fig. 1).

The present work addresses this important roadblock in image analysis research with a crowdsourcing-based approach involving both a Delphi process undertaken by a multidisciplinary consortium

of experts and a social-media campaign. It represents a comprehensive collection, visualization and detailed discussion of pitfalls, drawbacks and limitations regarding validation metrics that are 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, that is, image analysis problems that can be considered classification tasks at either the image, object or pixel level. Specifically, these encompass four problem categories: image-level classification, semantic segmentation, object detection and instance segmentation. Our contribution includes a dedicated profile for each metric (Supplementary 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 Tables 1–5) and enables researchers to quickly grasp whether a certain metric comes with pitfalls in a given use case.

Although our work grew out of image analysis research and practice in the field of biomedicine, which is highly complex and particularly high stakes owing to its direct impact on patient health, we believe the identified pitfalls are transferable to other application areas of imaging research. It should be noted that this work focuses on identifying, categorizing and illustrating metric pitfalls; the sister publication of this work gives specific recommendations on which metrics to use in certain circumstances⁵.

Information on metric pitfalls is largely inaccessible

Researchers and algorithm developers seeking to validate image analysis algorithms often face the problem of choosing adequate validation metrics while navigating a range of potential pitfalls. Common practice is often not the best option, as demonstrated by several recent publications^{1–4}. Making an informed decision 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.

The lack of accessibility is recognized as a major obstacle in image analysis validation¹. To illustrate this, we searched the literature for information on commonly used validation metrics. The search was conducted on the platform Google Scholar. We used search strings that combined various notations of metric names, including synonyms and acronyms, with terms indicating problems, such as ‘pitfall’ or ‘limitation.’ The mean and median number of hits for the metrics discussed in this work were 159,329 and 22,100, respectively, and ranged from 49 for the centerline dice similarity coefficient (cIDice) to 962,000 for sensitivity. Moreover, despite the valuable literature on individual relevant aspects (for example, refs. 3,4,6–10), we did not find a common point of entry for metric-related pitfalls in image analysis, whether in the form of a review paper or another credible source. We conclude that essential knowledge for making informed decisions and avoiding pitfalls related to the use of validation metrics is highly scattered and cannot be easily accessed by individuals.

Established practices are not always justified

To obtain initial insights 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, the Medical Image Computing and Computer Assisted Interventions Society and the Medical Imaging with Deep Learning foundation. The organizers were asked to provide their rationale for the choice of metrics used in these competitions. We analyzed 138 competitions held between 2018 and 2022 and found that metrics were frequently (in 24% of the competitions) chosen on the basis of common practices. We also found, however, that common practices are often not well-justified, and poor practices may even be passed across generations.

One remarkable example 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)^{11,12}. 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^{12,13}. Despite the popularity of mAP, a widely known challenge on cell instance segmentation (<https://www.kaggle.com/competitions/data-science-bowl-2018/overview/evaluation>) in 2018 introduced a new ‘mean average precision’ term. The task matches the task of the original ‘mean’ AP, object detection, but all terms in the newly proposed metric (mean, average and precision) refer to entirely different concepts. For instance, the common definition of precision from the literature, $TP/(TP + FP)$, was altered to $TP/(TP + FP + FN)$,

where TP, FP and FN refer to the cardinalities of the confusion matrix (that is, the true and false positives (TP and FP) and negatives (TN and FN)). The latter formula actually defines the intersection over union (IoU) metric. Despite this problem, the terminology was adopted by subsequent influential works^{14–17}, indicating its 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 metric definitions and limitations in biomedical image analysis, we formed an international multidisciplinary consortium comprising 62 experts in various fields related to biomedical image analysis. The consortium engaged in a multistage Delphi process^{18,19} to facilitate consensus building. The Delphi process involved multiple surveys, crafted by a coordinating team and completed by the remaining members of the consortium. On the basis of 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 work¹², 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 identified 37 distinct sources of pitfalls (Fig. 2). Notably, these pitfall sources (for example, class imbalances, uncertainties in the reference or poor image resolution) can occur irrespective of imaging modality or application. As a result, many pitfalls can be generalized 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 Supplementary 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 quickly identify which metrics are affected by which types of pitfalls. To achieve this, we developed a comprehensive taxonomy that categorizes the different pitfalls in a semantic manner. The taxonomy was 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 Figure 2, and the relations between the pitfall categories and individual metrics can be found in Extended Data Tables 1–5. We distinguish the following three main categories:

P1: pitfalls related to the inadequate choice of the problem category

A common pitfall occurs when metrics are applied to a problem category that 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 dataset (Fig. 3). In a cancer-monitoring application, they fail to measure instance progress, that is, the potential increase in the number of lesions (Fig. 1), which can have serious consequences. 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 Supplementary Note 2.1.

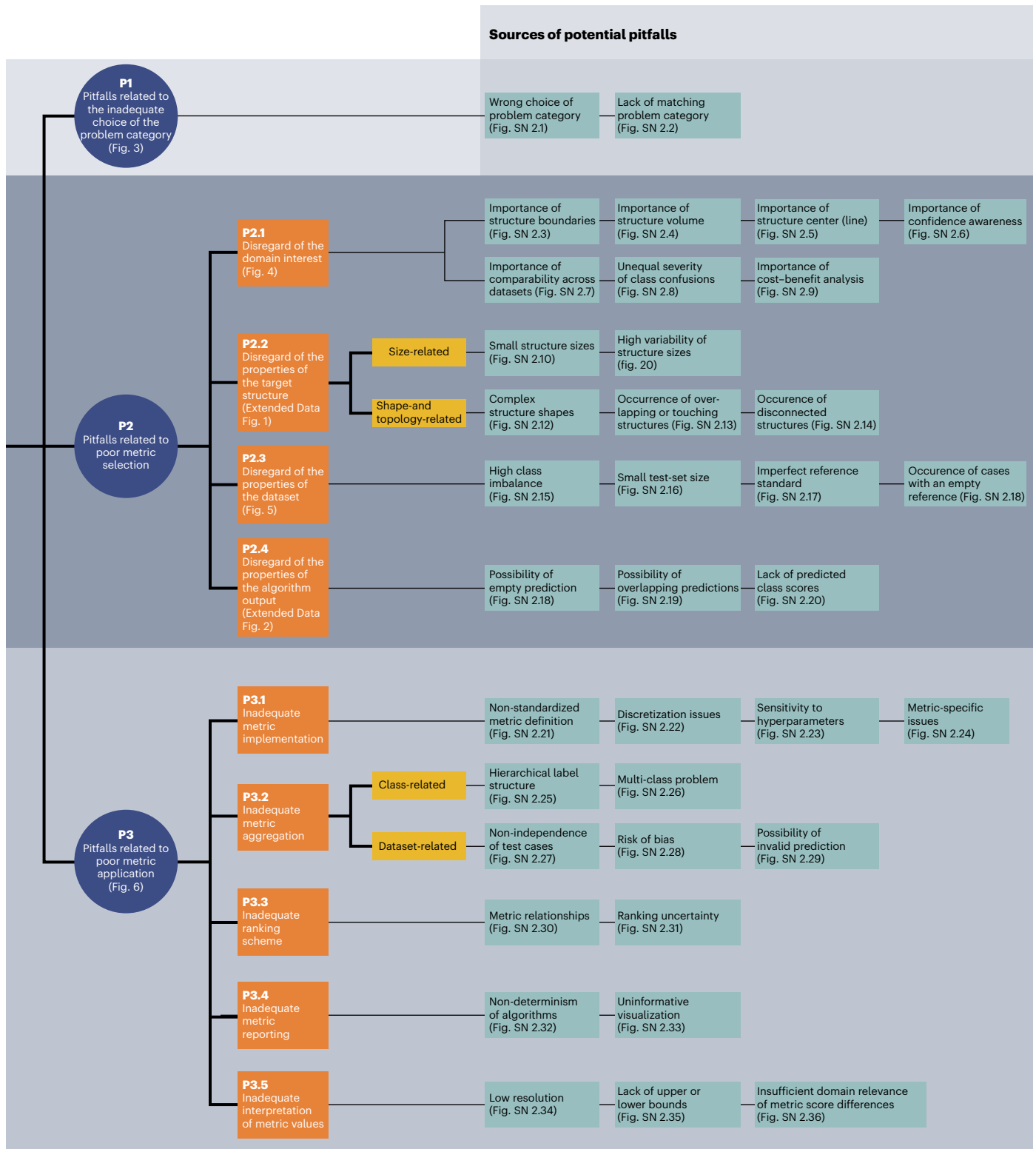


Fig. 2 | Overview of the taxonomy for metric-related pitfalls. Pitfalls can be grouped into three main categories: P1, pitfalls related to the inadequate choice of the problem category; P2, pitfalls related to poor metric selection; and P3, pitfalls related to poor metric application. P2 and P3 are further split into

subcategories. For all categories, pitfall sources are presented (turquoise), with references to corresponding illustrations of representative examples. The order in which the pitfall sources are presented does not correlate with importance.

P2: pitfalls related to poor metric selection

Pitfalls in this category occur when a validation metric is selected without considering specific properties of the given research problem or method that make this metric unsuitable in the particular context. P2 can be further divided into the following four subcategories:

P2.1: disregarding the domain interest. Frequently, several requirements stemming from the domain interest of the underlying research problem 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

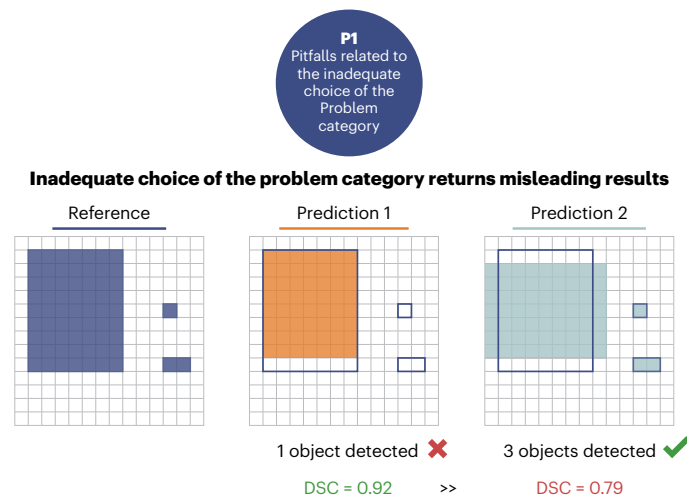


Fig. 3 | P1: pitfalls related to the inadequate choice of the problem category. The effect of using segmentation metrics for object-detection problems. The pixel-level DSC of a prediction recognizing every structure (Prediction 2) is lower than that of a prediction that only recognizes one of the three structures (Prediction 1).

not take the correctness of an object’s boundaries into account, as shown in Figure 4a. 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 could include an unequal severity of class confusions. This could be important in an ordinal grading use case, in which the severity of a disease is categorized by different scores. Predicting a low disease severity for a patient who actually has a severe disease should incur substantial penalties, a requirement not met by common classification metrics. An example is provided in Figure 4b. At the pixel level, this property relates to an unequal severity of over- versus 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 equally²⁰. Further pitfalls may occur if confidence awareness (Fig. SN 2.6), comparability across datasets (Fig. SN 2.7) or a cost–benefit analysis (Fig. SN 2.9) are of particular importance, as illustrated in Supplementary Note 2.2.1.

P2.2: disregard of the properties of the target structures. For problems that require local properties to be captured (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 can hugely impact the metric scores, as shown in Extended Data Figure 1a. Shape- and topology-related pitfalls can arise from the fact that common metrics disregard complex shapes (Extended Data Fig. 1b) or that boundary 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) can also influence metric values. We present further examples of P2.2 pitfalls in Supplementary Note 2.2.2.

P2.3: disregard of the properties of the dataset. Several properties of the dataset, such as class imbalances (Fig. 5a), small sample size (Fig. 5b) or reference-annotation quality, can directly affect metric values. Common metrics such as balanced accuracy (BA), for instance, might yield a very high score for a model that predicts many FP samples in an imbalanced setting (see Fig. 5a). When only small test datasets are used, common calibration metrics (which are typically biased

estimators) either underestimate or overestimate the true calibration error of a model (Fig. 5b)²¹. However, metric values can be impacted by reference annotations (Fig. SN 2.17). Spatial outliers in the reference might have a huge impact on distance-based metrics, such as the Hausdorff distance (HD) (Fig. 5c). Additional pitfalls can arise from the occurrence of cases with an empty reference (Extended Data Fig. 2b), causing division-by-zero errors. We present further examples of P2.3 pitfalls in Supplementary Note 2.2.3.

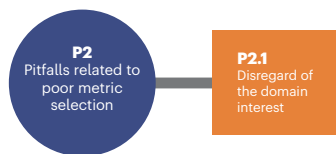
P2.4: disregard of the properties of the algorithm output. Reference-based metrics compare the algorithm output with a reference annotation to compute a metric score. Thus, the content and format of the prediction are of high importance when considering metric choice. Overlapping predictions in segmentation problems, for instance, can return misleading results. In Extended Data Figure 2a, 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 is empty predictions that can cause division-by-zero errors in metric calculations, as illustrated in Extended Data Figure 2b, or the lack of predicted class scores (Fig. SN 2.20). We present further examples of P2.4 pitfalls in Supplementary Note 2.2.3.

P3: pitfalls related to poor metric application

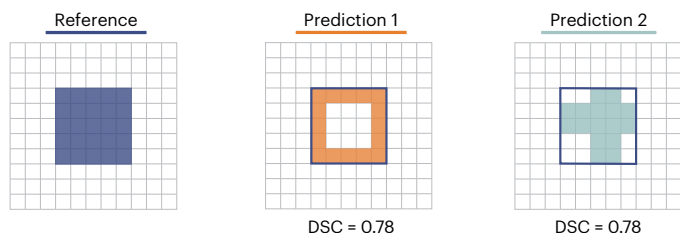
Selected metrics need to be applied to an image or an entire dataset. This step is not straightforward and comes with several pitfalls. For instance, when aggregating metric values over several images or patients, a common mistake is ignoring the hierarchical data structure, such as data from several hospitals or varying numbers of images from each patient. We present three examples of P3 pitfalls in Figure 6; for more pitfalls in this category, please refer to Supplementary Note 2.3. P3 can be further divided into five subcategories that are described in the following paragraphs.

P3.1: inadequate metric implementation. Metric implementation is, unfortunately, not standardized. As shown by ref. 22, different researchers typically employ various implementations for the same metric, which can yield a substantial variation in the metric scores. Although some metrics can be implemented in a straightforward way, others require more advanced techniques and offer different possibilities. In the following, we provide some examples for inadequate metric implementation:

- How identical confidence scores are handled in the computation of the AP metric might lead to substantial differences in metric scores. Microsoft Common Objects in Context (COCO)¹¹, for instance, processes each prediction individually, whereas CityScapes²³ processes all predictions with the same score in one joint step. Figure 6a provides an example in which two predictions have the same confidence score; the final metric scores differ depending on the chosen handling strategy for identical confidence scores. Similar issues can arise with other curve-based metrics, such as area under the receiver operating characteristic curve (AUROC), AP or free-response receiver operating characteristic scores (see, for example, ref. 24).
- Metric implementation might be subject to discretization issues, such as the chosen discretization of continuous variables, which can cause differences in the metric scores, as illustrated in Figure 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 challenge²⁵ and that implemented by Google DeepMind (<https://github.com/deepmind/surface-distance>) could produce different



a Overlap-based metrics disregard structure boundaries



b Common multi-class metrics ignore ordinal grading

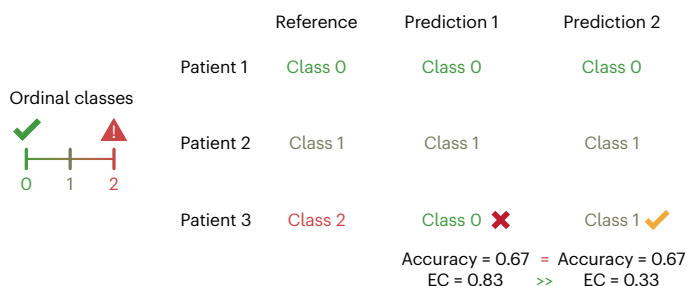


Fig. 4 | P2.1: disregard of the domain interest. **a**, Importance of structure boundaries. The predictions of two algorithms (Predictions 1 and 2) capture the boundary of the given structure in substantially different ways, but lead to the same DSC owing to the metric’s boundary unawareness. This pitfall is also relevant for other overlap-based metrics such as cDice, pixel-level F_{β} score and 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. **b**, Unequal severity of class confusions. When predicting the severity of a disease for three individuals in an ordinal classification problem, Prediction 1 assumes a much lower severity for Patient 3 than is actually observed. This critical issue is overlooked by common metrics (here, accuracy), which measure no difference with respect 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 BA, F_{β} score, positive likelihood ratio (LR+), Matthew’s correlation coefficient (MCC), net benefit (NB), negative predictive value (NPV), positive predictive value (PPV), sensitivity and specificity.

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 can also lead to metric scores that do not reflect the domain interest. For example, the choice of a threshold on a localization criterion (Fig. SN 2.23) or the chosen hyperparameter for the F_{β} score will heavily influence the subsequent metric scores²⁶.

More P3.1 pitfalls can be found in Supplementary 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 dataset and/or all classes. As detailed in Figure 6b and Supplementary Note 2.3.2, important information can be 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 dataset without considering their image or class of origin (<https://torchmetrics.readthedocs.io/en/stable/classification/dice.html?highlight=dice>). 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 might be suppressed by correctly segmented larger structures in other images (see for example, Fig. SN 2.26). An adequate aggregation scheme is also crucial for handling hierarchical class structure (Fig. SN 2.27), missing values (Fig. SN 2.29) and potential biases (Fig. SN 2.28) of the algorithm. Further P3.2 pitfalls are shown in Supplementary Note 2.3.2.

P3.3: inadequate ranking scheme. Rankings are often created to compare algorithm performance. 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 individually (Fig. SN 2.30). In this context, unawareness of metric synonyms can be equally misleading. Metrics can have different names; for instance, sensitivity and recall refer to the same mathematical formula. Despite this seemingly trivial fact, an analysis of 138 biomedical-image-analysis challenges³ found 3 challenges that unknowingly used two versions of the same metric to calculate their rankings. Moreover, rankings themselves can be unstable (Fig. SN 2.31)^{1,27}; rankings are highly sensitive to alterations of the metric aggregation operators, the underlying dataset or the general ranking method. Thus, if the robustness of rankings is disregarded, the winning algorithm might be identified by chance, rather than by true superiority.

P3.4: inadequate metric reporting. A thorough reporting of metric values and aggregates is important in terms of both transparency and interpretability. However, several pitfalls should be avoided. Notably, different types of visualization may vary substantially in terms of interpretability, as shown in Figure 6c. For example, although 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 s.d. or confidence intervals in the presented results, are common. Concrete examples of P3.4 pitfalls can be found in Supplementary 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 performance. However, several pitfalls can arise from interpretation. In rankings, for example, minor differences in metric scores might not be relevant from an application perspective but might 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 Supplementary Note 2.3.4.

An 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 aim of determining how many of the pitfalls that we identified could be located in existing resources. We were able to locate only a portion of the identified pitfalls

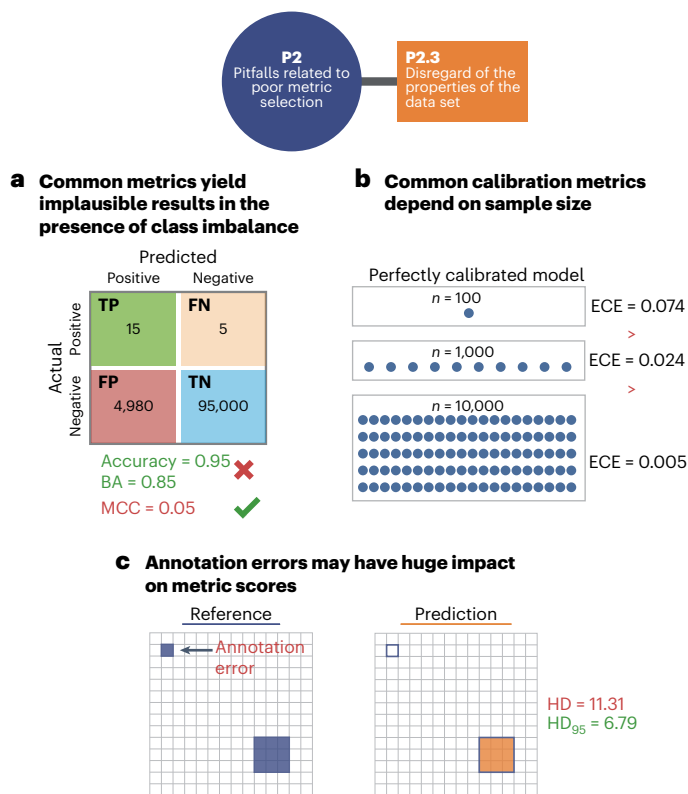


Fig. 5 | P2.3: disregard of the properties of the dataset. **a**, High class imbalance. In the case of underrepresented classes, common metrics may yield misleading values. In the given example, accuracy and BA have a high score despite the large amount of FP samples. The class imbalance is uncovered only by metrics considering predictive values (here, MCC). This pitfall is also relevant for other counting and multi-threshold metrics, such as AUROC, EC (depending on the chosen costs), LR+, NB, sensitivity, specificity and weighted Cohen's kappa. **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 sizes²¹. **c**, Imperfect reference standard. A single erroneously annotated pixel can lead to a large decrease in performance, especially in the case of the Hausdorff distance (HD) when applied to small structures. The HD 95th percentile (HD₉₅), however, was designed to deal with spatial outliers. This pitfall is also relevant for localization criteria such as box/approx IoU and point inside box/approx.

in existing literature (68%) or online resources, such as blog posts (11%; 8% were found in both). Only 27% of the pitfalls that we located were presented visually.

Our work now provides this key resource in a highly structured and easily understandable form. Supplementary 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 is the metric profiles presented in Supplementary Note 2, which, for each metric, summarize the most important information deemed of particular relevance by the Metrics Reloaded consortium that produced this publication's sister work⁵. The profiles provide the reader with a compact, at-a-glance overview of each metric and an explanation of the limitations and pitfalls identified in the Delphi process that we conducted.

Discussion

Flaws in the validation of algorithms for biomedical image analysis significantly impede the translation of methods into (clinical) practice and undermine the assessment of scientific progress in the field²⁸. They are often caused by poor choices in disregarding the specific properties

and limitations of individual validation metrics. This work represents a comprehensive collection of pitfalls and limitations 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 key information—which researchers can use when validating image analysis algorithms—that was previously largely scattered and inaccessible. Our work aims to disrupt the current common practice of choosing metrics on the basis of their popularity rather than their suitability to the underlying research problem. This practice, which, for instance, often manifests in the unreflected and inadequate use of the DSC, is concerningly prevalent, even among prestigious, high-quality biomedical-image-analysis competitions^{1–4,29–32}. The educational aspect of our work is complemented by dedicated 'metric profiles' that detail the definitions and properties of all the discussed metrics. Our work pioneers the examination of artificial intelligence (AI) validation pitfalls in the biomedical domain, in which they are arguably more critical than in many other areas. Flaws in biomedical algorithm validation can directly affect patient well-being and safety.

We propose that shortcomings in current common practice are characterized 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 examined in a reasonable time and with realistic effort, as well as the lack of a common point of entry to reliable metric information. Even when the specific pitfalls and related keywords uncovered by our consortium were known, only a fraction could be found in the literature, indicating the novelty and added value of our work.

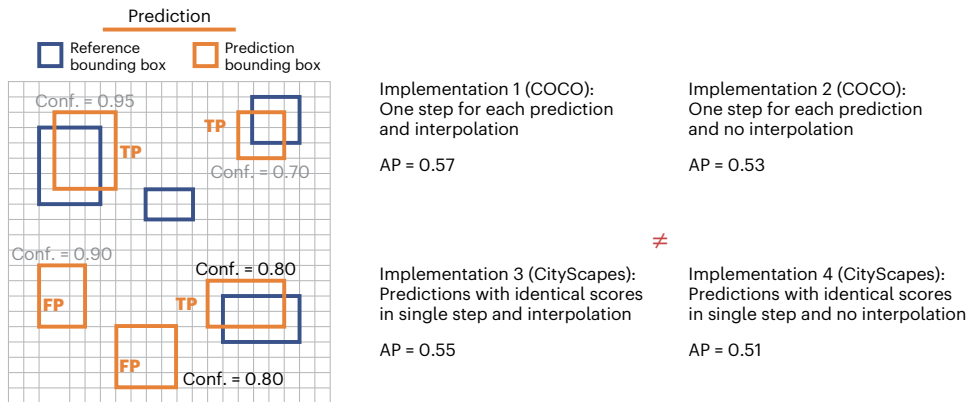
Several constraints regarding our literature search must be noted. First, the remarkably high number of search results inevitably includes duplicates of papers (for example, the same work in a conference paper and on arXiv) as well as results that are out of scope (for example, refs. 33,34); in the cited examples, for instance, this was due to a metric acronym (AUC) also 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. This could, in turn, account for missing publications that discuss pitfalls using 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 (for example, 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 and non-peer-reviewed online resources that mention a particular pitfall, because 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, if not impossible. In many cases, finding information on pitfalls seems feasible only if the specific pitfall and its related keywords are known exactly, which, of course, is usually not the case. Overall accessibility of such essential information, therefore, currently leaves much to be desired.

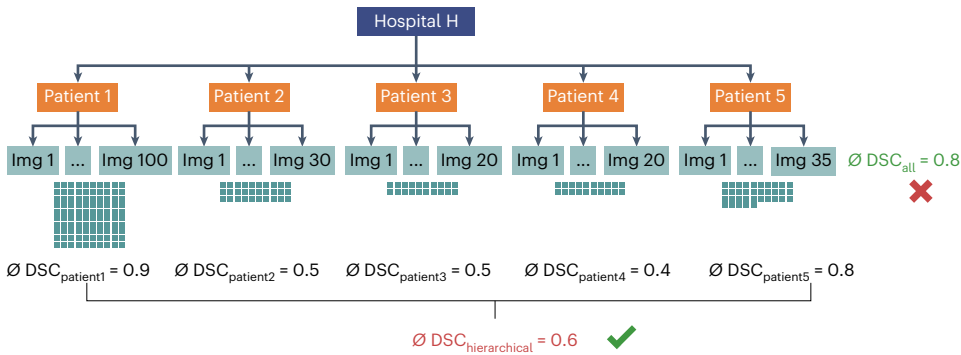
Compiling this information through a multistage Delphi process allowed us to leverage distributed knowledge from experts across different biomedical imaging domains and thus ensure that the resulting

P3
Pitfalls related to poor metric application

a Common metrics suffer from implementation ambiguities, in this case resulting from identical confidence scores



b Simple averaging disregards non-independence of test data



c Common visualization schemes hide relevant information

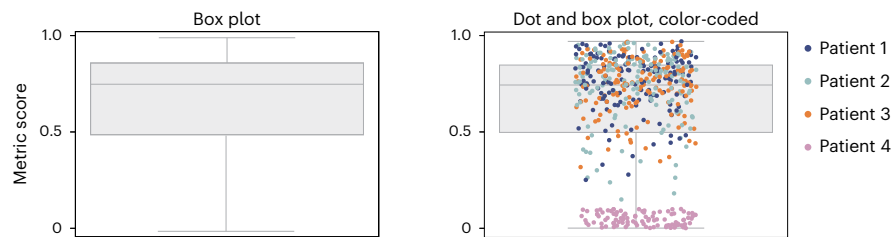


Fig. 6 | P3: pitfalls related to poor metric application. **a**, Non-standardized metric implementation. In the case of the AP metric and the construction of the PR curve, the strategy of how identical scores (here, a confidence (Conf.) score of 0.80 is present twice) are treated has a substantial impact on the metric scores. Microsoft COCO¹¹ and CityScapes²³ are used as examples. **b**, Non-independence of test cases. The number of images taken from Patient 1 is much higher compared than the numbers taken from Patients 2–5. Averaging over

all DSC values (∅) results in a high averaged score. Aggregating metric values per patient reveals much higher scores for Patient 1 than for 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- or shape-coding helps reveal data clusters.

illustrated collection of metric pitfalls and limitations is both comprehensive and of maximum practical relevance. We maintained a continuing connection of our work to practical applications by sharing the first results of this process as a dynamic preprint¹² with dedicated calls for feedback, as well as by 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 that 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 specific tasks and is introduced in a sister publication to this work⁵, 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 described in this work. Notably, 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 machine-learning-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 examined only pitfalls related to the tasks of image-level classification, semantic segmentation, instance segmentation and object detection, which can all be considered classification tasks at different levels (image, object or pixel) and hence share similarities in their validation. Although 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 that future work will address this. Moreover, our work focused on pitfalls related to reference-based metrics. Exploration of pitfalls pertaining to non-reference-based metrics, such as metrics that assess speed, memory consumption or carbon footprint, could be a future direction. Finally, although we aspired to be as comprehensive as possible in our compilation, there could be more pitfalls to take into account that the consortium and the 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 be updated continuously after release, will allow us to easily and transparently append missed pitfalls. This way, our work will remain a reliable point of access, reflecting the state of the art at any given moment in the future. In this context, we explicitly welcome feedback and further suggestions.

The expert consortium was primarily compiled in a way that covers the required expertise from various fields, but also consists of researchers of different countries, career stages, ages, roles and backgrounds (details can be found in the Supplementary Methods). The work 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 a 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 also to raise awareness on common issues and call into question flawed AI validation practice far beyond the boundaries of the field.

Data availability

No data were used in this study.

Code availability

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>.

References

- Maier-Hein, L. et al. Why rankings of biomedical image analysis competitions should be interpreted with care. *Nat. Commun.* **9**, 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 the Medical Image Computing and Computer Assisted Interventions.
- Gooding, M. J. et al. Comparative evaluation of autocontouring in clinical practice: a practical method using the Turing test. *Med. Phys.* **45**, 5105–5115 (2018).
- Kofler F. et al. Are we using appropriate segmentation metrics? Identifying correlates of human expert perception for CNN training beyond rolling the Dice coefficient. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.2103.06205> (2021).
- Vaassen, F. et al. Evaluation of measures for assessing time-saving of automatic organ-at-risk segmentation in radiotherapy. *Phys. Imaging Radiat. Oncol.* **13**, 1–6 (2020).
- Maier-Hein L. et al. Metrics reloaded: recommendations for image analysis validation. *Nat. Methods* <https://doi.org/10.1038/s41592-023-02151-z> (2024).
- Davide, C. & Giuseppe, J. The advantages of the Matthews correlation coefficient (MCC) over F_1 score and accuracy in binary classification evaluation. *BMC Genomics* **21**, 1–13 (2020).
- Davide, C., Niklas, T. & Giuseppe, J. The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData Min.* **14**, 1–22 (2021). The manuscript addresses the challenge of evaluating binary classifications. It compares MCC with other metrics, explaining their mathematical relationships and providing use cases where MCC offers more informative results.
- Grandini M., Bagli E. & Visani G. Metrics for multi-class classification: an overview. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.2008.05756> (2020).
- Taha, A. A. & Hanbury, A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Med. imaging* **15**, 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.
- Taha A. A., Hanbury A. & del Toro O. A J. A formal method for selecting evaluation metrics for image segmentation. In *2014 IEEE International Conference on Image Processing* 932–936 (IEEE, 2014).
- Lin T.-Y. et al. Microsoft COCO: common objects in context. In *European Conference on Computer Vision* 740–755 (Springer, 2014).
- Reinke A., et al. Common limitations of image processing metrics: a picture story. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.2104.05642> (2021).
- Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J. & Zisserman, A. The Pascal Visual Object Classes (VOC) challenge. *Int. J. Comput. Vis.* **88**, 303–338 (2010).
- Howard, A. et al. Sartorius—cell instance segmentation. *Kaggle* <https://www.kaggle.com/c/sartorius-cell-instance-segmentation> (2021).

15. Schmidt U., Weigert M., Broaddus C. & Myers G. Cell detection with star-convex polygons. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* 265–273 (Springer, 2018).
16. Stringer, C., Wang, T., Michaelos, M. & Pachitariu, M. Cellpose: a generalist algorithm for cellular segmentation. *Nat. methods* **18**, 100–106 (2021).
17. Hirling D. et al. Segmentation metric misinterpretations in bioimage analysis. *Nat. Methods* <https://doi.org/10.1038/s41592-023-01942-8> (2023).
18. Brown, B. B. *Delphi Process: A Methodology Used for the Elicitation of Opinions of Experts* (RAND Corporation, 1968).
19. Prashant Nasa, R. J. & Juneja, D. Delphi methodology in healthcare research: how to decide its appropriateness. *World J. Methodol.* **11**, 116 (2021).
20. Yeghiazaryan, V. & Voiculescu, I. D. Family of boundary overlap metrics for the evaluation of medical image segmentation. *J. Med. Imaging* **5**, 015006 (2018).
21. Gruber, S. & Buettner, F. Better uncertainty calibration via proper scores for classification and beyond. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.2203.07835> (2022).
22. Gooding, M. J., Boukerroui, D., Osorio, E. V., Monshouwer, R. & Brunenberg, E. Multicenter comparison of measures for quantitative evaluation of contouring in radiotherapy. *Phys. Imaging Radiat. Oncol.* **24**, 152–158 (2022).
23. Cordts, M. et al. The cityscapes dataset. In *CVPR Workshop on The Future of Datasets in Vision* (2015).
24. Muschelli, J. ROC and AUC with a binary predictor: a potentially misleading metric. *J. Classif.* **37**, 696–708 (2020).
25. Patrick Bilic, P. C. et al. The liver tumor segmentation benchmark (LITS). *Med. Image Anal.* **84**, 102680 (2023).
26. Tran, T. N. et al. Sources of performance variability in deep learning-based polyp detection. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.2211.09708> (2022).
27. Wiesenfarth, M. et al. Methods and open-source toolkit for analyzing and visualizing challenge results. *Sci. Rep.* **11**, 1–15 (2021).
28. Lennerz, J. K., Green, U., Williamson, D. F. K. & Mahmood, F. A unifying force for the realization of medical ai. *NPJ Digit. Med.* **5**, 172 (2022).
29. Correia, P. & Pereira, F. Video object relevance metrics for overall segmentation quality evaluation. *EURASIP J. Adv. Signal Process.* **2006**, 1–11 (2006).
30. Honauer, K., Maier-Hein, L. & Kondermann, D. The HCI stereo metrics: geometry-aware performance analysis of stereo algorithms. In *Proceedings of the IEEE International Conference on Computer Vision* 2120–2128 (IEEE, 2015).
31. Konukoglu, E., Glocker, B., Ye, D. H., Criminisi, A. & Pohl, K. M. Discriminative segmentation-based evaluation through shape dissimilarity. *IEEE Trans. Med. Imaging* **31**, 2278–2289 (2012).
32. Margolin, R., Zelnik-Manor, L. & Tal, A. How to evaluate foreground maps? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 248–255 (2014).
33. Carbonell, A., De la Pena, M., Flores, R. & Gago, S. 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 Res.* **34**, 5613–5622 (2006).
34. Di Sabatino, A. & Corazza, G. R. Nonceliac gluten sensitivity: sense or sensibility? *Ann. Intern. Med.* **156**, 309–311 (2012).
35. Roberts B. et al. Systematic gene tagging using CRISPR/Cas9 in human stem cells to illuminate cell organization. *Mol. Biol. Cell* **28**, 2854–2874 (2017).
36. Chen, J. et al. The Allen Cell and Structure Segmenter: a new open source toolkit for segmenting 3D intracellular structures in fluorescence microscopy images. Preprint at *bioRxiv* <https://doi.org/10.1101/491035> (2020).
37. Ounkomol, C., Seshamani, S., Malekar, M. M., Collman, F. & Johnson, G. R. Label-free prediction of three-dimensional fluorescence images from transmitted-light microscopy. *Nat. Methods* **15**, 917–920 (2018).
38. Viana, M. P. et al. Integrated intracellular organization and its variations in human IPS cells. *Nature* **613**, 345–354 (2023).

Acknowledgements

This work was initiated by the Helmholtz Association of German Research Centers in the scope of the Helmholtz Imaging Incubator (HI), the Medical Image Computing and Computer Assisted Interventions 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, NCI:U24CA279629 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, and the Natural Sciences and Engineering Research Council of Canada. F.B. was co-funded by the European Union (ERC, TAIPO, 101088594). The views and opinions expressed are those of only the authors 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) 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. Karargyris 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). T. Kurc was supported in part by 4UH3-CA225021- 03, 1U24CA180924-01A1, 3U24CA215109-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 P. 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 (Prof6 336449 funding program), University of Oulu strategic funding, Finnish Foundation for Cardiovascular Research, Wellbeing Services County of North Ostrobothnia (VTR project K62716) and the 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 P. Bankhead, G. S. Collins, R. Haase, F. Hamprecht, A. Karthikesalingam, H. Kenngott, P. Mattson, D. Moher, B. Stieltjes and M. Wiesenfarth for fruitful discussions on this work. We would like to thank S. Engelhardt, S. Koehler, M. A. Noyan, G. Polat, H. Rivaz, J. Schroeter, A. Saha, L. Sharan, P. Hirsch and M. Viana for suggesting additional illustrations that can be found in ref. 12.³⁵

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-led 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 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-led 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-led 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. T. Kooi 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. A. Kreshuk was an active member of the biomedical expert group, reviewed the manuscript and participated in surveys and workshops. T. Kurc 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. N. Rajpoot 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. N. Rieke 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-led 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 and suggested pitfalls. Z.R.Y. suggested pitfalls and participated in surveys. A. Karargyris, J.S.-R., C.I.S. and S.S. served on the expert Delphi panel and participated in workshops and surveys.

Competing interests

F.B. is an employee of Siemens (Munich, Germany). B.V.G. is a shareholder of Thirona (Nijmegen, the Netherlands). B.G. is an employee of HeartFlow (California, USA) and Kheiron Medical Technologies (London, UK). M.M.H. received an Nvidia GPU Grant. T. Kooi is an employee of Lunit (Seoul, South Korea). G.L. is on the advisory board of Canon Healthcare IT (Minnesota, USA) and is a shareholder of Aiosyn (Nijmegen, the Netherlands). N. Rajpoot is the founder and CSO of Histofy (New York, USA). N. Rieke is an employee of Nvidia (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, the Netherlands), 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.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41592-023-02150-0>.

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41592-023-02150-0>.

Correspondence should be addressed to Annika Reinke, Minu D. Tizabi, Paul F. Jäger or Lena Maier-Hein.

Peer review information *Nature Methods* thanks Pingkun Yan for their contribution to the peer review of this work. Primary Handling editor: Rita Strack, in collaboration with the *Nature Methods* team.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

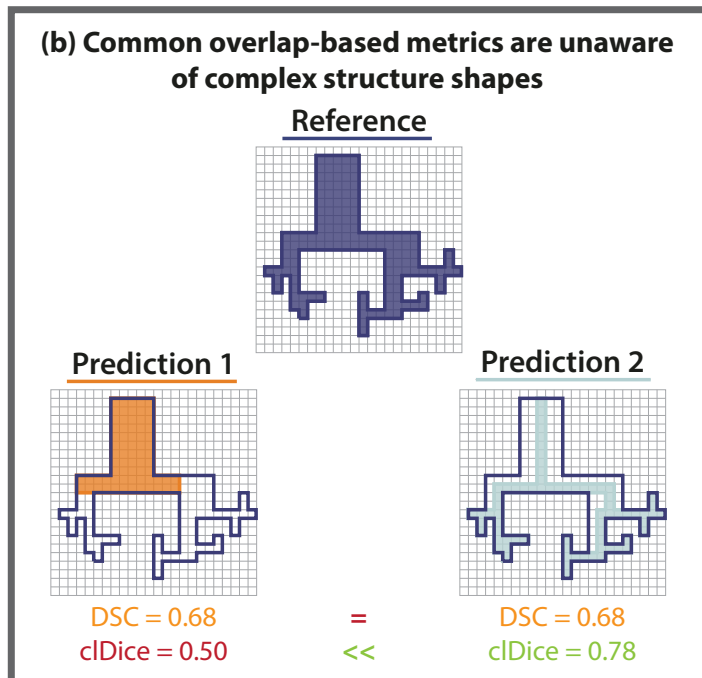
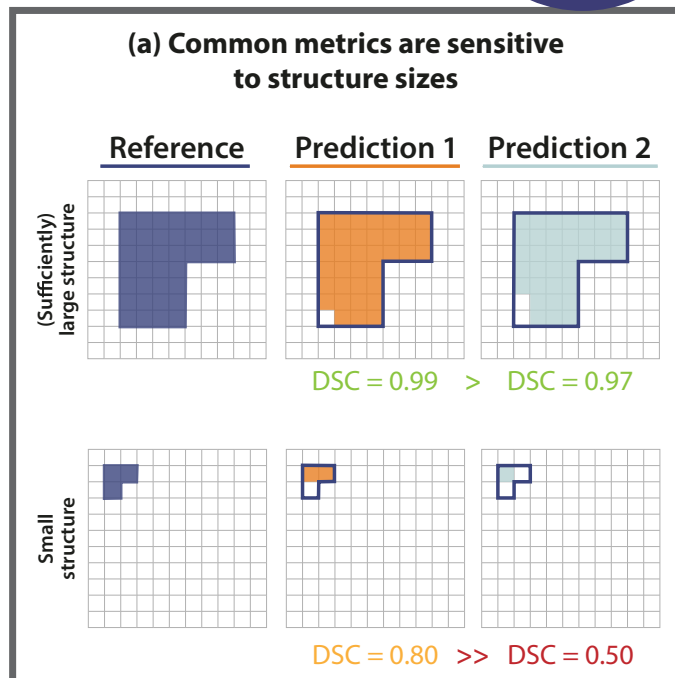
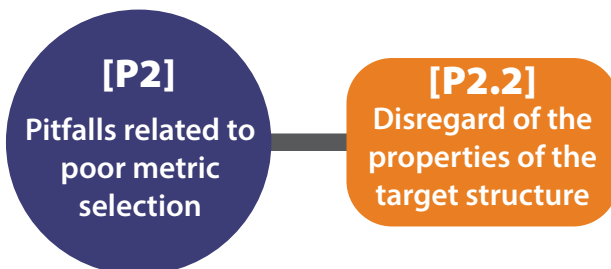
Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© Springer Nature America, Inc. 2024

Annika Reinke ^{1,2,3,86} , **Minu D. Tizabi** ^{1,4,86} , **Michael Baumgartner** ⁵, **Matthias Eisenmann**¹, **Doreen Heckmann-Nötzel**^{1,4}, **A. Emre Kavur** ^{1,5,6}, **Tim Rädtsch** ^{1,2}, **Carole H. Sudre**^{7,8}, **Laura Acion** ⁹, **Michela Antonelli** ^{8,10}, **Tal Arbel** ¹¹, **Spyridon Bakas** ^{12,13}, **Arriel Benis** ^{14,15}, **Florian Buettner**^{16,17,18,19,20}, **M. Jorge Cardoso** ⁸, **Veronika Cheplygina** ²¹, **Jianxu Chen** ²², **Evangelia Christodoulou**¹, **Beth A. Cimini** ²³, **Keyvan Farahani**²⁴, **Luciana Ferrer**²⁵, **Adrian Galdran**^{26,27}, **Bram van Ginneken**^{28,29}, **Ben Glocker** ³⁰, **Patrick Godau** ^{1,3,4}, **Daniel A. Hashimoto** ^{31,32}, **Michael M. Hoffman** ^{33,34,35,36}, **Merel Huisman**³⁷, **Fabian Isensee**^{5,6}, **Pierre Jannin** ^{38,39}, **Charles E. Kahn** ⁴⁰, **Dagmar Kainmueller**^{41,42}, **Bernhard Kainz**^{43,44}, **Alexandros Karargyris** ⁴⁵, **Jens Kleesiek** ⁴⁶, **Florian Kofler**⁴⁷, **Thijs Kooi**⁴⁸, **Annette Kopp-Schneider** ⁴⁹, **Michal Kozubek** ⁵⁰, **Anna Kreshuk** ⁵¹, **Tahsin Kurc**⁵², **Bennett A. Landman** ⁵³, **Geert Litjens** ⁵⁴, **Amin Madani**⁵⁵, **Klaus Maier-Hein**^{5,56}, **Anne L. Martel** ^{34,57}, **Erik Meijering** ⁵⁸, **Bjoern Menze** ⁵⁹, **Karel G. M. Moons**⁶⁰, **Henning Müller** ^{61,62}, **Brennan Nishchayporuk** ⁶³, **Felix Nickel**⁶⁴, **Jens Petersen**⁵, **Susanne M. Rafelski** ⁶⁵, **Nasir Rajpoot** ⁶⁶, **Mauricio Reyes**^{67,68}, **Michael A. Riegler** ^{69,70}, **Nicola Rieke** ⁷¹, **Julio Saez-Rodriguez** ^{72,73}, **Clara I. Sánchez**⁷⁴, **Shravya Shetty**⁷⁵, **Ronald M. Summers** ⁷⁶, **Abdel A. Taha**⁷⁷, **Aleksei Tiulpin** ^{78,79}, **Sotirios A. Tsafaris**⁸⁰, **Ben Van Calster**^{81,82}, **Gaël Varoquaux** ⁸³, **Ziv R. Yaniv** ⁸⁴, **Paul F. Jäger** ^{2,85,87}  & **Lena Maier-Hein** ^{1,2,3,4,73,87} 

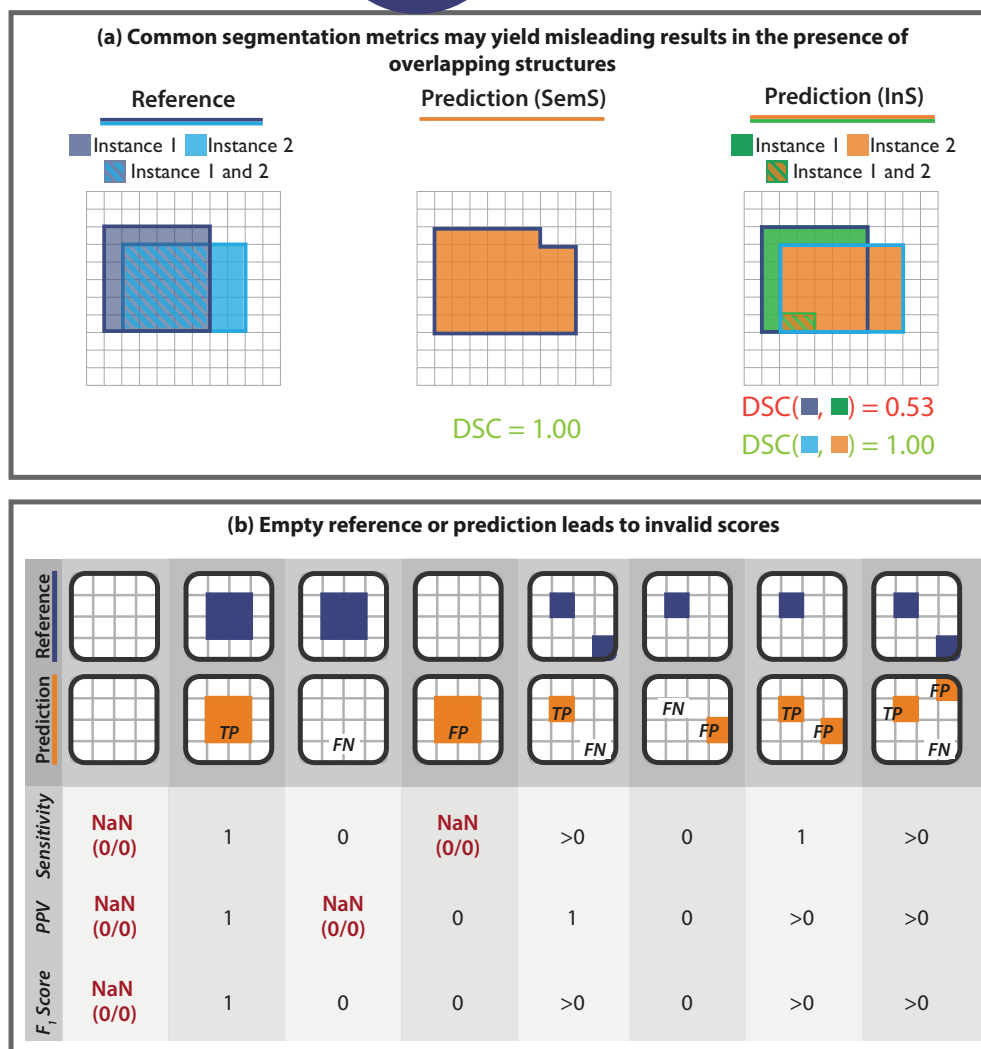
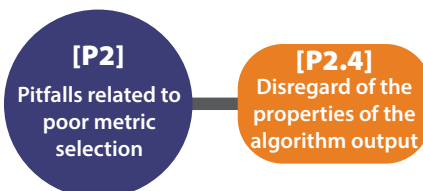
¹German Cancer Research Center (DKFZ) Heidelberg, Division of Intelligent Medical Systems, Heidelberg, Germany. ²German Cancer Research Center (DKFZ) Heidelberg, HI Helmholtz Imaging, Heidelberg, Germany. ³Faculty of Mathematics and Computer Science, Heidelberg University, Heidelberg, Germany. ⁴National Center for Tumor Diseases (NCT), NCT Heidelberg, a partnership between DKFZ and University Medical Center Heidelberg, Heidelberg, Germany. ⁵German Cancer Research Center (DKFZ) Heidelberg, Division of Medical Image Computing, Heidelberg, Germany. ⁶German Cancer Research Center (DKFZ) Heidelberg, HI Applied Computer Vision Lab, Heidelberg, Germany. ⁷MRC Unit for Lifelong Health and Ageing at UCL and Centre for Medical Image Computing, Department of Computer Science, University College London, London, UK. ⁸School of Biomedical Engineering and Imaging Science, King's College London, London, UK. ⁹Instituto de Cálculo, CONICET – Universidad de Buenos Aires, Buenos Aires, Argentina. ¹⁰Centre for Medical Image Computing, University College London, London, UK. ¹¹Centre for Intelligent Machines and MILA (Quebec Artificial Intelligence Institute), McGill University, Montréal, Quebec, Canada. ¹²Division of Computational Pathology, Dept of Pathology & Laboratory Medicine, Indiana University School of Medicine, Indianapolis, IN, USA. ¹³Center for Biomedical Image Computing and Analytics (CBICA), University of Pennsylvania, Philadelphia, PA, USA. ¹⁴Department of Digital Medical Technologies, Holon Institute of Technology, Holon, Israel. ¹⁵European Federation for Medical Informatics, Le Mont-sur-Lausanne, Switzerland. ¹⁶German Cancer Consortium (DKTK), partner site Frankfurt/Mainz, a partnership between DKFZ and UCT Frankfurt-Marburg, Frankfurt am Main, Germany. ¹⁷German Cancer Research Center (DKFZ) Heidelberg, Heidelberg, Germany. ¹⁸Goethe University Frankfurt, Department of Medicine, Frankfurt am Main, Germany. ¹⁹Goethe University Frankfurt, Department of Informatics, Frankfurt am Main, Germany. ²⁰Frankfurt Cancer Institute, Frankfurt am Main, Germany. ²¹Department of Computer Science, IT University of Copenhagen, Copenhagen, Denmark. ²²Leibniz-Institut für Analytische Wissenschaften – ISAS – e.V., Dortmund, Germany. ²³Imaging Platform, Broad Institute of MIT and Harvard, Cambridge, MA, USA. ²⁴Center for Biomedical Informatics and Information Technology, National Cancer Institute, Bethesda, MD, USA. ²⁵Instituto de Investigación en Ciencias de la Computación (ICC), CONICET-UBA, Ciudad Autónoma de Buenos Aires, Buenos Aires, Argentina. ²⁶Universitat Pompeu Fabra, Barcelona,

Spain. ²⁷University of Adelaide, Adelaide, South Australia, Australia. ²⁸Fraunhofer MEVIS, Bremen, Germany. ²⁹Radboud Institute for Health Sciences, Radboud University Medical Center, Nijmegen, the Netherlands. ³⁰Department of Computing, Imperial College London, South Kensington Campus, London, UK. ³¹Department of Surgery, Perelman School of Medicine, Philadelphia, PA, USA. ³²General Robotics Automation Sensing and Perception Laboratory, School of Engineering and Applied Science, University of Pennsylvania, Philadelphia, PA, USA. ³³Princess Margaret Cancer Centre, University Health Network, Toronto, Ontario, Canada. ³⁴Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada. ³⁵Department of Computer Science, University of Toronto, Toronto, Ontario, Canada. ³⁶Vector Institute for Artificial Intelligence, Toronto, Ontario, Canada. ³⁷Department of Radiology and Nuclear Medicine, Radboud University Medical Center, Nijmegen, the Netherlands. ³⁸Laboratoire Traitement du Signal et de l'Image – UMR_S 1099, Université de Rennes 1, Rennes, France. ³⁹INSERM, Paris, France. ⁴⁰Department of Radiology and Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA, USA. ⁴¹Max-Delbrück Center for Molecular Medicine in the Helmholtz Association (MDC), Biomedical Image Analysis and HI Helmholtz Imaging, Berlin, Germany. ⁴²University of Potsdam, Digital Engineering Faculty, Potsdam, Germany. ⁴³Department of Computing, Faculty of Engineering, Imperial College London, London, UK. ⁴⁴Department AIBE, Friedrich-Alexander-Universität (FAU), Erlangen-Nürnberg, Germany. ⁴⁵IHU Strasbourg, Strasbourg, France. ⁴⁶Translational Image-guided Oncology (TIO), Institute for AI in Medicine (IKIM), University Medicine Essen, Essen, Germany. ⁴⁷Helmholtz AI, Oberschleißheim, Germany. ⁴⁸Lunit, Seoul, South Korea. ⁴⁹German Cancer Research Center (DKFZ) Heidelberg, Division of Biostatistics, Heidelberg, Germany. ⁵⁰Centre for Biomedical Image Analysis and Faculty of Informatics, Masaryk University, Brno, Czech Republic. ⁵¹Cell Biology and Biophysics Unit, European Molecular Biology Laboratory (EMBL), Heidelberg, Germany. ⁵²Department of Biomedical Informatics, Stony Brook University, Health Science Center, Stony Brook, NY, USA. ⁵³Electrical Engineering, Vanderbilt University, Nashville, TN, USA. ⁵⁴Department of Pathology, Radboud University Medical Center, Nijmegen, the Netherlands. ⁵⁵Department of Surgery, University Health Network, Philadelphia, PA, USA. ⁵⁶Pattern Analysis and Learning Group, Department of Radiation Oncology, Heidelberg University Hospital, Heidelberg, Germany. ⁵⁷Physical Sciences, Sunnybrook Research Institute, Toronto, Ontario, Canada. ⁵⁸School of Computer Science and Engineering, University of New South Wales, UNSW Sydney, Kensington, New South Wales, Australia. ⁵⁹Department of Quantitative Biomedicine, University of Zurich, Zurich, Switzerland. ⁶⁰Julius Center for Health Sciences and Primary Care, UMC Utrecht, Utrecht University, Utrecht, the Netherlands. ⁶¹Information Systems Institute, University of Applied Sciences Western Switzerland (HES-SO), Sierre, Switzerland. ⁶²Medical Faculty, University of Geneva, Geneva, Switzerland. ⁶³MILA (Quebec Artificial Intelligence Institute), Montréal, Quebec, Canada. ⁶⁴Department of General, Visceral and Thoracic Surgery, University Medical Center Hamburg-Eppendorf, Hamburg, Germany. ⁶⁵Allen Institute for Cell Science, Seattle, WA, USA. ⁶⁶Tissue Image Analytics Laboratory, Department of Computer Science, University of Warwick, Coventry, UK. ⁶⁷ARTORG Center for Biomedical Engineering Research, University of Bern, Bern, Switzerland. ⁶⁸Department of Radiation Oncology, University Hospital Bern, University of Bern, Bern, Switzerland. ⁶⁹Simula Metropolitan Center for Digital Engineering, Oslo, Norway. ⁷⁰UiT The Arctic University of Norway, Tromsø, Norway. ⁷¹NVIDIA GmbH, München, Germany. ⁷²Institute for Computational Biomedicine, Heidelberg University, Heidelberg, Germany. ⁷³Faculty of Medicine, Heidelberg University Hospital, Heidelberg, Germany. ⁷⁴Informatics Institute, Faculty of Science, University of Amsterdam, Amsterdam, the Netherlands. ⁷⁵Google Health, Google, Palo Alto, CA, USA. ⁷⁶National Institutes of Health Clinical Center, Bethesda, MD, USA. ⁷⁷Institute of Information Systems Engineering, TU Wien, Vienna, Austria. ⁷⁸Research Unit of Health Sciences and Technology, Faculty of Medicine, University of Oulu, Oulu, Finland. ⁷⁹Neurocenter Oulu, Oulu University Hospital, Oulu, Finland. ⁸⁰School of Engineering, The University of Edinburgh, Edinburgh, Scotland. ⁸¹Department of Development and Regeneration and EPI-centre, KU Leuven, Leuven, Belgium. ⁸²Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, the Netherlands. ⁸³Parietal project team, INRIA Saclay-Île de France, Palaiseau, France. ⁸⁴National Institute of Allergy and Infectious Diseases, Bethesda, MD, USA. ⁸⁵German Cancer Research Center (DKFZ) Heidelberg, Interactive Machine Learning Group, Heidelberg, Germany. ⁸⁶These authors contributed equally: Annika Reinke, Minu D. Tizabi. ⁸⁷These authors jointly supervised this work: Paul F. Jäger, Lena Maier-Hein. ✉e-mail: a.reinke@dkfz-heidelberg.de; m.tizabi@dkfz-heidelberg.de; p.jaeger@dkfz-heidelberg.de; l.maier-hein@dkfz-heidelberg.de



Extended Data Fig. 1 | [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 Prediction 1 misses the fine-granular branches of the reference and favors Prediction 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.



Extended Data Fig. 2 | [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 F₁ Score are calculated. Some scenarios yield undefined values (Not a Number (NaN)).

Extended Data Table 1 | Overview of pitfall sources for image-level classification metrics ((a): counting metrics, (b): multi-threshold metrics) related to poor metric selection [P2]

(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 potential 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. 7a, SN 2.15)	⚠ (Fig. 7a)	⚠** (Fig. 7a)		⚠ (Fig. 7a)		⚠ (Figs. 7a, SN 2.15)	NPV: ⚠ (Figs. SN 2.15)	⚠ (Sens: Fig. 7a) ⚠ (Spec: Figs. 7a, SN 2.15)	⚠ (Figs. 7a, 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 score reflects 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. 1 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 potential pitfall	AP	AUROC
Importance of confidence awareness		⚠*
Importance of comparability across data sets	⚠ (Fig. SN 2.7)	
High class imbalance		⚠ (Fig. 7a)
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 score reflects 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. 1 and the metric profiles in Suppl. Note 3.2).

Pitfalls for semantic segmentation, object detection and instance segmentation are provided in Extended Data Tables 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 Supplementary Note 2. A comprehensive list of pitfalls is provided separately for each metric in the metrics cheat sheets (Supplementary 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). (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).

Extended Data Table 2 | Overview of pitfall sources for semantic segmentation metrics (a): overlap-based metrics, (b): boundary-based metrics) related to poor metric selection [P2]

(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 boundary	⚠ (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 (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 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)	
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. 2.a)	⚠ (Fig. SN 2.19, Extended Data Fig. 2.a)	⚠ (Fig. SN 2.19, Extended Data Fig. 2.a)	⚠ (Fig. SN 2.19, Extended Data Fig. 2.a)	⚠ (Fig. SN 2.19, Extended Data Fig. 2.a)	⚠ (Fig. SN 2.19, Extended Data Fig. 2.a)

* Can be mitigated by the choice of the percentile.

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 Supplementary Note 2. A comprehensive list of pitfalls is provided separately for each metrics in the metrics cheat sheets (Supplementary 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). (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).

Extended Data Table 3 | Overview of sources of pitfalls for object detection metrics ((a): detection metrics, (b): localization criteria) related to poor metric selection [P2]

(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 overlapping predictions	⚠ (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 for 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. 5a)	⚠ (Fig. 5a)	⚠ (Fig. 5a)	⚠ (Fig. 5a)
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. 2.10, Extended Data Fig. 1a)			
Complex structure shapes	⚠ (Fig. 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.

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 Supplementary Note 2. A comprehensive list of pitfalls is provided separately for each metrics in the metrics cheat sheets (Supplementary 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). (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).

Extended Data Table 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) Detection metrics. Considered counting metrics: F_{β} Score (Fig. SN. 3.43), Positive Predictive Value (PPV; Fig. SN 3.51), Panoptic Quality (PQ), 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	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 (Boundary 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)

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 Supplementary Note 2. A comprehensive list of pitfalls is provided separately for each metrics in the metrics cheat sheets (Supplementary 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), Panoptic Quality (Fig. SN 3.49) 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). (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).

Extended Data Table 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) Per instance segmentation overlap-based 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. ab)	⚠ (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 3.63), 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.

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 Supplementary 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: 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). (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 3.63), Mean Average Surface Distance (MASD) (Fig. SN 3.61) and Normalized Surface Distance (NSD) (Fig. SN 3.62).