nature methods

Perspective

https://doi.org/10.1038/s41592-023-02151-z

Metrics reloaded: recommendations for 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

Increasing evidence shows that flaws in machine learning (ML) algorithm validation are an underestimated global problem. In biomedical image analysis, chosen performance metrics often do not reflect the domain interest, and thus fail to adequately measure scientific progress and hinder translation of ML techniques into practice. To overcome this, we created Metrics Reloaded, a comprehensive framework guiding researchers in the problem-aware selection of metrics. Developed by a large international consortium in a multistage Delphi process, it is based on the novel concept of a problem fingerprint—a structured representation of the given problem that captures all aspects that are relevant for metric selection, from the domain interest to the properties of the target structure(s), dataset and algorithm output. On the basis of the problem fingerprint, users are guided through the process of choosing and applying appropriate validation metrics while being made aware of potential pitfalls. Metrics Reloaded targets image analysis problems that can be interpreted as classification tasks at image, object or pixel level, namely image-level classification, object detection, semantic segmentation and instance segmentation tasks. To improve the user experience, we implemented the framework in the Metrics Reloaded online tool. Following the convergence of ML methodology across application domains, Metrics Reloaded fosters the convergence of validation methodology. Its applicability is demonstrated for various biomedical use cases.

Automatic image processing with ML is gaining increasing traction in biological and medical imaging research and practice. Research has predominantly focused on the development of new image processing algorithms. The critical issue of reliable and objective performance assessment of these algorithms, however, remains largely unexplored. Algorithm performance in image processing is commonly assessed with validation metrics (not to be confused with distance metrics in the pure mathematical sense) that should serve as proxies for the domain interest. In consequence, the impact of validation metrics cannot be overstated; first, they are the basis for deciding on the practical (for example, clinical) suitability of a method and are thus *a* key component for translation into biomedical practice. In fact, validation that is not conducted according to relevant metrics could be one major reason for why many artificial intelligence (AI) developments in medical imaging fail to reach clinical practice[1,2]. In other words, the numbers presented in journals and conference proceedings do not reflect how successful a system will be when applied in practice. Second, metrics guide the scientific progress in the field; flawed metric use can lead to entirely futile resource investment and infeasible research directions while obscuring true scientific advancements.

Despite the importance of metrics, an increasing body of work shows that the metrics used in common practice often do not adequately reflect the underlying biomedical problems, diminishing the validity of the investigated methods[3–11]. This especially holds true for challenges, internationally respected competitions that have become the de facto standard for comparative performance

e-mail: l.maier-hein@dkfz-heidelberg.de; a.reinke@dkfz-heidelberg.de; p.jaeger@dkfz-heidelberg.de

https://doi.org/10.1038/s41592-023-02151-z

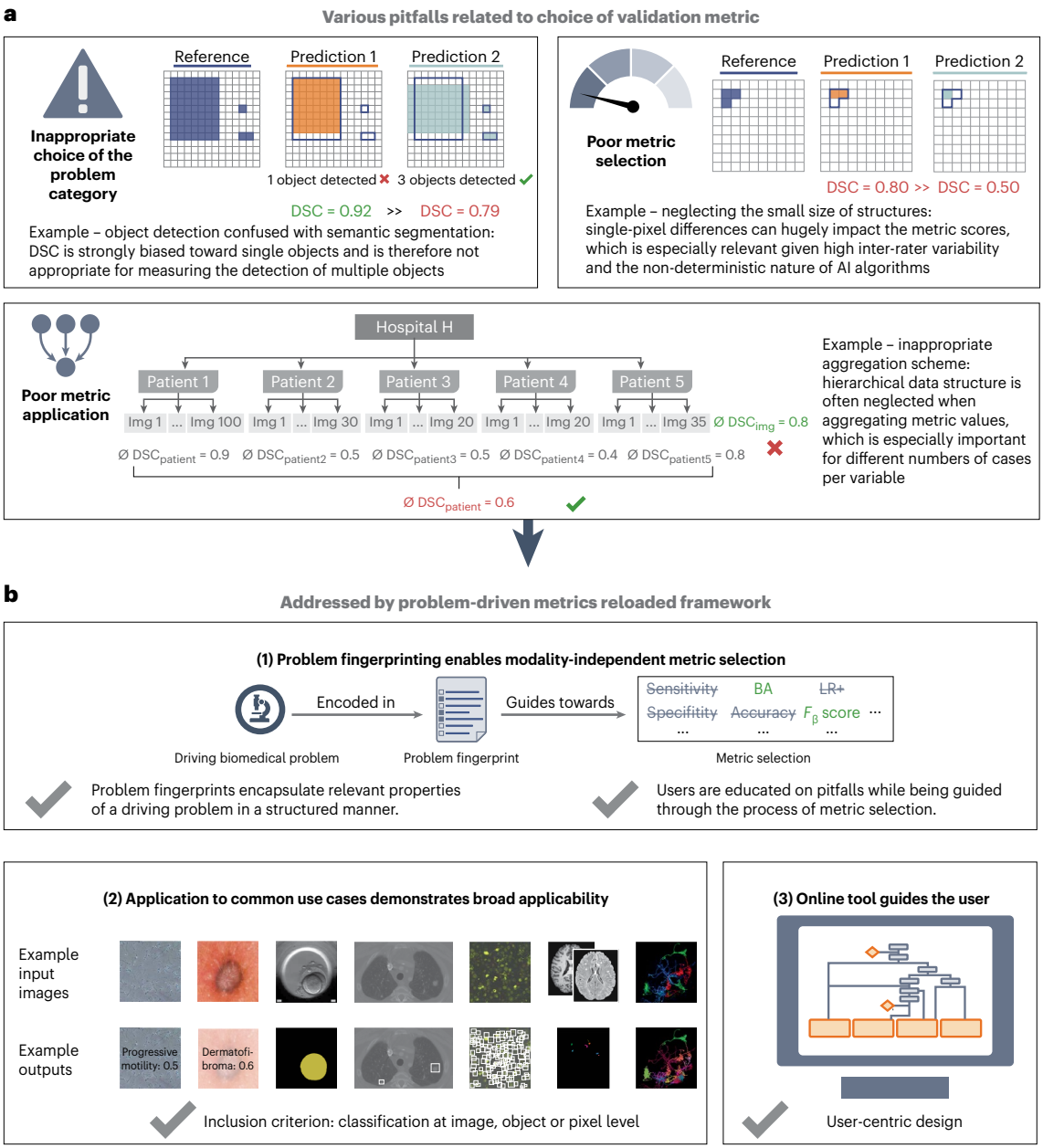


Fig. 1 | Contributions of the Metrics Reloaded framework. a, Motivation: Common problems related to metrics typically arise from inappropriate choice of the problem category (here: ObD confused with SemS; top left), poor metric selection (here: neglecting the small size of structures; top right) and poor metric application (here: inappropriate aggregation scheme; bottom). Pitfalls are highlighted in the boxes; Ø refers to the average DSC values. 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. Adapted from ref. 27 under a Creative Commons license CC BY 4.0. **b**, Metrics Reloaded addresses these pitfalls. (1) To enable the selection of metrics that match the domain interest, the framework is based on the new concept of problem fingerprinting, that is, the generation of a structured representation of the given biomedical problem that captures all properties that are relevant for metric selection. Based on the problem fingerprint, Metrics Reloaded guides the user through the process of metric selection and application while raising awareness of relevant pitfalls. (2) An instantiation of the framework for common biomedical use cases demonstrates its broad applicability. (3) A publicly available online tool facilitates application of the framework. Second input image reproduced from dermoscopedia (ref. 58) under a Creative Commons license CC BY 4.0; fourth input image reproduced with permission from ref. 59, American Association of Physicists in Medicine.

assessment of image processing methods. These challenges are often published in prestigious journals[12–14] and receive tremendous attention from both the scientific community and industry. Among a number of shortcomings in design and quality control that were recently unveiled by a multicenter initiative[8], the choice of inappropriate metrics stood out as a core problem. Compared to other areas of AI research, choosing the right metric is particularly challenging in image processing because the suitability of a metric depends on various factors. As a foundation for the present work, we identified three core categories related to pitfalls in metric selection (Fig. 1a):

Inappropriate choice of the problem category: The chosen metrics do not always reflect the biomedical need. For example, object detection (ObD) problems are often framed as segmentation tasks, resulting in the use of metrics that do not account for the potentially critical localization of all objects in the scene[15,16] (Fig. 1a).

Poor metric selection: Certain characteristics of a given biomedical problem render particular metrics inadequate. Mathematical metric properties are often neglected, for example, when using the Dice similarity coefficient (DSC) in the presence of particularly small structures (Fig. 1a).

Poor metric application: Even if a metric is well suited for a given problem in principle, pitfalls can occur when applying that metric to a specific dataset. For example, a common flaw pertains to ignoring hierarchical data structure, as in data from multiple hospitals or a variable number of images per patient (Fig. 1a), when aggregating metric values.

These problems are magnified by the fact that common practice often grows historically, and poor standards may be propagated between generations of scientists and in prominent publications. To dismantle such historically grown poor practices and leverage distributed knowledge from various subfields of image processing, we established the multidisciplinary Metrics Reloaded consortium. (We thank the Intelligent Medical Systems laboratory members N. Sautter, P. Vieten and T. Adler for the suggestion of the name, inspired by the Matrix movies.) This consortium comprises international experts from the fields of medical image analysis, biological image analysis, medical guideline development, general ML, different medical disciplines, statistics and epidemiology, representing a large number of biomedical imaging initiatives and societies.

The mission of Metrics Reloaded is to foster reliable algorithm validation through problem-aware, standardized choice of metrics with the long-term goal of (1) enabling the reliable tracking of scientific progress and (2) aiding to bridge the current chasm between ML research and translation into biomedical imaging practice.

Based on a kickoff workshop held in December 2020, the Metrics Reloaded framework (Figs. 1b and 2) was developed using a multistage Delphi process[17,18] for consensus building. Its primary purpose is to enable users to make educated decisions on which metrics to choose for a driving biomedical problem. The foundation of the metric selection process is the new concept of problem fingerprinting (Fig. 3). Abstracting from a specific domain, problem fingerprinting is the generation of a structured representation of the given biomedical problem that captures all properties relevant for metric selection. As depicted in Fig. 3, the properties captured by the fingerprint comprise domain interest-related properties, such as the particular importance of structure boundary, volume or center, target structure-related properties, such as the shape complexity or the size of structures relative to the image grid size, dataset-related properties, such as class imbalance, as well as algorithm output-related properties, such as the theoretical possibility of the algorithm output not containing any target structure.

Based on the problem fingerprint, the user is then, in a transparent and understandable manner, guided through the process of selecting an appropriate set of metrics while being made aware of potential pitfalls related to the specific characteristics of the underlying biomedical problem. The Metrics Reloaded framework currently supports problems in which categorical target variables are to be predicted based on a given n-dimensional input image (possibly enhanced with context information) at pixel, object or image level (Fig. 4). It thus supports problems that can be assigned to one of the following four problem categories: image-level classification (ImLC; image level), ObD (object level), semantic segmentation (SemS; pixel level) or instance segmentation (InS; pixel level). Designed to be imaging modality independent, Metrics Reloaded can be suited for application in various image analysis domains even beyond the field of biomedicine.

Here, we present the key contributions of our work in detail, namely (1) the Metrics Reloaded framework for problem-aware metric selection along with the key findings and design decisions that guided its development (Fig. 2), (2) the application of the framework to common biomedical use cases, showcasing its broad applicability (selection shown in Fig. 5) and (3) the open online tool that has been implemented to improve the user experience with our framework.

Metrics Reloaded framework

Metrics Reloaded is the result of a multistage Delphi process, comprising five international workshops, nine surveys, numerous expert group meetings and crowdsourced feedback processes, all conducted between 2020 and 2022. As a foundation of the recommendation framework, we identified common and rare pitfalls related to metrics in the field of biomedical image analysis using a community-powered process, detailed in this work's sister publication[19]. We found that common practice is often not well justified, and poor practices may even be propagated from one generation of scientists to the next. Importantly, many pitfalls generalize not only across the four problem categories that our framework addresses but also across domains (Fig. 4). This is because the source of the pitfall, such as class imbalance, uncertainties in the reference or poor image resolution, can occur irrespective of a specific modality or application.

Following the convergence of AI methodology across domains and problem categories, we therefore argue for the analogous convergence of validation methodology.

Cross-domain approach enables integration of distributed knowledge

To break historically grown poor practices, we followed a multidisciplinary cross-domain approach that enabled us to critically question common practice in different communities and integrate distributed knowledge in one common framework. To this end, we formed an international multidisciplinary consortium of 73 experts from various biomedical image analysis-related fields. Furthermore, we crowdsourced metric pitfalls and feedback on our approach in a social media campaign. Ultimately, a total of 156 researchers contributed to this work, including 84 mentioned in the acknowledgements. Consideration of the different knowledge and perspectives on metrics led to the following key design decisions for Metrics Reloaded:

Encapsulating domain knowledge: The questions asked to select a suitable metric are mostly similar regardless of image modality or application: Are the classes balanced? Is there a specific preference for the positive or negative class? What is the accuracy of the reference annotation? Is the structure boundary or volume of relevance for the target application? Importantly, while answering these questions requires domain expertise, the consequences in terms of metric selection can largely be regarded as domain independent. Our approach is

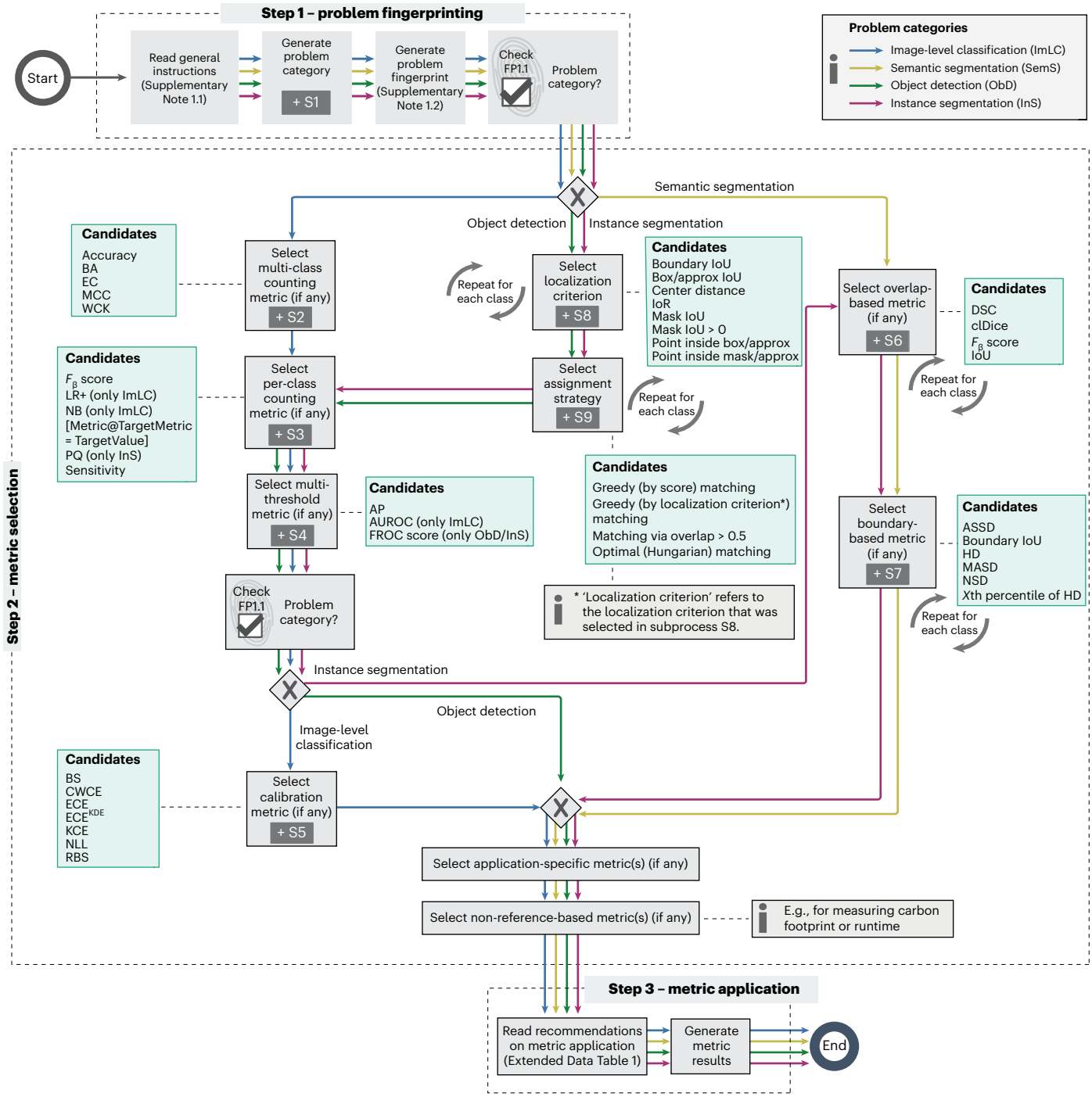
Fig. 2 | Metrics Reloaded recommendation framework from a user perspective. In step 1 – problem fingerprinting, the given biomedical image analysis problem is mapped to the appropriate image problem category, namely ImLC, SemS, ObD or InS; Fig. 4). The problem category and further characteristics of the given biomedical problem relevant for metric selection are then captured in a problem fingerprint (Fig. 3). In step 2 – metric selection, the user follows the respective colored path of the chosen problem category (ImLC →, SemS →, ObD → or InS →) to select a suitable pool of metrics from the Metrics Reloaded pools shown in green. When a tree branches, the fingerprint items determine which exact path to take. Finally, in step 3 – metric application, the user is supported in applying the metrics to a given dataset. During the traversal of the decision tree, the user goes through subprocesses, indicated by the plus sign, which are provided in Extended Data Figs. 1–9 and represent relevant steps in the metric selection process. Ambiguities related to metric selection are resolved via decision guides (Supplementary Note 2.7) that help users make an educated decision when multiple options are possible. A comprehensive textual description of the recommendations for all four problem categories as well as for the selection of corresponding calibration metrics (if any) is provided in Supplementary Notes 2.2–2.6. An overview of the symbols used in the process diagram is provided in Fig. SN 5.1. Condensed versions of the mappings for every category can be found in Supplementary Note 2.2 for ImLC, Supplementary Note 2.3 for SemS, Supplementary Note 2.4 for ObD and Supplementary Note 2.5 for InS.

https://doi.org/10.1038/s41592-023-02151-z

thus to abstract from the specific image modality and domain of a given problem by capturing the properties relevant for metric selection in a problem fingerprint (Fig. 3).

Exploiting synergies across classification scales: Similar considerations apply with regard to metric choice for classification, detection and segmentation tasks, as they can all be regarded as classification tasks at different scales (Fig. 4). The similarities between the categories, however, can also lead to problems when the wrong category is chosen (Fig. 1a). Therefore, we (1) address all four problem categories in one common framework (Fig. 2) and (2) cover the selection of the problem category itself in our framework (Extended Data Fig. 1).

Setting new standards: As the development and implementation of recommendations that go beyond the state of the art often requires critical mass, we involved stakeholders of various communities and



Abbreviations

AP Average precision
ASSD Average symmetric surface distance
AUROC Area under the receiver operating characteristic curve
BA Balanced accuracy
Boundary IoU Boundary intersection over union
Box/Approx Box/approximation
Box/Approx IoU Box/approximation intersection over union
BS Brier score
clDice Centerline Dice similarity coefficient
CWCE Class-wise calibration error
DSC Dice similarity coefficient
EC Expected cost
ECE Expected calibration error
ECE^{KDE} Expected calibration error kernel density estimate
FROC score Free-response receiver operating characteristic score
HD Hausdorff distance
ImLC Image-level classification
InS Instance segmentation
IoR Intersection over reference
IoU Intersection over union
KCE Kernel calibration error
LR+ Positive likelihood ratio
Mask IoU Mask intersection over union
MCC Matthews correlation coefficient
MASD Mean absolute surface distance
NB Net benefit
NLL Negative log likelihood
NSD Normalized surface distance
ObD Object detection
PQ Panoptic quality
RBS Root Brier score
SemS Semantic segmentation
WCK Weighted Cohen's κ

https://doi.org/10.1038/s41592-023-02151-z

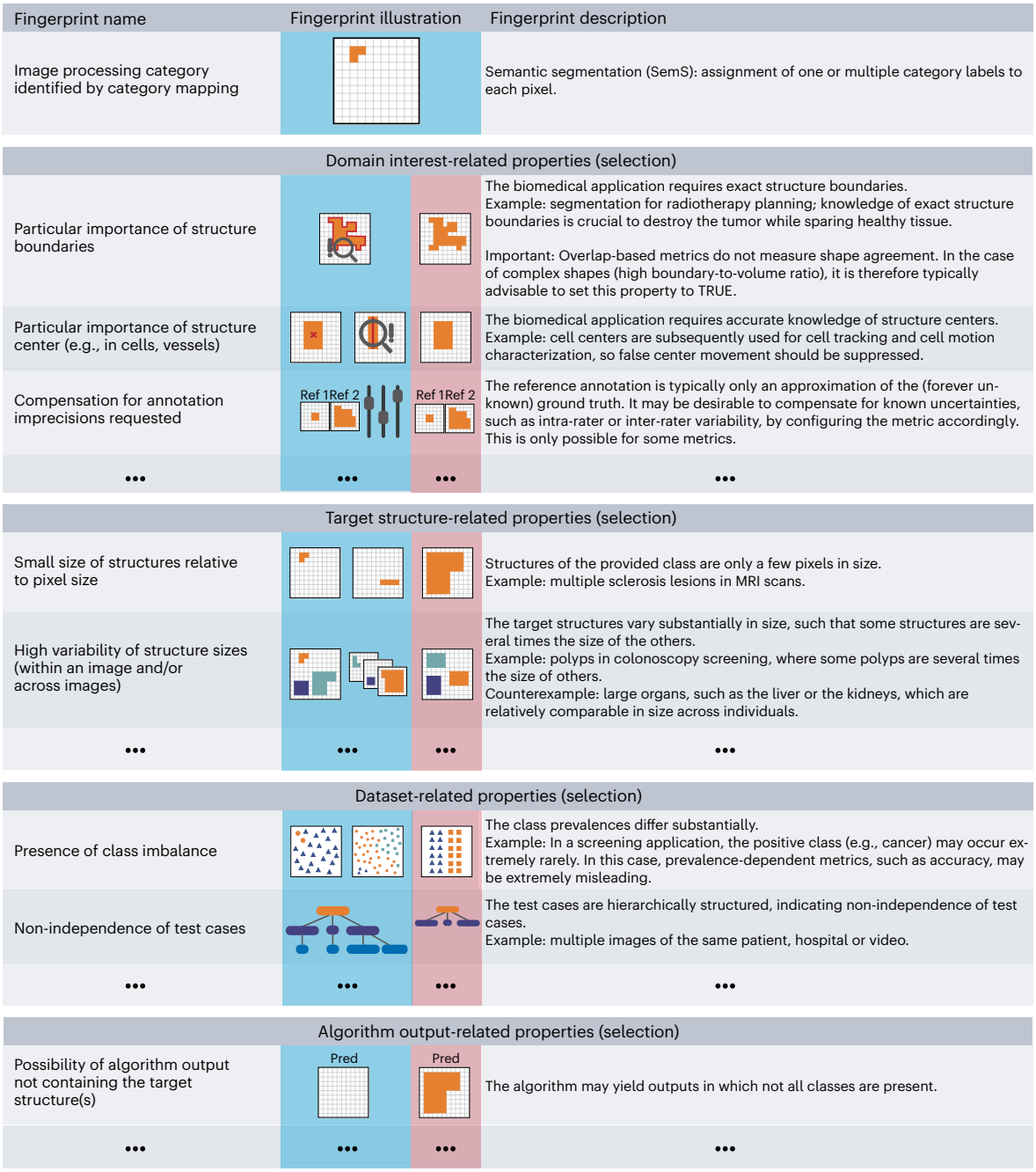


Fig. 3 | Relevant properties of a driving biomedical image analysis problem are captured by the problem fingerprint (selection for SemS shown here). The fingerprint comprises a set of items, each of which represents a specific property of the problem, is either binary or categorical, and must be instantiated by the user. Besides the problem category, the fingerprint comprises domain interest-related, target structure-related, dataset-related and algorithm output-related properties. A comprehensive version of the fingerprints for all problem categories can be found in Figs. SN 2.7–2.9 (ImLC), SN 2.10–2.11 (SemS), SN 2.12–2.14 (ObD) and SN 2.15–2.17 (InS). Pred, prediction; ref, reference.

societies in our consortium. Notably, our crowdsourcing-based approach led to a pool of metric candidates (Fig. SN 2.1) that includes not only commonly applied metrics, but also metrics that have to date received little attention in biomedical image analysis.

Abstracting from inference methodology: Metrics should be chosen based solely on the driving biomedical problem and not be affected by algorithm design choices. For example, the error functions applied in common neural network architectures do not justify the use of corresponding metrics (for example, validating with DSC to match the Dice loss used for training a neural network). Instead, the domain interest should guide the choice of metric, which, in turn, can guide the choice of the loss term.

Exploiting complementary metric strengths: A single metric typically cannot cover the complex requirements of the driving biomedical problem[20]. To account for the complementary strengths and weakness of metrics, we generally recommend the usage of multiple complementary metrics to validate image analysis problems. As detailed in our recommendations (Supplementary Note 2), we specifically recommend the selection of metrics from different families.

Validation by consensus building and community feedback: A major challenge for research on metrics is its validation, due to the lack of methods capable of quantitatively assessing the superiority of a given metric set over another. Following the spirit of large consortia formed to develop reporting guidelines (for example, CONSORT[21],

https://doi.org/10.1038/s41592-023-02151-z

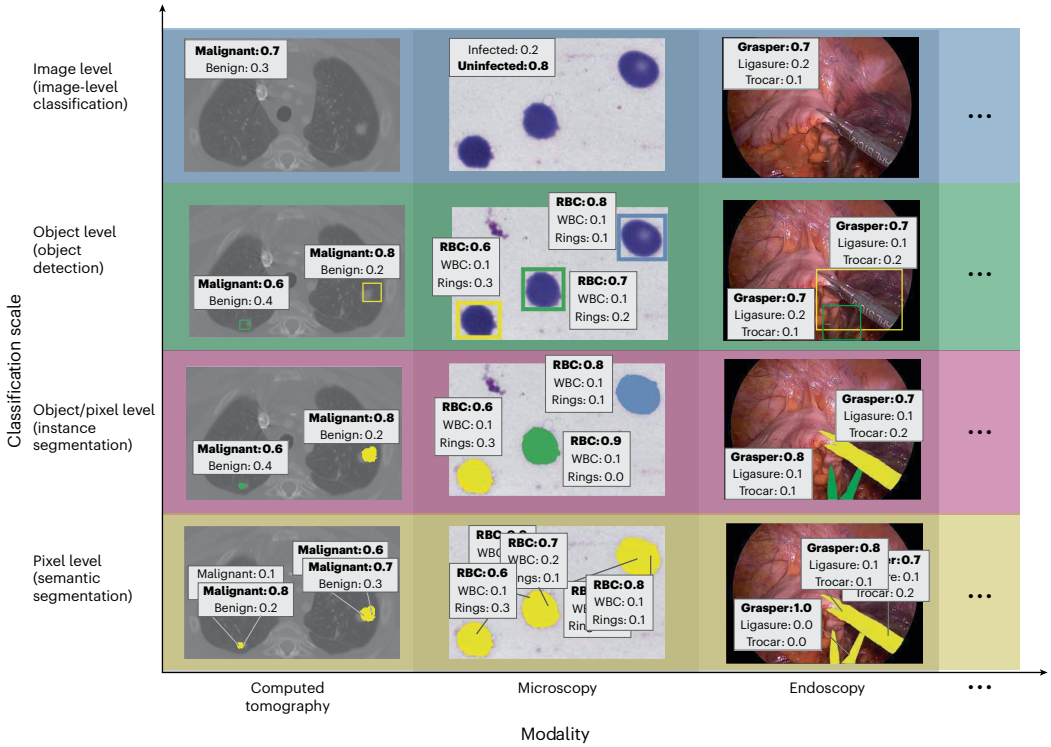


Fig. 4 | Metrics Reloaded fosters the convergence of validation methodology across modalities, application domains and classification scales. The framework considers problems in which categorical target variables are to be predicted at image, object and/or pixel level, resulting (from top to bottom) in ImLC, ObD, InS or SemS problems. These problem categories are relevant across modalities (here CT, microscopy and endoscopy) and application domains. From left to right: annotation of benign and malignant lesions in CT images[59], different cell types in microscopy images[60] and medical instruments in laparoscopy images[61]. Left, reproduced with permission from ref. 59, American Association of Physicists in Medicine; center, reproduced with permission from ref. 60, Springer Nature Limited; right, reproduced with permission from ref. 61, Springer Nature Limited. RBC, red blood cell; WBC, white blood cell.

TRIPOD[22] and STARD[23]), we built the validation of our framework on three main pillars: (1) Delphi processes to challenge and refine the proposals of the expert groups that worked on individual components of the framework; (2) community feedback obtained by broadcasting the framework via society mailing lists and social media platforms; and (3) and instantiation of the framework to a range of different biological and medical use cases.

Involving and educating users: Choosing adequate validation metrics is a complex process. Rather than providing a black box recommendation, Metrics Reloaded guides the user through the process of metric selection while raising awareness on pitfalls that may occur. In cases in which the trade-offs between different choices must be considered, decision guides (Supplementary Note 2.7) assist in deciding between competing metrics while respecting individual preferences.

Problem fingerprints encapsulate relevant domain knowledge

To encapsulate relevant domain knowledge in a common format and then enable a modality-agnostic metric recommendation approach that generalizes over domains, we developed the concept of problem fingerprinting (Fig. 3). As a foundation, we crowdsourced all properties of a driving biomedical problem that are potentially relevant for metric selection via surveys issued to the consortium (Supplementary Methods). This process resulted in a list of binary and categorical variables (fingerprint items) that must be instantiated by a user to trigger the Metrics Reloaded recommendation process. Common issues often relate to selecting metrics from the wrong problem category (Fig. 1a). To avoid such issues, problem fingerprinting begins with mapping a given problem with all its intrinsic and dataset-related properties to the corresponding problem category via the category mapping shown in Extended Data Fig. 1. The problem category is a fingerprint item itself.

In the following, we refer to all fingerprint items with the notation FP*X*.*Y*, where *Y* is a numerical identifier, and the index *X* represents one of the following families:

FP1 – Problem category refers to the problem category generated by S1 (Extended Data Fig. 1).

FP2 – Domain interest-related properties reflect user preferences and are highly dependent on the target application. A semantic image segmentation that serves as the foundation for radiotherapy planning, for example, would require exact contours (FP2.1 – particular importance of structure boundaries = TRUE). On the other hand, for a cell segmentation problem that serves as prerequisite for cell tracking, the object centers may be much more important (FP2.3 – particular importance of structure center(line) = TRUE). Both problems could be tackled with identical network architectures, but the validation metrics should be different.

FP3 – Target structure-related properties represent inherent properties of target structure(s) (if any), such as the size, size variability and the shape. Here, the term target structures can refer to any object/structure of interest, such as cells, vessels, medical instruments or tumors.

FP4 – Dataset-related properties capture properties inherent to the provided data to which the metric is applied. They primarily relate to class prevalences, uncertainties of the reference annotations and whether the data structure is hierarchical.

FP5 – Algorithm output-related properties encode properties of the output, such as the availability of predicted class scores.

Note that not all properties are relevant for all problem categories. For example, the shape and size of target structures is highly relevant for segmentation problems but irrelevant for image classification problems. The complete problem category-specific fingerprints are provided in Supplementary Note 1.3.

https://doi.org/10.1038/s41592-023-02151-z

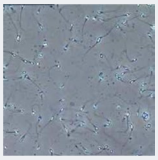
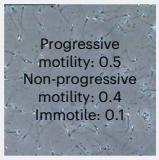
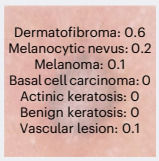
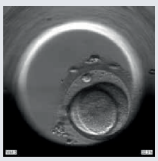
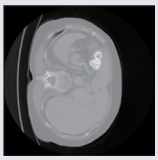
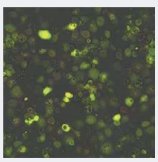
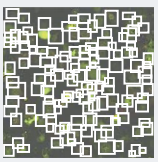
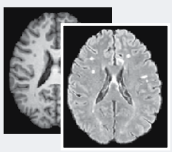
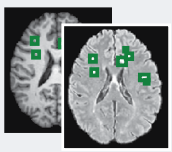
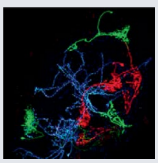
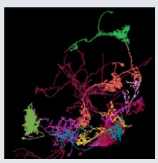
Problem description	ID	Scenario	Sample input image	Recommended ouput	Recommendation
Classification of images	ImLC-1	Frame-based sperm motility classification from microscopy time-lapse video of human spermatozoa			Problem category: Image-level classification Multi-class counting metric (S2): Balanced accuracy (BA) Per-class counting metric (S3): Positive likelihood ratio (LR+) Multi-threshold metric (S4): Area under the receiver operating characteristic curve (AUROC) Calibration metric (S5): Expected calibration error (ECE; top-label) and root Brier score (RBS)
	ImLC-2	Disease classification in dermoscopic images			
Segmentation of large objects	SemS-1	Embryo segmentation from microscopy images			Problem category: Semantic segmentation Overlap-based metric (S6): Dice similarity coefficient (DSC) Boundary-based metric (S7): Normalized surface distance (NSC) Specific property-related metric: Liver segmentation: absolute volume difference
	SemS-2	Liver segmentation in CT images			
Detection of multiple and arbitrarily located objects	ObD-1	Cell detection and tracking during the autophagy process in time-lapse microscopy videos			Problem category: Object detection Per-class counting metric (S3): FP per image (FPPI) @ sensitivity = 0.95 Multi-threshold metric (S4): Free-response receiver operating characteristic (FROC) score Localization criterion (S8): Box intersection over union (box IoU) Assignment strategy (S9): Greedy (by score) matching, set double assignments to FPs
	ObD-2	MS lesion detection in multimodal brain MRI images			
Segmentation and distinction of tubular objects	InS-1	Instance segmentation of neurons from the fruit fly in 3D multicolor light microscopy images			Problem category: Instance segmentation Per-class counting metric (S3): F_β score Multi-threshold metric (S4): average precision (AP) Overlap-based metric (S6): Centerline Dice similarity coefficient (clDice) Boundary-based metric (S7): NSD Localization criterion (S8): Neuron segmentation: mask IoU Instrument segmentation: boundary IoU Assignment strategy (S9): Greedy (by score) matching, set double assignments to FPs
	InS-2	Surgical instrument instance segmentation in colonoscopy videos			

Fig. 5 | Instantiation of the framework with recommendations for concrete biomedical questions. From top to bottom: (1) Image classification for the examples of sperm motility classification[62] and disease classification in dermoscopic images[63,58]. (2) SemS of large objects for the examples of embryo segmentation from microscopy[64] and liver segmentation in CT images[65,66]. (3) Detection of multiple and arbitrarily located objects for the examples of cell detection and tracking during the autophagy process[67,68] and MS lesion detection in multimodal brain MRI images[69,70]. (4) InS of tubular objects for the examples of InS of neurons from the fruit fly[71–73] and surgical instrument InS[61]. The corresponding traversals through the decision trees are shown in Supplementary Note 4. An overview of the recommended metrics can be found in Supplementary Note 3.1, including relevant information for each metric. ImLC-2, reproduced from dermoscopedia under a Creative Commons license CC BY 4.0; SemS-2, reproduced with permission from ref. 65, Springer Nature Limited; InS-1, reproduced with permission from ref. 61, Springer Nature Limited.

Metrics Reloaded addresses all three types of metric pitfalls

Metrics Reloaded was designed to address all three types of metric pitfalls identified in ref. 19 and illustrated in Fig. 1a. More specifically, each of the three steps shown in Fig. 2 addresses one type of pitfall:

Step 1– Fingerprinting. A user should begin by reading the general instructions of the recommendation framework (Supplementary Note 1.1). Next, the user should convert the driving biomedical problem to a problem fingerprint. This step not only is a prerequisite for applying the framework across application domains and classification scales, but also specifically addresses the inappropriate choice of the problem category via the integrated category mapping. Once the user's domain knowledge has been encapsulated in the problem fingerprint, the actual metric selection is conducted according to a domain-agnostic and modality-agnostic process.

Step 2 – Metric selection. A Delphi process yielded the Metrics Reloaded pool of reference-based validation metrics (Fig. SN 2.1). Notably, this pool contains metrics that are currently not widely known in some biomedical image analysis communities. A prominent example is the Net Benefit (NB)[24] metric, popular in clinical prediction tasks and designed to determine whether basing decisions on a method would do more good than harm. A diagnostic test, for example, may lead to early identification and treatment of a disease, but typically will also cause a number of individuals without disease to be subjected to unnecessary further interventions. NB allows the consideration of such trade-offs by putting benefits and harms on the same scale so that they can be directly compared. Another example is the expected cost (EC) metric[25], which can be seen as a generalization of accuracy with many desirable added features but is not well known in the biomedical image analysis communities[26]. Based on the Metrics Reloaded pool, the metric recommendation is performed with a business process model and notation (BPMN)-inspired flowchart (Fig. SN 5.1), in which conditional operations are based on one or multiple fingerprint properties (Fig. 2). The main flowchart has three substeps, each addressing the complementary strengths and weaknesses of common metrics. First, common reference-based metrics, which are based on the comparison of the algorithm output to a reference annotation, are selected. Second, the pool of standard metrics can be complemented with custom metrics to address application- specific complementary properties. Third, non-reference-based metrics assessing speed, memory consumption or carbon footprint, for example, can be added to the metric pool(s). In this paper, we focus on the step of selecting reference-based metrics, because this is where synergies across modalities and scales can be exploited.

These synergies are showcased by the substantial overlap between the different paths that, depending on the problem category, are taken through the mapping during metric selection. All paths comprise several subprocesses *S* (indicated by the ⊞ symbol), each of which holds a subsidiary decision tree representing one specific step of the selection process. Traversal of a subprocess typically leads to the addition of a metric to the problem-specific metric pool. In multi-class prediction problems, dedicated metric pools for each class may need to be generated as relevant properties may differ from class to class. A three-dimensional (3D) SemS problem, for example, could require the simultaneous segmentation of both tubular and non-tubular structures (for example, liver vessels and tissue). These require different metrics for validation. Although this is a corner case, our framework addresses this issue in principle. In ambiguous cases, that is, when the user can choose between two options in one step of the decision tree, a corresponding decision guide details the trade-offs that need to be considered (Supplementary Note 2.7). For example, the intersection over union (IoU) and the DSC are mathematically closely related. The concrete choice typically boils down to a simple user or community preference.

Figure 2 along with the corresponding subprocesses S1–S9 (Extended Data Figs. 1–9) captures the core contribution of this paper, namely the consensus recommendation of the Metrics Reloaded consortium according to the final Delphi process. For all ten components, the required Delphi consensus threshold (>75% agreement) was met. In all cases of disagreement, which ranged from 0% to 7% for Fig. 2 and S1–S9, each remaining point of criticism was respectively only raised by a single person. The following paragraphs present a summary of the four different colored paths through step 2 - Metric selection of the recommendation tree (Fig. 2) for the task of selecting reference-based metrics from the Metrics Reloaded pool of common metrics. More comprehensive textual descriptions can be found in Supplementary Note 2.

Image-level classification

ImLC is conceptually the most straightforward problem category, as the task is simply to assign one of multiple possible labels to an entire image (Supplementary Note 2.2). The validation metrics are designed to measure two key properties: discrimination and calibration.

Discrimination refers to the ability of a classifier to discriminate between two or more classes. This can be achieved by counting metrics that operate on the cardinalities of a fixed confusion matrix (that is, the true/false positives/negatives in the binary classification case). Prominent examples are sensitivity, specificity or F_1 score for binary settings and Matthews correlation coefficient (MCC) for multi-class settings. Converting predicted class scores to a fixed confusion matrix (in the binary case by setting a potentially arbitrary cutoff) can, however, be regarded as problematic in the context of performance assessment[27]. Multi-threshold metrics, such as area under the receiver operating characteristic curve (AUROC), are therefore based on varying the cutoff, which enables the explicit analysis of the trade-off between competing properties such as sensitivity and specificity.

While most research in biomedical image analysis focuses on the discrimination capabilities of classifiers, a complementary important property is the calibration of a model. An uncertainty-aware model should yield predicted class scores that represent the true likelihood of events[28], as detailed in Supplementary Note 2.6. Overoptimistic or underoptimistic classifiers can be especially problematic in prediction tasks where a clinical decision may be made based on the risk of the patient developing a certain condition. Metrics Reloaded hence provides recommendations for validating the algorithm performance both in terms of discrimination and calibration. We recommend the following process for classification problems (Fig. 2 and Supplementary Note 2.2):

1. Select multi-class metric (if any): Multi-class metrics have the unique advantage of capturing the performance of an algorithm for all classes in a single value. With the ability to take into account all entries of the multi-class confusion matrix, they provide a holistic measure of performance without the need for customized class-aggregation schemes. We recommend using a multi-class metric if a decision rule applied to the predicted class scores is available (FP2.6). In certain use cases, especially in the presence of ordinal data, there is an unequal severity of class confusions (FP2.5.2), meaning that different costs should be applied to different misclassifications reflected by the confusion matrix. In such cases, we generally recommend EC as a metric. Otherwise, depending on the specific scenario, accuracy, balanced accuracy (BA) and MCC may be viable alternatives. The concrete choice of metric depends primarily on the prevalences (frequencies) of classes in the provided validation set and the target population (FP4.1/2), as detailed in subprocess S2 (Extended Data Fig. 2) and the corresponding textual description in Supplementary Note 2.2.
   As class-specific analyses are not possible with multi-class metrics, which can potentially hide poor performance on individual classes, we recommend an additional validation with per-class counting metrics (optional) and multi-threshold metrics (always recommended).

2. Select per-class counting metric (if any): If a decision rule applied to the predicted class scores is available (FP2.6), a per-class counting metric, such as the F_β score, should be selected. Each class of interest is separately assessed, preferably in a 'one-versus-rest' fashion. The choice depends primarily on the decision rule (FP2.6) and the distribution of classes (FP4.2). Details can be found in subprocess S3 for selecting per-class counting metrics (Extended Data Fig. 3).
3. Select multi-threshold metric (if any): Counting metrics reduce the potentially complex output of a classifier (the continuous class scores) to a single value (the predicted class), such that they can work with a fixed confusion matrix. To compensate for this loss of information and obtain a more comprehensive picture of a classifier's discriminatory performance, multi- threshold metrics work with a dynamic confusion matrix reflecting a range of possible thresholds applied to the predicted class scores. While we recommend the popular, well-interpretable and prevalence-independent AUROC as the default multi-threshold metric for classification, average precision can be more suitable in the case of high-class balance because it incorporates predicted values, as detailed in subprocess S4 for selecting multi-threshold metrics (Extended Data Fig. 4).
4. Select calibration metric (if any): If calibration assessment is requested (FP2.7), one or multiple calibration metrics should be added to the metric pool as detailed in subprocess S5 for selecting calibration metrics (Extended Data Fig. 5).

Semantic segmentation

In SemS, classification occurs at pixel level. However, it is not advisable to simply apply the standard classification metrics to the entire collection of pixels in a dataset for two reasons. Firstly, pixels of the same image are highly correlated. Hence, to respect the hierarchical data structure, metric values should first be computed per image and then be aggregated over the set of images. Note in this context that the commonly used DSC is mathematically identical to the popular F_1 score applied at pixel level. Secondly, in segmentation problems, the user typically has an inherent interest in structure boundaries, centers or volumes of structures (FP2.1, FP2.2 and FP2.3). The family of boundary-based metrics (subset of distance-based metrics) therefore requires the extraction of structure boundaries from the binary segmentation masks as a foundation for segmentation assessment. Based on these considerations and given all the complementary strengths and weaknesses of common segmentation metrics[27], we recommend the following process for segmentation problems (Fig. 2 and Supplementary Note 2.3):

1. Select overlap-based metric (if any): In segmentation problems, counting metrics such as the DSC or IoU measure the overlap between the reference annotation and the algorithm prediction. As they can be considered the de facto standard for assessing segmentation quality and are well interpretable, we recommend using them by default unless the target structures are consistently small, relative to the grid size (FP3.1), and the reference may be noisy (FP4.3.1). Depending on the specific properties of the problems, we recommend the DSC or IoU (default recommendation), the F_β score (preferred when there is a preference for either false positive (FP) or false negative (FN)) or the centerline Dice similarity coefficient (clDice; for tubular structures). Details can be found in subprocess S6 for selecting overlap-based metrics (Extended Data Fig. 6).
2. Select boundary-based metric (if any): Key weaknesses of overlap-based metrics include shape unawareness and limitations when dealing with small structures or high size variability[27]. Our general recommendation is therefore to complement an overlap-based metric with a boundary-based metric. If annotation imprecisions should be compensated for (FP2.5.7), our default recommendation is the normalized surface distance (NSD). Otherwise, the fundamental user preference guiding metric selection is whether errors should be penalized by existence or distance (FP2.5.6), as detailed in subprocess S7 for selecting boundary-based metrics (Extended Data Fig. 7).

Object detection

ObD problems differ from segmentation problems in several key features with respect to metric selection. Firstly, they involve distinguishing different instances of the same class and thus require the step of locating objects and assigning them to the corresponding reference object. Secondly, the granularity of localization is comparatively rough, which is why no boundary-based metrics are required (otherwise the problem would be phrased as an InS problem). Finally, and crucially important from a mathematical perspective, the absence of true negatives (TNs) in ObD problems renders many popular classification metrics (for example, accuracy, specificity and AUROC) invalid. In binary problems, for example, suitable counting metrics can only be based on three of the four entries of the confusion matrix. Based on these considerations and taking into account all the complementary strengths and weaknesses of existing metrics[27], we propose the following steps for ObD problems (Fig. 2 and Supplementary Note 2.4):

1. Select localization criterion: An essential part of the validation is to decide whether a prediction matches a reference object. To this end, (1) the location of both the reference objects and the predicted objects must be adequately represented (for example, by masks, bounding boxes or center points), and (2) a metric for deciding on a match (for example, mask IoU) must be chosen. As detailed in subprocess S8 for selecting the localization criterion (Extended Data Fig. 8), our recommendation considers both the granularity of the provided reference (FP4.4) and the required granularity of the localization (FP2.4).
2. Select assignment strategy: As the localization does not necessarily lead to unambiguous matchings, an assignment strategy needs to be chosen to potentially resolve ambiguities that occurred during localization. As detailed in subprocess S9 for selecting the assignment strategy (Extended Data Fig. 9), the recommended strategy depends on the availability of continuous class scores (FP5.1) as well as on whether double assignments should be punished (FP2.5.8).

Select classification metric(s) (if any): Once objects have been located and assigned to reference objects, generation of a confusion matrix (without TN) is possible. The final step therefore simply comprises choosing suitable classification metrics for validation. Several subfields of biomedical image analysis have converged to choosing solely a counting metric, such as the F_β score, as the primary metric in ObD problems. We follow this recommendation when no continuous class scores are available for the detected objects (FP5.1). Otherwise, we disagree with the practice of basing performance assessment solely on a single, potentially suboptimal cutoff on the continuous class scores. Instead, we follow the recommendations for ImLC and propose complementing a counting metric (subprocess S3; Extended Data Fig. 3) with a multi-threshold metric (subprocess S4; Extended Data Fig. 4) to obtain a more holistic picture of performance. As multi-threshold metric, we recommend average precision or free-response receiver operating characteristic (FROC) score, depending on whether an easy interpretation (FROC score) or a standardized metric (average precision) is preferred. The choice of per-class counting metric depends primarily on the decision rule (FP2.6).

Note that the previous description implicitly assumed single-class problems, but generalization to multi-class problems is straightforward by applying the validation for each class. It is further worth mentioning that metric application is not trivial in ObD problems as the number of objects in an image may be extremely small, or even zero, compared to the number of pixels in an image. Special

considerations with respect to aggregation must therefore be made (Supplementary Note 2.4).

Instance segmentation

InS delivers the tasks of ObD and SemS at the same time. Thus, the pitfalls and recommendations for InS problems are closely related to those for segmentation and ObD[27]. This is directly reflected in our metric selection process (Fig. 2 and Supplementary Note 2.5):

1. Select ObD metric(s): To overcome problems related to instance unawareness (Fig. 1a), we recommend selection of a set of detection metrics to explicitly measure detection performance. To this end, we recommend almost the exact process as for ObD with two exceptions. Firstly, given the fine granularity of both the output and the reference annotation, our recommendation for the localization strategy differs, as detailed in subprocess S8 (Extended Data Fig. 8). Secondly, as depicted in S3 (Extended Data Fig. 3), we recommend panoptic quality[29] as an alternative to the F_β score. This metric is especially suited for InS, as it combines the assessment of overall detection performance and segmentation quality of successfully matched (true positive (TP)) instances in a single score.
2. Select segmentation metric(s) (if any): In a second step, metrics to explicitly assess the segmentation quality for the TP instances may be selected. Here, we follow the exact same process as in SemS (subprocesses S6 and S7; Extended Data Figs. 6 and 7). The primary difference is that the segmentation metrics are applied for each instance.

Importantly, the development process of the Metrics Reloaded framework was designed such that the pitfalls identified in the sister publication of this work[19] are comprehensively addressed. Table 1 makes the recommendations and design decisions corresponding to specific pitfalls explicit.

Once common reference-based metrics have been selected and, where necessary, complemented by application-specific metrics, the user proceeds with the application of the metrics to the given problem.

Step 3 - Metric application. Although the application of a metric to a given dataset may appear straightforward, numerous pitfalls can occur[27]. Our recommendations for addressing them are provided in Extended Data Table 1. Following the taxonomy provided in the sister publication of this work[19], they are categorized in recommendations related to metric implementation, aggregation, ranking, interpretation and reporting. While several aspects are covered in related work (for example, ref. 30), an important contribution of the present work is the metric-specific summary of recommendations captured in the metric cheat sheets (Supplementary Note 3.1). A further major contribution is our implementation of all Metrics Reloaded metrics in the open-source framework Medical Open Network for Artificial Intelligence (MONAI), available at https://github.com/Project-MONAI/MetricsReloaded/ (Supplementary Methods).

Metrics Reloaded is broadly applicable in biomedical image analysis

To validate the Metrics Reloaded framework, we used it to generate recommendations for common use cases in biomedical image processing (Supplementary Note 4). The traversal through the decision tree of our framework is detailed for eight selected use cases corresponding to the four different problem categories (Fig. 5):

ImLC (Figs. SN 5.5–5.8): frame-based sperm motility classification from time-lapse microscopy video of human spermatozoa (ImLC-1) and disease classification in dermoscopic images (ImLC-2).

SemS (Figs. SN 5.9 and 5.10): embryo segmentation in microscopy images (SemS-1) and liver segmentation in computed tomography (CT) images (SemS-2).

ObD (Figs. SN 5.6, 5.7, 5.11 and 5.12): cell detection and tracking during the autophagy process in time-lapse microscopy (ObD-1) and multiple sclerosis (MS) lesion detection in multimodal brain magnetic resonance imaging (MRI) images (ObD-2).

InS (Figs. SN 5.6, 5.7 and 5.9–5.12): InS of neurons from the fruit fly in 3D multicolor light microscopy images (InS-1) and surgical InS in colonoscopy videos (InS-2).

The resulting metric recommendations (Fig. 5) demonstrate that a common framework across domains is sensible. In the showcased examples, shared properties of problems from different domains result in almost identical recommendations. In the SemS use cases, for example, the specific image modality is irrelevant for metric selection. What matters is that a single object with a large size relative to the grid size should be segmented—properties that are captured by the proposed fingerprint. In Supplementary Note 4, we present recommendations for several other biomedical use cases.

The Metrics Reloaded online tool allows user-friendly metric selection

Selecting appropriate validation metrics while considering all potential pitfalls that may occur is a highly complex process, as demonstrated by the large number of figures in this paper. Some of the complexity, however, also results from the fact that the figures need to capture all possibilities at once. For example, many of the figures could be simplified substantially for problems based on only two classes. To leverage this potential and to improve the general user experience with our framework, we developed the Metrics Reloaded online tool (Supplementary Methods), which captures our framework in a user-centric manner and can serve as a trustworthy common access point for image analysis validation.

Discussion

Conventional scientific practice often grows through historical accretion, leading to standards that are not always well justified. This holds particularly true for the validation standards in biomedical image analysis.

The present work represents a comprehensive investigation and, importantly, constructive set of recommendations challenging the state of the art in biomedical image analysis algorithm validation with a specific focus on metrics. With the intention of revisiting—literally 're-searching'—common validation practices and developing better standards, we brought together experts from traditionally disjunct fields to leverage distributed knowledge. Our international consortium of more than 70 experts from the fields of biomedical image analysis, ML, statistics, epidemiology, biology and medicine, representing a large number of relevant biomedical imaging initiatives and societies, developed the Metrics Reloaded framework that offers guidelines and tools to choose performance metrics in a problem-aware manner. The expert consortium was primarily compiled in a way to cover the required expertise from various fields but also consisted of researchers of different countries, (academic) ages, roles and backgrounds (Supplementary Methods). Importantly, Metrics Reloaded comprehensively addresses all pitfalls related to metric selection (Table 1) and application (Extended Data Table 1) that were identified in this work's sister publication[19].

Metrics Reloaded is the result of a 2.5-year long process involving numerous workshops, surveys and expert group meetings. Many controversial debates were conducted during this time. Even deciding on the exact scope of the paper was anything but trivial. Our consortium eventually agreed on focusing on biomedical classification problems with categorical reference data and thus exploiting synergies across classification scales. Generating and handling fuzzy reference data (for example, from multiple observers) is a topic of its own[31,32] and was decided to be out of scope for this work. Furthermore, the inclusion of calibration metrics in addition to discrimination metrics was originally not intended because calibration is a complex topic, and the corresponding field is relatively young and currently highly dynamic.

Table 1 | Metrics Reloaded addresses common and rare pitfalls in metric selection, as compiled in ref. 19

Source of pitfall	Addressed in Metrics Reloaded by
Inadequate choice of the problem category	
Wrong choice of problem category	Problem category mapping (subprocess S1; Extended Data Fig. 1) as a prerequisite for metric selection.
Disregard of the domain interest	
Importance of structure boundaries	FP2.1 - Particular importance of structure boundaries; recommendation to complement common overlap-based segmentation metrics with boundary-based metrics (Fig. 2 and Supplementary Note 2.3) if the property holds.
Importance of structure volume	FP2.2 - Particular importance of structure volume; recommendation to complement common overlap-based and boundary-based segmentation metrics with volume-based metrics (Supplementary Note 2.3) if the property holds.
Importance of structure center(line)	FP2.3 - Particular importance of structure center(line); recommendation of the clDice as alternative to the common DSC or IoU in segmentation problems (subprocess S6; Extended Data Fig. 6) and recommendation of center point-based localization criterion in ObD (subprocess S8; Extended Data Fig. 8) if the property holds.
Importance of confidence awareness	FP2.7.1 - Calibration assessment requested; dedicated recommendations on calibration (Supplementary Note 2.6).
Importance of comparability across datasets	FP4.2 - Provided class prevalences reflect the population of interest; used in the subprocesses S2–S4 (Extended Data Figs. 2–4); general focus on prevalence dependency of metrics in the framework.
Unequal severity of class confusions	FP2.5 - Penalization of errors; recommendation of the so-far uncommon metric EC as a classification metric (subprocess S2; Extended Data Fig. 2); setting β in the F_β score according to preference for FP (oversegmentation) and FN (undersegmentation; see DG3.3 in Supplementary Note 2.7.2).
Importance of cost–benefit analysis	FP2.6 - Decision rule applied to predicted class scores: incorporation of a decision rule that is based on cost–benefit analysis; recommendation of the so-far uncommon metrics NB (Fig. SN 3.11) and EC (Fig. SN 3.6).
Disregard of target structure properties	
Small structure sizes	FP3.1 - Small size of structures relative to pixel size; recommendation to consider the problem an ObD problem (Supplementary Note 2.4); complementation of overlap-based segmentation metrics with boundary-based metrics in the case of small structures with noisy reference (subprocess S6; Extended Data Fig. 6); recommendation of lower ObD localization threshold in case of small sizes (see DG8.3 in Supplementary Note 2.7.7).
High variability of structure sizes	FP3.2 - High variability of structure sizes; recommendation of lower ObD localization threshold (see DG8.3 in Supplementary Note 2.7.7) and size stratification (Supplementary Note 2.4) in case of size variability.
Complex structure shapes	FP3.3 - Target structures feature tubular shape; recommendation of the clDice as alternative to the common DSC in segmentation problems (subprocess S6; Extended Data Fig. 6) and recommendation of point inside mask/box/approx as localization criterion in ObD if the property holds (subprocess S8; Extended Data Fig. 8).
Occurrence of overlapping or touching structures	FP3.5 - Possibility of overlapping or touching target structures; explicit recommendation to phrase problem as InS rather than SemS problem (Supplementary Note 2.3); recommendation of higher ObD localization threshold in case of small sizes (see DG8.3 in Supplementary Note 2.7.7).
Occurrence of disconnected structures	FP3.6 - Possibility of disconnected target structure(s); recommendation of appropriate localization criterion for ObD (DG8.2 in Supplementary Note 2.7.7).
Disregard of dataset properties	
High class imbalance	FP4.1 - High class imbalance and FP2.5.5 - compensation for class imbalances requested; compensation of class imbalance via prevalence-independent metrics such as EC and BA.
Small test set size	Recommendation of confidence intervals for all metrics.
Imperfect reference standard: noisy reference standard	FP4.3.1 - High inter-rater variability and FP2.5.7 - compensation for annotation imprecisions requested; default recommendation of the so-far rather uncommon metric NSD to assess the quality of boundaries.
Imperfect reference standard: spatial outliers in reference	FP4.3.2 - Possibility of spatial outliers in reference annotation and FP2.5.6 - handling of spatial outliers; recommendation of outlier-robust metrics, such as NSD in case no distance-based penalization of outliers is requested in segmentation problems.
Occurrence of cases with an empty reference	FP4.6 - Possibility of reference without target structure(s); recommendations for aggregation in the case of empty references according to Supplementary Note 2.4 and Extended Data Table 1.
Disregard of algorithm output properties	
Possibility of empty prediction	FP5.2 - Possibility of algorithm output not containing the target structure(s); selection of appropriate aggregation strategy in ObD (Supplementary Note 2.4).
Possibility of overlapping predictions	FP5.4 - Possibility of overlapping predictions; recommendation of an assignment strategy based on IoU > 0.5 if overlapping predictions are not possible and no predicted class scores are available.
Lack of predicted class scores	FP5.1 - Availability of predicted class scores; leveraging class scores for optimizing decision regions (FP2.6) and assessing calibration quality (FP2.7).

The first column lists all pitfall sources captured by the published taxonomy that relate to either the inadequate choice of the problem category or poor metric selection. The second column summarizes how Metrics Reloaded addresses these pitfalls. The notation FP*X.Y* refers to a fingerprint item (Supplementary Note 1.3).

This decision was reversed due to high demand from the community, expressed through crowdsourced feedback on the framework.

Extensive discussions also evolved around the inclusion criteria for metrics, considering the trade-off between established (potentially flawed) and new (not yet stress-tested) metrics. Our strategy for arriving at the Metrics Reloaded recommendations balanced this trade-off by using common metrics as a starting point and making adaptations where needed. For example, weighted Cohen's kappa, originally designed for assessing inter-rater agreement, is the state-of-the-art metric used in the medical imaging community when

handling ordinal data. Unlike other common multi-class metrics, such as (balanced) accuracy or MCC, it allows the user to specify different costs for different class confusions, thereby addressing the ordinal rating. However, our consortium deemed the (not widely known) metric EC generally more appropriate due to its favorable mathematical properties. Importantly, our framework does not intend to impose recommendations or act as a 'black box'; instead, it enables users to make educated decisions while considering ambiguities and trade-offs that may occur. This is reflected by our use of decision guides (Supplementary Note 2.7), which actively involve users in the decision-making process (for the example above, for instance, see DG2.1).

An important further challenge that our consortium faced was how to best provide recommendations in case multiple questions are asked for a single given dataset. For example, a clinician's ultimate interest may lie in assessing whether tumor progress has occurred in a patient. While this would be phrased as an ImLC task (given two images as input), an interesting surrogate task could be seen in a segmentation task assessing the quality of tumor delineation and providing explainability for the results. Metrics Reloaded addresses the general challenge of multiple different driving biomedical questions corresponding to one dataset pragmatically by generating a recommendation separately for each question. The same holds true for multi-label problems, for example, when multiple different types of abnormalities potentially co-occur in the same image/patient.

Another key challenge we faced was the validation of our framework due to the lack of ground truth 'best metrics' to be applied for a given use case. Our solution builds upon three pillars. Firstly, we adopted established consensus building approaches utilized for developing widely used guidelines such as CONSORT[21], TRIPOD[22] or STARD[23]). Secondly, we challenged our initial recommendation framework by acquiring feedback via a social media campaign. Finally, we instantiated the final framework to a range of different biological and medical use cases. Our approach showcases the benefit of crowdsourcing as a means of expanding the horizon beyond the knowledge peculiar to specific scientific communities. The most prominent change effected in response to the social media feedback was the inclusion of the aforementioned EC, a powerful metric from the speech recognition community. Furthermore, upon popular demand, we added recommendations on assessing the interpretability of model outputs, now captured by subprocess S5 (Extended Data Fig. 5).

After many highly controversial debates, the consortium ultimately converged on a consensus recommendation, as indicated by the high agreement in the final Delphi process (median agreement with the subprocesses: 93%). While some subprocesses (S1, S7 and S8) were unanimously agreed on without a single negative vote, several issues were raised by individual researchers. While most of them were minor (for example, concerning wording), a major debate revolved around calibration metrics. Some members, for example, questioned the value of stand-alone calibration metrics altogether. The reason for this view is the critically important misconception that the predicted class scores of a well-calibrated model express the true posterior probability of an input belonging to a certain class[33]—for example, a patient's risk for a certain condition based on an image. As this is not the case, several researchers argued for basing calibration assessment solely on proper scoring rules (such as the Brier score), which assess the quality of the posteriors better than the stand-alone calibration metrics. We have addressed all these considerations in our recommendation framework including a detailed rationale for our recommendations (Supplementary Note 2.6).

While we believe our framework covers the vast majority of biomedical image analysis use cases, suggesting a comprehensive set of metrics for every possible biomedical problem may be out of its scope. The focus of our framework lies in correcting poor practices related to the selection of common metrics. However, in some use cases, common reference-based metrics—as a matter of principle —be unsuitable. In fact, the use of application-specific metrics may be required in some cases. A prominent example are InS problems in which the matching of reference and predicted instances is infeasible, causing overlap-based localization criteria to fail. Metrics such as the Rand index[34] and variation of information[35] address this issue by avoiding one-to-one correspondence between predicted and reference instances. To make our framework applicable to such specific use cases, we integrated the step of choosing application-specific metrics in the main workflow (Fig. 2). Examples of such application-specific metrics can be found in related work[36,37].

Metrics Reloaded primarily provides guidance for the selection of metrics that measure some notion of the 'correctness' of an algorithm's predictions on a set of test cases. It should be noted that holistic algorithm performance assessment also includes other aspects. One of them is robustness. For example, the accuracy of an algorithm for detecting disease in medical scans should ideally be the same across different hospitals that may use different acquisition protocols or scanners from different manufacturers. Recent work, however, shows that even the exact same models with nearly identical test set performance in terms of predictive accuracy may behave very differently on data from different distributions[38].

Reliability is another important algorithmic property to be taken into account during validation. A reliable algorithm should have the ability to communicate its confidence and raise a flag when the uncertainty is high and the prediction should be discarded[39]. For calibrated models, this can be achieved via the predicted class scores, although other methods based on dedicated model outputs trained to express the confidence or on density estimation techniques are similarly popular. Importantly, an algorithm with reliable uncertainty estimates or increased robustness to distribution shift might not always be the best performing in terms of predictive performance[40]. For safe use of classification systems in practice, careful balancing of the trade-off between robustness and reliability over accuracy might be necessary.

So far, Metrics Reloaded focuses on common reference-based methods that compare model outputs to corresponding reference annotations. We made this design choice due to our hypothesis that reference-based metrics can be chosen in a modality-agnostic and application-agnostic manner using the concept of problem fingerprinting. As indicated by the step of choosing potential non-reference-based metrics (Fig. 2), however, it should be noted that validation and evaluation of algorithms should go far beyond purely technical performance[41,42]. In this context, Jannin introduced the global concept of 'responsible research' to encompass all possible high-level assessment aspects of a digital technology[43], including environmental, ethical, economic, social and societal aspects. For example, there are increasing efforts specifically devoted to the estimation of energy consumption and greenhouse gas emission of ML algorithms[44–46]. For these considerations, we refer the reader to available tools such as the Green Algorithms calculator[47] or Carbontracker[48].

It must further be noted that while Metrics Reloaded places a focus on the selection of metrics, adequate application is also important. Detailed failure case analysis[49] and performance assessment on relevant subgroups, for example, have been highlighted as critical components for better understanding when and where an algorithm may fail[50,51]. Given that learning-based algorithms rely on the availability of historical datasets for training, there is a real risk that any existing biases in the data may be picked up and replicated or even exacerbated when an algorithm makes predictions[52,53]. This is of particular concern in the context of systemic biases in healthcare, such as the scarcity of representative data from underserved populations and often higher error rates in diagnostic labels in particular subgroups[54,55]. Relevant meta information such as patient demographics, including biological sex and ethnicity, needs to be accessible for the test sets such that

potentially disparate performance across subgroups can be detected[56]. Here, it is important to make use of adequate aggregations over the validation metrics as disparities in minority groups might otherwise be missed.

Finally, it must be noted that our framework addresses metric choice in the context of technical validation of biomedical algorithms. For translation of an algorithm into, for example, clinical routine, this validation may be followed by a (clinical) validation step assessing its performance compared to conventional, non-algorithm-based care according to patient-related outcome measures, such as overall survival[57].

A key remaining challenge for Metric Reloaded is its dissemination such that it will substantially contribute to raising the quality of biomedical imaging research. To encourage widespread adherence to new standards, entry barriers should be as low as possible. While the framework with its vast number of subprocesses may seem very complex at first, it is important to note that from a user perspective only a fraction of the framework is relevant for a given task, making the framework more tangible. This is notably illustrated by the Metric Reloaded online tool, which substantially simplifies the metric selection procedure. As is common in scientific guideline and recommendation development, we intend to regularly update our framework to reflect current developments in the field, such as the inclusion of new metrics or biomedical use cases. This is intended to include an expansion of the framework's scope to further problem categories, such as regression and reconstruction. To accommodate future developments in a fast and efficient manner, we envision our consortium building consensus through accelerated Delphi rounds organized by the Metric Reloaded core team. Once consensus is obtained, changes will be implemented in both the framework and online tool and highlighted so that users can easily identify changes to the previous version, which will ensure full transparency and comparability of results. In this way, we envision the Metrics Reloaded framework and online tool as a dynamic resource reliably reflecting the current state of the art at any given time point in the future, for years to come[18].

Of note, while the provided recommendations originate from the biomedical image analysis community, many aspects generalize to imaging research as a whole. Particularly, the recommendations derived for individual fingerprints (for example, implications of class imbalance) hold across domains, although it is possible that for different domains the existing fingerprints would need to be complemented by further features that this community is not aware of.

In conclusion, the Metrics Reloaded framework provides biomedical image analysis researchers with systematic guidance on choosing validation metrics across different imaging tasks in a problem-aware manner. Through its reliance on methodology that can be generalized, we envision the Metrics Reloaded framework to spark a scientific debate and hopefully lead to similar efforts being undertaken in other areas of imaging research, thereby raising research quality on a much larger scale than originally anticipated. In this context, our framework and the process by which it was developed could serve as a blueprint for broader efforts aimed at providing reliable recommendations and enforcing adherence to good practices in imaging research.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

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

1. Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G. & King, D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med.* **17**, 195 (2019).
2. Shah, N. H., Milstein, A. & Bagley, S. C. Making machine learning models clinically useful. *JAMA* **322**, 1351–1352 (2019).
3. Correia, P. & Pereira, F. Video object relevance metrics for overall segmentation quality evaluation. *EURASIP J. Adv. Signal Process.* **2006**, 082195 (2006).
4. 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).
5. Honauer, K., Maier-Hein, L. and 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 (2015).
6. 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).
7. 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).
8. L. Maier-Hein, et al. Why rankings of biomedical image analysis competitions should be interpreted with care. *Nat. Comm.* **9**, 5217 (2018).
   With this comprehensive analysis of biomedical image analysis competitions (challenges), the authors initiated a shift in how such challenges are designed, performed, and reported in the biomedical domain. Its concepts and guidelines have been adopted by reputed organizations such as MICCAI.
9. Margolin, R., Zelnik-Manor, L., and Tal, A. How to evaluate foreground maps? In *Proceedings of the IEEE Conference On Computer Vision And Pattern Recognition* 248–255 (2014).
10. Tran, T. N. et al. Sources of performance variability in deep learning-based polyp detection. *Int. J. Comput. Assist. Radiol. Surg.* **18**, 1311–1322 (2023).
11. 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).
12. Chenouard, N. et al. Objective comparison of particle tracking methods. *Nat. Methods* **11**, 281–289 (2014).
13. Sage, D. et al. Quantitative evaluation of software packages for single-molecule localization microscopy. *Nat. Methods* **12**, 717–724 (2015).
14. Ulman, V. et al. An objective comparison of cell-tracking algorithms. *Nat. Methods* **14**, 1141–1152 (2017).
15. Carass, A. et al. Evaluating white matter lesion segmentations with refined Sørensen-Dice analysis. *Sci. Rep.* **10**, 8242 (2020).
16. Jäger, P. F. Challenges and opportunities of end-to-end learning in medical image classification. *Karlsruher Institut für Technologie* (2020).
17. Bernice B. B. Delphi process: a methodology used for the elicitation of opinions of experts. Technical report, The RAND Corporation (1968).
18. Nasa, P., Jain, R. & Juneja, D. Delphi methodology in healthcare research: how to decide its appropriateness. *World J. Methodol.* **11**, 116–129 (2021).
19. Reinke, A. et al. Understanding metric-related pitfalls in image analysis validation. *Nat. Methods* https://doi.org/10.1038/s41592-023-02150-0 (2023).
    Sister publication jointly submitted with this work.
20. Reinke, A. et al. How to exploit weaknesses in biomedical challenge design and organization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (eds. A. F. Frangi et al.) 388–395 (Springer, 2018).

21. Schulz, K. F., Altman, D. G., Moher, D. & CONSORT Group. Consort 2010 statement: updated guidelines for reporting parallel group randomized trials. *Ann. Intern. Med.* **152**, 726–732 (2010).
22. Moons, K. G. M. et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): explanation and elaboration. *Ann. Intern. Med.* **162**, W1–W73 (2015).
23. Bossuyt, P. M. et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the stard initiative. *Ann. Intern. Med.* **138**, 40–44 (2003).
24. Vickers, A. J., Van Calster, B. & Steyerberg, E. W. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ* **352**, i6 (2016).
25. van Leeuwen, D. A. & Brümmer, N. An introduction to application-independent evaluation of speaker recognition systems. In *Speaker classification I* (ed. C. Muller) 330–353 (Springer, 2007).
26. Ferrer, L. Analysis and comparison of classification metrics. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2209.05355 (2022).
The document discusses common performance metrics used in machine learning classification, and introduces the EC metric. It compares these metrics and argues that EC is superior due to its generality, simplicity and intuitive nature. Additionally, it highlights the potential of EC in measuring calibration and optimal decision-making using class posteriors.
27. 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).
28. Gruber, S. & Buettner, F. Better uncertainty calibration via proper scores for classification and beyond. *Adv. Neural Inform. Process Syst.* **35**, 8618–8632 (2022).
29. Kirillov, A., He, K., Girshick, R., Rother, C. and Dollár, P. Panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 9404–9413 (2019).
30. Wiesenfarth, M. et al. Methods and open-source toolkit for analyzing and visualizing challenge results. *Sci. Rep.* **11**, 2369 (2021).
31. Liu, X. et al. Baseline photos and confident annotation improve automated detection of cutaneous graft-versus-host disease. *Clin. Hematol. Int.* **3**, 108–115(2021).
32. Taha, A. A. & Hanbury, A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Med. Imaging* **15**, 29 (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.
33. Perez-Lebel, A., Le Morvan, M., and Varoquaux, G. Beyond calibration: estimating the grouping loss of modern neural networks. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2210.16315 (2023).
34. Rand, W. M. Objective criteria for the evaluation of clustering methods. *J. Am. Stat. Assoc.* **66**, 846–850 (1971).
35. Meilă, M. Comparing clusterings by the variation of information. In *Learning Theory and Kernel Machines* 173–187 (Springer, 2003).
36. Côté, M. A. et al. Tractometer: towards validation of tractography pipelines. *Medical Image Analysis* https://doi.org/10.1016/j.media.2013.03.009. (2013)
37. Ellis, D. G., Alvarez, C. M. and Aizenberg, M. R. Qualitative criteria for feasible cranial implant designs. In *Cranial Implant Design Challenge* 8–18 (Springer, 2021).
38. D'Amour, A. et al. Underspecification presents challenges for credibility in modern machine learning. *J. Mach. Learn. Res.* **23**, 10237–10297 (2022).
39. Schulam, P. & Saria, S. Can you trust this prediction? Auditing pointwise reliability after learning. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics* (eds. Chaudhuri, K. & Sugiyama, M.) Vol. 89, 1022–1031 (PMLR, 2019).
40. P. F. Jaeger, Carsten T. Lüth, Lukas Klein, and Till J. Bungert. A call to reflect on evaluation practices for failure detection in image classification. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2211.15259 (2023).
41. Université de Montréal. The Declaration - Montreal Responsible AI, 2017. https://declarationmontreal-iaresponsable.com/
42. The Institute for Ethical Ai and Machine Learning. https://ethical.institute/principles.html. Accessed 5/21/2022 (2018).
43. Jannin, P. Towards responsible research in digital technology for health care. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2110.09255 (2021).
44. Lacoste, A., Luccioni, A., Schmidt, V., and Dandres, T. Quantifying the carbon emissions of machine learning. Preprint at https://arxiv.org/abs/1910.09700 (2019).
45. Patterson, D., et al. Carbon emissions and large neural network training. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2104.10350 (2021).
46. Strubell, E., Ganesh, A., and McCallum, A. Energy and policy considerations for deep learning in NLP. Preprint at https://doi.org/10.48550/arXiv.1906.02243 (2019).
47. Lannelongue, L., Grealey, J. & Inouye, M. Green algorithms: quantifying the carbon footprint of computation. *Adv. Sci.* **8**, 2100707 (2021).
48. Anthony, L. F. W., Kanding, B., and Selvan, R. Carbontracker: tracking and predicting the carbon footprint of training deep learning models. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2007.03051 (2020).
49. Roß, T. et al. Beyond rankings: learning (more) from algorithm validation. *Med. Image Anal.* **86**, 102765 (2023).
50. Char, D. S., Shah, N. H. & Magnus, D. Implementing machine learning in health care - addressing ethical challenges. *N. Engl. J. Med.* **378**, 981–983 (2018).
51. Oakden-Rayner, L., Dunnmon, J., Carneiro, G., and Ré, C. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. *Proc. ACM Conf. Health Inference Learn* **2020**, 151–159 (2020).
52. Adamson, A. S. & Smith, A. Machine learning and health care disparities in dermatology. *JAMA Dermatol.* **154**, 1247–1248 (2018).
53. Geirhos, R. et al. Shortcut learning in deep neural networks. *Nat. Mach. Intell.* **2**, 665–673 (2020).
54. Ibrahim, H., Liu, X., Zariffa, N., Morris, A. D. & Denniston, A. K. Health data poverty: an assailable barrier to equitable digital health care. *Lancet Digit. Health* **3**, e260–e265 (2021).
55. Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **366**, 447–453 (2019).
56. McCradden, M. D. et al. A research ethics framework for the clinical translation of healthcare machine learning. *Am. J. Bioeth.* **22**, 8–22 (2022).
57. Park, S. H. et al. Methods for clinical evaluation of artificial intelligence algorithms for medical diagnosis. *Radiology* https://doi.org/10.1148/radiol.220182 (2023).
58. Usatine, R. & Manci, R. Dermoscopedia https://dermoscopedia.org/File:DF_chinese_dms.JPG (2021).

59. Armato, S. G. III et al. The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): a completed reference database of lung nodules on CT scans. *Med. Phys.* **38**, 915–931 (2011).
60. Ljosa, V., Sokolnicki, K. L. & Carpenter, A. E. Annotated high-throughput microscopy image sets for validation. *Nat. Methods* **9**, 637 (2012).
61. Maier-Hein, L. et al. Heidelberg colorectal data set for surgical data science in the sensor operating room. *Sci. Data* **8**, 101 (2021).
62. Haugen, T. B. et al. Visem: a multimodal video dataset of human spermatozoa. In *Proceedings of the 10th ACM Multimedia Systems Conference* 261–266 (2019).
63. Codella, N. et al. Skin lesion analysis toward melanoma detection 2018: a challenge hosted by the international skin imaging collaboration (ISIC). Preprint at *arXiv* https://doi.org/10.48550/arXiv.1902.03368 (2019).
64. Targosz, A., Przystałka, P., Wiaderkiewicz, R. & Mrugacz, G. Semantic segmentation of human oocyte images using deep neural networks. *Biomed. Eng. Online* **20**, 40 (2021).
65. Antonelli, M. et al. The medical segmentation decathlon. *Nat. Commun.* **13**, 4128 (2022).
66. Simpson, A. L. et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. Preprint at https://doi.org/10.48550/arXiv.1902.09063 (2019).
67. Nagao, Y., Sakamoto, M., Chinen, T., Okada, Y. & Takao, D. Robust classification of cell cycle phase and biological feature extraction by image-based deep learning. *Mol. Biol. Cell* **31**, 1346–1354 (2020).
68. Zhang, Y. et al. Deepphagy: a deep learning framework for quantitatively measuring autophagy activity in saccharomyces cerevisiae. *Autophagy* **16**, 626–640 (2020).
69. Commowick, O. et al. Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure. *Sci. Rep.* **8**, 13650 (2018).
70. Kofler, F. et al. blob loss: instance imbalance aware loss functions for semantic segmentation. In *International Conference on Information Processing in Medical Imaging* 755–767 (Springer Nature Switzerland, 2023).
71. Mais, L., Hirsch, P. and Kainmueller, D. Patchperpix for instance segmentation. In *European Conference on Computer Vision* 288–304 (Springer, 2020).
72. Meissner, G. et al. A searchable image resource of *Drosophila* GAL4-driver expression patterns with single neuron resolution. *eLife* **12**, e80660 (2023).
73. Tirian, L. & Dickson, B. J. The VT GAL4, Lexa, and split-GAL4 driver line collections for targeted expression in the *Drosophila* nervous system. Preprint at *bioRxiv* https://doi.org/10.1101/198648 (2017).
74. Brümmer, N. & Du Preez, J. Application-independent evaluation of speaker detection. *Comput. Speech Lang.* **20**, 230–275 (2006).

Acknowledgements

This work was initiated by the Helmholtz Association of German Research Centers in the scope of the Helmholtz Imaging Incubator (HI), the MICCAI Special Interest Group on biomedical image analysis challenges and the benchmarking working group of the MONAI initiative. It received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement no. 101002198, NEURAL SPICING). It was further supported in part by the Intramural Research Program of the National Institutes of Health (NIH) Clinical Center as well as by the National Cancer Institute (NCI) and the National Institute of Neurological Disorders and Stroke (NINDS) of the 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, the Natural Sciences and Engineering Research Council of Canada. F.B. was co-funded by the European Union (ERC, TAIPO, 101088594). Views and opinions expressed are those of the authors only and do not necessarily reflect those of the European Union or the ERC. Neither the European Union nor the granting authority can be held responsible for them. V.C. acknowledges funding from Novo Nordisk 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 (program grant no. 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 supported by the Ministry of Education, Youth and Sports of the Czech Republic (project LM2018129). T.K. was supported in part by 4UH3-CA225021-03, 1U24CA180924-01A1, 3U24CA215109-02 and 1UG3-CA225-021-01 grants from the NIH. G.L. receives research funding from the Dutch Research Council, the Dutch Cancer Association, HealthHolland, the ERC, the European Union and the Innovative Medicine Initiative. C.H.S. is supported by an Alzheimer's Society Junior Fellowship (AS-JF-17-011). M.R. is supported by Innosuisse (grant no. 31274.1) and Swiss National Science Foundation (grant no. 205320_212939). R.M.S. is supported by the Intramural Research Program of the NIH Clinical Center. A.T. acknowledges support from the Academy of Finland (Profi6 336449 funding program), University of Oulu strategic funding, Finnish Foundation for Cardiovascular Research, Wellbeing Services County of North Ostrobothnia (VTR project K62716) and 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 thank N. Sautter, P. Vieten and T. Adler for proposing the name for the project. We thank P. Bankhead, F. Hamprecht, H. Kenngott, D. Moher and B. Stieltjes for fruitful discussions on the framework. We thank S. Steger for the data protection supervision and A. Trotter for the hosting of the surveys. We thank L. Mais for instantiating the use case for InS of neurons from the fruit fly in 3D multicolor light microscopy images. We further thank the Janelia FlyLight Project Team for providing us with example images for this use case. We thank the following people for testing the metric mappings, reviewing the recommendations and performing metric-centric testing: T. Adler, C. Bender, A. B. Qasim, K. Dreher, N. Holzwarth, M. Hübner, D. Michael, L. -R. Müller, M. Rees, T. Rix, M. Schellenberg, S. Seidlitz, J. Sellner, A. Srivastava, F. Wolf, A. E. Yamlahi, S. D. Almeida, M. Baumgartner, D. Bounias, T. Bungert, M. Fischer, L. Klein, G. Köhler, B. Kovács, C. Lueth, T. Norajitra, C. Ulrich, T. Wald, I. Alekseenko, X. Liu, A. Marheim Storås and V. Thambawita. We thank the following people for taking our social media community survey and providing helpful feedback for improving the framework: Y. Akemi, R. Anteby, C. Arthurs, P. De Backer, H. Badgery, M. Baugh, J. Bernal, D. Bounias, F. C. Kitamura, J. Carse, C. Chen, I. Flipse, N. Gaggion, C. González, P. M. Gordaliza, T. Horeman, L. Joskowicz, A. Jose, A. Kamath, B. Kelly, Y. Kirchhoff, L. A. Kobelke, L. Krämer, M. Krendel, J. LaMaster, T. de Lange, J. L. Lavanchy, J. Li, C. Lüth, L. Mais, A. Marheim Storås, V. Nath, C. Scannell, C. Pape, M. P. Schijven, A. Selvanetti, B. S. Fadida, R. Staff, J. Tan, E. Tkaczyk, R. T. Calumby, A. Vlontzos, W. Zhang, C. Zhao and J. Zhu.

Author contributions

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, tested the online toolkit, and organized the social media campaign. 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, tested the metric mappings and the online toolkit, organized the social media campaign, and designed all figures. P.F.J. initiated and led the study, was a member of the Delphi core team, led the ObD and InS expert group, wrote and reviewed the manuscript, prepared and evaluated all surveys, organized all workshops, tested the metric mappings and the online toolkit, organized the social media campaign, and participated in surveys. P.G. led the ImLC expert group, was a member of the extended Delphi core team, wrote and reviewed the manuscript, prepared the BPMN diagrams, tested the online toolkit, and participated in surveys and workshops. M.D.T. was a member of the extended Delphi core team and wrote and reviewed the manuscript. F.B. led the calibration expert group, reviewed the manuscript, and participated in surveys. E.C. led the cross-topic expert group, was a member of the extended Delphi core team, and reviewed the manuscript. B.G. led the cross-topic expert group and was an active member of the SemS expert group, reviewed the manuscript, and participated in surveys and workshops. F.I. led the SemS expert group, reviewed the manuscript, tested the online toolkit, and participated in surveys and workshops. J.K. led the biomedical expert group, reviewed the manuscript, and participated in surveys and workshops. M.K. led the ObD and InS expert group, reviewed the manuscript, and participated in surveys and workshops. M.R. led the SemS expert group, reviewed the manuscript, and participated in surveys and workshops. M.A.R. led the ImLC expert group, reviewed the manuscript, tested the metric mappings, and participated in surveys and workshops. M.W. co-led the cross-topic expert group. A.E.K. implemented the online toolkit and was a member of the extended Delphi core team. C.H.S. implemented the reference implementations of all metrics in Python, was an active member of the ObD and InS expert group, reviewed the manuscript, and participated in surveys workshops. 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, tested the metric mappings and the online toolkit, and participated in surveys and workshops. M.E. was a member of the extended Delphi core team, prepared the BPMN diagrams, reviewed the document, assisted in survey preparation, tested the metric mappings and the online toolkit, and participated in surveys. D.H.-N. was a member of the extended Delphi core team and prepared all surveys. T.R. was a member of the extended Delphi core team, was an active member of the ObD and InS expert group, wrote and reviewed the document, assisted in survey preparation, tested the metric mappings and the online toolkit, 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 SemS expert group and participated in surveys and workshops. T.A. was an active member of the ObD and InS expert group, tested the metric mappings, reviewed the manuscript, and participated in surveys and workshops. S.B. co-led the SemS expert group, reviewed the manuscript, 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. M.B.B. triggered changes in the framework by responding to public questionnaire, reviewed the manuscript, and participated in surveys. M.J.C. was an active member of the 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. B.A.C. was an active member of the ObD and InS expert group, tested the metric mappings, reviewed the manuscript, 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. triggered changes in the framework by responding to public questionnaire, was an active member of the calibration expert group, reviewed the manuscript, and participated in surveys. A.G. triggered changes in the framework by responding to public questionnaire, was an active member of the calibration expert group, reviewed the manuscript, and participated in surveys. B.v.G. participated in surveys and workshops. R.H. triggered changes in the framework by responding to public questionnaire and participated in surveys. D.A.H. was an active member of the biomedical and cross-topic expert groups, reviewed the manuscript, 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. 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. triggered changes in the framework by responding to public questionnaire and participated in surveys. B.K. triggered changes in the framework by responding to public questionnaire, reviewed the manuscript, and participated in surveys. F.K. triggered changes in the framework by responding to public questionnaire 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. A.K. was an active member of the biomedical expert group, reviewed the manuscript, and 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 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. 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, and participated in surveys and workshops. H.M. was an active member of the ImLC expert group, tested the metric mappings, reviewed the manuscript, and participated in surveys and workshops. B.N. was an active member of the ObD and InS expert group, tested the metric mappings, reviewed the manuscript, and participated in surveys. 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 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, tested the metric mappings, 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 participated in surveys and workshops. G.S.C., A. Karthikesalingam, T.K., A.L.M., P.M., F.N., J.P., N. Rajpoot, J.S.-R., C.I.S., S.S. and M.v.S. served on the expert Delphi panel and participated in workshops and surveys.

Competing interests

We declare the following competing interests: Under terms of employment, M.B.B. is entitled to stock options in Mona.health,

a KU Leuven spinoff. F.B. is an employee of Siemens AG. F.B. reports funding from Merck. B.v.G. is a shareholder of Thirona. B.G. was an employee of HeartFlow and Kheiron Medical Technologies. M.M.H. received an Nvidia GPU grant. B.K. is a consultant for ThinkSono. G.L. is on the advisory board of Canon Healthcare IT and is a shareholder of Aiosyn BV. N. Rieke is an employee of NVIDIA. J.S.-R. reports funding from GSK, Pfizer and Sanofi and fees from Travere Therapeutics, Stadapharm, Astex Therapeutics, Pfizer and Grunenthal. R.M.S. receives patent royalties from iCAD, ScanMed, Philips, Translation Holdings and PingAn; the laboratory of R.M.S. received research support from PingAn through a Cooperative Research and Development Agreement. S.A.T. receives financial support from Canon Medical Research Europe. The remaining authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41592-023-02151-z.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-023-02151-z.

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

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.

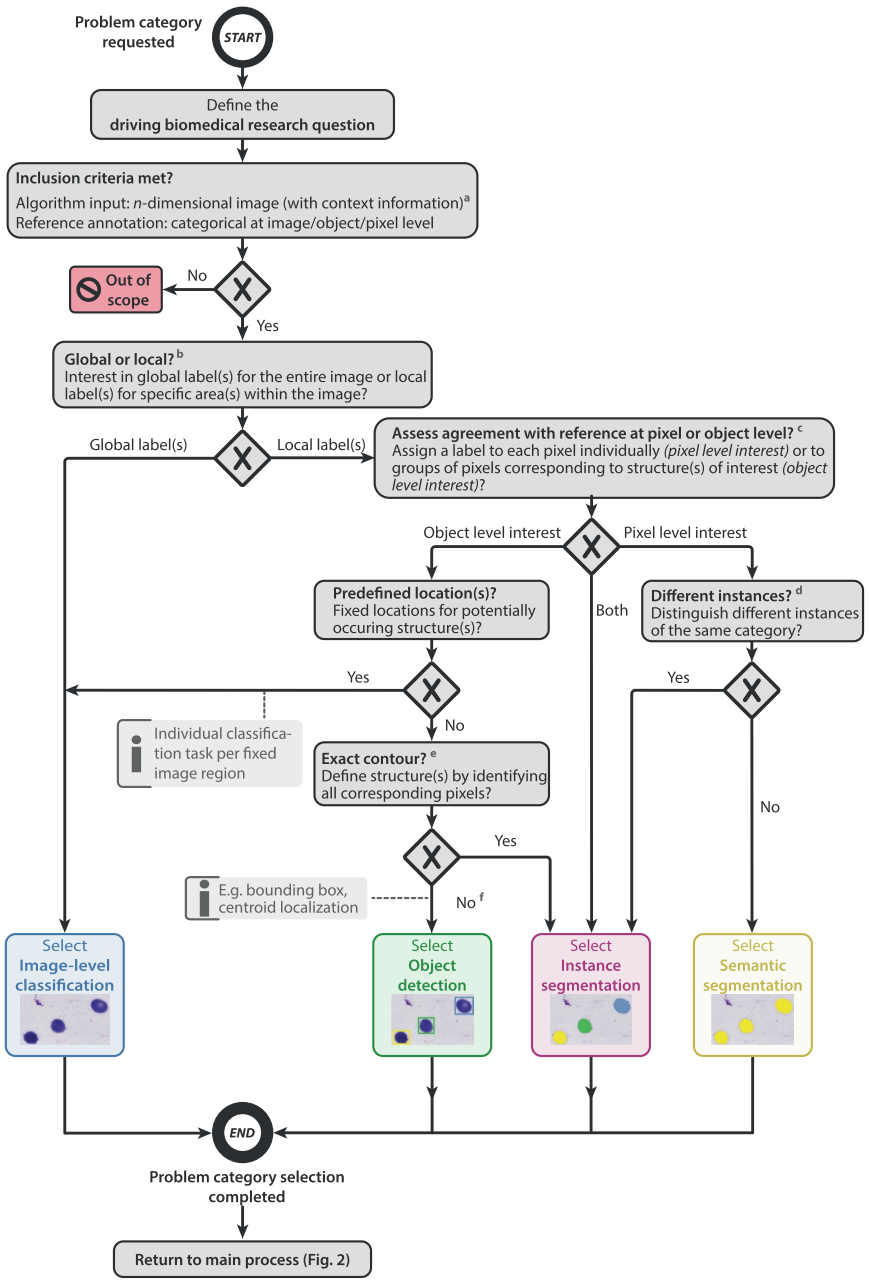
Lena Maier-Hein[1,2,3,4,5,90] ✉, Annika Reinke[1,2,3,90] ✉, Patrick Godau[1,3,5], Minu D. Tizabi[1,5], Florian Buettner[6,7,8,9,10], Evangelia Christodoulou[1], Ben Glocker[11], Fabian Isensee[12,13], Jens Kleesiek[14], Michal Kozubek[15], Mauricio Reyes[16,17], Michael A. Riegler[18,19], Manuel Wiesenfarth[20], A. Emre Kavur[1,12,13], Carole H. Sudre[21,22], Michael Baumgartner[12], Matthias Eisenmann[1], Doreen Heckmann-Nötzel[1,5], Tim Rädsch[1,2], Laura Acion[23], Michela Antonelli[22,24], Tal Arbel[25], Spyridon Bakas[26,27], Arriel Benis[28,29], Matthew B. Blaschko[30], M. Jorge Cardoso[22], Veronika Cheplygina[31], Beth A. Cimini[32], Gary S. Collins[33], Keyvan Farahani[34], Luciana Ferrer[35], Adrian Galdran[36,37], Bram van Ginneken[38,39], Robert Haase[40,41,89], Daniel A. Hashimoto[42,43], Michael M. Hoffman[44,45,46,47], Merel Huisman[48], Pierre Jannin[49,50], Charles E. Kahn[51], Dagmar Kainmueller[52,53], Bernhard Kainz[54,55], Alexandros Karargyris[56], Alan Karthikesalingam[57], Florian Kofler[58], Annette Kopp-Schneider[20], Anna Kreshuk[59], Tahsin Kurc[60], Bennett A. Landman[61], Geert Litjens[62], Amin Madani[63], Klaus Maier-Hein[12,64], Anne L. Martel[45,47,65], Peter Mattson[66], Erik Meijering[67], Bjoern Menze[68], Karel G. M. Moons[69], Henning Müller[70,71], Brennan Nichyporuk[72], Felix Nickel[73], Jens Petersen[12], Nasir Rajpoot[74], Nicola Rieke[75], Julio Saez-Rodriguez[76,77], Clara I. Sánchez[78], Shravya Shetty[79], Maarten van Smeden[69], Ronald M. Summers[80], Abdel A. Taha[81], Aleksei Tiulpin[82,83], Sotirios A. Tsaftaris[84], Ben Van Calster[85,86], Gaël Varoquaux[87] & Paul F. Jäger[2,88] ✉

[1]German Cancer Research Center (DKFZ) Heidelberg, Division of Intelligent Medical Systems, Heidelberg, Germany. [2]German Cancer Research Center (DKFZ) Heidelberg, HI Helmholtz Imaging, Heidelberg, Germany. [3]Faculty of Mathematics and Computer Science, Heidelberg University, Heidelberg, Germany. [4]Medical Faculty, Heidelberg University, Heidelberg, Germany. [5]National Center for Tumor Diseases (NCT), NCT Heidelberg, a partnership between DKFZ and University Medical Center Heidelberg, Heidelberg, Germany. [6]German Cancer Consortium (DKTK), partner site Frankfurt/Mainz, a partnership between DKFZ and UCT Frankfurt-Marburg, Frankfurt am Main, Germany. [7]German Cancer Research Center (DKFZ) Heidelberg, Heidelberg, Germany. [8]Department of Medicine, Goethe University Frankfurt, Frankfurt am Main, Germany. [9]Department of Informatics, Goethe University Frankfurt, Frankfurt am Main, Germany. [10]Frankfurt Cancer Insititute, Frankfurt am Main, Germany. [11]Department of Computing, Imperial College London, South Kensington Campus, London, UK. [12]German Cancer Research Center (DKFZ) Heidelberg, Division of Medical Image Computing, Heidelberg, Germany. [13]German Cancer Research Center (DKFZ) Heidelberg, HI Applied Computer Vision Lab, Heidelberg, Germany. [14]Institute for AI in Medicine, University Medicine Essen, Essen, Germany. [15]Centre for Biomedical Image Analysis and Faculty of Informatics, Masaryk University, Brno, Czech Republic. [16]ARTORG Center for Biomedical Engineering Research, University of Bern, Bern, Switzerland. [17]Department of Radiation Oncology, University Hospital Bern, University of Bern, Bern, Switzerland. [18]Simula Metropolitan Center for Digital Engineering, Oslo, Norway. [19]Department of Computer Science, UiT The Arctic University of Norway, Tromsø, Norway. [20]German Cancer Research Center (DKFZ) Heidelberg, Division of Biostatistics, Heidelberg, Germany. [21]MRC Unit for Lifelong Health and Ageing at UCL and Centre for Medical Image Computing, Department of Computer Science, University College London, London, UK. [22]School of Biomedical Engineering and Imaging Science, King's College London, London, UK. [23]Instituto de Cálculo, CONICET – Universidad de Buenos Aires, Buenos Aires, Argentina. [24]Centre for Medical Image Computing, University College London, London, UK. [25]Centre for Intelligent Machines and MILA (Québec Artificial Intelligence Institute), McGill University, Montréal, Quebec, Canada. [26]Division of Computational Pathology, Department of Pathology & Laboratory Medicine, Indiana University School of Medicine, IU Health Information and Translational Sciences Building, Indianapolis, IN, USA. [27]Center for Biomedical Image Computing and Analytics (CBICA), University of Pennsylvania, Philadelphia, PA, USA. [28]Department of Digital Medical Technologies, Holon Institute of Technology, Holon, Israel. [29]European Federation for Medical Informatics, Le Mont-sur-Lausanne, Switzerland. [30]Center for Processing Speech and Images, Department of Electrical Engineering, KU Leuven, Leuven, Belgium. [31]Department of Computer Science, IT University of Copenhagen, Copenhagen, Denmark. [32]Imaging Platform, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [33]Centre for Statistics in Medicine, University of Oxford, Nuffield Orthopaedic Centre, Oxford, UK. [34]Center for Biomedical Informatics and Information Technology,

National Cancer Institute, Bethesda, MD, USA. [35]Instituto de Investigación en Ciencias de la Computación (ICC), CONICET-UBA, Ciudad Autónoma de Buenos Aires, Buenos Aires, Argentina. [36]BCN Medtech, Universitat Pompeu Fabra, Barcelona, Spain. [37]Australian Institute for Machine Learning AIML, University of Adelaide, Adelaide, South Australia, Australia. [38]Fraunhofer MEVIS, Bremen, Germany. [39]Radboud Institute for Health Sciences, Radboud University Medical Center, Nijmegen, the Netherlands. [40]Technische Universität (TU) Dresden, DFG Cluster of Excellence 'Physics of Life', Dresden, Germany. [41]Center for Systems Biology, Dresden, Germany. [42]Department of Surgery, Perelman School of Medicine, Philadelphia, PA, USA. [43]General Robotics Automation Sensing and Perception Laboratory, School of Engineering and Applied Science, University of Pennsylvania, Philadelphia, PA, USA. [44]Princess Margaret Cancer Centre, University Health Network, Toronto, Ontario, Canada. [45]Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada. [46]Vector Institute for Artificial Intelligence, Toronto, Ontario, Canada. [47]Department of Computer Science, University of Toronto, Toronto, Ontario, Canada. [48]Department of Radiology and Nuclear Medicine, Radboud University Medical Center, Nijmegen, the Netherlands. [49]Laboratoire Traitement du Signal et de l'Image – UMR_S 1099, Université de Rennes 1, Rennes, France. [50]INSERM, Paris, France. [51]Department of Radiology and Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA, USA. [52]Max-Delbrück Center for Molecular Medicine in the Helmholtz Association (MDC), Biomedical Image Analysis and HI Helmholtz Imaging, Berlin, Germany. [53]Digital Engineering Faculty, University of Potsdam, Potsdam, Germany. [54]Department of Computing, Faculty of Engineering, Imperial College London, London, UK. [55]Department AIBE, Friedrich-Alexander-Universität (FAU), Erlangen-Nürnberg, Germany. [56]IHU Strasbourg, Strasbourg, France. [57]Google Health DeepMind, London, UK. [58]Helmholtz AI, Oberschleißheim, Germany. [59]Cell Biology and Biophysics Unit, European Molecular Biology Laboratory (EMBL), Heidelberg, Germany. [60]Department of Biomedical Informatics, Stony Brook University, Health Science Center, Stony Brook, NY, USA. [61]Electrical Engineering, Vanderbilt University, Nashville, TN, USA. [62]Department of Pathology, Radboud University Medical Center, Nijmegen, the Netherlands. [63]Department of Surgery, University Health Network, Philadelphia, PA, USA. [64]Pattern Analysis and Learning Group, Department of Radiation Oncology, Heidelberg University Hospital, Heidelberg, Germany. [65]Physical Sciences, Sunnybrook Research Institute, Toronto, Ontario, Canada. [66]Google, 1600 Amphitheatre Pkwy, Mountain View, CA, USA. [67]School of Computer Science and Engineering, University of New South Wales, UNSW Sydney, Kensington, New South Wales, Australia. [68]Department of Quantitative Biomedicine, University of Zurich, Zurich, Switzerland. [69]Julius Center for Health Sciences and Primary Care, UMC Utrecht, Utrecht University, Utrecht, the Netherlands. [70]Information Systems Institute, University of Applied Sciences Western Switzerland (HES-SO), Sierre, Switzerland. [71]Medical Faculty, University of Geneva, Geneva, Switzerland. [72]MILA (Québec Artificial Intelligence Institute), Montréal, Quebec, Canada. [73]Department of General, Visceral and Thoracic Surgery, University Medical Center Hamburg-Eppendorf, Hamburg, Germany. [74]Tissue Image Analytics Laboratory, Department of Computer Science, University of Warwick, Coventry, UK. [75]NVIDIA, München, Germany. [76]Institute for Computational Biomedicine, Heidelberg University, Heidelberg, Germany. [77]Faculty of Medicine, Heidelberg University Hospital, Heidelberg, Germany. [78]Informatics Institute, Faculty of Science, University of Amsterdam, Amsterdam, the Netherlands. [79]Google Health, Google, Palo Alto, CA, USA. [80]National Institutes of Health Clinical Center, Bethesda, MD, USA. [81]Institute of Information Systems Engineering, TU Wien, Vienna, Austria. [82]Research Unit of Health Sciences and Technology, Faculty of Medicine, University of Oulu, Oulu, Finland. [83]Neurocenter Oulu, Oulu University Hospital, Oulu, Finland. [84]School of Engineering, The University of Edinburgh, Edinburgh, Scotland. [85]Department of Development and Regeneration and EPI-centre, KU Leuven, Leuven, Belgium. [86]Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, the Netherlands. [87]Parietal project team, INRIA Saclay-Île de France, Palaiseau, France. [88]German Cancer Research Center (DKFZ) Heidelberg, Interactive Machine Learning Group, Heidelberg, Germany. [89]Present address: Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI), Leipzig University, Leipzig, Germany. [90]These authors contributed equally: Lena Maier-Hein, Annika Reinke. ✉e-mail: l.maier-hein@dkfz-heidelberg.de; a.reinke@dkfz-heidelberg.de; p.jaeger@dkfz-heidelberg.de

https://doi.org/10.1038/s41592-023-02151-z

+S1



[a] Context data: For example, medical images may be processed along with clinical data; video frames may be processed along with preceeding video snippets.

[b] If the interest is global, a single predicted class score for the entire image is compared to a global reference; otherwise, predicted class scores per pixel or object are compared to the corresponding reference.

[c] If validation at object level is desired, a single predicted score for an entire group of pixels (corresponding to an object) is compared to a single reference label for this object.

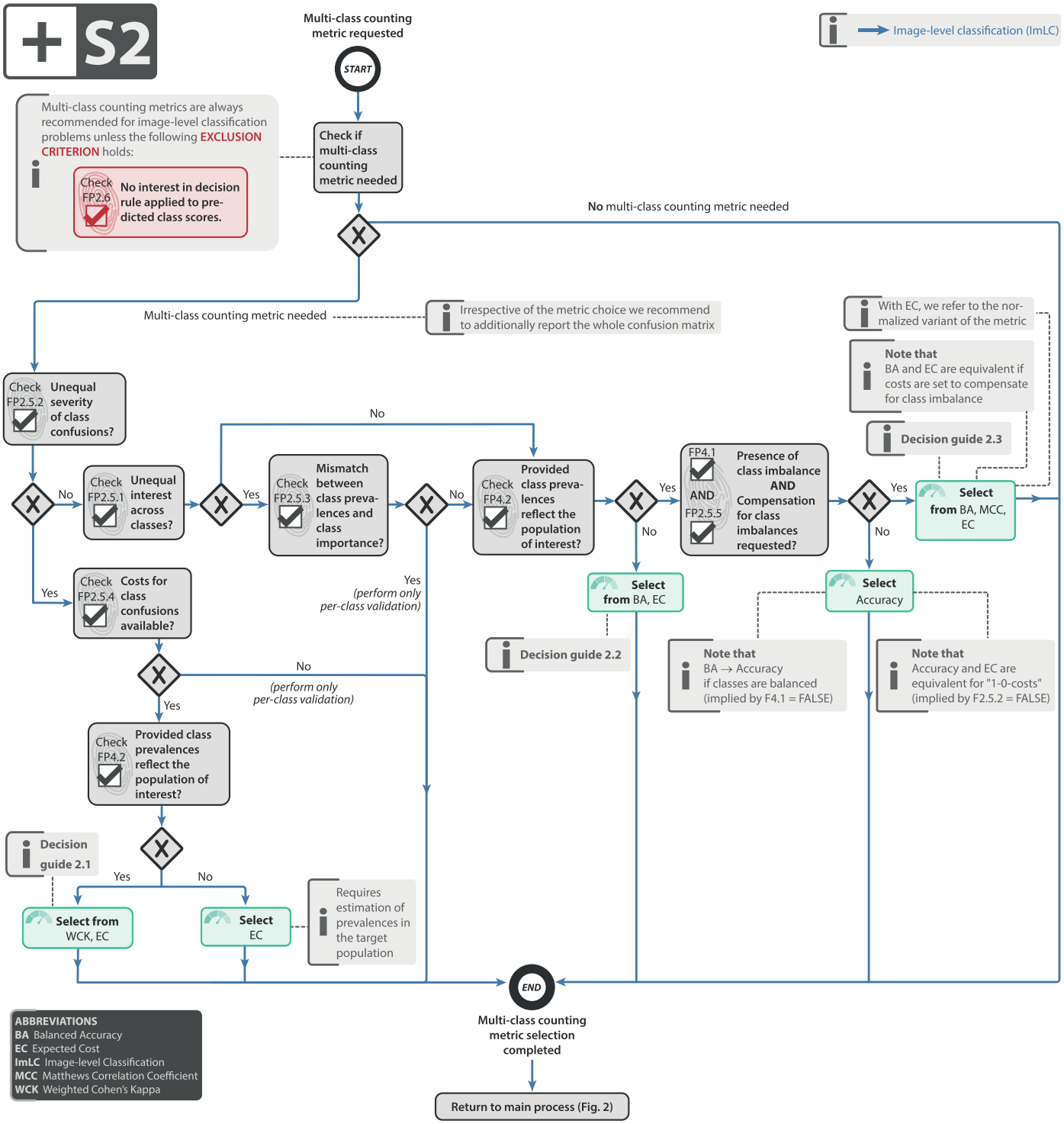
[d] If multiple structures of the same type can be seen in the same image and structure boundaries are important (FP2.1), we recommend setting this property to TRUE to avoid issues with boundary-based metrics resulting from comparing a given structure boundary to the boundary of the wrong reference instance (Fig. SN 1.2).

[e] if a substantial fraction of objects is small, we recommend framing the problem as an object detection problem ("no") to avoid brittle overlap-based localization criteria.

[f] If there is predefined fixed number of structures per category and image, the task would be considered a regression problem and thus defined as **out of scope**.

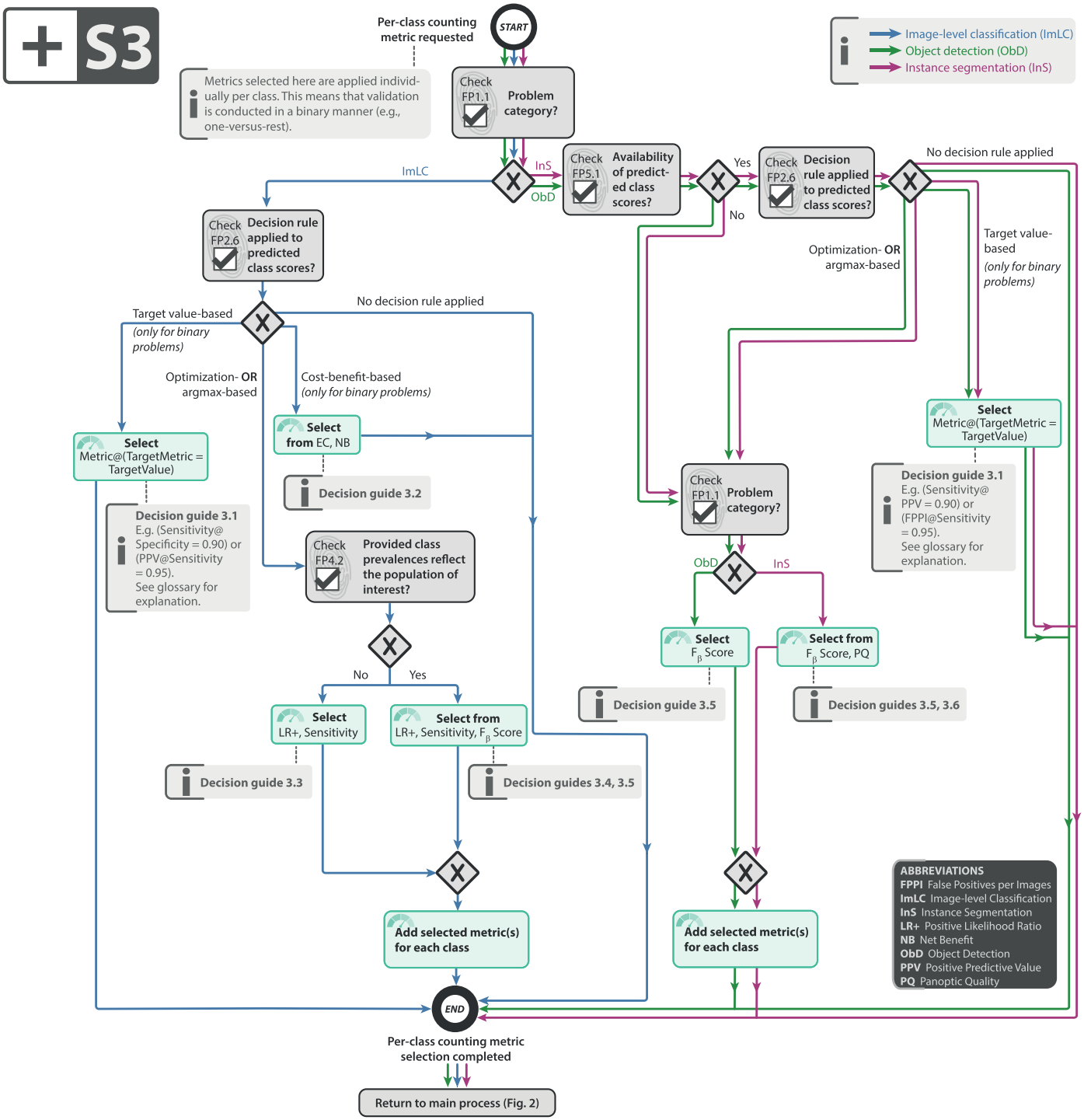
Extended Data Fig. 1 | Subprocess S1 for selecting a problem category. The Category Mapping maps a given research problem to the appropriate problem category with the goal of grouping problems by similarity of validation. The leaf nodes represent the categories: image-level classification, object detection, instance segmentation, or semantic segmentation. FP2.1 refers to fingerprint 2.1 (see Fig. SN 1.10). An overview of the symbols used in the process diagram is provided in Fig. SN 5.1.

https://doi.org/10.1038/s41592-023-02151-z



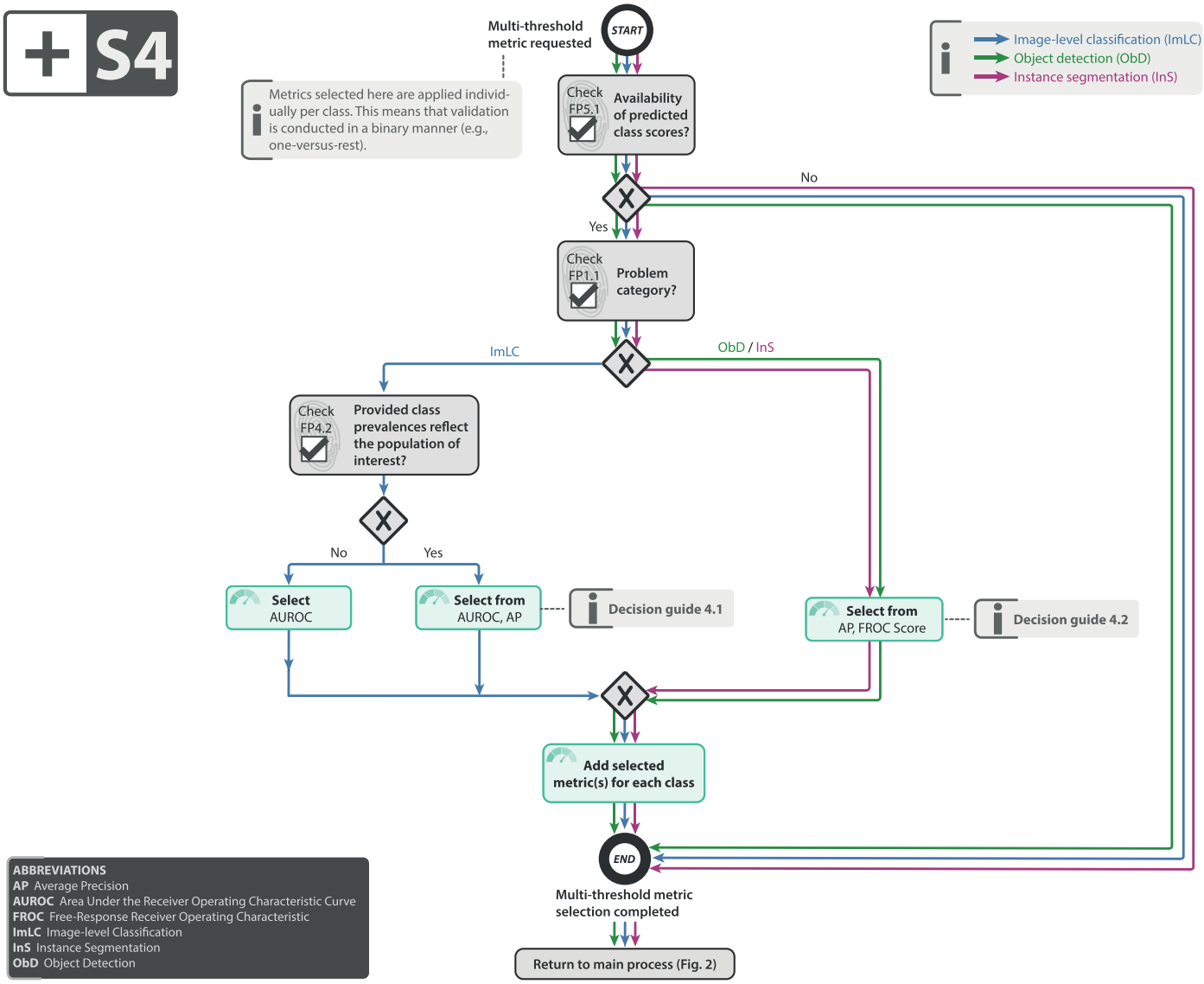
Extended Data Fig. 2 | Subprocess S2 for selecting multi-class metrics (if any). Applies to: image-level classification (ImLC). In the case of presence of class imbalance and no compensation of class imbalance being requested, one should follow the 'No' branch. Decision guides are provided in Supplementary Note 2.7.1. A detailed description of the subprocess is given in Supplementary Note 2.2.

https://doi.org/10.1038/s41592-023-02151-z



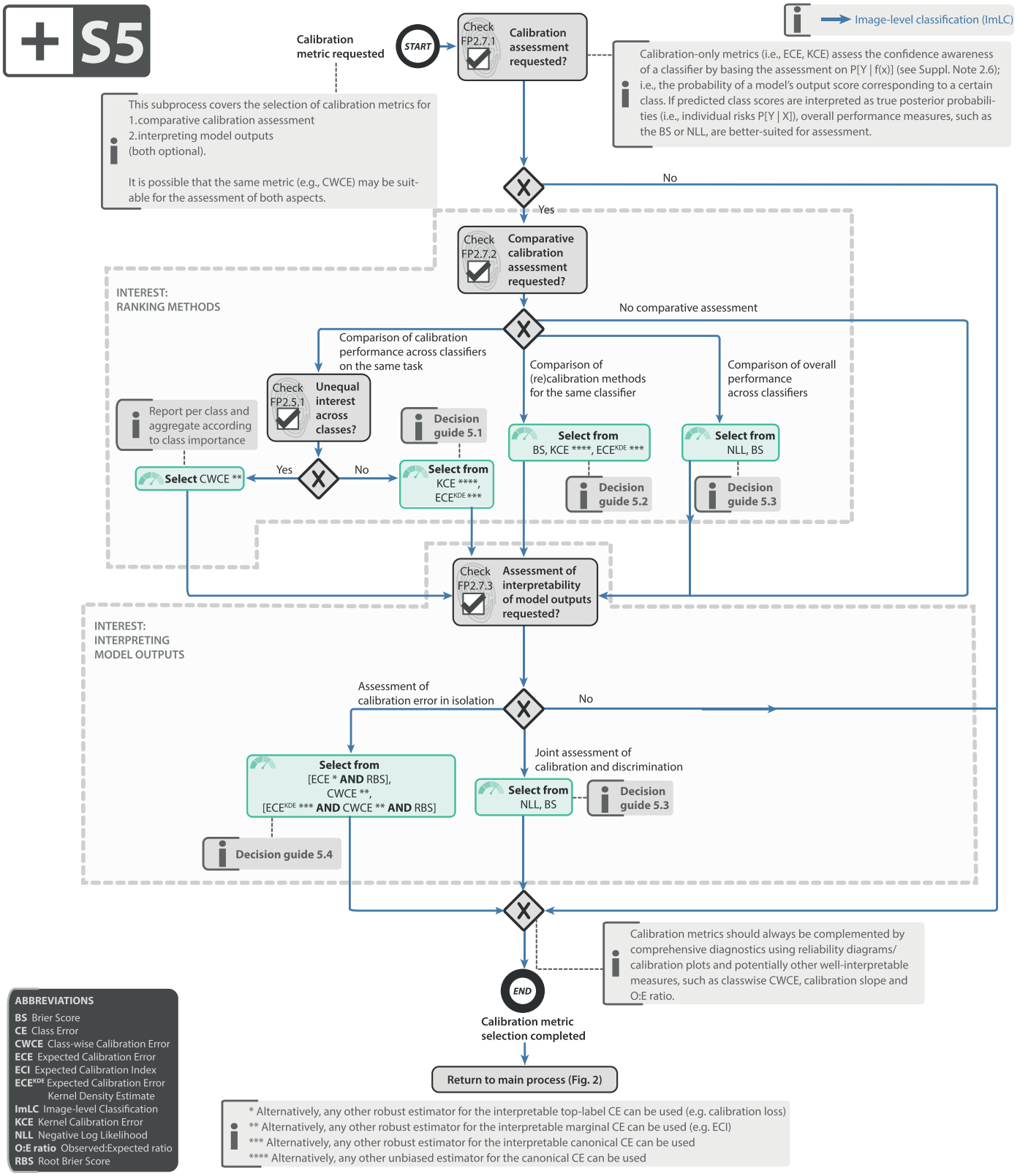
Extended Data Fig. 3 | Subprocess S3 for selecting a per-class counting metric (if any). Applies to: image-level classification (ImLC), object detection (ObD), and instance segmentation (InS). Decision guides are provided in Supplementary Note 2.7.2. A detailed description of the subprocess is given in Supplementary Notes 2.2, 2.4, and 2.5.

https://doi.org/10.1038/s41592-023-02151-z



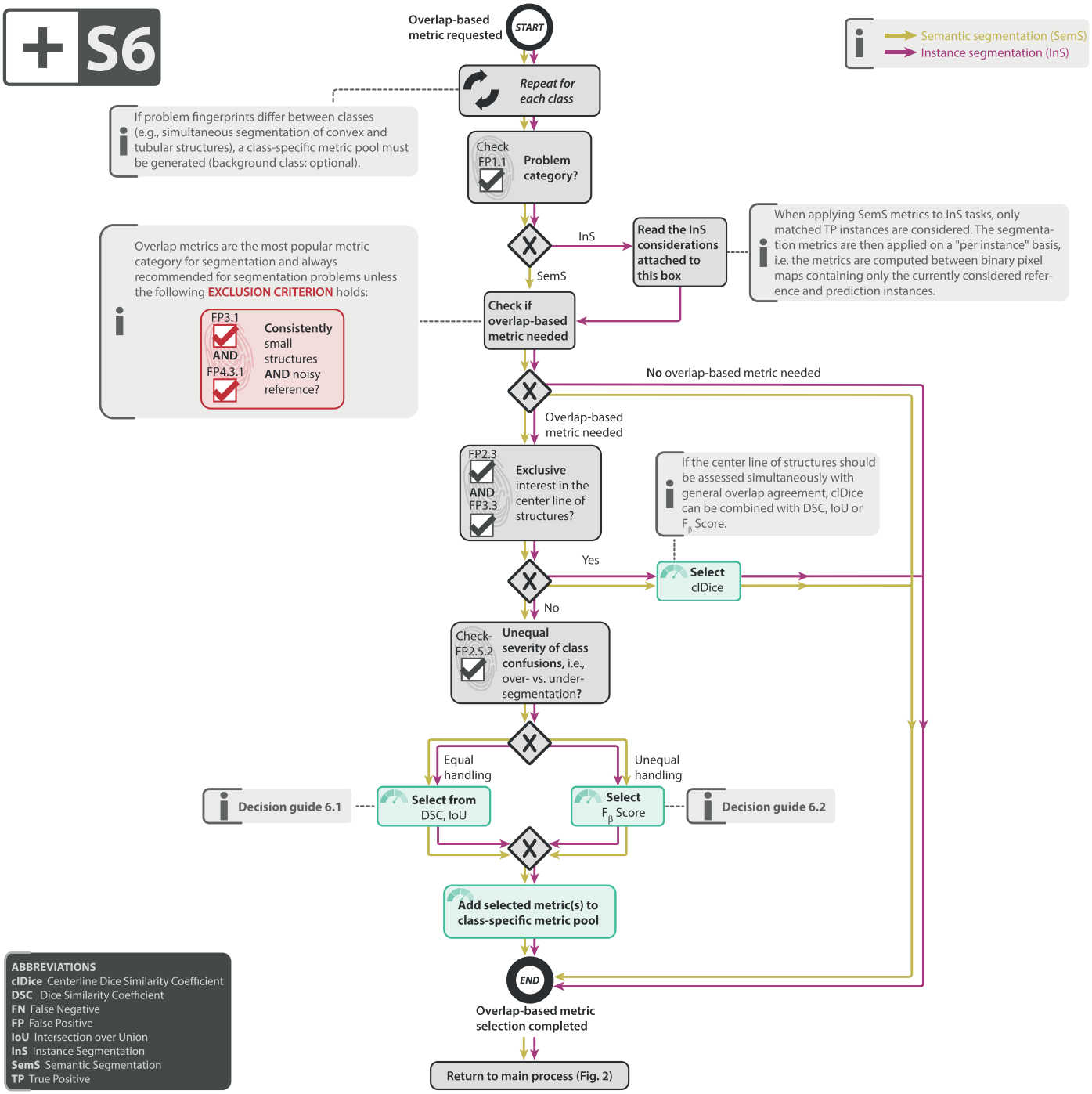
Extended Data Fig. 4 | Subprocess S4 for selecting a multi-threshold metric (if any). Applies to: image-level classification (ImLC), object detection (ObD), and instance segmentation (InS). Decision guides are provided in Supplementary Note 2.7.3. A detailed description of the subprocess is given in Supplementary Notes 2.2, 2.4, and 2.5.

https://doi.org/10.1038/s41592-023-02151-z



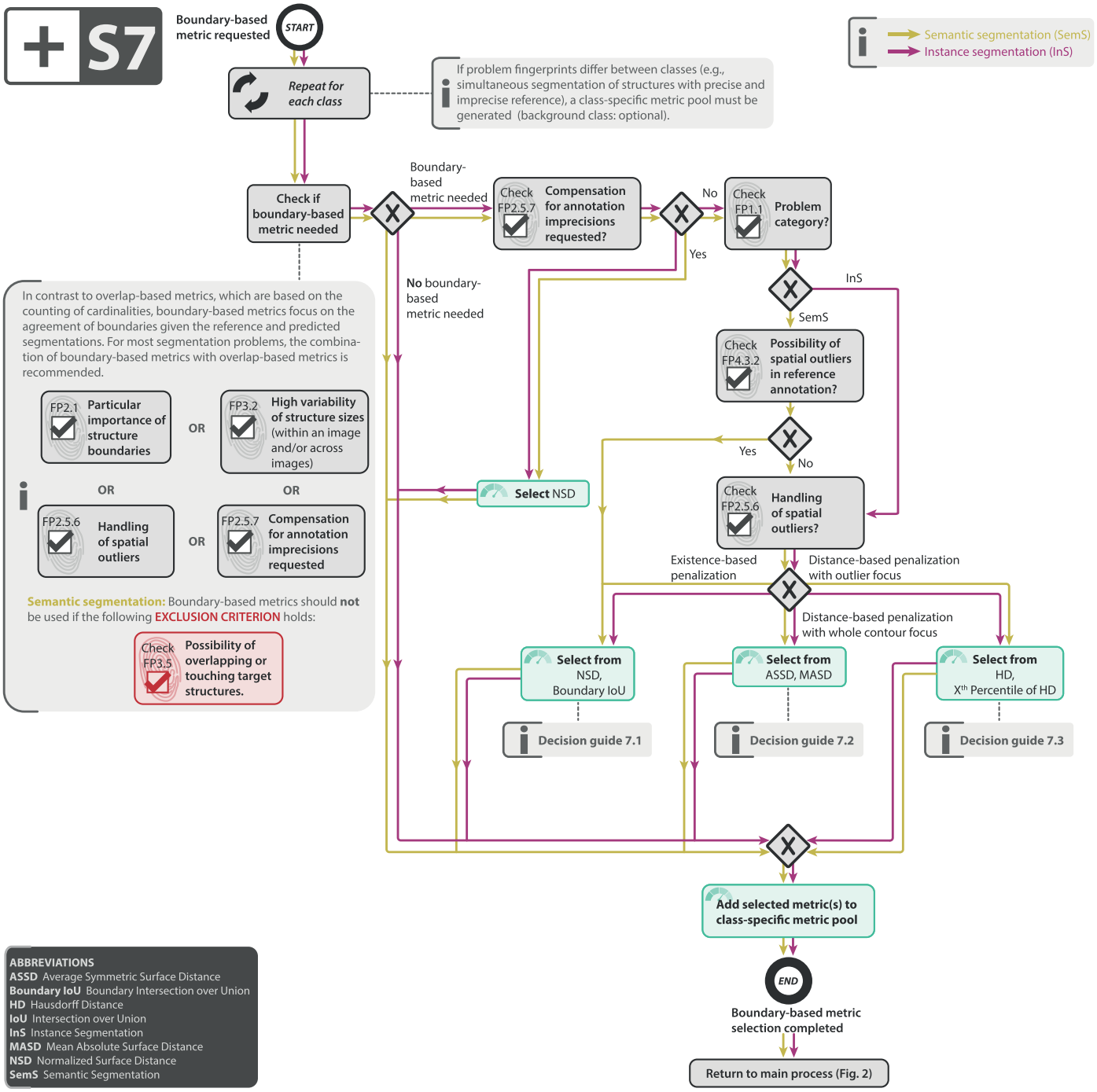
Extended Data Fig. 5 | Subprocess S5 for selecting a calibration metric (if any). Applies to: image-level classification (ImLC). Decision guides are provided in Supplementary Note 2.7.4. A detailed description of the subprocess is given in Supplementary Note 2.6. Further suggested calibration metrics include the calibration loss[74], calibration slope[46], Expected Calibration Index (ECI)[24] and Observed:Expected ratio (O:E ratio)[49].

https://doi.org/10.1038/s41592-023-02151-z



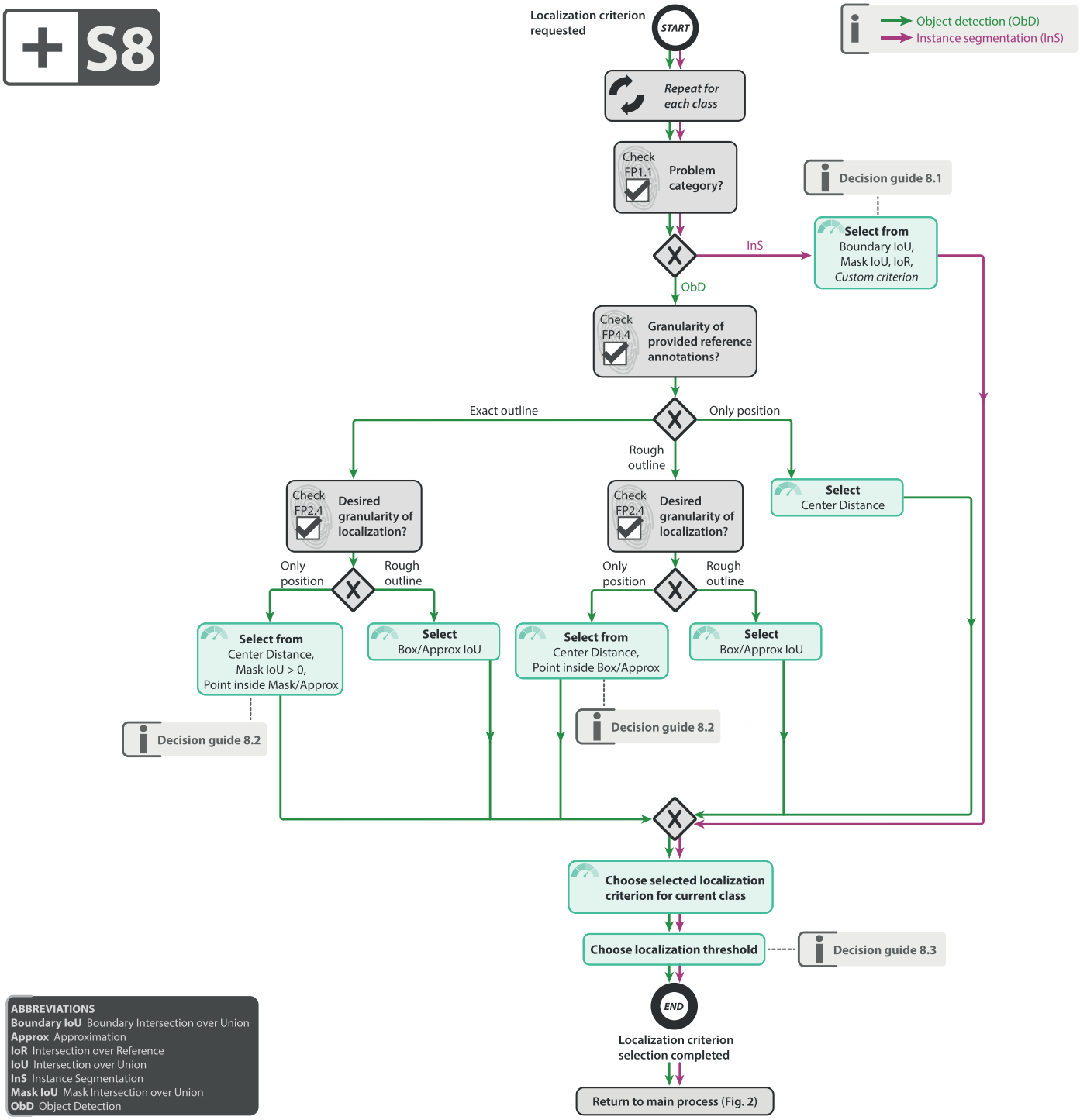
Extended Data Fig. 6 | Subprocess S6 for selecting overlap-based segmentation metrics (if any). Applies to: semantic segmentation (SemS) and instance segmentation (InS). Decision guides are provided in Supplementary Note 2.7.5. A detailed description of the subprocess is given in Supplementary Notes 2.3 and 2.5.

https://doi.org/10.1038/s41592-023-02151-z



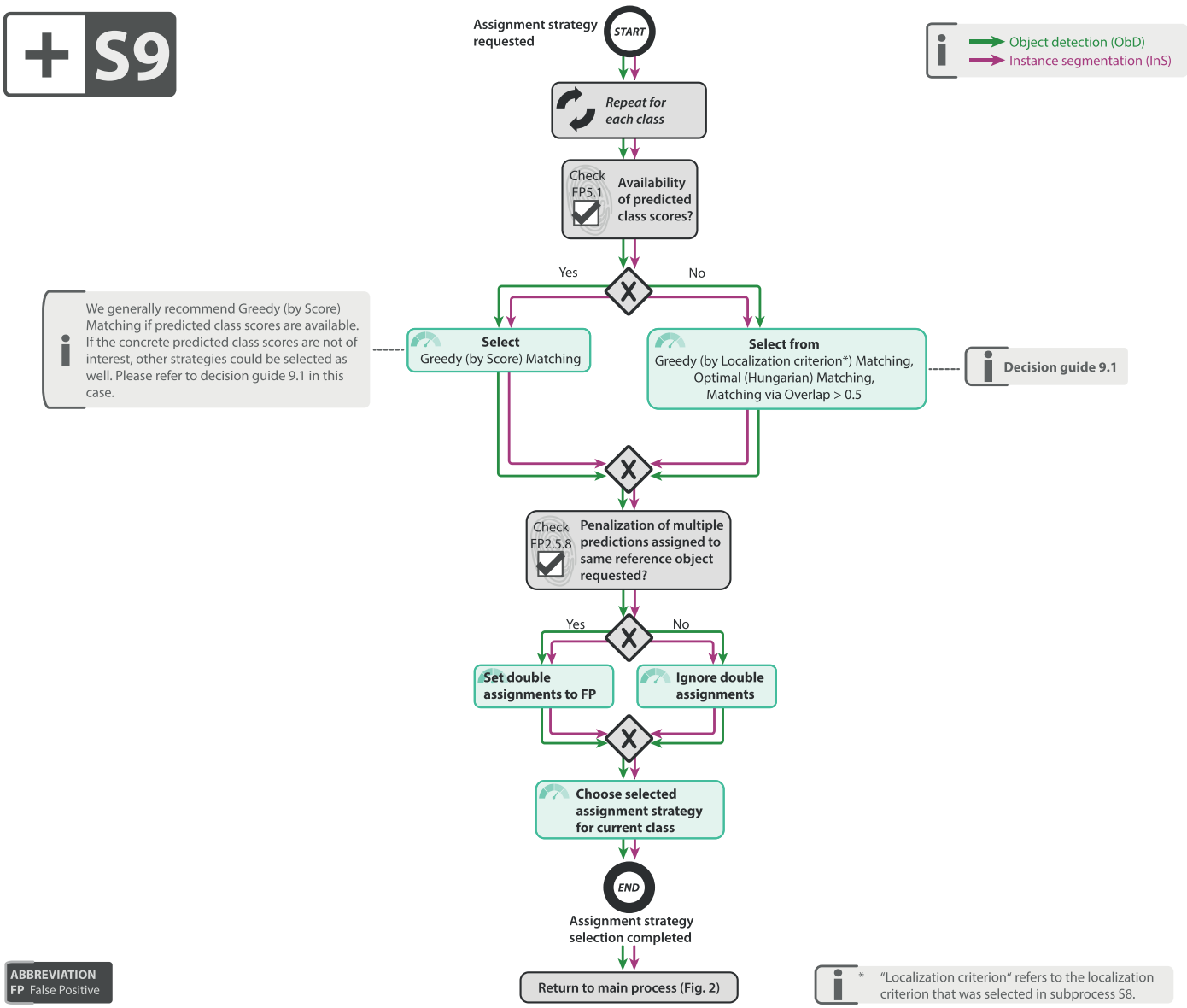
Extended Data Fig. 7 | Subprocess S7 for selecting a boundary-based segmentation metric (if any). Applies to: semantic segmentation (SemS) and instance segmentation (InS). Decision guides are provided in Supplementary Note 2.7.6. A detailed description of the subprocess is given in Supplementary Notes 2.3 and 2.5.

https://doi.org/10.1038/s41592-023-02151-z



Extended Data Fig. 8 | Subprocess S8 for selecting the localization criterion. Applies to: object detection (ObD) and instance segmentation (InS). Definitions of the localization criteria can be found in[19]. Decision guides are provided in Supplementary Note 2.7.7. A detailed description of the subprocess is given in Supplementary Notes 2.4 and 2.5.

https://doi.org/10.1038/s41592-023-02151-z



Extended Data Fig. 9 | Subprocess S9 for selecting the assignment strategy. Applies to: object detection (ObD) and instance segmentation (InS). Assignment strategies are defined in[19]. Decision guides are provided in Supplementary Note 2.7.8. A detailed description of the subprocess is given in Supplementary Notes 2.4 and 2.5.

Extended Data Table 1 | Recommendations for metric application addressing the pitfalls collected in ref. 19

Source of Pitfall	Recommendation
Metric implementation	
Non-standardized metric definition and undefined corner cases	Use reference implementations provided at https://github.com/Project-MONAI/MetricsReloaded
Discretization issues	Use unbiased estimates of properties of interest if possible (Suppl. Note 2.6).
Metric-specific issues including sensitivity to hyperparameters	Read metric-specific recommendations in the cheat sheets (Suppl. Note 3.1).
Aggregation	
Hierarchical label/class structure	Address the potential correlation between classes when aggregating [Kang & Sukthankar, 2006].
Multi-class problem	Complement validation with multi-class metrics such as Expected Cost (EC) or Matthews Correlation Coefficient (MCC) with per-class validation (Fig. 2); perform weighted class aggregation if *FP2.5.1 Unequal interest across classes* holds.
Non-independence of test cases (FP4.5)	Respect the hierarchical data structure when aggregating metrics [Liang & Zeger, 1986].
Risk of bias	Leverage metadata (e.g. on imaging device/protocol/center) to reveal potential algorithmic bias [Badgeley et al., 2019].
Possibility of invalid prediction (FP5.3)	Follow category-specific aggregation strategy detailed in Suppl. Note 2.
Ranking	
Metric relationships	Avoid combining closely related metrics (see Fig. SN 2.1) when choosing metrics to be used in algorithm ranking.
Ranking uncertainties	Provide information beyond plain tables that make possible uncertainties in rankings explicit as detailed in [30].
Reporting	
Non-determinism of algorithms	Consider multiple test set runs to address the variability of results resulting from non-determinism [Khan et al., 2019, Summers & Dinneen, 2021].
Uninformative visualization	Include a visualization of the raw metric values [30] and report the full confusion matrix unless *FP2.6 = no decision rule applied* holds.
Interpretation	
Low resolution	Read metric-related recommendations to obtain awareness of the pitfall (Suppl. Note 3.1).
Lack of lower/upper bounds	Read metric-related recommendations to obtain awareness of the pitfall (Suppl. Note 3.1).
Insufficient domain relevance of metric score differences	Report on the quality of the reference (e.g. intra-rater and inter-rater variability) [Kottner et al., 2011]. Choose the number of decimal places such that they reflect both relevance and uncertainties of the reference. More than one decimal number is often not useful given the typically high inter-rater variability.

[Kang & Sukthankar, 2006] Kang, F., Jin, R., & Sukthankar, R. (2006, June). Correlated label propagation with application to multi-label learning. In 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06) (Vol. 2, pp. 1719-1726). IEEE.

[Liang & Zeger, 1986] Liang, K. Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. Biometrika, 73(1), 13-22.

[Badgeley et al., 2019] Badgeley, M. A., Zech, J. R., Oakden-Rayner, L., Glicksberg, B. S., Liu, M., Gale, W., ... & Dudley, J. T. (2019). Deep learning predicts hip fracture using confounding patient and healthcare variables. NPJ digital medicine, 2(1), 31.

[Khan et al., 2019] Khan, D. A., Li, L., Sha, N., Liu, Z., Jimenez, A., Raj, B., & Singh, R. (2019). Non-Determinism in Neural Networks for Adversarial Robustness. arXiv preprint arXiv:1905.10906.

[Summers & Dinneen, 2021] Summers, C., & Dinneen, M. J. (2021, July). Nondeterminism and instability in neural network optimization. In International Conference on Machine Learning (pp. 9913-9922). PMLR.

[Kottner et al., 2011] Kottner, J., Audigé, L., Brorson, S., Donner, A., Gajewski, B. J., Hróbjartsson, A., ... & Streiner, D. L. (2011). Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. International journal of nursing studies, 48(6), 661-671.

Recommendations for metric application addressing the pitfalls collected in ref. 19. The first column comprises all sources of pitfalls captured by the published taxonomy that relate to the application of (already selected) metrics. The second column provides the Metrics Reloaded recommendation. The notation FPX.Y refers to a fingerprint item (Supplementary Note 1.3).

nature portfolio

Corresponding author(s): Lena Maier-Hein, Annika Reinke, Paul F. Jäger

Last updated by author(s): Sep 12, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☒	☐	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No data was used in this study (no experiments conducted)
Data analysis	No data was used in this study (no experiments conducted)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

No data was used in this study (no experiments conducted)

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	NA (no data was used in this study/no experiments conducted)
Reporting on race, ethnicity, or other socially relevant groupings	NA (no data was used in this study/no experiments conducted)
Population characteristics	NA (no data was used in this study/no experiments conducted)
Recruitment	NA (no data was used in this study/no experiments conducted)
Ethics oversight	NA (no data was used in this study/no experiments conducted)

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	NA
Data exclusions	NA
Replication	NA
Randomization	NA
Blinding	NA

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging