#	Parameter name	Description	Representative instantiations
	Challenge organization		
1	Challenge name	Full name of the challenge with year	Example: MICCAI Endoscopic Vision Challenge 2015
2	Challenge acronym	Acronym of the challenge (if any).	Example: EndoVis15
3	Challenge website	URL of challenge website (if any).	- URL to challenge website - Private link to website under construction - No website
4	Organizing institutions and contact person	Information on the organizing team including contact person and other team members.	Should include: - Contact person with affiliation - Team members with affiliations
5	Lifecycle type	Submission cycle of the challenge. Not every challenge closes after the submission deadline (one time event). Sometimes it is possible to submit results after the deadline (open call) or the challenge is repeated with some modifications (repeated event). *Example 1 - Brain tumor segmentation: One time event*	- One time event - Repeated event - Open call
6	Challenge schedule	Timetable for the challenge which includes the release of training and test cases, the submission dates, possibly associated workshop days, release of results and other important dates.	Should include: - Training data release(s) - Test data release(s) - Submission deadline - Conference day (if any)
7	Ethics approval	Information on ethics approval, preferably Institutional Review Board, location, date and number of the ethics approval. *Example 1 - Brain tumor segmentation: <URL to ethics approval>*	- No ethics needed (due to in silico validation) - URL to ethics approval document - No ethics required (data downloaded from a public database)
	Participation conditions		
8	Interaction level policy	Allowed user interaction of the algorithms assessed. *Example 1 - Brain tumor segmentation: Both automatic and semi-automatic algorithms can participate in the challenge.*	- Fully interactive - Semi-automatic - Fully automatic
9	Organizer participation policy	Participation policy for members of the organizers' institutes. *Example 1 - Brain tumor segmentation: Members of the organizers' institutes may participate but they are not eligible for awards.*	- Members of the organizers' institutes may participate but they are not eligible for awards and they will not be listed in the leaderboard. - Members of the organizers' institutes may not participate.

10	Training data policy	Policy on the usage of training data. The data used to train algorithms may, for example, be restricted to the data provided by the challenge or to publicly available data including (open) pre-trained nets. *Example 1 - Brain tumor segmentation: The challenge training data may be complemented by other publicly available data.*	- No policy as no training data is required - No additional data allowed - Publicly available data may be added - Private data may be added - Docker Container
11	Submission format	Method that is used for result submission. *Example 1 - Brain tumor segmentation: Participants send the algorithm output to the organizers via email.*	- Docker - Cloud - Upload whole code - Upload executable - Send algorithm output to organizers - API - Evaluation Platform
	Validation objective		
12	Field(s) of application	Medical or biological application that the algorithm was designed for. *Example 1 - Brain tumor segmentation: Diagnosis*	- Training - Intervention planning - Intervention follow-up - Diagnosis - Screening - Assistance (e.g. tracking tasks) - Research (e.g. cell tracking) - Cross-phase - Education - Prognosis - Prevention - Medical data management
13	Task category(ies)	Category(ies) of the algorithms assessed. *Example 1 - Brain tumor segmentation: Segmentation*	- Segmentation - Classification - Tracking - Retrieval - Detection - Localization - Registration - Reconstruction - Modeling - Simulation - Regression - Stitching - Restoration - Prediction - Denoising

14	Target cohort	Description of subjects/objects from whom the data would be acquired in the final application. Remark: A challenge could be designed around the task of medical instrument tracking in robotic kidney surgery. While the validation (see parameter validation cohort) could be performed ex vivo in a laparoscopic training environment with porcine organs, the final application (i.e. robotic kidney surgery) would be targeted on real patients with certain characteristics defined by inclusion criteria such as restrictions regarding gender or age. *Example 1 - Brain tumor segmentation: Patients diagnosed with glioblastoma that got MRI scans for diagnosis including T1-weighted 3D acquisitions, T1-weighted contrast-enhanced (gadolinium contrast) 3D acquisitions and T2-weighted FLAIR 3D acquisitions.*	- Healthy volunteers that undergo screening - Patients that undergo laparoscopic surgery - Patients that get an abdominal CT - Patients of a particular database - Patients referred for early Barrett's esophagus cancer without visible abnormalities - Patients attending a state-of-the-art cardiac MRI diagnostic center - Healthy volunteers that are recruited for a certain study - Patients that get chemotherapy - Men with clinical suspicion of having prostate cancer - Standardized cancer cell lines (such as HeLa) - Physicians that use a da Vinci Si for surgical training in an *ex vivo* setting - OR team (surgeons, nurses, ...) during liver transplantation - Specific journals with an oncology focus (for retrieval tasks)
15	Algorithm target(s)	Structure/subject/object component that the algorithm focuses on. *Example 1 - Brain tumor segmentation: Glioblastoma*	- Glioblastoma - Hepatocellular carcinoma (HCC) - Vessels - Liver - Tool tip - (Any) Tumor - Surgeon - Nurse - Specific cell type - Operating room - Specular reflections - Fiber pathway
16	Data origin	Region(s)/part(s) of subjects/objects from which the data would be acquired in the final application. *Example 1 - Brain tumor segmentation: Brain*	- Abdomen - Liver - Thorax - Whole body - Whole operating room - Cortical gray matter - Specific journal (for retrieval tasks) - Blood obtained from forearm
17	Assessment aim(s)	Property(ies) of the algorithms aimed to be optimized. Remark: Ideally, the metrics used in the study assess the properties of the algorithm as defined by the parameter *assessment aim(s)*. For example, an assessment aim could be targeted on the accuracy of segmentation algorithms. Possible metrics to assess the accuracy include the Dice Similarity Coefficient (DSC) and the Hausdorff Distance (HD). *Example 1 - Brain tumor segmentation: Accuracy of enhancing tumor/necrosis/edema segmentation*	- Accuracy - Robustness - Reliability - Precision - Sensitivity - Specificity - Consistency - Runtime - Applicability - Feasibility - Complexity - Usability - User satisfaction - Criteria linked to ergonomics - Integration in (clinical) workflow - Hardware requirements

	Study conditions		
18	Validation cohort	Subject(s)/object(s) from whom/which the data was acquired used to validate the algorithm. Remark: While a challenge is typically targeted on humans, validation may exclusively involve porcine models or phantoms. *Example 1 - Brain tumor segmentation: Patients with glioblastoma (retrospective analysis)*	- Specific mouse model - Porcine model - Physical phantom - Patients under controlled conditions - Patients in clinical routine - Porcine liver *(in vitro)* - *In silico* data - Healthy volunteers
19	Center(s)	Center(s) or institute(s) in which the data was acquired. *Example 1 - Brain tumor segmentation: National Center for Tumor Diseases (NCT) Heidelberg*	- Centers involved in the xy study - University Clinic xy - Centers that are part of the xy consortium
20	Imaging modality(ies)	Imaging technique(s) applied for training/test data acquisition. *Example 1 - Brain tumor segmentation: MRI*	- Magnetic Resonance Imaging (MRI) - Computed Tomography (CT) - Ultrasound (US) - 3D US - Intravascular US (IVUS) - Positron Emission Tomography (PET) - Light Microscopy (LM) - Electron Microscopy (EM) - X-ray - Optical Coherence Tomography (OCT) - Endomicroscopy (w/ or w/o dye) - SPECT - Video - Fluoroscopy - Thermography
21	Context information	Additional information given along with the images. The information may correspond directly to the image data (e.g. tumor volume), to the patient in general (e.g. gender, medical history) or to the acquisition process (e.g. medical device data during endoscopic surgery, calibration data for an image modality). *Example 1 - Brain tumor segmentation: Clinical patient data: {age, gender, ...}*	- No additional information - Genetic information - Age - Gender - Pathology - Clinical diagnoses - Patient number - Medical record - Weight - BMI - Race - Cancer (sub)type - Cancer/disease stage - Body weight/height - Smoking status - Clinical treatment details - Lab data - Clinical history - OR device data - Free text, such as the radiology report, the operation report or histopathology report

	Validation Datasets		
22	Distribution of training and test cases	Describes how training and test data were split and for what reason this division was chosen. This should include information (1) on why a specific proportion of training/test data was chosen, (2) why a certain total amount of cases was chosen and (3) why certain characteristics were chosen for the training/test set (e.g. class distribution according real-world distribution vs equal class distribution). *Example 1 - Brain tumor segmentation: 80% training data and 20% test data according to common practice in machine learning.*	- Not applicable as no training data is provided - Randomly distributed - Balanced false and negative cases - 80% training data, 20% test data as recommended by [ref]
23	Category of training data generation method	Method for determining the desired algorithm output for the training data. Possible methods include manual image annotation, in silico ground truth generation and annotation by automatic methods, and no training data generated. *Example 1 - Brain tumor segmentation: Hybrid: Initiation by algorithm and refinement/correction by expert physician*	- Ground truth from simulation (exact) - Reference from algorithm - Reference from single human rater - Reference from multiple human raters - Hybrid: Initiation by algorithm, refinement by expert physician - Reference derived from clinical practice (diagnosis/disease code etc.) - Crowdsourced annotations
24	Number of training cases	Number of cases that can be used for algorithm training and parameter optimization. A case encompasses all data that is processed to produce one result (e.g. one segmentation) as well as the corresponding reference result. *Example 1 - Brain tumor segmentation: 400*	- No training data provided - 100 images - 100 raw endoscopic video sequences with a total of 1,000 fully annotated frames
25	Characteristic of training cases	Additional information on the training cases describing their nature, such as the level of detail of the annotations (e.g. fully vs weakly annotated). *Example 1 - Brain tumor segmentation: Pixel-level segmentation of the structures of interest and additional clinical information as described in context information.*	- No training data provided - Full annotation (pixel level) - Weak annotation (image level): tumor volume, disease stage - Mixed annotation: 1,000 fully annotated images, 100 weakly annotated images - 100 endoscopic video images with 10 fully annotated training images
26	Annotation policy for the training cases	Instructions given to the annotators prior to training case annotation. This may include description of a training phase with the software. *Example 1 - Brain tumor segmentation: The annotator was instructed to segment the edema using the T2 and FLAIR images. The enhancing tumor was subsequently to be segmented on the T1 contrast-enhanced modality. Finally, the necrotic core was to be outlined using the T1 and contrast-enhanced T1 image. The annotations were to be performed in axial slices. The undergraduate student received training on 5 cases (by the radiologist) to extract the weak labels (see parameter: context information).*	- Challenge-specific detailed instructions - e.g. should an annotation be performed along a tumor boundary or including a safety zone? Is it allowed to guess a boundary if not clearly visible - URL to annotation instructions - What tissue would you resect? - Where would you take a (small) biopsy?

27	Annotator(s) of training data	Details on the subjects/algorithms who/which annotated the training data. *Example 1 - Brain tumor segmentation: Weak annotation (parameter context information) extracted from medical reports by undergraduate medical student; full image annotation performed by radiologist.*	- No training data provided - Surgeon who has done >100 cases of a specific type of surgery - Undergraduate physician (third year) - Engineer who developed the software - Physician with no prior experience in usage of the software - Crowd - Algorithm xy
28	Annotation aggregation method(s) for training cases	Method(s) used to merge multiple annotations for one case. *Example 1 - Brain tumor segmentation: (only one observer)*	- No aggregation - Simultaneous Truth and Performance Level Estimation (STAPLE) - Majority vote - An additional annotator resolves conflicts - Average - Selective and Iterative Method for Performance Level Estimation (SIMPLE) - Level-set based approach maximizing the a posteriori probability (LSML) - Strict combination (positive if and only if all annotators agree) - No training data is required
29	Category of reference generation method	Method for determining the reference (i.e. the desired algorithm result, also referred to as gold standard) which is used for assessing the participants' algorithms. Possible methods include manual image annotation, in silico ground truth generation and annotation by automatic methods. *Example 1 - Brain tumor segmentation: Manual annotation*	- Ground truth from simulation (exact) - Reference from algorithm - Reference from single human rater - Reference from multiple human raters - Acquired through previously validated methods according to [ref] - Reference derived from clinical practice (diagnosis/disease code etc.) - Crowdsourced annotations - Hybrid methods (e.g. initiation by algorithm, refinement by expert physician)
30	Number of test cases	Number of cases used to assess the performance of an algorithm. A case encompasses all data that is processed to produce one result as well as the corresponding reference result (typically not provided to the participants). *Example 1 - Brain tumor segmentation: 100*	- 100 images - 100 raw endoscopic video sequences with a total of 1,000 fully annotated frames
31	Characteristic of test cases	Additional information on the test cases describing their nature, such as the level of detail of the annotations (e.g. fully vs weakly annotated). *Example 1 - Brain tumor segmentation: Pixel-level segmentation of the structures of interest and additional clinical information as described in context information*	- Full annotation (pixel level) - Weak annotation (image level): tumor volume, disease stage - Mixed annotation: 1,000 fully annotated images, 100 weakly annotated images - 100 endoscopic video images with 10 fully annotated test images

32	Annotation policy for test cases	Instructions given to the annotators prior to test case annotation. This may include description of a training phase with the software. *Example 1 - Brain tumor segmentation: The annotator was instructed to segment the edema using the T2 and FLAIR images. The encancing tumor was subsequently to be segmented on the T1 contrast-enhanced modality. Finally, the necrotic core was to be outlined using the T1 and contrast-enhanced T1 image. The annotations were to be performed in axial slices. The undergraduate student had received training on extracting the weak labels when annotating the training images.*	- Challenge-specific detailed instructions - e.g. should an annotation be performed along a tumor boundary or including a safety zone? Is it allowed to guess a boundary if not clearly visible? - URL to annotation instructions - What tissue would you resect? - Where would you take a (small) biopsy?
33	Annotator(s) of test cases	Details on the subjects/algorithms who/which annotated the test data. *Example 1 - Brain tumor segmentation: Radiologist with 5 years of experience*	- Surgeon who has done >100 cases of a specific type of surgery - Undergraduate physician (third year) - Engineer who developed the software - Physician with no prior experience in usage of the software - Crowd - Algorithm xy
34	Annotation aggregation method(s) for test cases	Method(s) used to merge multiple annotations for one case (if any). *Example 1 - Brain tumor segmentation: No merging*	- No aggregation (ranking provided for each annotator) - STAPLE - Majority vote - An additional annotator resolves conflicts - Average - SIMPLE - LSML - Strict combination (positive if and only if all annotators agree)
	Assessment method		
35	Metric(s)	Function(s) to assess a property of an algorithm. These functions should reflect the validation objective (see parameter *assessment aim(s)*). *Example 1 - Brain tumor segmentation: 95% HD and precision applied separately to necrosis, enhancing tumor and edema.*	- Hausdorff distance (HD) - Dice Similarity Coefficient (DSC) - Jaccard index - Computation time - Recall - Precision - Area under curve (AUC) - Root mean square error (RMSE) - Absolute volume difference - True positive rate - Computational complexity - Average Symmetric Surface Distance (ASSD) - F1-Score - Specificity - Intraclass correlation coefficient - Concordance index - MAP

36	Justification of metrics	Justification why the metric(s) were chosen, preferably with reference to the clinical application. *Example 1 - Brain tumor segmentation: 95% Hausdorff distance as opposed to standard HD: Try to avoid that outliers have too much weight. All other metrics are commonly used in segmentation assessment (cf. ref. xy).*	- According to best practice recommendations [ref] - According to paper [ref]
37	Rank computation method	Method used to compute a rank for all participants based on the generated results on the test data. It may include methods for aggregating over all test cases and/or for determining a final rank from multiple single metric-based ranks. It also includes the ranking order for tied positions. *Example 1 - Brain tumor segmentation: For each participant pi and each test case cj: Compute the metric values for the 95% Hausdorff distance and precision. For each participant pi and each test case cj, determine the rank corresponding to both metrics (i.e. R(precision, pi, cj): descending order for precision, R(HD, pi, cj) ascending for 95% HD). For each participant pi and each test case cj, compute the average rank R(pi, cj) over both metric ranks. Finally, compute the average over all case-specific ranks to get one final rank for each participant pi.*	Example 1: 0. Initialization: For each participant pi and each test case cj: compute metric values M1(pi,cj) and M2(pi,cj) for metrics M1 and M2. 1. Metric-based aggregation: For each participant pi compute the median over all cases cj for each metric M1(pi) and M2(pi). 2. For each participant pi, compute the sum over the two metrics as M1(pi) + M2(pi). 3. Build rank for each participant by sorting the values M1(pi) + M2(pi) for each participant. Example 2: 0. Initialization: For each participant pi and each test case cj: compute metric values M1(pi,cj), M2(pi,cj) and M3(pi,cj) for metrics M1-M3. 1. Case-based aggregration: For each partipant pi and each case cj, determine the performance score sj on case cj: sij := 1/3 (M1(pi,cj) + M2(pi,cj) + M3(pi,cj)). 2. For each partipant pi and each case cj, determine the rank R(pi,cj) for case cj according to score sj. 3. Compute the average over all case-specific ranks for each participant pi si := 1/N*Sum_j (R(pi,cj)) to obtain the final rank.
38	Interaction level handling	Methods to handle any diversity in the level of user interaction when generating the performance ranking. *Example 1 - Brain tumor segmentation: Weighting function (automatic methods are ranked higher)*	- Indication in ranking - Separate ranking for fully-automatic methods - Only automatic methods allowed
39	Missing data handling	Methods used to manage submissions with missing results on test cases. *Example 1 - Brain tumor segmentation: In case of missing data for participant pi and case cj, the case-based ranks for all metrics m R(m, pi, cj) are set to the maximum.*	- Missing data not allowed (incomplete submissions not evaluated) - Missing data ignored - Missing data handled as in [ref]

40	Uncertainty handling	Method(s) used to make uncertainties in ranking explicit. *Example 1 - Brain tumor segmentation: Test the sensitivity of the ranking with bootstrapping according to [ref].*	- Test sensitivity of the ranking by * Leaving out test data * Bootstrapping approaches * Changes in rank computation details * Changes in reference annotation
41	Statistical test(s)	Statistical test(s) used to compare the results of challenge participants. *Example 1 - Brain tumor segmentation: T-test used to test the stability of the first three ranks as described in [ref].*	Quantities on which the hypothesis is taken: - Stability of the ranking - See whether the best results have statistically significant differences Tests: - Wilcoxon–Mann–Whitney test - t-test (paired, unpaired, one-sided, two-sided) - Saphiro-Wilk test