Appendix A: BIAS Reporting Guideline

Section/ Topic	Parameter name	Item No	Checklist Item	Reported on page No
TITLE, ABSTRACT, KEYWORDS	**Title**	**1**	Use the title to convey the essential information on the **challenge mission**. The title should ... • … identify the paper as biomedical image analysis challenge. • … indicate the image modality(ies) applied with a commonly used term in the title. • … indicate the task and/or task category (e.g. classification, segmentation; see parameter 18) with a commonly used term in the title. • … (optionally) include information on the biomedical target application. • … (optionally) include the year for repeated challenges with fixed cycle.	
	Abstract	**2**	Provide a **summary** of the challenge purpose, design and results and report the main conclusion(s).	
	Keywords	**3**	List the primary **keywords** that characterize the challenge.	
INTRO-DUCTION	**Challenge motivation and objective**	**4a**	Provide a general introduction to the topic from a **biomedical point of view**. This should include the envisioned biomedical impact (short-term and/or long-term).	
		4b	Provide a general introduction to the topic from a **technical point of view**. This should include an overview of the state of the art along the envisioned technical/methodological impact.	
		4c	Based on the biomedical and technical motivation, provide a concise statement of the **primary challenge objective**. This should include a **statement of the task**.	
METHODS Challenge organi-zation	**Challenge name**	**5a**	Provide a **representative name** of the challenge. *Example:* MICCAI Endoscopic Vision Challenge 2015	
		5b	Provide the **acronym** of the challenge (if any). *Example:* EndoVis15	
	Organizing team	**6**	Provide information on the **organizing team** (names and affiliations).	
	Life cycle type	**7**	Define the intended **submission cycle** of the challenge. Include information on whether/how the challenge has been/will be continued after the present study. *Examples:* • One-time event with fixed submission deadline	

		• Open call • Repeated event with annual fixed submission deadline
Challenge venue and platform	**8a**	Report the **event** (e.g. conference) that was **associated** with the challenge (if any).
	8b	Report the **platform** (e.g. grand-challenge.org) used to run the challenge.
	8c	Provide the **URL** for the challenge website (if any).
Participation policies	**9a**	Define the **allowed user interaction** of the algorithms assessed (e.g. only (semi-) automatic methods allowed).
	9b	Define the policy on the **usage of training data**. The data used to train algorithms may, for example, have been restricted to the data provided by the challenge or to publicly available data including (open) pre-trained nets.
	9c	Define the **participation policy for members of the organizers' institutes**. For example, members of the organizers' institutes could participate in the challenge but were not eligible for awards.
	9d	Define the **award policy**. In particular, provide details with respect to challenge prizes.
	9e	Define the policy for **results announcement**. *Examples:* • Top three performing methods were announced publicly. • Participating teams could choose whether the performance results will be made public.
	9f	Define the **publication policy**. In particular, provide details on ... • … who of the participating teams/the participating teams' members qualified as author • … whether the participating teams could publish their own results separately, and (if so) • … whether an embargo time was defined (so that challenge organizers can publish a challenge paper first).
Submission method	**10a**	Describe the method used for result submission. If available, provide a link to the **submission instructions**. *Examples:* • Docker container on the Synapse platform. Link to submission instructions: <URL> • Algorithm output was sent to organizers via e-mail. Submission instructions were sent by e-mail.
	10b	Provide information on the possibility for participating teams to **evaluate** their **algorithms before submitting** final results. For example, many

			challenges allow submission of multiple results, and only the last run is officially counted to compute challenge results.
	Challenge schedule	**11**	Provide a **timetable** for the challenge. Preferably, this should include • the release date(s) of the training cases (if any) • the registration date/period • the release date(s) of the test cases and validation cases (if any) • the submission date(s) • associated workshop days (if any) • the release date(s) of the results
	Ethics approval	**12**	Indicate whether **ethics approval** was necessary for the data. If yes, provide details on the ethics approval, preferably institutional review board, location, date and number of the ethics approval (if applicable). Add the URL or a **reference to the document** of the ethics approval (if available).
	Data usage agreement	**13**	Clarify how the data can be used and distributed by the teams that participate in the challenge and by others. This should include the explicit **listing of the license** applied. *Examples:* • CC BY (Attribution) • CC BY-SA (Attribution-ShareAlike) • CC BY-ND (Attribution-NoDerivs) • CC BY-NC (Attribution-NonCommercial) • CC BY-NC-SA (Attribution-NonCommercial-ShareAlike) • CC BY-NC-ND (Attribution-NonCommercial-NoDerivs)
	Code availability	**14a**	Provide information on the **accessibility of the organizers' evaluation software** (e.g. code to produce rankings). Preferably, provide a link to the code and add information on the supported platforms.
		14b	In an analogous manner, provide information on the **accessibility of the participating teams' code.**
	Conflicts of interest	**15**	Provide information related to conflicts of interest. In particular provide information related to **sponsoring/funding** of the challenge. Also, state explicitly who had **access to the test case labels** and when.
	Author contributions	**16**	List the **contributions** of all authors to the paper (preferably in the appendix).
METHODS Mission of the challenge	**Field(s) of application**	**17**	State the **main field(s) of application** that the participating algorithms target. *Examples:* • Diagnosis • Education • Intervention assistance

		• Intervention follow-up • Intervention planning • Prognosis • Research • Screening • Training • Cross-phase
Task category(ies)	**18**	State the **task category(ies)**. *Examples:* • Classification • Detection • Localization • Modeling • Prediction • Reconstruction • Registration • Retrieval • Segmentation • Tracking
Cohorts	We distinguish between the *target cohort* and the *challenge cohort*. For example, a challenge could be designed around the task of medical instrument tracking in robotic kidney surgery. While the challenge could be based on *ex vivo* data obtained from a laparoscopic training environment with porcine organs (challenge cohort), the final biomedical 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 (target cohort).	
	19a	Describe the **target cohort**, i.e. the subjects/objects from whom/which the data would be acquired in the final biomedical application.
	19b	Describe the **challenge cohort**, i.e. the subject(s)/object(s) from whom/which the challenge data was acquired.
Imaging modality(ies)	**20**	Specify the **imaging technique(s)** applied in the challenge.
Context information	Provide additional **information given along with the images**. The information may correspond ...	
	21a	… directly to the **image data** (e.g. tumor volume). If necessary, differentiate between target and challenge cohort.
	21b	… to the **patient** in general (e.g. gender, medical history). If necessary, differentiate between target and challenge cohort.
	21c	… to the **acquisition process** (e.g. medical device data during endoscopic surgery, calibration data for an image modality). If necessary, differentiate between target and challenge cohort.
Target entity(ies)	**22a**	Describe the **data origin**, i.e. the region(s)/part(s) of subject(s)/object(s) from whom/which the image data would be acquired in the final

			biomedical application (e.g. brain shown in computed tomography (CT) data, abdomen shown in laparoscopic video data, operating room shown in video data, thorax shown in fluoroscopy video). If necessary, differentiate between target and challenge cohort.
		22b	Describe the **algorithm target**, i.e. the structure(s)/subject(s)/object(s)/component(s) that the participating algorithms have been designed to focus on (e.g. tumor in the brain, tip of a medical instrument, nurse in an operating theater, catheter in a fluoroscopy scan). If necessary, differentiate between target and challenge cohort.
	Assessment aim(s)	**23**	Identify the **property(ies) of the algorithms to be optimized** to perform well in the challenge. If multiple properties were assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (parameter 29), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties. • *Example 1:* Find liver segmentation algorithm for CT images that processes CT images of a certain size in less than a minute on a certain hardware with an error that reflects inter-rater variability of experts. • *Example 2:* Find lung tumor detection algorithm with high sensitivity and specificity for mammography images. Corresponding metrics are listed below (parameter 29).
METHODS Challenge data sets	**Data source(s)**	**24a**	Specify the **device(s)** used to acquire the challenge data. This includes details on the device(s) used to acquire the imaging data (e.g. manufacturer) as well as information on additional devices used for performance assessment (e.g. tracking system used in a surgical setting).
		24b	Describe relevant details on the imaging process/**data acquisition** for each acquisition device (e.g. image acquisition protocol(s)).
		24c	Specify the **center(s)/institute(s)** in which the data was acquired and/or the **data providing platform/source** (e.g. previous challenge). If this information is not provided (e.g. for anonymization reasons), specify why.
		24d	Describe relevant **characteristics** (e.g. level of expertise) **of the subjects** (e.g. surgeon)/objects (e.g. robot) involved in the data acquisition process (if any).
	Training and test case characteristics	**25a**	State what is meant by one **case** in this challenge. A case encompasses all data that is processed to produce one result that is then compared to the corresponding reference result (i.e. the desired algorithm output).

		Examples: • Training and test cases both represented a CT image of a human brain. Training cases had a weak annotation (tumor present or not and tumor volume (if any)) while the test cases were annotated with the tumor contour (if any). • A case refers to all information that is available for one particular patient in a specific study. This information always includes the image information as specified in *data source(s)* (parameter 24) and may include context information (parameter 21). Both training and test cases were annotated with survival (binary) 5 years after (first) image was taken.
	25b	State the **total number** of cases as well as the number of training, validation and test cases separately.
	25c	Explain **why a total number** of cases and **the specific proportion** of training, validation and test cases was chosen.
	25d	Mention **further important characteristics** of the training, validation and test cases (e.g. class distribution in classification tasks chosen according to real-world distribution vs. equal class distribution) and justify the choice.
Annotation characteristics	**26a**	Describe the **method for determining the reference annotation**, i.e. the desired algorithm output. Provide the information separately for the training, validation and test cases if necessary. Possible methods include *manual image annotation*, *in silico ground truth generation* and *annotation by automatic methods*. If human annotation was involved, state the **number of annotators**.
	26b	Provide the **instructions given to the annotators** (if any) prior to the annotation. This may include description of a training phase with the software. Provide the information separately for the training, validation and test cases if necessary. Preferably, provide a link to the **annotation protocol**.
	26c	Provide **details on the subject(s)/algorithm(s) that annotated** the cases (e.g. information on **level of expertise** such as number of years of professional experience, medically-trained or not). Provide the information separately for the training, validation and test cases if necessary.
	26d	Describe the **method(s) used to merge multiple annotations** for one case (if any). Provide the information separately for the training, validation and test cases if necessary.
Data pre-processing method(s)	**27**	Describe the **method(s) used for pre-processing** the raw training data before it is provided to the participating teams. Provide the information separately for the training, validation and test cases if necessary.

	Sources of error	**28a**	Describe the most relevant **possible error sources related to the image annotation**. If possible, **estimate the magnitude** (range) of these errors, using inter-and intra-annotator variability, for example. Provide the information separately for the training, validation and test cases, if necessary.
		28b	In an analogous manner, describe and quantify **other relevant sources of error**.
METHODS Assess-ment methods	**Metric(s)**	**29a**	Define the **metric(s) to assess a property of an algorithm**. These metrics should reflect the desired algorithm properties described in *assessment aim(s)* (parameter 21). State which metric(s) were used to compute the ranking(s) (if any). ● *Example 1:* Dice Similarity Coefficient (DSC) and run-time ● *Example 2:* Area under curve (AUC)
		29b	**Justify why** the metric(s) was/were chosen, preferably with reference to the biomedical application.
	Ranking method(s)	**30a**	Describe the **method used to compute a performance rank** for all submitted algorithms based on the generated metric results on the test cases. Typically the text will describe how results obtained per case and metric are aggregated to arrive at a final score/ranking.
		30b	Describe the method(s) used to manage **submissions with missing results** on test cases.
		30c	**Justify why** the described ranking scheme(s) was/were used.
	Statistical analyses	**31a**	Provide **details for all statistical methods** used in the scope of the challenge analysis. This may include ● description of the **missing data handling**, ● details about the assessment of **variability of rankings**, ● description of any method used to assess **whether the data met the assumptions**, required for the particular statistical approach, or ● indication of any **software product** that was used for data analysis.
		31b	**Justify why** the described statistical method(s) was/were used.
RESULTS Challenge outcome	**Challenge submissions**	Provide **summarizing information** on ...	
		32a	… the **number of registrations**.
		32b	… the **number of participating teams** that provided valid submissions (if applicable in each phase).
		32c	… the **number of participating teams** that the **paper** refers to (with **justification**).
	Information on selected	Provide the following information for the participating teams that are included in the paper:	
		33a	**Team identifier**.

	participating teams	**33b**	A **method description** including parameter instantiation and/or a reference/URL to a document containing this information.
	Metric values	**34**	Provide raw and/or **aggregated metric values** (including measure of variability) for all participating teams and each metric (if applicable) as well as the **numbers of test set submissions** (the last one was used to compute metric(s)) for each participating team.
	Ranking(s)	**35a**	Report the **ranking(s)** (if any) including the number of test set submissions for each participating team.
		35b	Provide the results of the **statistical analyses**.
	Further Analyses	**36**	Present results of further analyses (if applicable), e.g. related to ● **combining algorithms** via ensembling, ● **inter-algorithm variability**, ● **common problems/biases** of the submitted methods, or ● **ranking variability**.
DISCUS-SION	**Summary**	**37**	Summarize the **main results** of the challenge.
	Impact	**38a**	Describe the (expected) **biomedical impact** of the challenge in the context of the state of the art with reference to the challenge motivation (parameter 4a).
		38b	Describe the (expected) **technical impact** of the challenge in the context of the state of the art with reference to the challenge motivation (parameter 4b).
	Discussion of challenge results	**39a**	Provide a detailed discussion and conclusion whether the **task is now solved** in a satisfactory way (e.g. the remaining errors are comparable to inter-annotator variability).
		39b	Provide a detailed **analysis of individual cases**, in which the majority of algorithms performed poorly (if any).
		39c	Provide a **discussion on advantages and disadvantages of the submitted methods**. Include time and memory consumption comparison if time and memory were not among the metrics.
	Limitations of the challenge	**40**	Discuss **limitations** related to the challenge design and execution.
	Future work	**41**	Provide **recommendations for future work** and maintenance plans for the challenge and its website (if any).
	Conclusions	**42**	Provide a **concise conclusion** based on the results of the study.