BIAS Reporting Guideline

The Biomedical Image Analysis Challenges (BIAS) statement [1] has been officially approved by the Enhancing the QUAlity and Transparency Of health Research (EQUATOR) initiative as a guideline for reporting challenge results (click here for details). The present checklist represents a light version, comprising the parameters that represent the main body of the paper. **Please report the page number(s) (or N/A) for each parameter, such that the reviewers can quickly find relevant information and assess the comprehensiveness of the report.** Please cite [1] when referring to/using the BIAS guideline and/or the light version of the checklist as basis of your paper.

[1] Maier-Hein, L. et al. BIAS: Transparent reporting of biomedical image analysis challenges. Medical image analysis, 66, 101796 (2020).

Section/ Topic	No	Checklist Item	Reported on page
Introduction			
	4a	General introduction to the topic from a **biomedical point of view**.	
	4b	General introduction to the topic from a **technical point of view**.	
	4c	Concise statement of the **primary challenge objective**, including a **statement of the task**.	
Methods			
Challenge organization	5	**Representative name** of the challenge including an **acronym** (if any).	
	6	Information on the **organizing team** (names and affiliations).	
	7	Intended **submission cycle** of the challenge, including information on whether/how the challenge has been/will be continued after the present study. *Examples:* One-time event with fixed submission deadline; repeated event with annual fixed submission deadline	
	8a	**Event** (e.g. conference) that was **associated** with the challenge (if any).	
	8b	**Platform** (e.g. grand-challenge.org) used to run the challenge.	
	8c	**URL** for the challenge website (if any).	
	9a	**Allowed user interaction** of the algorithms (e.g. only (semi-) automatic methods).	
	9b	Policy on the **usage of training data**. *Examples:* Training data have been restricted to the data provided by the challenge or to publicly available data including (open) pre-trained nets.	
	9c	**Participation policy for members of the organizers' institutes**. *Example:* Members of the organizers' institutes could participate but were not eligible for awards.	
	9d	**Award policy**, including details with respect to challenge prizes.	
	9e	Policy for **results announcement**. *Example:* The top three performing methods were announced publicly	
	9f	**Publication policy**, including details on who of the participating teams' members qualified as author, whether participating teams could publish their own results separately, and whether an embargo time was defined.	
	10a	Method used for result submission, including a link to the **submission instructions** (if any). *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	Information on the possibility for participating teams to **evaluate** their **algorithms before submitting** final results.	
	11	**Timetable** for the challenge, including the release date(s) of the training/validation/test cases, registration date/period, submission date(s), associated workshop days, release date(s) of the results.	

	12	Indication whether **ethics approval** was necessary for the data. If yes, details on the ethics approval and URL or **reference to the document** of the ethics approval (if available).	
	13	Information on how the data can be used and distributed by the participating teams as well as others, including the explicit **listing of the license**. *Examples:* CC BY (Attribution); CC BY-SA (Attribution-ShareAlike)	
	14a	Information on the **accessibility of the organizers' evaluation software**, preferably, including a link to the code.	
	14b	Information on the **accessibility of the participating teams' code** in an analogous manner.	
	15	Information related to conflicts of interest, including information related to **sponsoring/funding** of the challenge and an explicit statement who had **access to the test case labels** and when.	
	16	**Contributions** of all authors to the paper (preferably in the appendix).	
Mission of the challenge	17	**Main field(s) of application** that the participating algorithms target. *Examples:* Diagnosis; intervention planning; screening; training	
	18	**Task category/-ies**. *Examples:* Classification; detection; segmentation; reconstruction	
	19a	**Target cohort**, i.e. the subjects/objects from whom/which the data would be acquired in the final biomedical application.	
	19b	**Challenge cohort**, i.e. the subject(s)/object(s) from whom/which the challenge data was acquired.	
	20	**Imaging technique(s)** applied in the challenge.	
	Additional information given along with the images, corresponding to the...		
	21a	... **image data** (e.g. tumor volume).	
	21b	... **patient** in general (e.g. gender, medical history).	
	21c	... **acquisition process** (e.g. calibration data for an image modality).	
	22a	**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.	
	22b	**Algorithm target**, i.e. the structure(s)/subject(s)/object(s)/ component(s) that the participating algorithms have been designed to focus on.	
	23	**Property/-ies of the algorithms to be optimized** to perform well in the challenge. Should be reflected in the metrics (parameter 29) and ranking (parameter 30) applied. • *Example 1:* Find liver segmentation algorithm that processes CT images of a certain size in less than a minute with an error that reflects inter-rater variability of experts. • *Example 2:* Find lung tumor detection algorithm with high sensitivity for mammography images.	
Challenge data sets	24a	**Device(s)** used to acquire the challenge data, including details on the device(s) used to acquire the imaging data as well as information on additional devices used for performance assessment.	
	24b	Relevant details on the imaging process/**data acquisition** for each acquisition device (e.g. image acquisition protocol(s)).	
	24c	**Center(s)/institute(s)** in which the data was acquired and/or **data providing platform/source** (e.g. previous challenge).	
	24d	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).	
	25a	Information on the meaning of 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). *Example:* 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) while the test cases were annotated with the tumor contour.	
	25b	**Total number** of cases as well as the number of training, validation and test cases.	
	25c	**Justification why a total number** of cases and **the specific proportion** of training, validation and test cases was chosen.	
	25d	**Further important characteristics** of the training, validation and test cases (e.g. class distribution in classification tasks) and justification of choice.	

	26a	**Method for determining the reference annotation**. If human annotation was involved, state the **number of annotators**.
	26b	**Instructions given to the annotators** (if any) prior to the annotation, preferably, including a link to the **annotation protocol**.
	26c	**Details on the subject(s)/algorithm(s) that annotated** the cases (e.g. information on **level of expertise** such as number of years of experience, medically-trained or not).
	26d	**Method(s) used to merge multiple annotations** for one case (if any).
	27	**Method(s) used for pre-processing** the raw training data before it is provided to the participating teams.
	28a	Most relevant **possible error sources related to the image annotation**, preferably including an **estimate of the magnitude** (range) of these errors, using inter-and intra-annotator variability, for example.
	28b	In an analogous manner, **other relevant sources of error**.
Assessment method	29a	**Metric(s) used to assess a property of an algorithm**. Should reflect the desired properties described in *assessment aim(s)* (parameter 23). Which metric(s) were used to compute the ranking(s) (if any).
	29b	**Justification why** the metric(s) was/were chosen, preferably with reference to the biomedical application.
	30a	**Method used to compute a performance rank** for all submitted algorithms, including how results were obtained per case and metrics are aggregated to get a final ranking.
	30b	Method(s) used to manage **submissions with missing results** on test cases.
	30c	**Justification why** the described ranking scheme(s) was/were used.
	31a	**Details for all statistical methods**, including description of the **missing data handling**, details about assessment of **variability of rankings**, or indication of any **software product** used for data analysis.
	31b	**Justification why** the described statistical method(s) was/were used.
Results		
Challenge outcome	32a	Summarizing information on the **number of registrations**.
	32b	Summarizing information on the **number of participating teams** that provided valid submissions.
	32c	Summarizing information on the **number of participating teams** that the **paper** refers to (with **justification**).
	33a	**Team identifiers** for the participating teams that are included in the paper.
	33b	**Method description** including parameter instantiation and/or a reference/URL to a document containing this information for the participating teams that are included in the paper.
	34	Raw and/or **aggregated metric values** (including measure of variability) for all participating teams and each metric and the **numbers of test set submissions** for each participating team.
	35a	**Ranking(s)** (if any) including the number of test set submissions for each team.
	35b	Results of the **statistical analyses**.
	36	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**.
Discussion		
	37	**Main results** of the challenge.
	38	(Expected) **Biomedical and technical impact** of the challenge in the context of the state of the art with reference to the challenge motivation (parameter 1).
	39a	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).
	39a	Detailed **analysis of individual cases**, in which the majority of algorithms performed poorly (if any).
	39c	**Discussion on advantages and disadvantages of the submitted methods**. Include time and memory consumption comparison if time and memory were not among the metrics.
	40	**Limitations** related to the challenge design and execution.

41	**Recommendations for future work** and maintenance plans for the challenge and its website (if any).	
42	**Concise conclusion** based on the results of the study.	