Medical Image Analysis 66 (2020) 101796

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

# BIAS: Transparent reporting of biomedical image analysis challenges

Lena Maier-Hein[a,*], Annika Reinke[a], Michal Kozubek[b], Anne L. Martel[c,d], Tal Arbel[e], Matthias Eisenmann[a], Allan Hanbury[f,g], Pierre Jannin[h], Henning Müller[i,j], Sinan Onogur[a], Julio Saez-Rodriguez[k,l,m], Bram van Ginneken[n], Annette Kopp-Schneider[o], Bennett A. Landman[p]

[a] Division of Computer Assisted Medical Interventions (CAMI), German Cancer Research Center (DKFZ), Im Neuenheimer Feld 223, Heidelberg 69120, Germany
[b] Centre for Biomedical Image Analysis, Masaryk University, Botanická 68a, Brno 60200, Czech Republic
[c] Physical Sciences, Sunnybrook Research Institute, 2075 Bayview Avenue, Rm M6-609, Toronto ON M4N 3M5, Canada
[d] Department Medical Biophysics, University of Toronto, 101 College St Suite 15-701, Toronto, ON M5G 1L7, Canada
[e] Centre for Intelligent Machines, McGill University, 3480 University Street, McConnell Engineering Building, Room 425, Montreal QC H3A 0E9, Canada
[f] Institute of Information Systems Engineering, Technische Universität (TU) Wien, Favoritenstraße 9-11/194-04, Vienna 1040, Austria
[g] Complexity Science Hub Vienna, Josefstädter Straße 39, Vienna 1080, Austria
[h] Laboratoire Traitement du Signal et de l'Image (LTSI) – UMR_S 1099, Université de Rennes 1, Inserm, Rennes, Cedex 35043, France
[i] University of Applied Sciences Western Switzerland (HES-SO), Rue du Technopole 3, Sierre 3960, Switzerland
[j] Medical Faculty, University of Geneva, Rue Gabrielle-Perret-Gentil 4, Geneva 1211, Switzerland
[k] Institute of Computational Biomedicine, Heidelberg University, Faculty of Medicine, Im Neuenheimer Feld 267, Heidelberg 69120, Germany
[l] Heidelberg University Hospital, Im Neuenheimer Feld 267, Heidelberg 69120, Germany
[m] Joint Research Centre for Computational Biomedicine, Rheinisch-Westfälische Technische Hochschule (RWTH) Aachen, Faculty of Medicine, Aachen 52074, Germany
[n] Department of Radiology and Nuclear Medicine, Medical Image Analysis, Radboud University Center, Nijmegen 6525 GA, The Netherlands
[o] Division of Biostatistics, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 581, Heidelberg, 69120, Germany
[p] Electrical Engineering, Vanderbilt University, Nashville, Tennessee TN 37235-1679, USA

ARTICLE INFO

*Article history:*
Received 8 April 2019
Revised 12 June 2020
Accepted 27 July 2020
Available online 21 August 2020

*Keywords:*
Biomedical challenges
Good scientific practice
Biomedical image analysis
Guideline

ABSTRACT

The number of biomedical image analysis challenges organized per year is steadily increasing. These international competitions have the purpose of benchmarking algorithms on common data sets, typically to identify the best method for a given problem. Recent research, however, revealed that common practice related to challenge reporting does not allow for adequate interpretation and reproducibility of results. To address the discrepancy between the impact of challenges and the quality (control), the **B**iomedical **I**mage **A**nalysis Challenge**S** (BIAS) initiative developed a set of recommendations for the reporting of challenges. The BIAS statement aims to improve the transparency of the reporting of a biomedical image analysis challenge regardless of field of application, image modality or task category assessed. This article describes how the BIAS statement was developed and presents a checklist which authors of biomedical image analysis challenges are encouraged to include in their submission when giving a paper on a challenge into review. The purpose of the checklist is to standardize and facilitate the review process and raise interpretability and reproducibility of challenge results by making relevant information explicit.



## 1. Introduction

The importance of data science techniques in almost all fields of biomedicine is increasing at an enormous pace (Esteva et al., 2019; Topol, 2019). This holds particularly true for the field of biomedical image analysis, which plays a crucial role in many areas including tumor detection, classification, staging and progression modeling (Ayache and Duncan, 2016; Hosny et al., 2018; Litjens et al., 2017) as well as automated analysis of cancer cell images acquired using microscopy (Bora et al., 2017; Bychkov et al., 2018; Yu et al., 2016).

While clinical trials are the state of the art methods to assess the effect of new medication in a comparative manner

* Corresponding author.
*E-mail address:* l.maier-hein@dkfz-heidelberg.de (L. Maier-Hein).

https://doi.org/10.1016/j.media.2020.101796

(Meldrum, 2000), benchmarking in the field of image analysis is performed by so-called *challenges*. Challenges are international competitions, typically hosted by individual researchers, institutes, or societies, that aim to assess the performance of multiple algorithms on identical data sets and encourage benchmarking (Kozubek, 2016). They are often published in prestigious journals (Chenouard et al., 2014; Maier-Hein et al., 2017; Menze et al., 2015; Sage et al., 2015; Setio et al., 2017; Zheng et al., 2017), are associated with significant amounts of prize money (up to €1 million on platforms like Kaggle (2010)) and receive a huge amount of attention, indicated by the number of downloads, citations and views. A recent comprehensive analysis of biomedical image analysis challenges, however, revealed a huge discrepancy between the impact of a challenge and the quality (control) of the design and reporting standard. It was shown that (1) "common practice related to challenge reporting is poor and does not allow for adequate interpretation and reproducibility of results", (2) "challenge design is very heterogeneous and lacks common standards, although these are requested by the community" and (3) "challenge rankings are sensitive to a range of challenge design parameters, such as the metric variant applied, the type of test case aggregation performed and the observer annotating the data" (Maier-Hein et al., 2018). The authors conclude that "journal editors and reviewers should provide motivation to raise challenge quality by establishing a rigorous review process."

The Enhancing the QUAlity and Transparency Of health Research (EQUATOR) network is a global initiative with the aim of improving the quality of research publications and research itself. A key mission in this context is to achieve accurate, complete and transparent reporting of health research studies to support reproducibility and usefulness. A core activity of the network is to assist in the development, dissemination and implementation of robust reporting guidelines, where a guideline is defined as "a checklist, flow diagram or structured text to guide authors in reporting a specific type of research" (TheEQUATORNetwork (2008)). Between 2006 and 2019, more than 400 reporting guidelines have been published under the umbrella of the equator network. A well-known guideline is the CONSORT statement (Moher et al., 2001; Schulz et al., 2010) developed for reporting of randomized controlled trials. Prominent journals, such as Lancet, JAMA or the British Medical Journal require the CONSORT checklist to be submitted along with the actual paper when reporting results of a randomized controlled trial.

Inspired by this success story, the **B**iomedical **I**mage **A**nalysis Challenge**S** (BIAS) initiative was founded by the challenge working group of the Medical Image Computing and Computer Assisted Intervention (MICCAI) Society board with the goal of bringing biomedical image analysis challenges to the next level of quality.

As a first step towards better scientific practice, this paper of the initiative presents a guideline to standardize and facilitate the writing and reviewing process of biomedical image analysis challenges and help readers of challenges interpret and reproduce results by making relevant information explicit.

Please note that we do not want to put unnecessary restrictions on researchers. For this reason, the template for challenge papers as proposed in the following sections merely serves as guidance, and authors are free to arrange the relevant information in any way they want. What we regard as important is for the information in the paper to be complete, such that transparency and reproducibility can be guaranteed. For this reason, we encourage authors of challenge papers to submit the checklist presented in this manuscript (Appendix A; Maier-Hein et al., 2020) along with their paper such that reviewers can easily verify whether the information on challenge design and results is comprehensive. If information is missing (represented by "n/a" in the column *Reported on page No* of the checklist) it is up to the reviewers to request adding it.

Section 2 introduces the terminology used to describe challenges and describes the process applied to generate this guideline document. Section 3 gives recommendations on how to report the design and results of a biomedical image analysis challenge. The paper then closes with a brief discussion in section 4.

## 2. Methods

In this paper, we define a *biomedical image analysis challenge* as an open competition on a specific scientific problem in the field of biomedical image analysis (Maier-Hein et al., 2018). A challenge may encompass multiple competitions related to multiple *tasks*, whose participating teams may differ and for which separate rankings/leaderboards/results are generated. For example, a challenge may target the problem of anatomical structure segmentation in computed tomography (CT) images, where one task may refer to the segmentation of the liver and a second task may refer to the segmentation of the kidney. We use the term *case* to refer to a data set for which a participating algorithm produce one result (e.g. a segmentation or classification). Each case must include at least one image of a biomedical imaging modality.

*Metrics* are used to compute the performance of an algorithm for a given case and should reflect the property(ies) of the algorithms to be optimized. Note that we do not use the term *metric* in the strict mathematical sense. Metrics are usually computed by comparing the results of the participating team with a *reference* annotation. We prefer the term *reference* (alternatively: *gold standard*) to *ground truth* because reference annotations are typically only approximations of the (forever unknown) truth (Jannin et al., 2006).

Typically, a challenge has a *training phase* of several weeks or months, at the beginning of which the challenge organizers release training cases with corresponding reference annotations. These annotations help the participating teams develop their method (e.g. by training a machine learning algorithm). Alternatively, the training data is not directly released but participating teams may submit their algorithms to the challenge platform (using Docker containers, for example (Guinney and Saez-Rodriguez (2018))). Note that the official training phase may be preceded by a *dry run phase*. During this phase, the challenge organizers may themselves work with the data to determine the level of difficulty of the task(s), for example. In the *test phase*, participating teams either upload their algorithms, or they get access to the test cases without the reference annotations and submit the results of their algorithms on the test cases to the challenge organizers. This procedure may be replaced or complemented by an on-site challenge event in which participating teams receive a set of test cases and are asked to produce the corresponding results on-site (typically on the same day).

For many challenges a ranking of the participating teams is produced based on the metric values computed for the test cases. Note that some challenges additionally include a *validation phase* between training and test phase, in which initial rankings (so-called *leaderboards*) are generated to show participating teams how well their methods generalize. Insights in this step may be used for final parameter tuning. A glossary of some of the terms used in this paper is provided in Appendix B.

The procedure to generate this guideline document was heavily based on a previous study related to the critical analysis of common practice in challenge organization (Maier-Hein et al., 2018) and is summarized in the following paragraphs:

**Challenge capture**

To analyze the state of the art in the field, the publicly available data on biomedical image analysis challenges was acquired. To

capture a challenge in a structured manner, a list of 53 challenge parameters was compiled by a group of 49 scientists from 30 institutions worldwide. These parameters include information on the challenge organization and participation conditions, the mission of the challenge, the challenge data sets (e.g. number of training/test cases, information on imaging protocols), the assessment method (e.g. metrics and ranking scheme) and challenge outcome (e.g. rankings). Analysis of websites hosting and presenting challenges, such as grand-challenge.org, dreamchallenges.org and kaggle.com yielded a list of 150 biomedical image analysis challenges with 549 tasks performed in a time span of 12 years (Maier-Hein et al., 2018). Between 2004 and 2016, most challenges were organized in the scope of international conferences, primarily the MICCAI conference (48%) and the International Symposium on Biomedical Imaging (ISBI) (24%). More recently, an increasing number of challenges are hosted on platforms like (Kaggle (2010)), Synapse (for the DREAM challenges (DREAM, 2006; Saez-Rodriguez et al., 2016)) and crowdAI (2018) (for the ImageCLEF challenges). Details on the challenge characteristics (e.g. imaging modalities applied, algorithm categories investigated, number of training/test cases) can be found in Maier-Hein et al. (2018).

**Analysis of challenge reporting**

It was found that reports on biomedical challenges covered only a median of 62% of the 53 challenge parameters identified as relevant by the international consortium. The list of parameters often not reported include some that are crucial for interpretation of results such as information on how metrics are aggregated to obtain a ranking, whether training data provided by challenge organizers may have been supplemented by other data, and how the reference annotation was performed and by whom. It was further found that challenge design is highly heterogeneous, as detailed in Maier-Hein et al. (2018).

**Prospective structured challenge capture**

To address some of the issues, a key conclusion of Maier-Hein et al. (2018) was to publish the complete challenge design before the challenge by instantiating the list of parameters proposed. To test the applicability of this recommendation, the MICCAI board challenge working group initiated the usage of the parameter list for structured submission of challenge proposals for the MICCAI conferences 2018 and 2019. The submission system required a potential MICCAI 2018/2019 challenge organizer to instantiate at least 90% of a reduced set of 40 parameters (cf. Tab. 1 in Maier-Hein et al. (2018)) that were regarded as essential for judging the quality of a challenge design proposal. The median percentage of parameters instantiated was 100% (min: 94%) (16/25 submitted challenges in 2018/2019).

**Finalization of checklist** Based on the prospective challenge capture, the parameter list was revised by the MICCAI board challenge working group to improve clarity. A questionnaire was then sent to all co-authors to acquire final feedback on the parameters. Each author had to independently assess every single parameter ($n = 48$) of the list by answering the following questions:

1. I agree with the name (yes/sort of/no).
2. I agree with the explanation (yes/sort of/no).
3. If you do not agree with the name or the explanation, please provide constructive feedback.
4. Please rate the importance of the checklist item. If you think that it is absolutely essential for challenge result interpretation and/or challenge participation put *absolutely essential*. Otherwise choose between *should be included* and *may be omitted*.
5. Please indicate whether the checklist item(s) is (are) essential for challenge review (yes/no).

To identify missing information, participants were also asked to add further relevant checklist items that were not covered and to add any other issue important to compile the checklist. The MICCAI board challenge working group then developed a proposal to address all the comments and points of criticism raised in the poll. In a final conference call with the co-authors of this paper, remaining conflicts were resolved, and the checklist was finalized, resulting in a list of 42 main parameters and 79 sub-parameters.

The following sections describe the authors' recommendations on how to report the design and outcome of individual tasks of a biomedical image analysis challenge based on this parameter list. The corresponding reporting guideline is provided in Appendix A and was uploaded to Zenodo (2013) to ensure version control (Maier-Hein et al., 2020).

## 3. Guideline for challenge reporting

Following standard scientific writing guidelines, we propose dividing a challenge paper into the sections *Introduction, Methods, Results* and *Discussion*, where the *Methods* section corresponds to the challenge design and the *Results* section corresponds to the challenge outcome. These sections are preceded by a concise title and abstract as well as a list of representative keywords to summarize the challenge mission and outcome. The following sections give basic recommendations on how to structure and write the individual sections. Appendix A (Maier-Hein et al., 2020) serves as a structured summary of this section.

### 3.1. Title, Abstract and Keywords

The *title* should convey the essential information on the challenge mission. In particular, it should identify the paper as a biomedical image analysis challenge and indicate the image modality(ies) applied as well as the task category (e.g. classification, segmentation) corresponding to the challenge. The *abstract* should serve as a high-level summary of the challenge purpose, design and results and report the main conclusions. The *keywords* should comprise the main terms characterizing the challenge.

### 3.2. Introduction: Research Context

The first section should provide the *challenge motivation and objectives* from both a *biomedical* and *technical* point of view. It should summarize the most important related work and clearly outline the expected impact of the challenge compared to previous studies. The task to be solved/addressed by the challenge should be explicitly stated, and the section should clarify whether the challenge mainly focuses on comparative benchmarking of existing solutions or whether there is a necessity of improving existing solutions.

### 3.3. Methods: Reporting of Challenge Design

The challenge design parameters to be reported are classified in four categories related to the topics *challenge organization, mission of the challenge, challenge data sets*, and *assessment method*. The following paragraphs summarize the information that should be provided in the corresponding subsections.

#### 3.3.1. Challenge organization

This section should include all of the relevant information regarding challenge organization and participation conditions. This information can either be reported in the main document or be provided as supplementary information (e.g. using the form provided in *Suppl 1*). It should include the *challenge name* (including *acronym* (if any)) as well as information on the *organizing team* and the intended challenge *life cycle type*. Note that not every challenge closes after the submission deadline (one-time

event). Sometimes it is possible to submit results after the deadline (open call/continuous benchmarking) or the challenge is repeated with some modifications (repeated event). Information on *challenge venue and platform* should include the event (e.g. conference, if any) that the challenge was associated with, the platform that was applied to run the challenge as well as a link to the *challenge website* (if any).

Comprehensive information about *participation policies* should be related to the *interaction level policy* (e.g. only fully-automatic methods allowed), the *training data policy* (indicating which data sets could be used to complement the data sets provided by the challenge (if any)), the *award policy*, which typically refers to challenge prizes and the *results announcements* (e.g. only names of top 3 performing teams will be publicly announced). It should also contain information about the *organizer participation policy*. A policy related to this aspect may be, for example, that members of the organizers' institutes could participate in the challenge but are not eligible for awards and are not listed in the leaderboard. Crucially, annotators of the test data should generally not be allowed to annotate additional training data that is exclusively provided to only one/some of the participating teams[1]. Finally, details on the *publication policy* should be provided: Do all participating teams automatically qualify as co-authors? Or only the top performing ones (theoretically, this could prevent people from participating with an arbitrary method just for the sake of being an author of a highly cited paper)? Who of the participating teams' members qualifies as an author (e.g. fixed maximum number per team? All team members? First author and supervisor (if any)?)? Can the participating teams publish their results separately? If so: After an embargo?

The section should further contain information on the *submission method*, preferably including a link to the *instructions* that the participating teams received. It should also include information on the procedure for *evaluating the algorithms before the best runs/the final method were submitted* for final performance assessment.

Information on the *challenge schedule* should focus on the date(s) of training, validation (if any) and test data release as well as on the submission of algorithm results on test data, the associated workshop days (if any) and the release date of the results. In some challenges, a post-competition collaborative phase can take place, where e.g. top teams are brought together to further improve on solutions. This should be explicitly mentioned in the schedule.

Crucially, information related to the challenge organization should include information on the *ethics approval* (if applicable) and the *data usage agreement* (indicating who may use the data for which purposes under which conditions). Similarly, information on *code availability* should be provided explicitly relating to both the *organizers'* and the *participating teams'* software. To make *conflicts of interest* transparent, the section should also list the funding/sponsoring associated with the challenge and should explicitly state who had access to the test case(s) labels and when. Finally, the *author contributions* should be explicitly listed in the supplementary material.

### 3.3.2. Mission of the challenge

This paragraph should state the biomedical application (*field of application*, e.g. diagnosis, screening, intervention planning) and the *task category* (e.g. segmentation, classification, retrieval, detection) that the participating teams' algorithms were designed for. To refer to the subjects (e.g. patients)/objects (e.g. physical phantoms) from whom/which the image data was acquired, we use the term *cohort*. The paper should explicitly distinguish between the *target cohort*, which refers to the subject(s)/object(s) from whom/which the data would be acquired in the final biomedical application (e.g. healthy subjects who undergo screening) and the *challenge cohort*, defined as the subject(s)/object(s) from whom/which the challenge data was acquired (e.g. white male healthy subjects from Germany who participated in a voluntary study X). Note that this differentiation is crucial to understand the potential "domain gap" when transferring challenge results to the actual application. Important differences may be related to various aspects including the subject(s)/object(s) from whom/which the image data is acquired (e.g. cancer patients in real application vs. porcine models in a challenge), the source of image data (e.g. various CT scanners in real application vs. a specific scanner in the challenge) and the characteristics of data (e.g. fraction of malignant cases in real world vs. equal number of malignant and benign cases in the challenge).

To describe the cohorts in detail, the section should also include information on the *imaging modalities*, and additional *context information* (e.g. clinical data) acquired. Most challenges performed to date are based solely on images and corresponding reference annotations (e.g. tumor labels), yet an increasing number of competitions provide further information on the patients, such as general information (age, gender), or laboratory results.

The section should further state the *target entity(ies)* which includes the *data origin*, i.e. the region from which the image data is acquired (e.g. scan of the head, video of the whole operating theater) and the *algorithm target* defined as the structure (e.g. tumor in the brain), object (e.g. robot), subject (e.g. nurse) or component (e.g. tip of a medical instrument) that the participating algorithms focus on.

Finally, it should provide a concise statement of the *assessment aim(s)* (e.g. finding the most sensitive lesion detection algorithm vs. identifying the fastest algorithm that provides a median detection accuracy below a certain threshold). The metric(s) and ranking scheme chosen (parameters 29 and 30 in *assessment method, Appendix A*, Maier-Hein et al., 2020) should reflect the assessment aims as closely as possible, i.e. optimizing the metrics will ideally optimize the properties of the algorithm that are important according to the assessment aim. Note that it is necessary to make the assessment aim explicit, as it may not be straightforward to find an appropriate metric for certain properties to be optimized.

### 3.3.3. Challenge data sets

While the information contained in the challenge mission section should refer to both the target cohort and the challenge cohort, this section is exclusively dedicated to the challenge cohort. It should start with a description on the *data source(s)*. This should include information on specific *acquisition devices* (e.g. the specific type of magnetic resonance imaging (MRI) scanner used), *acquisition protocols* (e.g. the specific MRI imaging protocol applied) as well as *centers/data providing source(s)* and *operators* that were involved in the data acquisition (e.g. a specific robot in a specific university clinic). If the centers involved cannot be mentioned due to requirements for anonymity, this should be made explicit. Information on the operators should focus on the *relevant* information in the challenge context. It may, for example, be irrelevant to state the years of experience of the person acquiring an MRI image according to an established protocol whereas, for data derived from a complex surgical procedure, it may be crucially important to explicitly list the level of expertise of the surgeon.

The section should further provide information on the *training and test case characteristics*. It should begin with stating explicitly *what data encompasses a single case*, i.e. which data is meant to

[1] Remark of the authors: Such a case of intentional or unintentional "cheating" has occurred in the past.

be processed to produce one result that is compared to the corresponding reference result. Information on the cases should further include information on the *number of training/test cases* as well as on *why* a specific proportion of training/test data was chosen, why a certain total number of cases was chosen and why certain *characteristics* were chosen for the training/test set (e.g. class distribution according to real-world distribution vs. equal class distribution).

Information on the *annotation characteristics* should begin with describing the *general approach to training/test case annotation* (e.g. annotation of the test data by a medical expert vs. annotation of the training data via crowdsourcing such as in Maier-Hein et al. (2014)). It should include the *instructions* given to the annotators prior to the annotation, 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) and the *method(s) used to merge multiple annotations* for one case. All information should be provided separately for the training, validation and test cases if necessary.

*Data pre-processing methods* (if any) used to process the raw data should also be well-described and justified. Crucially, potential *sources of errors* related to the *annotation* but also the *data acquisition* should be comprehensively described. Sources of error related to the data acquisition may, for example, refer to calibration errors of the image modality, tracking errors related to pose computation of surgical instruments or errors resulting from substantial motion during image acquisition. Preferably, a quantitative analysis (e.g. using the concept of intra-annotator and inter-annotator variability) should be performed to estimate the magnitude of the different error sources.

#### 3.3.4. Assessment method

Information on the assessment method are related to the metric(s) applied, the ranking scheme as well the statistical analyses.

The *metric(s)* to assess a property of an algorithm should be well-explained including the parameters used (if any) and preferably with reference to a paper. The metrics should be justified in the context of the challenge objective (parameter *assessment aim(s)*) and the biomedical application. For example, the Dice similarity coefficient (DSC) is a well-suited metric for assessing segmentation accuracy for large structures, such as organs in images, but is not well-suited for quantifying segmentation accuracy in the case of small pathologies.

If one or multiple rankings were generated for the challenge, the *ranking method(s)* should be specified by describing how metric values are aggregated/used to generate a final ranking (if any). It should also provide information on the rank value in case of tied positions, as well as on methods used to manage submissions with missing results on test cases (*missing data handling*) and to handle any diversity in the level of user interaction when generating the performance ranking (interaction level handling). The section should further make explicit how the ranking chosen matches the assessment aim.

Details for all the statistical methods used to analyze the data should be provided. If results on test cases were entered as missing values, it should be described how these were handled in the *statistical analyses*. Further, details about the assessment of the robustness of the ranking should be provided. If statistical hypothesis tests were used to compare, e.g. participating teams, details about the statistical method should be provided including a description of any method used to assess whether the data met the assumptions required for the particular statistical approach. For all data analysis methods, the software product used should be mentioned. Preferably, the code should also be released along with the paper.

*Further analyses* performed should also be reported in this section. This includes experiments based on combining individual algorithms via ensembling and experiments on inter-algorithm variability, for example.

### 3.4. Results: Reporting of Challenge Outcome

We suggest subdividing the outcome section into five categories: *Challenge submission, information on selected participating teams, metric values, rankings* and *further analyses.*

At first, information on the *submissions* received should be summarized. This includes the *number of registrations*, the *number of (valid) submissions* (if applicable: in each phase) and the *number of participating teams* that the challenge paper is about (selected participating teams) with *justification* why these participating teams were chosen (e.g. top *n* performing teams; teams with a metric value above a threshold; *m*% top performing teams). Depending on the number of participating teams, *information on selected participating teams* can be provided in the main document or in the appendix. Information on those teams referred to in the results section should include a *team identifier* (as name of the team) as well as a *description of the method*. The latter can for example be a brief textual summary plus a link to a document that provides detailed information not only on the basic method but also on the specific parameters/optimizations performed for the challenge. Ideally, this reference document should also provide information on complexity analysis with respect to time and memory consumption, hardware/OS requirements and reference to the source code.

Depending on the number of test cases and participating teams, raw metric values (i.e. metric values for each test case) and/or aggregated *metric values* should be provided for all participating teams/the selected teams. Parts of these results may be moved to the appendix.

The *ranking(s)* (if any) should be reported including the results on robustness analyses (e.g. bootstrapping results (Wiesenfarth et al., 2019)) and other *statistical analyses*. Again, depending on the number of participating teams, the paper may refer to only the top performing teams (referred to as selected participating teams above). Depending on the number of participating teams, full (if necessary partially anonymized) ranking(s) should be provided in the main document, as supplementary material or in another citable document.

The results of *further analyses* performed (if any) should also be reported in this section. This includes analyses of common problems/biases of the methods.

### 3.5. Discussion: Putting the Results into Context

The final section should provide a concise *summary* of the challenge outcome and discuss the findings of the challenge thoroughly in the context of the state of the art. It should clearly distinguish between the *technical and biomedical impact*. Current performance of the best methods should be *discussed* and conclusions drawn about *whether the task is already solved* in a satisfactory way (e.g. the remaining errors are comparable to inter-annotator variability). Furthermore, an *analysis of individual cases*, in which the majority of algorithms performed poorly (if any), should be included. Also, *advantages and disadvantages of the participating methods* should be discussed. In this context, it should be made explicit whether an algorithm with clearly superior performance could be identified or if more than one algorithm is well-suited for the specific task. Furthermore, *limitations of the challenge* should be made explicit (design *and* execution). Finally, concrete *recommendations for future work* should be provided and a *conclusion* drawn.

## 4. Discussion

As a first step to address the discrepancy between the impact of biomedical image analysis challenges and the quality (control),

the BIAS initiative aims to improve the transparency of the *reporting*. This article describes how the BIAS statement was developed and presents a checklist which authors of biomedical image analysis challenges are encouraged to include in their submission when giving a challenge paper into review. By making relevant information explicit, the checklist has the purpose to standardize and facilitate the reviewing/editorial process and raise interpretability and reproducibility of challenge results.

The checklist generated in the scope of this article relies heavily on the challenge parameter list published in our previous work (Maier-Hein et al., 2018). In the meantime, this parameter list has been instantiated with more than 500 tasks from more than 150 challenges, both retrospectively (Maier-Hein et al., 2018) and prospectively in the scope of the structured challenge submission system used for MICCAI 2018 and 2019. According to our experience, the (updated) list presented in this work should be appropriate to capture the relevant information on current challenge design and organization. It is worth noting, however, that an update of the checklist may be required at a later point in time. For example, ongoing research is investigating the generation of *probabilistic output* for a whole range of algorithm categories (Kohl et al., 2018); rather than providing a single contour as result for a segmentation task. For instance, such methods produce a whole range of *plausible* solutions via sampling. It is currently unknown how such output may be efficiently handled in the design of future challenges.

An increasingly relevant problem is that it typically remains unknown which specific feature of an algorithm actually makes it better than competing algorithms (Maier-Hein et al., 2018). For example, many researchers are convinced that the method for data augmentation often has a much bigger influence on the performance of a deep learning algorithm than the network architecture itself. For this reason, a structured description (e.g. using ontologies) not only of the challenge but also of the participating algorithms may be desirable. Due to the lack of common software frameworks and terminology, however, this is not trivial to implement at this stage (Maier-Hein et al., 2018).

It is worth mentioning that our guideline has explicitly been developed for reporting the design and results for *one task* of a challenge. If a challenge includes multiple tasks, the results should preferably be reported in separate publications. If this is not desirable (i.e. a single paper refers to multiple tasks with substantial overlap between tasks such as tasks sharing the same data sets), a separate checklist for each task should be generated. Alternatively, a single checklist may be provided in which some items are common to all tasks and other items contain separate parts for each task.

It should also be noted that challenges could in theory focus on collaboration rather than competition. In such collaborative challenges, the participating teams would work jointly on a dedicated problem, and the focus would be on solving a problem together rather than benchmarking different methods. We have not explicitly addressed such collaborative challenges with the checklist.

We believe that the work invested to improve challenge reporting could also be valuable in guiding challenge design. For this reason, we have converted the reviewer checklist into a document that can be used to comprehensively report the envisioned design of a challenge and could thus be used to review a challenge before it is organized (*Suppl 2*). Based on this document, the MICCAI society and MICCAI 2020 organizing team decided to introduce the concept of challenge registration. Similar to how clinical trials have to be registered before they are started, the complete design of accepted MICCAI challenges had to be put online before challenge execution. This was achieved with Zenodo (2013), a general-purpose open-access repository that allows researchers to deposit data sets, software, and other research-related items. Such stored items are citable, because a persistent digital object identifier (DOI) is generated for each submission. As Zenodo also supports version control, changes to a challenge design (e.g. to the metrics or ranking schemes applied) can be made transparent. These changes must be communicated to the MICCAI society and well-justified. To date (June 2020), 8 out of the 28 challenges committed changes to the designs, originally uploaded in April 2020. Most of them were changes to the schedule, which can be attributed to the COVID-19 outbreak. We believe that the transparency and quality control that comes along with challenge registration is a big step towards higher quality of biomedical challenges.

Challenges are becoming increasingly important in various fields, ranging from protein structure, to systems biology, text mining, and genomics, thanks to initiatives such as CASP (Critical Assessment of Techniques for Protein Structure Prediction) (Moult et al., 2017), BioCreative (Critical Assessment of Information Extraction in Biology (Doğan et al., 2019)), DREAM (Dialogue for Reverse Engineering Assessment and Methods (Saez-Rodriguez et al., 2016)), and CAGI (Critical Assessment of Genome Interpretation (Daneshjou et al., 2017)). The checklist and the challenge design document could be adapted to these research areas and thus contribute substantially to better scientific practice related to challenges in general.

In conclusion, this document is the first to provide a guideline for the reporting of a biomedical image analysis challenge regardless of field of application, image modality or algorithm category assessed. We hope that the checklist provided will help editors of journals in the field of biomedical image analysis and beyond to establish a rigorous review process with the mid-term goal of increasing interpretability and reproducibility of results and raising the quality of challenge design in general.

## Declaration of Competing Interest

Anne Martel is a co-founder, CSO of Pathcore, Toronto, Canada. The remaining authors declare no competing interests.

## Acknowledgments

This project was conducted in the scope of the Helmholtz Imaging Platform (HIP) funded by the Helmholtz Association of German Research Centres. We acknowledge support from the European Research Council (ERC) under the New Horizon Framework Programme grant agreement ERC-2015-StG-37960 (ERC starting grant COMBIOSCOPY) as well as the Natural Science and Engineering Research Council (NSERC) (RGPIN-2016-06283), the Canadian Cancer Society (#705772), the Ministry of Education, Youth and Sports of the Czech Republic (Projects LTC17016 and CZ.02.1.01/0.0/0.0/16_013/0001775), the National Science Foundation 1452485 and the National Institutes of Health R01-EB017230-01A1.

## Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.media.2020.101796.

## References

Ayache, N., Duncan, J., 2016. 20th anniversary of the medical image analysis journal (MedIA). Med. Image Anal. 33, 1–3. doi:10.1016/j.media.2016.07.004.

Bora, K., Chowdhury, M., Mahanta, L.B., Kundu, M.K., Das, A.K., 2017. Automated classification of pap smear images to detect cervical dysplasia. Comput. Methods Programs Biomed. 138, 31–47. doi:10.1016/j.cmpb.2016.10.001.

Bychkov, D., Linder, N., Turkki, R., Nordling, S., Kovanen, P.E., Verrill, C., Walliander, M., Lundin, M., Haglund, C., Lundin, J., 2018. Deep learning based tissue analysis predicts outcome in colorectal cancer. Sci. Rep. 8 (1), 3395. doi:10.1038/s41598-018-21758-3.

Chenouard, N., Smal, I., De Chaumont, F., Maška, M., Sbalzarini, I.F., Gong, Y., Cardinale, J., Carthel, C., Coraluppi, S., Winter, M., Cohen, A.R., Godinez, W.J., Rohr, K., Kalaidzidis, Y., Liang, L., Duncan, J., Shen, H., Xu, Y., Magnusson, K.E.G., Jaldé, Blau, H.M., Paul-Gilloteaux, P., Roudot, P., Kervrann, C., Waharte, F., Tinevez, J.-Y., Shorte, S.L., Willemse, J., Celler, K., van Wezel, G.P., Dan, H.-W., Tsai, Y.-S., Ortiz de Solórzano, C., Olivo-Marin, J.-C., Meijering, E., 2014. Objective comparison of particle tracking methods. Nat. Methods 11 (3), 281–289. doi:10.1038/nmeth.2808.

crowdAI, 2018. crowdAI. www.crowdai.org/. Accessed: 2019-04-01.

Daneshjou, R., Wang, Y., Bromberg, Y., Bovo, S., Martelli, P.L., Babbi, G., Lena, P.D., Casadio, R., Edwards, M., Gifford, D., Jones, D.T., Sundaram, L., Bhat, R.R., Li, X., Kundu, K., Yin, Y., Moult, J., Jiang, Y., Pejaver, V., Pagel, K.A., Li, B., Mooney, S.D., Radivojac, P., Shah, S., Carraro, M., Gasparini, A., Leonardi, E., Giollo, M., Ferrari, C., Tosatto, S.C.E., Bachar, E., Azaria, J.R., Ofran, Y., Unger, R., Niroula, A., Vihinen, M., Chang, B., Wang, M.H., Franke, A., Petersen, B.S., Prooznia, M., Zandi, P., McCombie, R., Potash, J.B., Altman, R.B., Klein, T.E., Hoskins, R.A., Repo, S., Brenner, S.E., Morgan, A.A., 2017. Working toward precision medicine: Predicting phenotypes from exomes in the critical assessment of genome interpretation (CAGI) challenges. Hum. Mut. 38 (9), 1182–1192. doi:10.1002/humu.23280.

Doğan, R.I., Kim, S., Chatr-aryamontri, A., Wei, C.-H., Comeau, D.C., Antunes, R., Matos, S., Chen, Q., Elangovan, A., Panyam, N.C., Verspoor, K., Liu, H., Wang, Y., Liu, Z., Altinel, B., Hsnbeyi, Z.M., Zgr, A., Fergadis, A., Wang, C.-K., Dai, H.-J., Tran, T., Kavuluru, R., Luo, L., Steppi, A., Zhang, J., Qu, J., Lu, Z., 2019. Overview of the BioCreative VI precision medicine track: mining protein interactions and mutations for precision medicine. Database (Oxford) 2019. doi:10.1038/10.1093/database/bay147.

DREAM, 2006. Dream challenges. dreamchallenges.org/. Accessed: 2019-04-01.

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., Dean, J., 2019. A guide to deep learning in healthcare. Nat. Med. 25 (1), 24. doi:10.1038/s41591-018-0316-z.

Guinney, J., Saez-Rodriguez, J., 2018. Alternative models for sharing confidential biomedical data. Nat. Biotechnol. 36 (5), 391. doi:10.1038/nbt.4128.

Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L.H., Aerts, H.J., 2018. Artificial intelligence in radiology. Nat. Rev. Cancer 18, 500. doi:10.1038/s41568-018-0016-5.

Jannin, P., Grova, C., Maurer, C.R., 2006. Model for defining and reporting reference-based validation protocols in medical image processing. Int. J. CARS 1 (2), 63–73. doi:10.1007/s11548-006-0044-6.

Kaggle, 2010. Your home for data science. www.kaggle.com/. Accessed: 2019-02-12.

Kohl, S., Romera-Paredes, B., Meyer, C., De Fauw, J., Ledsam, J.R., Maier-Hein, K., Eslami, S.A., Rezende, D.J., Ronneberger, O., 2018. A probabilistic U-net for segmentation of ambiguous images. In: Adv. Neural Inf. Process. Syst. (NIPS 2018), 31, pp. 6965–6975.

Kozubek, M., 2016. Challenges and benchmarks in bioimage analysis. In: Focus Bio Image Inform.. Springer, pp. 231–262. doi:10.1007/978-3-319-28549-8_9.

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. Med. Image Anal. 42, 60–88. doi:10.1016/j.media.2017.07.005.

Maier-Hein, K.H., Neher, P.F., Houde, J.-C., Côté, M.-A., Garyfallidis, E., Zhong, J., Chamberland, M., Yeh, F.-C., Lin, Y.-C., Ji, Q., Reddick, W.E., Glass, J.O., Qixiang Chen, D., Feng, Y., Gao, C., Wu, Y., Ma, J., Renjie, H., Li, Q., Westin, C.-F., Deslauriers-Gauthier, S., Omar Ocegueda González, J., Paquette, M., St-Jean, S., Girard, G., Rheault, F., Sidhu, J., Tax, C.M.W., Guo, F., Mesri, H.Y., Dávid, S., Froeling, M., Heemskerk, A.M., Leemans, A., Boré, A., Pinsard, B., Bedetti, C., Desrosiers, M., Brambati, S., Doyon, J., Sarica, A., Vasta, R., Cerasa, A., Quattrone, A., Yeatman, J., Khan, A.R., Hodges, W., Alexander, S., Romascano, D., Barakovic, M., Auría, A., Esteban, O., Lemkaddem, A., Thiran, J.-P., Cetingul, H.E., Odry, B.L., Mailhe, B., Nadar, M.S., Pizzagalli, F., Prasad, G., Villalon-Reina, J.E., Galvis, J., Thompson, P.M., De Santiago Requejo, F., Laguna, P.L., Lacerda, L.M., Barrett, R., Dell'cqua, F., Catani, M., Petit, L., Caruyer, E., Daducci, A., Dyrby, T.B., Holland-Letz, T., Hilgetag, C.C., Bram, S., Descoteaux, M., 2017. The challenge of mapping the human connectome based on diffusion tractography. Nat. Commun. 8 (1), 1349. doi:10.1038/s41467-017-01285-x .

Maier-Hein, L., Eisenmann, M., Reinke, A., Onogur, S., Stankovic, M., Scholz, P., Arbel, T., Bogunovic, H., Bradley, A.P., Carass, A., Feldmann, C., Frangi, A.F., Full, P.M., van Ginneken, B., Hanbury, A., Honauer, K., Kozubek, M., Landman, B.A., Mrz, K., Maier, O., Maier-Hein, K., Menze, B.H., Müller, H., Neher, P.F., Niessen, W., Rajpoot, N., Sharp, G.C., Sirinukunwattana, K., Speidel, S., Stock, C., Stoyanov, D., Taha, A.A., van der Sommen, F., Wang, C.-W., Weber, M.-A., Zheng, G., Jannin, P., Kopp-Schneider, A., 2018. Why rankings of biomedical image analysis competitions should be interpreted with care. Nat. Commun. 9 (1), 5217. doi:10.1038/s41467-019-08563-w.

Maier-Hein, L., Mersmann, S., Kondermann, D., Stock, C., Kenngott, H.G., Sanchez, A., Wagner, M., Preukschas, A., Wekerle, A.-L., Helfert, S., Bodenstedt, S., Speidel, S., 2014. Crowdsourcing for reference correspondence generation in endoscopic images. In: Int. Conf. Med. Image Comput. Comp. Assis. Interv.. Springer, pp. 349–356. doi:10.1007/978-3-319-10470-6_44.

Maier-Hein, L., Reinke, A., Kozubek, M., Martel, A.L., Arbel, T., Eisenmann, M., Hanbury, A., Jannin, P., Müller, H., Onogur, S., Saez-Rodriguez, J., van Ginneken, B., Kopp-Schneider, A., Landman, B.A., 2020. Biomedical Image Analysis Challenges (BIAS) Reporting Guideline. Zenodo doi:10.5281/zenodo.4008953.

Meldrum, M.L., 2000. A brief history of the randomized controlled trial: From oranges and lemons to the gold standard. Hematol. Oncol. Clin. North Am. 14 (4), 745–760. doi:10.1016/S0889-8588(05)70309-9.

Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., Lanczi, L., Gerstner, E., Weber, M.-A., Arbel, T., Avants, B.B., Ayache, N., Buendia, P., Collins, D.L., Cordier, N., Corso, J.J., Criminisi, A., Das, T., Delingette, H., Demiralp, C., Durst, C.R., Dojat, M., Doyle, S., Festa, J., Forbes, F., Geremia, E., 2015. The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans. Med. Imaging 34 (10), 1993–2024. doi:10.1109/TMI.2014.2377694.

Moher, D., Schulz, K.F., Altman, D.G., 2001. The consort statement: revised recommendations for improving the quality of reports of parallel group randomized trials. BMC Med. Res. Methodol. 1 (1), 2.

Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., Tramontano, A., 2017. Critical assessment of methods of protein structure prediction (CASP) – Round XII. Proteins 86, 7–15. doi:10.1002/prot.25415.

Saez-Rodriguez, J., Costello, J.C., Friend, S.H., Kellen, M.R., Mangravite, L., Meyer, P., Norman, T., Stolovitzky, G., 2016. Crowdsourcing biomedical research: leveraging communities as innovation engines. Nat. Rev. Genet. 17 (8), 470–486. doi:10.1038/nrg.2016.69.

Sage, D., Kirshner, H., Pengo, T., Stuurman, N., Min, J., Manley, S., Unser, M., 2015. Quantitative evaluation of software packages for single-molecule localization microscopy. Nat. Methods 12 (8), 717–724. doi:10.1038/nmeth.3442.

Schulz, K.F., Altman, D.G., Moher, D., 2010. Consort 2010 statement: updated guidelines for reporting parallel group randomised trials. BMC Med. 8 (1), 18.

Setio, A.A.A., Traverso, A., De Bel, T., Berens, M.S., van den Bogaard, C., Cerello, P., Chen, H., Dou, Q., Fantacci, M.E., Geurts, B., van der Gugten, R., Heng, P.A., Jansen, B., de Kaste, M.M.J., Kotov, V., Lin, J.Y.-H., Manders, J.T.M.C., Sóñora Mengana, A., García-Naranjo, J.C., Papavasileiou, E., Prokop, M., Saletta, M., Schaefer-Prokop, C.M., Scholten, E.T., Scholten, L., Snoeren, M.M., Torres, E.L., Vandemeulebroucke, J., Walasek, N., Zuidhof, G.C.A., van Ginneken, B., Jacobs, C., 2017. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge. Med. Image Anal. 42, 1–13. doi:10.1016/j.media.2017.06.015.

TheEQUATORNetwork, 2008. The EQUATOR network – Enhancing the QUAlity and Transparency Of health Research. http://www.equator-network.org. Accessed: 2019-09-12.

Topol, E.J., 2019. High-performance medicine: the convergence of human and artificial intelligence. Nat. Med. 25 (1), 44. doi:10.1038/s41591-018-0300-7.

Wiesenfarth, M., Reinke, A., Landman, B.A., Cardoso, M.J., Maier-Hein, L., Kopp-Schneider, A., 2019. Methods and open-source toolkit for analyzing and visualizing challenge results. arXiv preprint arXiv:1910.05121 Submitted for publication.

Yu, K.-H., Zhang, C., Berry, G.J., Altman, R.B., Ré, C., Rubin, D.L., Snyder, M., 2016. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. Nat. Commun. 7, 12474. doi:10.1038/ncomms12474.

Zenodo, 2013. https://zenodo.org/(accessed: June 2020).

Zheng, G., Chu, C., Belavỳ, D.L., Ibragimov, B., Korez, R., Vrtovec, T., Hutt, H., Everson, R., Meakin, J., Andrade, I.L.A., Glocker, B., Chen, H., Dou, Q., Heng, P.-A., Qang, C., Forsberg, D., Neubert, A., Fripp, J., Urschler, M., Stern, D., Wimmer, M., Novikov, A.A., Cheng, H., Armbrecht, G., Felsenberg, D., Li, S., 2017. Evaluation and comparison of 3D intervertebral disc localization and segmentation methods for 3D T2 MR data: A grand challenge. Med. Image Anal. 35, 327–344. doi:10.1016/j.media.2016.08.005.