nature portfolio

Corresponding author(s): Lena Maier-Hein

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

Reporting Summary

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

Statistics

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

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

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

Software and code

Policy information about availability of computer code

Data collection	The following softwares have been utilized for data collection: Annotations were obtained via Amazon Mechanical Turk (www.mturk.com) and understand.ai (www.understand.ai), Google Forms with informed consent.
Data analysis	The following open source libraries have been utilized for the data analysis pipeline (all of them are python (python version 3.8.10). Packages to be installed via "pip install <package-name>"): cmcrameri==1.4, cycler==0.11.0, fonttools==4.33.3, kiwisolver==1.4.3, matplotlib==3.4.3, numpy==1.22.4, packaging==21.3, pandas==1.4.2, Pillow==9.1.1, pyparsing==3.0.9, python-dateutil==2.8.2, pytz=2022.1, scipy=1.8.1, seaborn==0.11.2, six==1.16.0. For the two-part mixed models: R version 4.0.2/4.0.3 (package brms version 2.16.0, lme4 version 1.1.27.1, glmmTMB version 1.1.2). The code repository is publicly available at https://github.com/IMSY-DKFZ/labeling_instructions_matter.

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

Data

Policy information about availability of data

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

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

Six data sets were utilized during the current study: DS1: Captured biomedical competition design documents from publicly available sources (2020-2021). DS2: Reference annotations for the Heidelberg colorectal data set for surgical data science in the sensor operating room. DS3: Captured annotations from professional annotators and Amazon Mechanical Turk crowdworkers. DS4: Individual professional annotator responses to survey "Labeling Instructions Survey". DS5: Individual MTurk crowdworker responses to optional work characteristics survey. These questions are a subset of the DS4 questions. DS6: Dice Similarity Coefficient scores between DS2, DS3 and the existing output of six algorithms from a recent medical instrument instance segmentation challenge. For DS1, the individual challenge design documents are freely available at MICCAI (https://miccai.org/index.php/special-interest-groups/challenges/miccai-registered-challenges/). A reporting summary for the evaluation is available as Supplementary Material. DS2 is freely available from Synapse (www.synapse.org/#!Synapse:syn18779624/wiki). DS3, DS4, DS5 and DS6 are available from the corresponding author L.M.-H. upon reasonable request.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	-/-
Population characteristics	-/-
Recruitment	-/-
Ethics oversight	-/-

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

Field-specific reporting

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

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

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

Life sciences study design

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

Sample size	International professional annotator survey: The survey was distributed among five internationally operating annotation companies in best-cost countries that exclusively employ professional annotators and 363 entries were obtained. MICCAI competition analysis: We included all MICCAI registered competitions which were published until the end of 2021. This resulted in a list of 53 competitions with 96 competition tasks (a competition can have n discrete tasks). Annotations: Half of the reference annotations were selected by their category, which had the largest impact on Machine Learning models (Ross et al. 2021). The other half was selected randomly. Over 14,000 instance-aware segmentated images were obtained in the process. Annotators: The number and diversity of annotation providers exceeds all previous studies (acc. to our knowledge) with five different providers with a total of 156 professional annotators and 708 crowdworkers. The professional annotation companies operate internationally and their annotators are located in best-cost countries. The selected companies had a proven track record in large-scale industry annotation projects. The annotators were assigned by the companies to mimic a regular project conducted with professional annotation companies. Participating crowdworkers on Amazon Mechanical Turk (MTurk) had to fulfill the following quality requirements: a) 98 % accepted Human Intelligence Tasks (HITs) and b) a minimum of 5,000 accepted HITs. Our quality requirements thus far surpassed the quality requirements of researchers working with MTurk, which normally require 95 % accepted HITs with a minimum of 100 accepted HITs (Kennedy et al. 2020, Heim et al. 2020, Bragg et al. 2018, Cheplygina et al. 2016). To capture a representative sample of the population on MTurk, we spread our HITs across 40 days and all times of day. We followed Litman et al. to ensure a fair worker compensation (Litman et al. 2015). The existing output of six instance segmentation algorithms includes 234 instance segmentation masks per algorithm, resulting in 85,644 mask comparisons for the labeling for validation/testing.
Data exclusions	International professional annotator survey: To increase the statistical validity of the submitted entries (n = 363), we employed a twofold filtering strategy of the entries: a) a control question pair and b) an instructional manipulation check, as recommended by Oppenheimer et al. (Oppenheimer et al. 2009). The filtering resulted in 298 remaining entries. MICCAI competition analysis: 20 out of the 96 competition tasks were excluded as not applicable in the process, as they provided a valid reasoning why labeling instructions cannot be provided. An example is the Medical Out-of-Distribution Analysis Challenge (Zimmerer et al. 2021), where the publication of the labeling instruction would enable cheating, because it contains information about the placement of out-of-distribution objects in the competition data. The selection process of

	the tasks is reported in the standardized PRISMA statement (see Supplementary Material). Annotations: No data was excluded from the analyses. In the rare event of an obvious spammer, the annotations were declined via MTurk upfront and not entered in the pipeline. Existing instance segmentation algorithms output: We excluded the annotation masks from one team which did not adhere to the challenge rules (Ross et al. 2021), resulting in six included algorithmic outputs.
Replication	The analysis of the data (survey, competition analysis and annotation analysis) was repeated by 3 different people on 5 different computing systems and obtained the same results following the instructions in the repository/Methods section.
Randomization	All annotators were randomly selected within their respectice group. Each annotator is associated with a single predefined annotation provider. No worker selection was performed for either professional annotators or Amazon Mechanical crowdworkers. For the existing algorithms output, the originating algorithm was modelled as a random factor in the two part mixed model.
Blinding	Blinding is not relevant. The annotation providers are predefined (for recruitment information: see above) and all data is processed by the same data pipeline. No annotation information extending the written labeling instructions was shared with any annotation provider.

Reporting for specific materials, systems and methods

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

Materials & experimental systems

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

Methods

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