Online Resource 5: TRIPOD+AI checklist

Article: When Does Supervision Beat Reconstruction? Magnification Shift and the Robustness of AI Quality Control in Digital Pathology

Journal: Journal of Medical Systems

Authors: Konstantinos Moutselos, Ilias Maglogiannis

Corresponding author: Konstantinos Moutselos, Department of Digital Systems, University of Piraeus, Piraeus, Greece; kmouts@unipi.gr

Reference numbers in square brackets refer to the reference list of the article. Sections, tables and figures without the prefix S are those of the article.

The checklist items below are reproduced from the TRIPOD+AI statement [26] (Collins GS et al., BMJ 2024;385:e078378, https://doi.org/10.1136/bmj-2023-078378; CC BY 4.0), without its footnotes. D, development; E, evaluation; D;E, both. Locations are given by section, table and figure of the article and by Online Resource (OR), because page and line numbers change with typesetting.

This study develops artifact detectors from annotated public whole-slide images and evaluates them internally and externally. The outcome is the presence of an artifact at each position of the image, not a patient outcome, and the units are slides or images rather than participants; items about participants are answered for slides.

Topic	Item	D/E	Checklist item	Reported in / response
Title	1	D;E	Identify the study as developing or evaluating the performance of a multivariable prediction model, the target population, and the outcome to be predicted	Title. The study develops and evaluates artifact detectors (quality control) for H&E whole-slide images; the predicted outcome is pixel-level artifact presence, not a patient outcome.
Abstract	2	D;E	See TRIPOD+AI for Abstracts checklist	Abstract (structured: Purpose, Methods, Results, Conclusions). No registration (item 18d).
Background	3a	D;E	Explain the healthcare context (including whether diagnostic or prognostic) and rationale for developing or evaluating the prediction model, including references to existing models	Introduction, paragraphs 1–4: quality control before review and analysis; supervised and reconstruction detectors, with references. The models are neither diagnostic nor prognostic; they screen image quality.
Background	3b	D;E	Describe the target population and the intended purpose of the prediction model in the context of the care pathway, including its intended users (eg, healthcare professionals, patients, public)	Introduction, paragraphs 1 and 4; OR2, A2 (intended use: research quality-control step for H&E slides before human review or computational analysis; not for clinical decision-making).
Background	3c	D;E	Describe any known health inequalities between sociodemographic groups	Not applicable: the models assess image quality, and the public datasets carry no sociodemographic data.
Objectives	4	D;E	Specify the study objectives, including whether the study describes the development or validation of a prediction model (or both)	Introduction, last two paragraphs and contributions. Both development (probes, adapted diffusion detectors) and evaluation (internal and external).
Data	5a	D;E	Describe the sources of data separately for the	Section 2.1 (development data: TCGA slides

Topic	Item	D/E	Checklist item	Reported in / response
			development and evaluation datasets (eg, randomised trial, cohort, routine care or registry data), the rationale for using these data, and representativeness of the data	with AIRAQc annotations; external data: SlideGuard, SlideInspect, Fırat); Table S1 (OR3); OR2, Part B. Representativeness: Discussion, Limitations.
Data	5b	D;E	Specify the dates of the collected participant data, including start and end of participant accrual; and, if applicable, end of follow-up	Retrospective public datasets used as released (Section 2.1, with references); acquisition dates are those of the source datasets. No follow-up.
Participants	6a	D;E	Specify key elements of the study setting (eg, primary care, secondary care, general population) including the number and location of centres	Section 2.1: TCGA (multi-centre); SlideGuard cohorts (TCGA, Dartmouth-Hitchcock lung, Maria Skłodowska-Curie National Research Institute of Oncology lymphatic tissue, mouse skin); SlideInspect (several organs); Fırat (veterinary teaching slides, five organs). Table S1 (OR3).
Participants	6b	D;E	Describe the eligibility criteria for study participants	Section 2.1 (eight AIRAQc slides excluded: no usable patches or too little tissue; clean controls; Fırat clean and fold images); Table S1 (OR3); OR1, S1.
Participants	6c	D;E	Give details of any treatments received, and how they were handled during model development or evaluation, if relevant	Not applicable (no treatments).
Data preparation	7	D;E	Describe any data pre-processing and quality checking, including whether this was similar across relevant sociodemographic groups	Section 2.4 (read resolution, resolution guard, controlled resampling); Section 2.6 (scored tissue regions); OR1, S6–S7. Sociodemographic groups: not applicable.
Outcome	8a	D;E	Clearly define the outcome that is being predicted and the time horizon, including how and when assessed, the rationale for choosing this outcome, and whether the method of outcome assessment is consistent across sociodemographic groups	Section 2.1: the reference is the released pixel annotation (four artifact types), compared on a grid of 8 × 8 µm cells; edge and no-tissue cells; secondary reference with teal ink. No time horizon.
Outcome	8b	D;E	If outcome assessment requires subjective interpretation, describe the qualifications and demographic characteristics of the outcome assessors	Annotations were produced by the dataset authors before this study (Section 2.1; OR2, Part B); their procedures are reported in the dataset publications.
Outcome	8c	D;E	Report any actions to blind assessment of the outcome to be predicted	SlideGuard annotations are held by the organizers and were scored through their server (Sections 2.1, 2.6); all other references predate this study.
Predictors	9a	D	Describe the choice of initial predictors (eg, literature, previous models, all available predictors) and any pre-selection of predictors before model building	Section 2.2: image content only (1,024-pixel patches; UNI2-h class tokens for the probe). No clinical predictors and no predictor selection.
Predictors	9b	D;E	Clearly define all predictors, including how and when they were measured (and any actions to blind assessment of predictors for the outcome and other predictors)	Sections 2.2 and 2.4 (patches at about 1.0 µm per pixel; tokens of 256 pixels; 1,536-dimensional class token); OR1, S2 and S4.
Predictors	9c	D;E	If predictor measurement requires subjective interpretation, describe the qualifications and demographic characteristics of the predictor assessors	Not applicable (features are computed automatically).
Sample size	10	D;E	Explain how the study size was arrived at (separately for development and evaluation), and justify that the study size was sufficient to answer the research question. Include details of any sample size calculation	Fixed by the available datasets; no sample size calculation. Kish effective sample size reported (Section 2.6; Discussion, Limitations).
Missing data	11	D;E	Describe how missing data were handled. Provide reasons for omitting any data	Section 2.1 (excluded slides, with reasons); Section 2.2 (tokens excluded from probe labels). No SlideGuard or Fırat file was rejected by the resolution guard. Two Fırat fold images with empty masks were excluded from fold endpoints (OR1, S1.3; OR2, Part B).

Topic	Item	D/E	Checklist item	Reported in / response
Analytical methods	12a	D	Describe how the data were used (eg, for development and evaluation of model performance) in the analysis, including whether the data were partitioned, considering any sample size requirements	Section 2.1: 16 training slides in four slide-level folds; 24-slide benchmark; 2 held-out slides used only to monitor training; external sets used only for evaluation.
Analytical methods	12b	D	Depending on the type of model, describe how predictors were handled in the analyses (functional form, rescaling, transformation, or any standardisation)	Section 2.2: standardized class-token features for the probe; noise-prediction error per latent position for the diffusion detectors; OR1, S2 and S4.
Analytical methods	12c	D	Specify the type of model, rationale, all model building steps, including any hyperparameter tuning, and method for internal validation	Sections 2.2–2.3; OR1, S2–S5 (low-rank adaptation of PixCell-1024; L2-regularized logistic regression with regularization chosen by slide-grouped cross-validation; combination rules selected by leave-one-slide-out and frozen).
Analytical methods	12d	D;E	Describe if and how any heterogeneity in estimates of model parameter values and model performance was handled and quantified across clusters (eg, hospitals, countries). See TRIPOD-Cluster for additional considerations	Performance is reported per cohort, organ and dataset (Sections 3.2–3.3; Table 2, Fig. 4, Table S3); intervals from a bootstrap over slides (Section 2.6). No formal heterogeneity model.
Analytical methods	12e	D;E	Specify all measures and plots used (and their rationale) to evaluate model performance (eg, discrimination, calibration, clinical utility) and, if relevant, to compare multiple models	Section 2.6 (per-type sensitivity, clean false positives, F1, precision, Dice, SlideGuard metrics, FP-matched comparison, reliability tables, ECE, Brier score, paired bootstrap); Section 2.5 (AUROC of the shift monitors).
Analytical methods	12f	E	Describe any model updating (eg, recalibration) arising from the model evaluation, either overall or for particular sociodemographic groups or settings	Section 2.2 (scale-augmented probe, Section 2.4). No recalibration.
Analytical methods	12g	E	For model evaluation, describe how the model predictions were calculated (eg, formula, code, object, application programming interface)	Predictions are computed with the released code, adapters and probe parameters (Section 2.7; Code availability).
Class imbalance	13	D;E	If class imbalance methods were used, state why and how this was done, and any subsequent methods to recalibrate the model or the model predictions	Probe training used 13,902 positive and 17,445 negative tokens (OR1, S4). Evaluation stratified the bootstrap by clean and artifact status and compared detectors at matched false-positive rates (Section 2.6).
Fairness	14	D;E	Describe any approaches that were used to address model fairness and their rationale	Patient-level fairness is not applicable (no sociodemographic data). Performance is reported across organs, cohorts, scanners and species (Sections 3.2–3.3; OR2, A3).
Model output	15	D	Specify the output of the prediction model (eg, probabilities, classification). Provide details and rationale for any classification and how the thresholds were identified	Section 2.3: probe probability map thresholded at $\tau = 0.5$; diffusion error map thresholded by rule R0 (label-free); combination rules PC, PR and BF; binary masks on the grid.
Training versus evaluation	16	D;E	Identify any differences between the development and evaluation data in healthcare setting, eligibility criteria, outcome, and predictors	Section 2.1 (external sets from other institutions, scanners or cameras; SlideInspect distributed at 5×; TCGA exposure of the pretrained models); Table S1 (OR3); Section 3.3.
Ethical approval	17	D;E	Name the institutional research board or ethics committee that approved the study and describe the participant informed consent or the ethics committee waiver of informed consent	Statements and Declarations, Ethics approval: public, de-identified image datasets; no approval required.
Funding	18a	D;E	Give the source of funding and the role of the funders for the present study	Statements and Declarations, Funding: source of funding; the funder had no role in the study design, data analysis, decision to publish or preparation of the manuscript.
Conflicts of interest	18b	D;E	Declare any conflicts of interest and financial disclosures for all authors	Statements and Declarations, Competing interests.

Topic	Item	D/E	Checklist item	Reported in / response
Protocol	18c	D;E	Indicate where the study protocol can be accessed or state that a protocol was not prepared	No study protocol was prepared or published. Analyses added after the SlideInspect results are labelled post hoc (Section 2.6; OR1, S10).
Registration	18d	D;E	Provide registration information for the study, including register name and registration number, or state that the study was not registered	The study was not registered.
Data sharing	18e	D;E	Provide details of the availability of the study data	Statements and Declarations, Data availability.
Code sharing	18f	D;E	Provide details of the availability of the analytical code	Statements and Declarations, Code availability (Zenodo, https://doi.org/10.5281/zenodo.23448122).
Patient and public involvement	19	D;E	Provide details of any patient and public involvement during the design, conduct, reporting, interpretation, or dissemination of the study or state no involvement	No patient or public involvement.
Participants	20a	D;E	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful	Section 2.1 (50 AIRAQc slides, 8 excluded: 16 training, 24 benchmark with 11 clean and 13 artifact slides, 2 held out; external sets with clean and artifact counts); Fig. 1; Table S1 (OR3).
Participants	20b	D;E	Report the characteristics overall and, where applicable, for each data source or setting, including the key dates, key predictors (including demographics), treatments received, sample size, number of outcome events, follow-up time, and amount of missing data. A table may be helpful. Report any differences across key demographic groups	Table S1 (OR3); OR2, Part B. Demographics are not available.
Participants	20c	E	For model evaluation, show a comparison with the development data of the distribution of important predictors (demographics, predictors, and outcome)	Section 2.1 and Table S1 (OR3) describe the differences; the embedding distance of each set from the training slides is measured by monitor M3 (Sections 2.5, 3.7).
Model development	21	D;E	Specify the number of participants and outcome events in each analysis (eg, for model development, hyperparameter tuning, model evaluation)	Section 2.1; Tables 1–5 and Table S1 give the slides, cases or images in each analysis.
Model specification	22	D	Provide details of the full prediction model (eg, formula, code, object, application programming interface) to allow predictions in new individuals and to enable third party evaluation and implementation, including any restrictions to access or reuse (eg, freely available, proprietary)	OR2, A1; the Zenodo deposit (adapters, probe parameters, code). Restrictions: Statements and Declarations, Data availability (CC BY-NC 4.0; UNI2-h and Phikon-v2 licences).
Model performance	23a	D;E	Report model performance estimates with confidence intervals, including for any key subgroups (eg, sociodemographic). Consider plots to aid presentation	Sections 3.1–3.7 with 95% bootstrap intervals; Tables 1–5; Figs. 2, 4 and 6; subgroups by cohort and organ (Sections 3.2–3.3; Table S3).
Model performance	23b	D;E	If examined, report results of any heterogeneity in model performance across clusters. See TRIPOD-Cluster for additional details	Per cohort (Table 2, Fig. 4) and per organ (Table S3); not modelled formally.
Model updating	24	E	Report the results from any model updating, including the updated model and subsequent performance	Section 3.4 (scale-augmented probe) and Section 3.5 (its calibration).
Interpretation	25	D;E	Give an overall interpretation of the main results, including issues of fairness in the context of the objectives and previous studies	Discussion.
Limitations	26	D;E	Discuss any limitations of the study (such as a non-representative sample, sample size, overfitting, missing data) and their effects on any biases, statistical uncertainty, and	Discussion, Limitations.

Topic	Item	D/E	Checklist item	Reported in / response
			generalisability	
Usability of the model in the context of current care	27a	D	Describe how poor quality or unavailable input data (eg, predictor values) should be assessed and handled when implementing the prediction model	Section 2.4 (resolution guard); Discussion, four practical checks; OR2, A5 (monitoring).
Usability of the model in the context of current care	27b	D	Specify whether users will be required to interact in the handling of the input data or use of the model, and what level of expertise is required of users	OR2, A2 and A5: the models run without user interaction; users verify the read resolution and monitor the probe's flag rate on known-clean slides.
Usability of the model in the context of current care	27c	D;E	Discuss any next steps for future research, with a specific view to applicability and generalisability of the model	Discussion (benefit of scale augmentation under real scanner differences remains to be shown; Limitations).