INPLASY

Artificial intelligence for predicting immunohistochemical biomarkers from haematoxylin and eosin histopathology in solid tumours: a systematic review and meta-analysis

INPLASY2026110020
doi: 10.37766/inplasy2026.11.0020
Received: 6 November 2026
Published: 6 November 2026

Chen, YC; Sun, PE; Chen, LY; Wang, YC; Yang, CY.

Corresponding author:
Chih-Yu Yang

cyyang3@vghtpe.gov.tw

Author Affiliation:
National Yang Ming Chiao Tung University.

ADMINISTRATIVE INFORMATION

Support - No support.

Review stage at the time of initial registration - Preliminary searches.

Conflicts of interest - None declared.

INPLASY registration number: INPLASY2026110020

Amendments - This protocol was registered with the International Platform of Registered Systematic Review and Meta-Analysis Protocols (INPLASY) on 6 November 2026 and was last updated on 6 November 2026.

INTRODUCTION

Review question / Objective To evaluate the accuracy of artificial intelligence models that predict tissue immunohistochemical biomarker states from haematoxylin and eosin (H&E) histopathology in human solid tumours. The review synthesises sensitivity, specificity and area under the receiver operating characteristic curve (AUROC) within compatible tumour and biomarker definitions, and examines variation by validation setting, analysis unit, specimen type and reference method.

Rationale H&E-based biomarker prediction has been investigated across diverse solid tumours, but studies differ in their target definitions, specimens, evaluation units and validation designs. These differences limit direct comparison of reported performance. A systematic review with target-specific meta-analysis can establish the available evidence for each biomarker and characterise sources of variation in accuracy. The review includes conventional immunohistochemistry (IHC) and tissue immunofluorescence references, while preserving the clinical and research definitions used by the original studies.

Condition being studied Protein expression, protein loss and antibody-defined biomarker classifications in human solid tumours, including treatment-associated markers, diagnostic and lineage markers, proliferation markers and research phenotypes. Historical and research-defined cutoffs are eligible alongside definitions used in current clinical practice.

METHODS

Participant or population Human solid tumours, including central nervous system and paediatric tumours, assessed in biopsy or resection specimens from primary or metastatic disease. Separable eligible subgroups of mixed populations are included. Haematological malignancies, non-human studies and cell-line-only studies are excluded.

Chen et al. INPLASY protocol 2026110020. doi:10.37766/inplasy2026.11.0020 Downloaded from https://inplasy.com/inplasy-2026-11-0020/

Intervention Artificial intelligence models using digitised H&E histopathology as the sole predictive input at inference. IHC, immunofluorescence or molecular information may provide training supervision. Multimodal studies are eligible when an H&E-only model can be evaluated separately. Virtual staining studies require an evaluated downstream biomarker classification endpoint; image similarity alone is insufficient.

Comparator Clinical or research IHC; tissue immunofluorescence, including multiplex platforms such as CODEX and CyCIF; documented antibody-assay-derived cohort labels; or IHC-anchored composite definitions, such as HER2 IHC with in situ hybridisation adjudication. Mixed IHC/molecular references are evaluated under their reported composite definitions. Purely molecular targets without an eligible antibody-defined outcome are excluded. Reference cutoffs and AI decision thresholds are recorded separately.

Study designs to be included Original model development or validation studies with held-out testing or cross-validation, including prospective and retrospective designs and eligible reader or workflow evaluations. Complete journal articles, full conference papers and full preprints are eligible. Reports from trials are eligible when they contain a relevant model evaluation.

Eligibility criteria Full-section, tissue microarray (TMA), fixed and frozen specimens are eligible. Patient-, slide-, core/region-of-interest-, patch/tile- and tissue-cell-level classification results are included at their reported units. Studies remain eligible for the systematic review when numerical reporting is insufficient for pooling. Missing descriptive metadata alone does not determine exclusion.

Reviews, editorials without eligible original results, registrations without results, abstract-only reports, single-case reports and descriptive case series with fewer than 10 patients are excluded. This case-series restriction does not impose a minimum sample size on an AI validation cohort. The eligibility framework has no general publication-year or language restriction.

Training-only performance and demonstrable same-patient training/test leakage are not accepted as validation evidence; a separable valid evaluation remains eligible. Unreported patient separation in an otherwise documented held-out or cross-validation design is assessed through risk of bias. Inseparable multimodal inference, actual IHC/IF images as predictive inputs, and unrelated outcomes are excluded. Reference-assisted region selection or calibration is assessed for applicability. Related publications and versions are linked, with distinct eligible results retained.

Information sources MEDLINE/PubMed, Embase, Web of Science Core Collection, IEEE Xplore, ACM Digital Library and Cochrane CENTRAL, supplemented by reference-list and forward-citation searches. Full articles, supplementary materials, appendices and relevant dataset documentation inform eligibility and extraction.

Main outcome(s) Sensitivity and specificity, with 95% confidence intervals, for each tumour-specific biomarker definition and comparator at the reported evaluation unit and AI decision threshold. True-positive, false-negative, true-negative and false-positive counts are extracted or reconstructed when supported by the reported data. Synthesis is target-specific; unrelated cancers and biomarkers are not combined into a single accuracy estimate.

Additional outcome(s) Model AUROC and its uncertainty, analysed separately from threshold-specific accuracy. Other reported outcomes include multiclass or ordinal classification, calibration, concordance, reader comparisons, selective-output coverage and errors, and observed time or cost. These outcomes are summarised where available using the original definitions and denominators. Empirical results and author-reported simulations are distinguished.

Data management Bibliographic records are deduplicated and linked across versions and related publications. Screening decisions, reasons for exclusion, source locations and extracted data are maintained in a structured dataset. New or unresolved records will be assessed independently by two reviewers against the eligibility criteria, and included-study eligibility will be verified against the original reports.
One reviewer will check source extractions and a second will verify the reference definition, cohort, analysis unit, selected result, denominators and numerical derivations. Disagreements will be resolved by discussion and source review, with further adjudication when needed. Extracted characteristics include tumour and specimen type, assay and cutoff, model and validation design, patient separation, analysis unit, sample sizes, performance measures and uncertainty. Reported, reconstructed and approximate values are distinguished. Unretrieved reports are recorded separately from eligibility exclusions.

Chen et al. INPLASY protocol 2026110020. doi:10.37766/inplasy2026.11.0020 Downloaded from https://inplasy.com/inplasy-2026-11-0020/

Quality assessment / Risk of bias analysis Risk of bias and applicability are assessed using QUADAS-3 version 1.2 at the relevant question and result level. The domains are Participants, Index Test, Target Condition and Analysis. Two reviewers will independently assess risk of bias and applicability and resolve disagreements by discussion.
Assessment considers participant and region selection, reference ascertainment, patient separation, model or threshold optimisation on test data, inherited labels and unaccounted clustering. Internal validation, TMA use and patch-level evaluation are assessed in context rather than assigned an automatic high-risk rating. Domain judgments inform interpretation and sensitivity analyses without a total quality score or automatic study exclusion. Evidence is discussed in terms of bias, consistency, precision and applicability.

Strategy of data synthesis Analyses are grouped by tumour setting, reference-defined target and comparator. Operationally equivalent definitions may be combined; materially different targets and mixed reference definitions are represented separately. Differences in reporting detail alone do not define separate outcomes. Evaluation unit and specimen type are examined through subgroup and sensitivity analyses.

One representation of known overlapping test material is selected per question and metric. Among otherwise comparable results, patient-level estimates are preferred, followed by slide, core/region-of-interest and patch/tile estimates. Selection favours the author-designated final model and completeness of reporting, with presentation order resolving equivalent choices. Selection is not based on the highest observed accuracy. Test-optimised selections are flagged for bias assessment. Distinct non-overlapping cohorts may contribute separately; AUROC and threshold-specific analyses may use different data.

Counts are taken from reported tables, predictions or unambiguous figures, or reconstructed from matching class sizes and performance measures. Reconstructions preserve the cohort, unit, model, threshold and outcome direction and are identified for sensitivity analysis.

Random-effects meta-analysis is used when at least two independent test-material groups provide compatible, estimable data. Two-group estimates are exploratory; single-group results are presented individually. Sensitivity and specificity are synthesised using bivariate binomial models without routine zero-cell corrections. When reliable joint estimation is not possible, separate binomial models are used; an estimable component is retained if the other cannot be estimated. Results include 95% confidence intervals and forest plots. SROC curves and prediction regions are presented when supported by a joint model.

AUROCs are synthesised on the logit scale using restricted maximum likelihood and modified Hartung–Knapp intervals no narrower than conventional intervals. Standard errors are obtained from reported SEs with identified scales or confidence intervals with known levels. Reported uncertainty with incompletely described construction is assessed in sensitivity analyses; fold/run standard deviations are not treated automatically as patient-sampling SEs. Hanley–McNeil approximations use matching independent positive and negative case counts. Boundary AUROCs and estimates without usable uncertainty remain descriptive.

Cluster-adjusted uncertainty is preferred where available. Otherwise, analyses based on reported unit-level counts use a working-independence assumption, with its implications for precision acknowledged. Mixed-unit estimates are interpreted at the included units. Conditional or selective-output performance is analysed separately from all-case classification. Other results are summarised in structured tables or narrative synthesis. Analyses use R with reproducible scripts.

Subgroup analysis Subgroups prioritise validation setting, evaluation unit, and TMA versus full-section specimens. Biopsy versus resection and reference-method differences within compatible targets may also be examined when data permit.
Exploratory meta-regression is considered when at least 10 independent test-material groups and sufficient representation across covariate categories and model programmes support an interpretable model. One justified covariate is examined at a time, with residual confounding considered. Where these conditions are not met, subgroup or descriptive summaries are used.

Sensitivity analysis Sensitivity analyses examine approximate or reconstructed inputs, possible cohort overlap, alternative eligible results, risk-of-bias restrictions, test-optimised or unclear result selection, unreported patient separation and material reference-definition differences. Unit or specimen restrictions and leave-one-cohort-out analyses are used where informative. The choice of analyses depends on the evidence available for each target.

Chen et al. INPLASY protocol 2026110020. doi:10.37766/inplasy2026.11.0020 Downloaded from https://inplasy.com/inplasy-2026-11-0020/

Deeks' asymmetry test is considered for diagnostic-accuracy syntheses with at least 10 independent test-material groups and informative variation in study size. Reporting-bias limitations are described when statistical testing is not supported.

Country(ies) involved Taiwan.

Keywords Artificial intelligence; digital pathology; haematoxylin and eosin; immunohistochemistry; immunofluorescence; solid tumours; diagnostic accuracy; meta-analysis.

Contributions of each author

Author 1 - Yu-Chuan Chen
Email: ycchen.md08@nycu.edu.tw
Author 2 - Pei-En Sun
Email: spn110101162.md10@nycu.edu.tw
Author 3 - Lian-Yuan Chen
Email: thomaslychen.md12@nycu.edu.tw
Author 4 - Yun-Chu Wang
Email: yunzuwang.md08@nycu.edu.tw
Author 5 - Chih-Yu Yang
Email: cyyang3@vghtpe.gov.tw

Chen et al. INPLASY protocol 2026110020. doi:10.37766/inplasy2026.11.0020 Downloaded from https://inplasy.com/inplasy-2026-11-0020/