Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

MammoTrace Academic for teaching critical appraisal of mammography AI Software development and technical verification with worked educational activities

Anderson Díaz-Pérez

anderson.diaz@campusuninunez.edu.co

Corporación Universitaria Rafael Nuñez

Research Article

Keywords: medical education, artificial intelligence literacy, mammography, research software, data leakage, reproducibility, critical appraisal, uncertainty

Posted Date: November 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11543238/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

MammoTrace Academic for teaching critical appraisal of mammography AI

Software development and technical verification with worked educational activities

Anderson Díaz-Pérez

Programa de Medicina, Facultad de Ciencias de la Salud, Corporación Universitaria Rafael Núñez, Barranquilla, Colombia

Correspondence: anderson.diaz@campusuninunez.edu.co

ORCID: 0000-0003-2448-0953

Abstract

Background

Medical students and health-profession trainees need to judge whether claims about artificial intelligence (AI) are supported by the data and procedures behind them. We aimed to develop and technically verify MammoTrace Academic v4.1.0 and translate its documented outputs into supervised critical-appraisal activities using mammography as the learning context. This developer-led software study combined a retained audit of public mammography records, execution of versioned automated tests, direct characterization of selected report texts, and documentary comparison with Orange and 3D Slicer. New tests used the archived public release and were interpreted separately from historical data-dependent findings. Constructive alignment linked five appraisal objectives to activities and expected explanations; the teaching materials were developed after the technical audit.

Results

The educational package comprises six activities, 27 questions, worked answers, a facilitator guide and a proposed formative rubric. Learners are asked to attribute sources, identify the unit of independence, distinguish annotations from model outputs, interpret report context and justify uncertainty. An evidence-only route requires neither software installation nor image downloads; optional application exercises demonstrate documented checks. A final synthetic case asks learners to apply the reasoning to a new problem. In the retained

audit, 31 subjects overlapped between development and test sets; excluding 83 development annotations preserved all 704 test annotations. Model fitting was blocked when only one of 3020 metadata-eligible full-image series was unambiguously indexed locally. Direct report characterization exposed punctuation-sensitive confirmation of a historical category. New public-release execution produced 281 passing and 59 skipped core tests and 11 passing and two skipped application checks, without failures. Documentary comparison showed that established tools share core capabilities; MammoTrace combines mammography-specific checks, report examples and appraisal tasks. Comparative performance was not tested. No learners were recruited, and educational outcomes and scoring reliability were not evaluated.

Conclusions

MammoTrace provides a technically documented resource for supervised teaching of AI critical appraisal in medicine. Its contribution is the explicit connection between inspectable evidence and worked reasoning tasks. Comparative learner evaluation is needed to establish effects on appraisal competence, retention and transfer.

Keywords: medical education; artificial intelligence literacy; mammography; research software; data leakage; reproducibility; critical appraisal; uncertainty

Background

Critical appraisal of artificial intelligence (AI) requires learners to distinguish what a computational result shows from what it leaves untested. Competency development and needs assessments in medical education identify understanding data, recognizing limitations and exercising professional judgment as relevant learning needs [1,2]. These needs become concrete when a learner must decide whether image metadata establish data availability, whether overlapping subjects compromise evaluation, or whether an annotation is evidence of a model's performance.

Practical radiology curricula have combined conceptual instruction with hands-on analysis. Van Kooten and colleagues described a curriculum that included Orange-based laboratories and reported a small pilot evaluation of perceived knowledge and confidence [3]. Li and colleagues surveyed faculty radiologists, residents and medical students to identify requirements for AI-supported radiology education [4]. These studies provide a context for practical resources and learner-centered design. They also leave room for instructional materials in

which the evidence supporting an AI claim can be inspected step by step, with explanations assessed separately from confidence or interface operation.

In a scoping review of laparoscopic training, Vega Ochoa and colleagues mapped generally favorable learner perceptions while distinguishing acceptance from demonstrated competence [5]. Their findings support examining teaching conditions and assessment alongside learner experience, with cautious transfer between professional groups.

Mammography offers useful appraisal problems. Subject overlap must be checked after datasets are combined, because file-level separation alone may not establish independent evaluation [6]. A reference lesion mask and a model explanation also support different inferences; visually convincing saliency is not sufficient evidence of valid localization or clinical reasoning [7,8]. Constructive alignment provides a way to connect such problems to observable learning objectives, activities and assessment [9].

MammoTrace Academic v4.1.0 [10] organizes five evidence distinctions: source records versus locally available pixels; annotation rows versus independent subjects; reference labels versus experimental model responses; historical versus current report statements; and completed technical checks versus clinical or educational conclusions. The objective was to develop and technically verify these functions and translate their documented outputs into supervised critical-appraisal activities for medical students and health-profession trainees. The educational resource was designed after the engineering audit. The study asks what the software demonstrably does and how its documented outputs can support assessable teaching tasks; it does not test an educational intervention.

Implementation

Design, sources and verification sequence

This single-project development and technical verification study used retained project records and an additional execution of the public release. The original audit, dated 13 September 2026, used four CBIS-DDSM case-description files, a National Biomedical Imaging Archive manifest, a repository digest and a partial download log. CBIS-DDSM is distributed through The Cancer Imaging Archive (TCIA) [11,12,13]. Eight input artifacts, including the digest spreadsheet and its CSV export, were tracked by SHA-256. These hashes identify the

retained snapshot; the full source corpus was not reacquired for the present verification. Selected CLAIM reporting principles informed the description of provenance and data handling [14].

Three records are kept distinct: the original development audit, the published public-release verification and a new execution on 21 October 2026. The new execution used an isolated copy of the immutable v4.1.0 source commit, c460f85ab73895d3ceba954427b608a70a02eb3b. Python 3.12.14 and Streamlit 1.55.0 were used in a virtual environment with inherited scientific packages and the required pinned base dependencies. This differed from the original Python 3.13.5 development environment. Bulk CBIS inputs, historical model weights, PyTorch and Keras were absent. The source implementation and release tag were unchanged.

The new workflow verified the release manifest, generated the supplied synthetic examples, executed the core and Streamlit AppTest suites [15], and collected JUnit results and statement coverage for mammoapp and mammotrace. The first attempt encountered a missing OpenCV dependency; the unchanged suites completed after installation of the repository-pinned headless package. A manuscript-to-source discrepancy prompted separate direct characterization of five constructed Spanish texts, involving seven category-confirmation attempts. This purposive characterization was not a language-performance benchmark, and version 4.0 was not reexecuted. Additional file 4 contains the commands, environment, logs, literal inputs and outputs, and a script that checks the parser's hash before execution.

The work was developer-led and not preregistered. Generative AI assistance was used for code execution, evidence organization, literature retrieval and manuscript preparation; the additional verification is not an independent human audit. No learners, educators or patients were recruited. No institutional clinical records, participant-level educational outcomes or new clinical assessments were collected.

Software architecture and instructional design

MammoTrace uses Python and Streamlit. The mammoapp package provides data preparation, report handling, educational guidance and interface components; mammotrace retains research infrastructure. The public PUBLICA_SIN_PESOS edition includes source, tests, selected audit reports and synthetic-example generation. Historical neural-network weights, bulk third-party metadata and the DICOM corpus are excluded. Figure 1 separates the public-image route, constructed-report route and experimental model components.

The intended instructional setting is a supervised seminar for medical students or health-profession trainees with introductory knowledge of mammography terminology and research design. The companion material links each objective to an evidence artifact, a learner action and an explanation that a facilitator can assess (Table 1; Figure 2). An evidence-only route uses aggregate tables and worked outputs without requiring images or installation. An optional application route demonstrates the corresponding checks after the facilitator verifies local installation. Teacher-led explanation, feedback and scoring are external to the application.

The six activities comprise source attribution, overlap and partitioning, metadata versus pixels, annotations versus model outputs, report-context interpretation, and transfer to a wholly synthetic example. They contain 27 questions, a separate facilitator answer key and a proposed five-dimension rubric scored 0–2 per dimension. The illustrative total of 0–10 has no validated pass threshold. A proposed 90-minute schedule is a design estimate, not an observed completion time. The learner workbook and facilitator guide are supplied separately as Additional files 3a and 3b.

Table 1 Alignment of appraisal objectives and companion activities

Objective	Inspectable material and task	Expected explanation
Attribute sources	Activity 1: identify the origin and version of software, dataset, audit output and fictional report.	State what each source supports and separate software and dataset citations.
Recognize leakage	Activity 2: reconcile subject memberships, compute overlap and justify exclusions.	Use the subject as the independence unit; distinguish subjects from annotations.
Distinguish evidence	Activities 3–4: compare availability records, reference masks, labels and experimental responses.	Explain why metadata are not pixels and an annotation is not a prediction.
Interpret report context	Activity 5: compare literal negation and temporal strings with observed outputs.	Identify the current statement and recognize a punctuation-sensitive parser limitation.
Communicate uncertainty	Activities 3–6: state a supported conclusion, an unsupported claim and the next evidence needed.	Explain a blocked operation or ambiguous output without assigning unsupported clinical meaning.

Activities and scoring are proposed teaching materials. Expected explanations are answer-key targets, not observed learner responses.

Data preparation and subject separation

An annotation row represented a lesion-view record; patient_id identified a subject, SeriesInstanceUID a series and SOPInstanceUID an image instance. A provisional lesion key combined subject, laterality, lesion family and abnormality identifier. It did not establish longitudinal lesion identity. CSV parsing honored quoted paths

containing embedded line breaks. The original BENIGN, BENIGN_WITHOUT_CALLBACK and MALIGNANT categories were retained. For a separate binary candidate task, the first two mapped to 0 and MALIGNANT to 1. BENIGN_WITHOUT_CALLBACK was not relabeled as normal.

Subject sets were compared within and across the calcification and mass files. Global overlap was rejected by default. An explicit holdout_priority policy preserved official test annotations and removed development annotations from any subject present in test. Remaining development subjects were sorted by the hexadecimal SHA-256 digest of 20260913:patient_id; the first 243 were assigned to validation. This deterministic allocation was not stratified. For a single-label image candidate, mixed binary-label image groups were excluded from eligibility; disagreement across views was recorded separately. These are task-specific rules rather than corrections to the source pathology labels.

Metadata eligibility and local pixel availability were assessed separately. Manifest identifiers, digest records and a partial download log were reconciled before local indexing and training preflight. A bounded repair handled a recorded decimal-comma defect in the partial log, with other malformed structures rejected. The DICOM reader accepted only the inspected single-frame monochrome, uncompressed little-endian integer-pixel profile and checked pixel length and subject linkage. Its format restrictions do not establish complete de-identification under DICOM confidentiality profiles [16]. Detailed repair rules, image dimensions and the separate October illustration check are retained in Additional file 1.

Experimental models and report interpretation

Historical ResNet50 infrastructure used labels derived from random indicators, preventing interpretation of its scores as clinical probabilities. Its weights are absent from the public edition. A 49-region occlusion laboratory represents an experimental numerical response rather than validated lesion localization. The separate CBIS candidate comprises three convolutional blocks with new initialization and 224-pixel input. Fitting depends on successful preflight. Reference labels and experimental model scores do not feed the report-guidance rules.

The report route accepts TXT, PDF and DOCX and separates text extraction from user confirmation of a current BI-RADS category. Rule-based negation and temporal context are used to restrict confirmation. The new characterization attempted three negated categories and both categories in each of two temporally equivalent

Spanish strings that differed in punctuation. English glosses describe the examples but were not tested parser inputs. Constructed reports are unrelated to the public mammogram subjects; human confirmation is not treated as clinical adjudication.

Results

Subject overlap and partitioning

The four case files contained 3568 lesion-view annotations from 1566 unique subjects, with 2050 provisional lesion keys and 3103 full-image series. Annotation categories were BENIGN (1429), BENIGN_WITHOUT_CALLBACK (682) and MALIGNANT (1457). Each lesion family separately had no training–test subject overlap. Combining the families revealed 31 shared subjects: 13 between calcification test and mass development and 18 between mass test and calcification development.

The overlap provides a complete worked appraisal example. The original development and test sets contained 1248 and 349 subjects. Their sum, 1597, counts the 31 shared subjects twice: $1248 + 349 - 31 = 1566$ unique subjects. Relative to test membership, overlap was $31/349 = 8.9\%$. The holdout-priority policy excluded 83 development annotations and preserved all 704 test annotations. Remaining development records comprised 2240 annotations from 974 training subjects and 541 from 243 validation subjects (Table 2; Figure 3). All final subject-set intersections were empty. The educational task asks the learner to explain the independence unit, denominator and cost of preserving the official test set; the worked arithmetic is observed evidence, whereas learner performance remains unmeasured.

Seventeen image groups had mixed binary labels and three provisional lesion keys were discordant across views. After candidate-specific exclusions, 3020 full-image series were metadata-eligible: 1927 training, 454 validation and 639 test. These exclusions concern the proposed single-label candidate and do not imply that different lesions within an image share one biological outcome.

Table 2 Subject partitions and full-image eligibility

Partition	Annotations	Subjects	Full-image series	Eligible series
Training	2240	974	1934	1927
Validation	541	243	460	454
Official test	704	349	645	639

Excluded from development	83	31	64	0

The 31 subjects excluded from development remain in test. They are not additional subjects. Eligibility is based on metadata and does not establish pixel availability.

Availability and training preflight

Case references, manifest and digest agreed on 6775 series identifiers. The digest declared 10239 instances, while the partial working download log contained 544 series and 837 declared instances. Its 251 full-image series related to 293 annotation rows. All 544 malformed log rows matched the bounded repair pattern, and rereading the normalized export required no further repair.

The original integration read three DICOM instances from two series and one subject: a full image, a crop and a reference mask. Of the 3020 eligible full-image series, only one test series was unambiguously available in the original local index; 3019 were missing or ambiguous. The preflight therefore blocked fitting. No candidate predictions, discrimination estimates or calibration results were produced. The companion activity requires learners to identify why a consistent manifest cannot substitute for a usable pixel inventory.

Reproduction of public-release tests

All 94 entries in the immutable release SHA256SUMS.json matched the archived source files. The new execution reproduced the published counts: 281 core tests passed and 59 were skipped; 11 AppTest tests passed and two were skipped, with no failures or errors (Table 3). Core statement coverage was 957/1738 (55.06%). This denominator and resource profile differ from the historical development report, which recorded 1226/1732 covered statements (70.79%); no coverage trend is inferred.

The 59 core skips comprised all 53 tests in the CBIS module and six historical-model checks. A module-wide guard also skipped the report regressions in that module, although their constructed texts did not require CBIS pixels. Consequently, the 281 passing tests alone do not verify the report examples. The separate direct characterization addressed that gap. The 11 passing AppTest checks included initial startup, eight fictional scenarios, confirmation and reset behavior, and hiding the experimental classifier in the comprehension view. AppTest is a native application test harness rather than a browser, usability or educational evaluation.

Table 3 Verification records with distinct resource profiles

Record	Core tests: passed / skipped	AppTest: passed / skipped	Statement coverage
Development, 13 Sep 2026; archived report	340 / 0	Unavailable	1226/1732 (70.79%)
Public release; published verification [17]	281 / 59	11 / 2	Not reported in the cited release record
New public-profile execution, 21 Oct 2026	281 / 59	11 / 2	957/1738 (55.06%)

The new run had zero failures and errors. Development results were not reexecuted with their original data/model resources. Counts are not pooled. Coverage is statement coverage for the specified packages in the core run, not branch or whole-project coverage.

Literal report behavior and its boundary

The three selected negation strings rejected confirmation of their negated categories. In the colon-delimited temporal string, category 2 was rejected and category 5 was permitted. Replacing the context-label colons with periods permitted confirmation of both categories separately (Table 4; Figure 4). This punctuation-sensitive result was not apparent from the overall passing-test count. It is a limitation of the released parser, not a code correction or evidence of general language understanding.

Activity 5 presents both temporal strings and their recorded outputs. Learners are asked to identify the unchanged intended distinction between the earlier and current report, explain the discrepancy between meaning and parser behavior, and specify an appropriate human review step. This makes an observed software limitation inspectable rather than concealing it behind a successful regression example.

Table 4 Direct characterization of literal Spanish inputs in v4.1.0

Constructed input	Category attempted	Observed confirmation
Sin criterios para BI-RADS 4.	4	Rejected
No hay hallazgos BI-RADS 5.	5	Rejected
Se descarta la categoría BI-RADS 4A.	4A	Rejected
Estudio anterior: BI-RADS 2. Actual: BI-RADS 5.	2; then 5	2 rejected; 5 permitted
Estudio anterior. BI-RADS 2. Actual. BI-RADS 5.	2; then 5	Both permitted independently

Five texts and seven attempts were executed on 21 October 2026. The first three negate a category; the last two distinguish a previous category 2 from a current category 5. These selected inputs do not estimate sensitivity, specificity or a general success rate. No version 4.0 comparison was executed.

Companion educational material

The completed material comprises six activities with evidence cards, 27 questions, a worked answer key, common-error explanations, a proposed formative rubric and a separate critical-error log. Activity 6 uses wholly synthetic counts to ask learners to transfer the reasoning to a new appraisal problem; these counts are explicitly separated from the MammoTrace results. The workbook can be reviewed without installing software. It provides a concrete instructional design output, while its difficulty, usability, timing, scoring reliability and effect on learning remain to be evaluated.

Discussion

Educational contribution and existing tools

MammoTrace links technical checks to questions that learners can answer with inspectable evidence. Its contribution is the integration of CBIS-specific provenance and partition checks, selected report-context examples and worked instructional material. A blocked training run can illustrate the difference between metadata and available pixels; an overlap correction can expose the relevant unit of independence; and a punctuation-sensitive report can show why plausible output still requires scrutiny. The intended outcome is a justified appraisal explanation, which is different from successful operation of the interface.

Established software already supports substantial parts of this work. Orange offers visual analysis workflows and is used in hands-on radiology AI teaching [3,18]. Its current Test and Score documentation supports separate test datasets and cross-validation by a categorical metadata feature [19]. 3D Slicer provides image visualization and quantitative analysis [20]. MammoTrace does not introduce those general capabilities. Table 5 compares documented functions relevant to the proposed tasks and identifies where instructor-designed materials are needed. Comparator software was not installed or benchmarked, and unassessed functions must not be interpreted as absent.

Table 5 Documented capabilities relevant to the appraisal tasks

Task	MammoTrace 4.1.0	Orange [18,19]	3D Slicer [20]
Inspect sources and preprocessing	CBIS-specific source, eligibility and availability records.	Visual, reusable data-analysis workflows.	Image import, inspection and processing workflows.

Control evaluation separation	Subject-overlap checks and explicit holdout-priority exclusions.	Separate test data; cross-validation by a categorical metadata feature.	Not assessed for this partitioning task.
Inspect image roles	Restricted reader; reference-image roles documented.	Not assessed for DICOM role verification.	Image visualization and quantitative analysis.
Interpret selected report context	Directly characterized Spanish negation and temporal examples.	Not assessed.	Not assessed.
Elicit an appraisal explanation	Companion workbook, worked answers and proposed rubric.	Teaching tasks and assessment must be specified for the selected workflow.	Teaching tasks and assessment must be specified for the selected workflow.

This is a task-oriented documentary comparison, not an exhaustive feature audit, performance benchmark or demonstration of educational superiority. MammoTrace evidence comes from the identified release and recorded executions; comparator evidence comes from publications and official documentation.

Why the evidence boundaries matter

The combined-file overlap illustrates a general methodological point: independence should be examined at the relevant unit after data are merged. It does not invalidate the original lesion-specific partitions or establish leakage in another study. The chosen test-priority policy sacrifices development annotations to preserve the official test set. Deterministic assignment makes the operation reproducible but does not establish balanced case mix or external validity.

Similarly, a reference mask, dataset label and occlusion response answer different questions. Model cards, datasheets and FAIR principles provide established approaches to documenting intended use, provenance and reuse conditions [21,22,23]. The workbook turns these distinctions into questions with worked explanations. This organization can support instructional planning, but neither a written warning nor architectural separation establishes that learners will avoid overinterpreting a score.

The punctuation result is especially informative for teaching appraisal. The overall test suite passed its executed checks, yet a narrowly changed input exposed a limitation outside that coverage. Passing tests establish behavior under specified conditions, not comprehensive semantic correctness. Preserving exact inputs, skipped-test reasons and environment details makes this boundary reviewable and prevents historical and current verification records from being conflated.

Clinical cyberbioethics adds a complementary reason for making these distinctions teachable. The encounter-level model proposed by Díaz-Pérez and Pérez specifies responsible actors, documented reasons, uncertainty communication and triggers for pausing or escalating AI-supported decisions [24]. In MammoTrace, the report-context exercise asks the learner to identify the evidentiary problem and justify the human review required. This is a pedagogical connection to a conceptual model, whose feasibility and clinical effects remain untested.

Evaluation needed to establish educational value

Evaluation should begin with clinical and educational review of task relevance, response-process interviews and usability testing with the intended learners. The proposed rubric requires evidence of content validity and scoring reliability before high-stakes use. Critical errors should be recorded separately from the total score so that a superficially high score does not conceal an unsupported clinical inference. Following the distinction emphasized by Vega Ochoa and colleagues [5], learner experience should be measured separately from appraisal performance, retention and transfer, with facilitator preparation and feedback documented.

A subsequent study should compare MammoTrace-supported instruction with an active, time-matched approach using the same cases and learning objectives, for example conventional teaching materials or an established tool with equivalent prompts. A prespecified primary endpoint could assess transfer to previously unseen appraisal tasks at a delayed follow-up, using baseline adjustment and blinded scoring. The protocol should justify sample size, handle clustering and missing outcomes, and distinguish instructor contributions from software functions. Additional file 1 provides a prospective outline; no effect size or learner data are available from this study.

Responsible implementation also requires clear governance of teaching materials, personal information and professional responsibility [25,26]. The evidence-only route avoids requiring local patient files for these activities. Any future learner study or clinical-data deployment needs its own institutional review, appropriate permissions and a defined human oversight process.

Limitations

The project and its evaluation are developer-led, concern one software version and rely partly on retained input snapshots. The new execution used Python 3.12 with inherited packages rather than a clean installation of the

documented Python 3.13 profile. It reproduced public-release test counts but did not reproduce the original data-dependent audit or acquire the complete image corpus. Statement coverage was incomplete, and a broad module guard skipped relevant linguistic tests. Direct characterization therefore remains essential to interpreting the selected examples.

DICOM observations were limited to selected files and a restricted reader profile. Spanish report characterization was purposive and small; the temporal punctuation limitation remains in the release. Full model training, external diagnostic validation, validated explanation methods, accessibility, concurrent-user performance, live external-AI transport and complete security testing were not evaluated. Screen-film mammography and the candidate's 224-pixel input also constrain transfer to modern imaging tasks.

The teaching resource was constructed after the technical work without learner testing, expert content validation or a comparative educational study. Its rubric has no established validity, reliability or pass threshold. Learning gains, retention, transfer, usability and clinical benefit remain undetermined. The software comparison is documentary and cannot demonstrate a significant empirical advantage. Some original local artifacts are not publicly deposited, and the Dataverse ZIP was not independently compared byte for byte. The institutional ethics status of the historical secondary use is stated in the declarations.

Conclusions

MammoTrace Academic v4.1.0 provides traceable examples for teaching how claims about mammography AI are constructed and limited. Subject-overlap checks, a blocked training preflight and a punctuation-sensitive report case connect observed software behavior to worked appraisal activities. The present evidence establishes those technical observations and makes the instructional design inspectable. A comparative study with learners is needed to test its educational contribution.

Availability and requirements

Project: MammoTrace Academic 4.1.0, PUBLICA_SIN_PESOS edition. Public source and versioned release: https://github.com/Yolanda2727/MammoTrace-Academic/releases/tag/v4.1.0 [17]. Software deposit: Harvard Dataverse, DOI 10.7910/DVN/LXGDZZ, deposit version 1.0 [10]. Software version 4.1.0 and deposit version

1.0 are distinct. GitHub publication was 13 September 2026; the deposit became available on 20 September 2026.

Language and platform: Python; Linux was used for the recorded executions. The public profile targets Python 3.13 and specifies dependencies in requirements.txt and requirements-base.txt. Optional experimental dependencies are separated in requirements-ml.txt. The additional execution used Python 3.12.14. Completed Windows or macOS acceptance is not demonstrated.

License and permissions: Permission is granted, free of charge, to download, install and run the original MammoTrace Academic code for anonymous editorial review and noncommercial teaching and research. No individual permission or material transfer agreement is required for these uses. This permission applies only to code whose rights are held by Anderson Díaz-Pérez. Third-party dependencies, datasets and images remain subject to their respective terms. The dated notice AUTHOR_PERMISSION.txt accompanies the source archive. This specific authorization does not designate a standard open-source license or alter the original release or deposit terms. Historical model weights and bulk third-party imaging inputs are not redistributed.

Abbreviations

AI, artificial intelligence; BI-RADS, Breast Imaging Reporting and Data System; CBIS-DDSM, Curated Breast Imaging Subset of Digital Database for Screening Mammography; CLAIM, Checklist for Artificial Intelligence in Medical Imaging; CSV, comma-separated values; DICOM, Digital Imaging and Communications in Medicine; DOI, digital object identifier; FAIR, findable, accessible, interoperable and reusable; TCIA, The Cancer Imaging Archive.

Declarations

Ethics approval and consent to participate

No people were recruited, and no prospective educational or clinical intervention or animal experiment was conducted. The work reused public CBIS-DDSM metadata and human-derived mammographic images. No new participant consent was obtained. No institutional approval, exemption or waiver is documented in the project record for this secondary use. Public accessibility is not treated as proof of institutional ethical exemption or complete de-identification.

Consent for publication

Not applicable. The manuscript and figures contain aggregate technical information, diagrams and constructed report texts, with no identifiable participant narratives or individual clinical images. The study did not obtain new publication consent for the source mammograms, which are not reproduced in the current submission.

Availability of data and materials

The source release and selected technical reports are available through GitHub [17]. The same source commit is supplied in Additional file 5, with the author's permission for anonymous review and noncommercial teaching and research stated in AUTHOR_PERMISSION.txt. Harvard Dataverse identifies the software deposit [10], whereas CBIS-DDSM is an external dataset with a separate TCIA DOI [12]. Official DataCite metadata link the software deposit to the same source revision and directory as the GitHub release; this supports bibliographic correspondence but not byte-level identity of the uninspected deposit ZIP. Additional files provide the evidence inventory, aggregate counts, teaching materials and execution records. The source archive excludes bulk imaging inputs, historical model weights and some local development artifacts.

Competing interests

ADP is the developer and author of MammoTrace Academic. This relationship is disclosed because the evaluation is developer-led. The author reports no financial conflicts or additional competing interests.

Funding

This work received no funding and had no sponsor.

Authors' contributions

ADP: conceptualization, methodology, software development, data curation, analysis, technical verification, visualization, and manuscript drafting and revision. Generative assistance is described below. The author approved the manuscript and accompanying submission materials on 21 October 2026 and accepts responsibility for their content.

Acknowledgements

Not applicable.

Generative AI assistance

ChatGPT (OpenAI) assisted programming, test execution, analysis organization, literature searching, drafting, translation, instructional-material preparation and editorial review. A complete historical model-version and prompt log was not retained. Historical counts were distinguished from the additional execution records. No participant observations, learning outcomes or clinical validation were generated. The author approved the resulting manuscript and assumes responsibility for its content. AI assistance does not constitute independent human review.

Additional files

Additional file 1. Material-Suplementario-S1-S10.docx. Technical sources, partitioning, availability, DICOM observations, verification boundaries, reproduction details and a prospective educational evaluation outline.

Additional file 2. Additional_file_2_aggregate_evidence.csv. Aggregate engineering counts with their source and verification scope, including the separate 21 October 2026 execution.

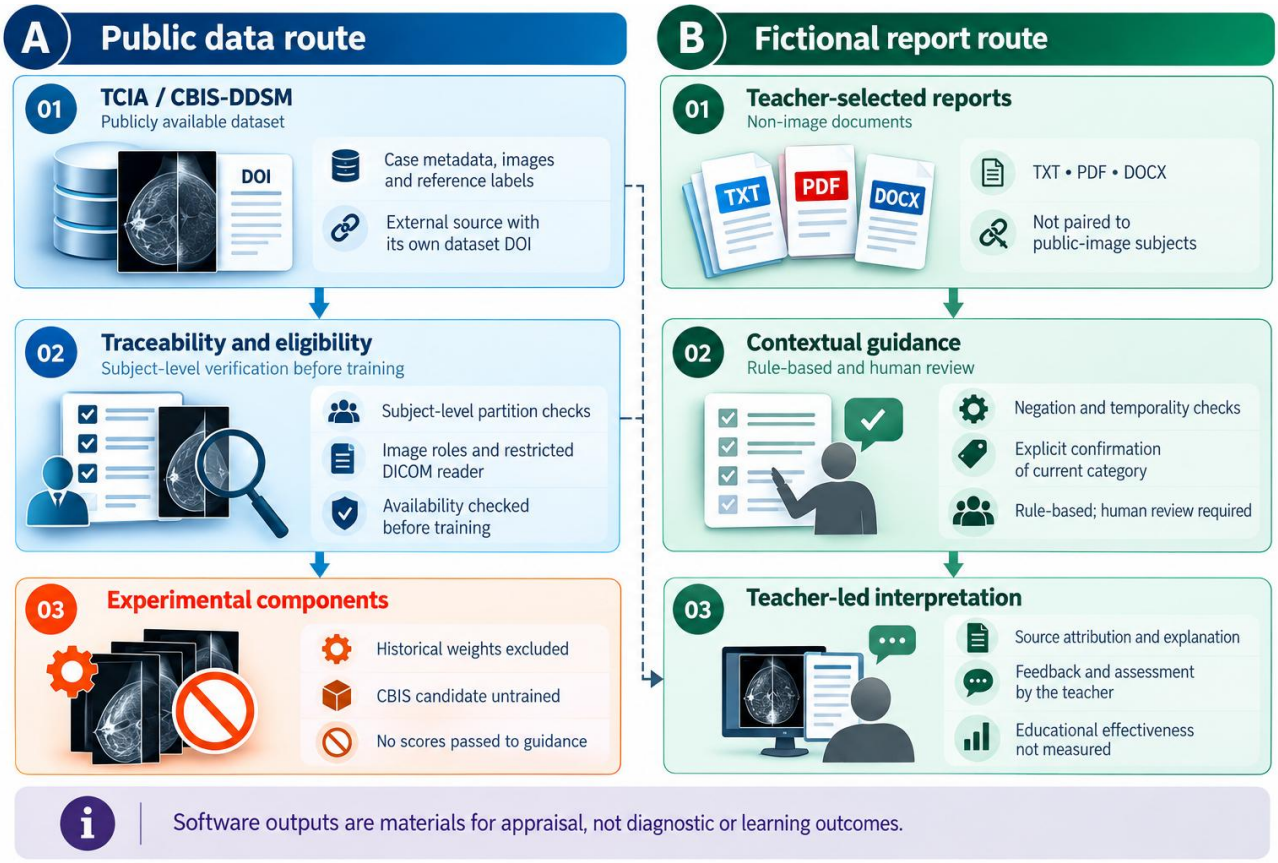
Additional file 3a. Additional_file_3a_Learner_workbook.docx. Six proposed appraisal activities, evidence cards and 27 learner questions; no answer key.

Additional file 3b. Additional_file_3b_Facilitator_guide.docx. Worked answers, explanations of common errors, proposed formative rubric and debrief guidance. The rubric is not validated.

Additional file 4. Additional_file_4_technical_reproduction.zip. Public-profile execution logs, JUnit and coverage outputs, environment, source-integrity verification and reproducible literal parser characterization; no patient images or historical model weights.

Additional file 5. Additional_file_5_MammoTrace_v4_1_0_source.zip. Unmodified source archive of commit c460f85ab73895d3ceba954427b608a70a02eb3b, corresponding to the published v4.1.0 release. Its original 94-entry content manifest was verified. AUTHOR_PERMISSION.txt accompanies the archive as the author's dated software-use notice.

Figure legends



383

384 Figure 1. Software architecture and evidence routes. Public images and metadata, fictional reports and
385 experimental AI are separate inputs. Solid connectors show implemented flow; dashed connectors show
386 proposed teacher-led use. Experimental model scores do not feed educational guidance. The diagram describes
387 architecture and an instructional proposal, not diagnostic safety or learning effects.

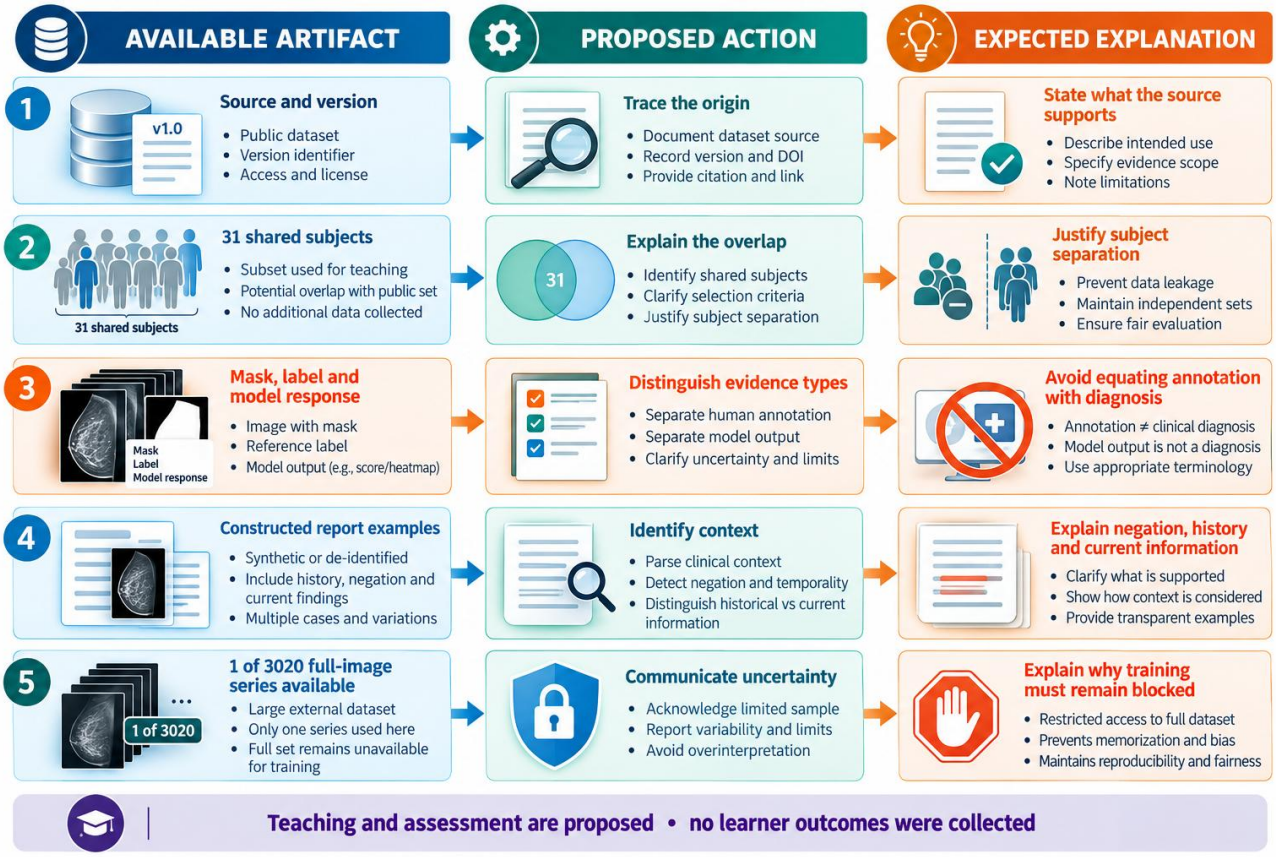
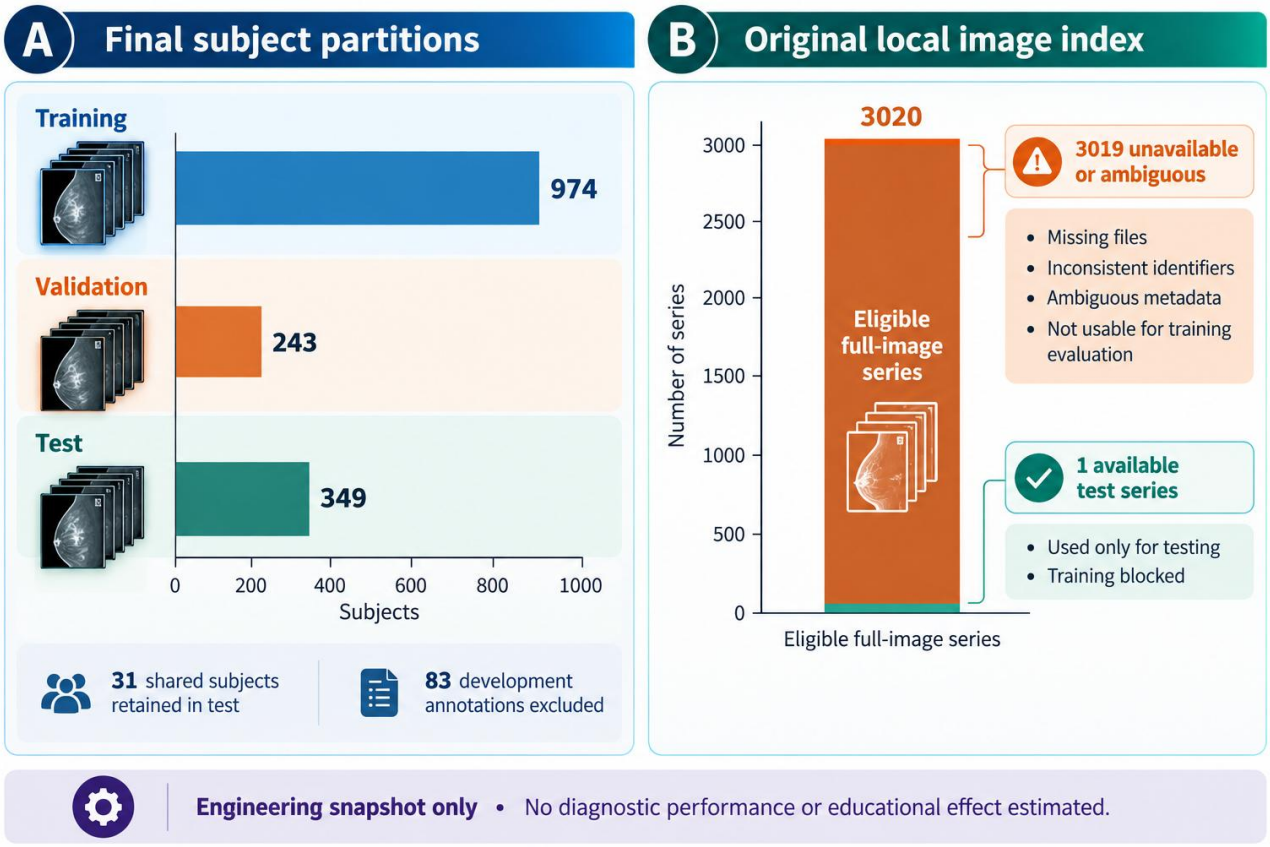


Figure 2. Proposed alignment of five appraisal objectives. Artifacts are connected to learner actions and expected explanations. Teaching, feedback and scoring are external activities, with six detailed companion activities covering the five objectives. No learner outcomes were collected.



392

393 Figure 3. Subject separation and original local image availability. Panel A shows final subject counts; the 31
394 subjects excluded from development remain in test. Panel B distinguishes 3020 metadata-eligible full-image
395 series from one unambiguously available test series and 3019 unavailable or ambiguous series. Training
396 remained blocked. These are engineering counts from the retained September snapshot.

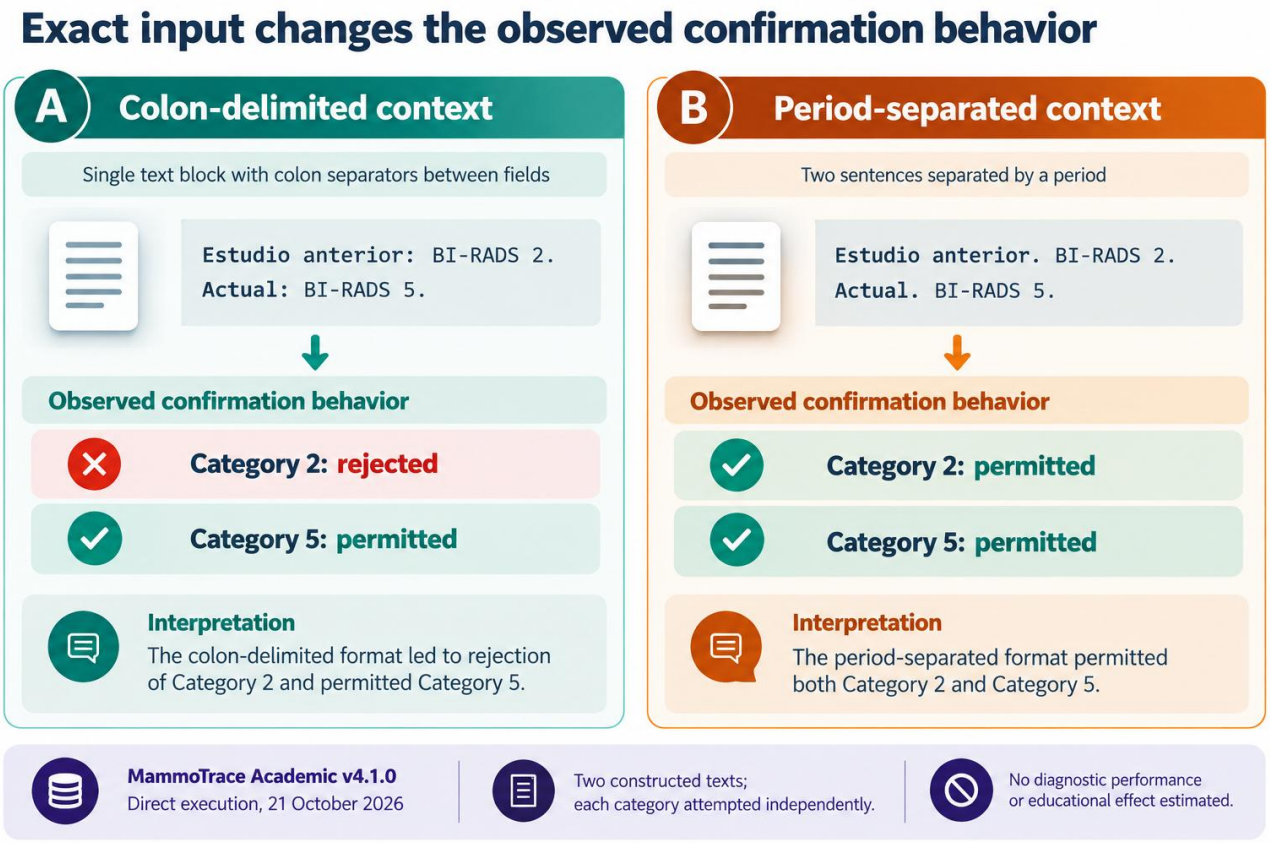


Figure 4. Punctuation-sensitive temporal interpretation. Exact Spanish inputs and observed confirmation behavior from direct execution of v4.1.0 on 21 October 2026. Replacing colons after the temporal context labels with periods allowed confirmation of the historical category 2. Categories were attempted independently. English explanatory labels are not tested parser inputs; no diagnostic or linguistic performance rate is inferred.

References

1. Çalışkan SA, Demir K, Karaca O. Artificial intelligence in medical education curriculum: an e-Delphi study for competencies. PLoS One. 2022;17(7):e0271872. https://doi.org/10.1371/journal.pone.0271872.

2. Civaner MM, Uncu Y, Bulut F, Chalil EG, Tatli A. Artificial intelligence in medical education: a cross-sectional needs assessment. BMC Med Educ. 2022;22:772. https://doi.org/10.1186/s12909-022-03852-3.

3. van Kooten MJ, Tan CO, Hofmeijer EIS, van Ooijen PMA, Noordzij W, Lamers MJ, et al. A framework to integrate artificial intelligence training into radiology residency programs: preparing the future radiologist. Insights Imaging. 2024;15:15. https://doi.org/10.1186/s13244-023-01595-3.

4. Li R, Liu G, Zhang M, Rong D, Su Z, Shan Y, et al. Integration of artificial intelligence in radiology education: a requirements survey and recommendations from faculty radiologists, residents, and medical students. BMC Med Educ. 2025;25:380. https://doi.org/10.1186/s12909-025-06859-8.

5. Vega Ochoa AD, Pinto Bautista HL, Weber Ipuana WJ, González Cuello MA, Díaz Pérez A. Student perceptions of technology-enhanced simulation resources in laparoscopic training: a scoping review and implications for surgical instrumentation. BMC Med Educ. 2026;26:1679. https://doi.org/10.1186/s12909-026-10009-z.

6. Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. Patterns. 2023;4(9):100804. https://doi.org/10.1016/j.patter.2023.100804.

7. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health. 2021;3(11):e745–e750. https://doi.org/10.1016/S2589-7500(21)00208-9.

8. Arun N, Gaw N, Singh P, Chang K, Aggarwal M, Chen B, et al. Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. Radiol Artif Intell. 2021;3(6):e200267. https://doi.org/10.1148/ryai.2021200267.

9. Biggs J. Enhancing teaching through constructive alignment. High Educ. 1996;32:347–364. https://doi.org/10.1007/BF00138871.

10. Díaz Pérez A. MammoTrace Academic v4.1.0: Traceable CBIS-DDSM Integration, Experimental AI Separation and Educational Guidance [software on the Internet]. Harvard Dataverse; 2026 [cited 2026 Oct 21]. Deposit version 1.0. Available from: https://doi.org/10.7910/DVN/LXGDZZ.

11. Lee RS, Gimenez F, Hoogi A, Miyake KK, Gorovoy M, Rubin DL. A curated mammography data set for use in computer-aided detection and diagnosis research. Sci Data. 2017;4:170177. https://doi.org/10.1038/sdata.2017.177.

12. Sawyer-Lee R, Gimenez F, Hoogi A, Rubin D. Curated Breast Imaging Subset of Digital Database for Screening Mammography (CBIS-DDSM) [dataset on the Internet]. The Cancer Imaging Archive; 2016 [cited 2026 Oct 21]. Available from: https://doi.org/10.7937/K9/TCIA.2016.7O02S9CY.

13. Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. J Digit Imaging. 2013;26(6):1045–1057. https://doi.org/10.1007/s10278-013-9622-7.

14. Tejani AS, Klontzas ME, Gatti AA, Mongan JT, Moy L, Park SH, et al. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 update. Radiol Artif Intell. 2024;6(4):e240300. https://doi.org/10.1148/ryai.240300.

15. Streamlit. st.testing.v1.AppTest [Internet]. https://docs.streamlit.io/develop/api-reference/app-testing/st.testing.v1.apptest. Accessed 21 Oct 2026.

16. DICOM Standards Committee. DICOM PS3.15 Security and system management profiles. Annex E Attribute confidentiality profiles [Internet]. https://dicom.nema.org/medical/dicom/current/output/chtml/part15/chapter_E.html. Accessed 21 Oct 2026.

17. Díaz Pérez A. MammoTrace Academic v4.1.0 public research release [software release notes on the Internet]. GitHub; 2026 Sep 13. https://github.com/Yolanda2727/MammoTrace-Academic/releases/tag/v4.1.0. Accessed 21 Oct 2026.

18. Demšar J, Curk T, Erjavec A, Gorup Č, Hočevar T, Milutinovič M, et al. Orange: Data Mining Toolbox in Python. J Mach Learn Res. 2013;14:2349–2353. Available from: https://www.jmlr.org/papers/v14/demsar13a.html.

19. Orange Data Mining. Test and Score [Internet]. https://orangedatamining.com/widget-catalog/evaluate/testandscore/. Accessed 21 Oct 2026.

20. Fedorov A, Beichel R, Kalpathy-Cramer J, Finet J, Fillion-Robin JC, Pujol S, et al. 3D Slicer as an image computing platform for the Quantitative Imaging Network. Magn Reson Imaging. 2012;30(9):1323–1341. https://doi.org/10.1016/j.mri.2012.05.001.

21. Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B, et al. Model cards for model reporting. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. New York: ACM; 2019. p. 220–229. https://doi.org/10.1145/3287560.3287596.

22. Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, Daumé H 3rd, et al. Datasheets for datasets. Commun ACM. 2021;64(12):86–92. https://doi.org/10.1145/3458723.

23. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3:160018. https://doi.org/10.1038/sdata.2016.18.

24. Díaz-Pérez A, Pérez WJA. Clinical cyberbioethics for AI-mediated clinical decision-making: a critical narrative review and conceptual proposal for an encounter-level deliberative model. BMC Med Ethics. 2026 Sep 12. https://doi.org/10.1186/s12910-026-01596-z. Online ahead of print.

25. World Health Organization. Ethics and governance of artificial intelligence for health: WHO guidance. Geneva: World Health Organization; 2021. https://www.who.int/publications/i/item/9789240029200.

26. Masters K. Ethical use of artificial intelligence in health professions education: AMEE Guide No. 158. Med Teach. 2023;45(6):574–584. https://doi.org/10.1080/0142159X.2023.2186203.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- MaterialSuplementarioS1S10.docx
- Additionalfile2aggregateevidence.csv
- Additionalfile3aLearnerworkbook.docx
- Additionalfile3bFacilitatorguide.docx
- Additionalfile4technicalreproduction.zip
- AUTHORPERMISSION.txt
- Additionalfile5MammoTracev4.1.0source.zip