STUDY PROTOCOL WITH STATISTICAL ANALYSIS PLAN

Title: AI-Assisted Acute Myeloid Leukemia Evaluation With the Leukemia End-to-End Analysis Platform (LEAP) Versus Clinician-Only Assessment

ClinicalTrials.gov ID	NCT07203885
Protocol Version	1.0
Sponsor	Harvard Medical School (HMS and HSDM)
Principal Investigator	Kun-Hsing Yu, MD, PhD Associate Professor of Biomedical Informatics Harvard Medical School Email: Kun-Hsing_Yu@hms.harvard.edu Phone: 617-432-2144
IRB Approval	Harvard Human Research Protection Program Approval Number: IRB23-0403

Table of Contents

1. PROTOCOL SUMMARY

1.1 Study Title

AI-Assisted Acute Myeloid Leukemia Evaluation With the Leukemia End-to-End Analysis Platform (LEAP) Versus Clinician-Only Assessment

1.2 Study Phase

Not Applicable (N/A)

1.3 Study Rationale

Acute promyelocytic leukemia (APL) is a rare but life-threatening subtype of acute myeloid leukemia (AML) that requires urgent diagnosis and treatment initiation. The diagnosis of APL relies on microscopic examination of bone marrow aspirate smears and molecular confirmation. However, inter-observer variability in microscopic evaluation has been reported, and molecular profiling requires additional time and resources. Artificial intelligence (AI) has the potential to augment diagnostic accuracy and efficiency in detecting APL from whole-slide images (WSIs). This study evaluates whether AI assistance through the Leukemia End-to-End Analysis Platform (LEAP) improves clinicians' diagnostic performance compared to unaided review.

1.4 Study Design

This is a prospective, randomized, crossover study, with AI assistance serve as the intervention. Clinicians review de-identified Wright-Giemsa-stained bone marrow WSIs under three conditions: (1) Unaided Review, (2) AI as Double-Check, and (3) AI as First Look. The order of unaided and AI-assisted reviews is randomized to minimize bias and learning effects.

1.5 Study Population

Participants (Readers): Board-certified pathologists and hematologists who routinely perform hematopathology diagnoses. Target enrollment: 10 participants (actual enrollment: 10 participants).

Cases (Pathology Slides): De-identified Wright-Giemsa-stained bone marrow aspirate WSIs with confirmed diagnoses. Each participant reviews 102 slides (34 slides per condition, stratified by APL status).

1.6 Study Duration

Study Start Date: September 9, 2025 (Actual)
Study Completion Date: October 10, 2025 (Actual)

1.7 Primary Objective

To evaluate the diagnostic performance of clinicians in detecting APL from bone marrow WSIs with and without AI assistance, measured by accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV).

1.8 Secondary Objectives

- To assess the time required to reach a diagnosis under different conditions
- To measure inter-observer variability across conditions
- To determine concordance between AI predictions and clinicians' diagnoses

- To evaluate decision-change rates after clinicians reviewed AI diagnostic results
- To quantify the net benefit of AI assistance on diagnostic accuracy
- To assess clinician confidence levels across conditions

1.9 Primary Endpoint

Diagnostic performance metrics (accuracy, sensitivity, specificity, PPV, NPV) for APL detection under unaided and AI-assisted conditions.

2. STUDY OBJECTIVES

2.1 Primary Objective

To evaluate and compare the diagnostic performance of clinicians in detecting acute promyelocytic leukemia (APL) from Wright-Giemsa-stained bone marrow whole-slide images under three conditions:

- **Condition X (Unaided Review):** Clinicians review slides without AI assistance.
- **Condition Y1 (AI as Double-Check):** Clinicians provide initial diagnosis, then review AI prediction and may revise their decision. Diagnoses made before and after AI exposure were recorded.
- **Condition Y2 (AI as First Look):** Clinicians review slides with AI prediction visible from the start.

2.2 Secondary Objectives

1. **Time efficiency:** To assess the average time required to reach a diagnostic decision under each condition.
2. **Inter-observer agreement:** To measure the consistency of diagnoses among clinicians across different conditions using inter-rater reliability metrics.
3. **AI-clinician concordance:** To determine the proportion of cases where AI predictions match clinicians' final diagnoses in each condition.
4. **Decision revision patterns:** To quantify how often clinicians change their initial diagnosis after viewing AI predictions in the "AI as Double-Check" condition.
5. **Net diagnostic benefit:** To calculate the overall improvement in diagnostic accuracy attributable to AI assistance.
6. **Confidence assessment:** To evaluate clinicians' self-reported diagnostic confidence across conditions and examine the relationship between confidence and accuracy.

3. STUDY DESIGN

3.1 Overall Design

This is a prospective, randomized, crossover study evaluating AI diagnostic assistance as the intervention. Each participating clinician (reader) reviews the same total number of cases (102

WSIs) but under three different conditions, with cases allocated such that no individual case is reviewed twice by the same reader.

3.2 Study Type

Study Type: Interventional
Study Design: Crossover assignment
Primary Purpose: To evaluate the effectiveness of AI diagnostic assistance
Study Phase: Not applicable (N/A)
Allocation: Randomized
Masking: Triple (Patients, Participating Clinicians, and Investigators)

3.3 Study Arms and Randomization

Participants are randomized into one of two study arms, which differ in the order in which diagnostic conditions are presented:

Arm A: Unaided Review First, Then AI-Assisted Review (X → Y)

- **Block X (Unaided Review):** Participants review each slide in subset SX (34 slides) without AI assistance.
- **Block Y (AI-Assisted Review):** Participants review two subsets with AI assistance:
    - **Sub-block Y1 (AI as Double-Check):** Subset SY1 (34 slides) - Participants review each slide, provide an initial diagnosis, then view the AI-generated prediction and may revise their diagnosis if desired.
    - **Sub-block Y2 (AI as First Look):** Subset SY2 (34 slides) – Participants review each slide with the AI prediction visible from the beginning.

Arm B: AI-Assisted Review First, Then Unaided Review (Y → X)

- **Block Y (AI-Assisted Review):** Participants review two subsets with AI assistance, following the same procedure described for Block Y above.
    - **Sub-block Y1 (AI as Double-Check):** Subset SY1 (34 slides)
    - **Sub-block Y2 (AI as First Look):** Subset SY2 (34 slides)
- **Block X (Unaided Review):** Participants review each slide in subset SX (34 slides) without AI assistance, following the same procedure described for Block X above.

3.4 Slide Allocation

For each participant, the 102 WSIs are randomly divided into three disjoint subsets (SX, SY1, SY2), each containing 34 slides. The allocation is stratified by APL status to ensure balanced representation of positive and negative cases across all three conditions.

Key principles:

- No slide is shown to the same reader more than once.
- Slide allocation is unique to each reader.
- Stratification ensures a comparable APL prevalence across all three subsets for each reader.

- Randomization minimizes bias from case difficulty or other confounders.

3.5 Study Workflow

Each participant completes a study session with the following structure:

1. **Randomization:** Participant is randomly assigned to Arm A or Arm B.
2. **Block sequence setting:** Participant completes blocks in the assigned order (X→Y or Y→X), depending on their Arm assignment.
3. **Slide sequence setting:** Participant review slides according to a predefined randomized order. The sequence is different for each participant.
4. **Pre-study briefing:** Participants receive instructions on the review platform; They also review two mock slides, one unaided, one aided, to familiarize themselves with the testing platform.
5. **Data recording:** For each slide, the system automatically records:
   - Diagnosis (APL vs. non-APL)
   - Confidence score (1-5 scale)
   - Time to decision (in seconds)
   - For Y1 (AI as Double-Check): both pre-AI and post-AI diagnoses and confidence scores

3.6 AI System (LEAP) Implementation

The LEAP AI system generates diagnostic predictions independently for all 102 WSIs before the study session begins. During the study:

- **Condition X (Unaided):** AI predictions are not shown to the clinician
- **Condition Y1 (AI as Double-Check):** AI prediction is revealed only after the clinician submits their initial diagnosis
- **Condition Y2 (AI as First Look):** AI prediction is visible throughout the review

AI predictions are presented as binary classifications (APL vs. non-APL).

3.7 Masking and Blinding

The study employs triple masking:

- **Patients:** Patients are blinded to the study arm to which their samples are assigned.
- **Care Provider:** Clinicians are blinded to the ground-truth diagnoses during all review sessions.
- **Investigator:** Study coordinators administering the session are blinded to the ground-truth diagnoses and to participants' arm assignments. In addition, statistical analysts are blinded to participant identities and arm assignments before results are finalized.

3.8 Ground Truth Determination

The reference standard diagnosis for each case is established using flow cytometry, immunohistochemical staining of trephine needle BM biopsies, cytogenetic analysis, and genomic testing.

4. STUDY POPULATION

4.1 Participating Clinicians (Readers)

4.1.1 Target Enrollment

Planned: 10 participants
Actual: 10 participants

4.1.2 Inclusion Criteria for Readers

- Board-certified pathologists who routinely interpret hematopathology specimens in clinical practice, OR board-certified or board-eligible hematologists who routinely make hematopathology diagnoses in clinical practice.
- Willingness and ability to complete both unaided and AI-assisted review sessions.
- Familiarity with microscopic interpretation of Wright-Giemsa-stained bone marrow aspirate smears.

4.2 Patients

4.2.1 Sample Size

The number of cases reviewed per participating clinician: 102 Wright-Giemsa-stained bone marrow aspirate WSIs from 102 patients
Distribution per condition: 34 slides for Condition X, 34 slides for Condition Y1, 34 slides for Condition Y2

4.2.2 Inclusion Criteria for Cases

- Wright-Giemsa-stained bone marrow aspirate smear digitized as a whole-slide image.
- Adequate specimen quality and cellularity for diagnostic interpretation.
- Final diagnosis confirmed through molecular testing and established ground-truth.

4.2.3 Exclusion Criteria for Cases

- Poor-quality or technically inadequate slides (e.g., inadequate staining, crush artifact).
- Unreadable or corrupted digital images.
- Cases used in the training or internal validation of the LEAP AI model.

4.3 Recruitment and Consent

4.3.1 Recruitment Strategy

Participants are recruited by invitation from Harvard Medical School, with recruitment targeting the following institutions to ensure geographical diversity:

- Brigham and Women's Hospital, Boston, MA, USA
- Division of Hematopathology, Mayo Clinic, Rochester, Minnesota, USA
- Pennsylvania State University, Hummelstown, PA, USA
- Hospital of the University of Pennsylvania, PA, USA
- Taipei Veterans General Hospital, Taipei, Taiwan
- Northern Jiangsu People's Hospital, Yangzhou, Jiangsu province, China
- Medical University of Vienna, Department of Pathology, Vienna, Austria
- Metropolis Healthcare Ltd. Indore, Madhya Pradesh, India

5. STUDY PROCEDURES

5.1 Pre-Study Procedures

5.1.1 AI Prediction Generation

Prior to participant enrollment, the LEAP AI system generates diagnostic predictions for all 102 study cases. AI predictions are stored securely and revealed to participants only according to the assigned study condition.

5.1.2 Randomization Procedures

- Participants are randomized to Arm 1 (X→Y) or Arm 2 (Y→X) using a computer-generated randomization sequence with size 2 and size 4 blocks.
- For each participant, slides are randomly allocated to subsets SX, SY1, and SY2 with stratification by APL status
- All randomization is performed using validated software and documented in the study database.

5.2 Study Session Procedures

5.2.1 Reader Orientation

At the beginning of the study session, readers receive:

- Instructions on using the digital slide review platform
- Explanation of the three review conditions (X, Y1, Y2)
- Descriptions of the confidence rating scale (1-5)
- Opportunity to ask questions and practice with example slides not included in the study set

5.2.2 Slide Review Process

For each slide, readers:

1. Review the digitized WSI using the study platform with zoom and navigation capabilities
2. Formulate a diagnosis (APL vs. non-APL)
3. Assign a confidence score (1 = Random Guess; 2 = Very Doubtful; 3 = Unsure; 4 = Mostly Certain; 5 = Absolutely Certain)

4. Submit their decision (or initial decision in Condition Y1)
5. **For Condition Y1 only:** After submission, view AI prediction and optionally revise diagnosis and confidence rating
6. Proceed to the next slide

5.2.3 Data Capture

The study platform automatically records for each slide:

- Reader ID (de-identified)
- Slide ID (de-identified)
- Study condition (X, Y1, or Y2)
- Diagnosis (APL vs. non-APL)
- Confidence score (1-5)
- Time from slide presentation to diagnosis submission (in seconds)
- For Condition Y1 only:
  - Pre-AI diagnosis and confidence score
  - Post-AI diagnosis and confidence score
- Timestamp of review

5.2.4 Session Duration

Participants complete all 102 slides in a single session. The expected duration is approximately 2-3 hours, with optional breaks permitted. The study platform allows participants to pause between slides but not during the review of an individual slide.

5.3 Study Completion

Actual Study Start Date: September 9, 2025
Actual Study Completion Date: October 10, 2025
Actual Number of Participants: 10
Actual Number of Slides per Participant: 102 (34/34/34)
Actual Number of Total Diagnostic Decisions in This Study: 1,020

6. OUTCOME MEASURES

6.1 Primary Outcome Measure

6.1.1 Diagnostic Performance of APL Detection

Measurement Time Frame: Periprocedural (at the time of slide review)

Description: The primary outcome is the diagnostic performance of clinicians in detecting APL, evaluated separately for each of the three study conditions (X, Y1, Y2). Performance will be measured using the following metrics:

- **Accuracy:** Proportion of correct diagnoses (both APL and non-APL cases)
  - Formula: (TP + TN) / (TP + TN + FP + FN)
- **Sensitivity:** Proportion of true APL cases correctly identified as APL

    - Formula: TP / (TP + FN)
- **Specificity:** Proportion of true non-APL cases correctly identified as non-APL
    - Formula: TN / (TN + FP)
- **Positive Predictive Value (PPV):** Proportion of cases diagnosed as APL that are truly APL
    - Formula: TP / (TP + FP)
- **Negative Predictive Value (NPV):** Proportion of cases diagnosed as non-APL that are truly non-APL
    - Formula: TN / (TN + FN)

Where: TP = True Positive, TN = True Negative, FP = False Positive, FN = False Negative

6.2 Secondary Outcome Measures

6.2.1 Time to Diagnosis

Measurement Time Frame: Periprocedural (at the time of slide review)

Description: Time in seconds required to reach a diagnostic decision for each case. This will be calculated as the elapsed time from when the slide is displayed to when the participant submits their diagnosis. For Condition Y1, both pre-AI and post-AI (total) review times will be recorded. Time will be compared across the three conditions to assess whether AI assistance affects diagnostic efficiency.

6.2.2 Inter-Observer Variability

Measurement Time Frame: Periprocedural (at the time of slide review)

Description: Agreement among clinicians across the three conditions, measured using inter-rater reliability metrics:

- Cohen's kappa (for pairwise agreement)

6.2.3 Concordance Between AI Predictions and Clinicians' Diagnoses

Measurement Time Frame: Periprocedural (at the time of slide review)

Description: The proportion of cases in which AI predictions match clinicians' final diagnoses in each study condition. This will be calculated as the percentage agreement between LEAP predictions and clinician diagnoses. Concordance will be evaluated separately for Conditions X, Y1, and Y2.

6.2.4 Decision-Change Rates

Measurement Time Frame: Periprocedural (at the time of slide review)

Description: In Condition Y1 (AI as Double-Check), the proportion of cases in which a clinician's initial diagnosis is revised after viewing the AI prediction. Decision changes will be categorized as:

- Correct-to-incorrect (C→I): Initial correct diagnosis changed to incorrect after viewing the AI prediction
- Incorrect-to-correct (I→C): Initial incorrect diagnosis corrected after viewing the AI prediction

The net decision-change rate will be calculated, along with the proportion of changes that improved accuracy.

6.2.5 Net Benefit After AI Exposure

Measurement Time Frame: Periprocedural (at the time of slide review)

Description: The overall change in diagnostic accuracy attributable to AI assistance, calculated as the difference in accuracy between unaided and aided Conditions:

6.2.6 Clinician Confidence Level

Measurement Time Frame: Periprocedural (at the time of slide review)

Description: Self-reported diagnostic confidence recorded for each case using a 5-point Likert scale:

- 5 = Absolutely Certain
- 4 = Mostly Certain
- 3 = Unsure
- 2 = Very Doubtful
- 1 = Random Guess

Confidence scores will be analyzed to:

- Compare confidence levels across the three conditions
- Assess the relationship between confidence and diagnostic accuracy
- Evaluate whether AI assistance affects clinician confidence (both for correct and incorrect diagnoses)

7. STATISTICAL ANALYSIS PLAN

7.1 General Considerations

7.1.1 Analysis Populations

Intention to Treat (ITT): Participating clinicians will be analyzed according to their originally assigned study arm, regardless of any subsequent switch between arms.

Per-Protocol (PP): Participating clinicians will be analyzed according to the review protocol they actually followed. In cases of arm switching, standard causal inference methods will be applied to adjust for baseline covariates associated with switching behavior.

Primary Analysis Population: The ITT framework will be used for all primary and secondary analyses. Although arm switching is not anticipated, a supplementary PP analysis will be conducted if any switches occur.

7.1.2 Level of Significance

A two-sided alpha level of 0.05 will be used for all comparisons.

7.1.3 Handling of Missing Data

Missing data are not anticipated in this study design, as all diagnostic decisions are captured electronically in real-time.

7.2 Primary Analysis

7.2.1 Primary Endpoint Analysis

Objective: To compare diagnostic performance (accuracy, sensitivity, specificity, PPV, NPV) across the three study conditions.

Diagnostic Performance Metrics:

For each reader and condition, the following metrics will be calculated:

- **Accuracy:** Overall proportion of correct diagnoses
- **Balanced Accuracy:** Average of sensitivity and specificity, accounting for class imbalance
- **Sensitivity:** Proportion of APL cases correctly identified
- **Specificity:** Proportion of non-APL cases correctly identified
- **Positive Predictive Value (PPV):** Proportion of APL predictions that are correct
- **Negative Predictive Value (NPV):** Proportion of non-APL predictions that are correct
- **Weighted F1 Score:** Harmonic mean of precision and recall, weighted by class prevalence

Statistical Method: Friedman test (a non-parametric test for repeated measures across multiple groups) followed by a Wilcoxon signed-rank test.

Rationale: The Friedman test is the appropriate statistical method for this study design because:

- **Within-subject design:** Each reader evaluates cases under all three conditions, creating paired data
- **Non-parametric approach:** Does not assume normal distribution of performance metrics, which is appropriate for bounded proportions (0-1 range)
- **Accounts for reader variability:** Controls for baseline differences among readers by treating each reader as a block, focusing on within-reader differences across conditions
- **Multiple conditions:** Simultaneously tests for differences among three related groups (X, Y1, Y2) rather than requiring multiple pairwise tests
- **Robust to outliers:** Uses rank-based methods that are less sensitive to extreme values

Analysis Procedure:

1. **Global test:** For each performance metric, conduct a Friedman test with the null hypothesis that the distributions of the metric are identical across all conditions.
2. **Significance threshold:** Two-sided alpha = 0.05.
3. **Post-hoc pairwise comparisons:** If the Friedman test yields $p < 0.05$ for a given metric, conduct pairwise Wilcoxon signed-rank tests for all condition pairs.

Wilcoxon Signed-Rank Test Rationale: This paired non-parametric test is appropriate for two-sample comparisons because it preserves the within-reader pairing structure and does not assume normality of the differences between conditions.

7.3 Secondary Analyses

7.3.1 Time to Diagnosis

Statistical Method: Friedman test (a non-parametric test for repeated measures across multiple groups) followed by a Wilcoxon signed-rank test.

Rationale: The Friedman test is appropriate because:

- **Within-reader design:** Each reader reviews slides under all three conditions, creating paired data at the reader level
- **Non-parametric approach:** Does not assume normal distribution of review times
- **Accounts for reader variability:** Controls for baseline differences among readers by treating each reader as a block
- **Appropriate for repeated measures**: Tests within-reader differences across conditions

Analysis Procedure:

1. **Data aggregation:** For each reader and condition, calculate the median review time across all slides reviewed in that condition (reader-level aggregation: n = 10 readers × 3 conditions = 30 observations total).

2. **Outlier detection:** Prior to aggregation, identify and remove extreme outlier values at the slide level using the 3-sigma rule (values exceeding mean + 3 standard deviations within each condition). The large outliers likely represent cases where participants left the computer with the slide review running. Analyses will be conducted both with and without the detected outliers.

3. **Global test:** Conduct a Friedman test with the null hypothesis that the distributions of the time are identical across all conditions.

4. **Significance threshold:** Two-sided alpha = 0.05.

5. **Post-hoc pairwise comparisons:** If the Friedman test yields $p < 0.05$, conduct pairwise Wilcoxon signed-rank tests for all condition pairs.

Wilcoxon Signed-Rank Test Rationale: This paired non-parametric test is appropriate for two-sample comparisons because it preserves the within-reader pairing structure (each reader's median review time is compared between conditions of interest) and does not assume normality of the differences between conditions.

7.3.2 Inter-Observer Variability

Statistical Method:

- Cohen's kappa will be calculated to assess overall inter-rater agreement.

7.3.3 Concordance Between AI Predictions and Clinicians' Diagnoses

Statistical Method: Permutation tests comparing AI predictions to clinician decisions.

Rationale:

- **Distribution-free:** Avoids large-sample and normality assumptions by creating an empirical null distribution through random permutations
- **Appropriate for proportion comparison:** Tests whether the proportion of cases where clinicians agree with AI differs significantly between conditions
- **Robust to small samples:** Does not rely on asymptotic approximations
- **Exact p-values**: Provides exact statistical inference rather than relying on theoretical distributions
- **Slide-level analysis**: Treats each clinician's diagnosis on each slide as an independent observation when comparing concordance rates between conditions

Analysis Procedure:

1. For each condition and each slide, determine the AI prediction (APL vs. non-APL).
2. For each clinician's diagnosis, determine whether it matches the AI prediction (concordant vs. discordant).
3. Calculate the overall concordance rate for each condition: proportion of all clinician-slide pairs where the diagnosis matches the AI prediction.
4. Conduct pairwise permutation tests comparing concordance rates between conditions:
    - Calculate the observed difference in concordance rates between two conditions
    - Randomly permute condition assignments while preserving the total number of observations per condition
    - Recalculate the difference in concordance rates for each permutation (n=10,000 permutations)
    - Compute two-sided p-value as the proportion of permuted differences as extreme or more extreme than the observed difference

7.3.4 Decision-Change Rates

Statistical Method: Series of exact binomial tests to evaluate decision changes and their clinical impact in Condition Y1 (AI-generated predictions shown after participants recorded their initial diagnoses).

Rationale:

- **Distribution-free:** Does not assume normality or large-sample approximations
- **Appropriate for binary outcomes:** Tests proportions of categorical outcomes (changed/unchanged, beneficial/harmful)
- **Exact inference:** Provides exact p-values based on the binomial distribution rather than asymptotic approximations
- **Multiple hypotheses**: Evaluates whether (1) changes occur, (2) changes are beneficial, and (3) overall improvement is achieved

Analysis Procedure:

1. For Condition Y1 only, identify the diagnosis before (Y1_PRE) and after the AI-generated prediction is displayed (Y1_POST).
2. Calculate overall decision-change rate: proportion of all Y1 cases with diagnostic revision.
3. Classify each decision change according to alignment with ground truth:
   - Beneficial (I→C): Incorrect initial diagnosis corrected after viewing AI
   - Harmful (C→I): Correct initial diagnosis changed to incorrect after viewing AI
4. Statistical Testing:
   - Test 1 - Overall Change Rate:
     1. Null hypothesis: Decision change rate = 0 (no changes occur)
     2. Test: One-sided exact binomial test
     3. Evaluates: Whether clinicians revise their diagnoses after viewing AI
   - Test 2 - Direction of Changes (among changes only):
     1. Null hypothesis: The likelihood of beneficial changes = The likelihood of harmful changes
     2. Test: One-sided exact binomial test comparing number of beneficial vs. harmful changes
     3. Evaluates: Whether changes tend to improve or worsen diagnostic accuracy
   - Test 3 - Overall Improvement Rate (all Y1 cases):
     1. Null hypothesis: Improvement rate = 0 (AI provides no benefit)
     2. Test: One-sided exact binomial test
     3. Evaluates: Proportion of all Y1 cases where AI led to correction of an initially incorrect diagnosis

7.3.5 Diagnostic Confidence

Objective: To assess whether AI assistance affects clinician confidence in their diagnoses, particularly in Condition Y1 where pre-AI and post-AI confidence can be compared.

7.3.5.1 Overall Distribution Comparison

Statistical Method: Wilcoxon signed-rank test.

Rationale:

- **Paired data:** Compares pre-AI and post-AI confidence for the same clinician on the same slide, preserving the within-subject pairing structure

- **Ordinal data:** Confidence ratings (1-5 scale) are ordinal in nature; Wilcoxon uses rank-based methods that appropriately handle ordinal measurements without assuming interval properties or normal distributions
- **Non-parametric approach:** Does not require assumptions about the underlying distribution of confidence scores
- **Sensitive to magnitude:** Considers both the direction and magnitude of changes in confidence, not just the presence or absence of change

Analysis Procedure:

1. **Data structure:** For each participant-slide pair in Condition Y1, extract the confidence rating before (Y1_PRE) and after (Y1_POST) viewing the AI prediction.
2. **Unit of analysis:** Individual slide-level assessments (participant × slide pairs), treating each diagnostic decision as an independent paired observation. This slide-level approach directly evaluates confidence changes at the natural unit of measurement (the diagnostic decision).
3. **Global test:** Conduct a Wilcoxon signed-rank test with the null hypothesis that the distribution of confidence scores is identical between Y1_PRE and Y1_POST.
4. **Significance threshold:** Two-sided alpha = 0.05.

7.3.5.2 Level-Specific Confidence Changes

Statistical Method: McNemar test for paired pathology evaluation with and without AI.

Rationale: The McNemar test is appropriate because:

- **Paired data:** Compares pre-AI and post-AI confidence for the same clinician on the same cases
- **Binary classification:** At each confidence level (1-5), each case can be classified as either at that confidence level or not
- **Focus on discordant pairs:** Tests whether changes in confidence at a specific level differ significantly between pre-AI and post-AI assessments
- **Accounts for within-subject dependence:** Properly handles the matched nature of pre- and post-AI assessments
- **No distributional assumptions:** Non-parametric test suitable for ordinal confidence ratings

Analysis Procedure:

1. For each confidence level (1 = Random Guess, 2 = Very Doubtful, 3 = Unsure, 4 = Mostly Certain, 5 = Absolutely Certain) in Condition Y1 (AI as Double-Check):

    - Create binary indicators: 1 if confidence equals that level, 0 otherwise
    - Construct 2×2 contingency table comparing pre-AI vs. post-AI classifications
    - Conduct McNemar exact test with null hypothesis that the proportion of cases at that confidence level is unchanged by AI exposure

2. Calculate the proportion of cases at each confidence level before and after AI review
3. Compute the change in proportion (post-AI minus pre-AI) for each confidence level
4. Report both increases and decreases in confidence at each level following AI review.

8. DATA MANAGEMENT AND QUALITY ASSURANCE

8.1 Data Collection and Management

8.1.1 Electronic Data Capture

All study data are collected electronically through a secure, purpose-built web-based platform. The platform automatically captures:

- Participant responses (diagnoses, confidence scores)
- Timestamps for all actions
- Case and participant code (sequential; de-identified)
- Study condition assignments
- AI predictions (displayed only at the time defined by the study protocol)

8.1.2 Data Storage and Security

- All data are stored on secure, encrypted servers maintained by Harvard Medical School
- Access is restricted to authorized study personnel using password-protected accounts
- Regular backups are performed and stored securely

8.1.3 Data Quality Control

Quality assurance procedures include:

- Automated validation checks during data entry
- Real-time monitoring of data completeness
- Regular review of captured data for anomalies
- Verification that all 102 cases were completed by each participant

8.2 Data Monitoring

8.2.1 Monitoring Plan

This study does not employ an independent Data Monitoring Committee (DMC) as it is a minimal-risk AI user study. However, the principal investigator and study coordinators monitor:

- Participant enrollment and completion rates
- Data quality and completeness
- Technical issues with the study platform
- Any adverse events or participant concerns

9. ETHICAL CONSIDERATIONS

9.1 Ethical Conduct

This study is conducted in accordance with:

- Declaration of Helsinki (2013 version)
- International Council for Harmonisation (ICH) Good Clinical Practice (GCP) guidelines
- U.S. Code of Federal Regulations Title 45 Part 46 (Common Rule)
- Health Insurance Portability and Accountability Act (HIPAA)
- Institutional policies of Harvard Medical School and collaborating institutions

9.2 Institutional Review Board (IRB) Approval

IRB: Harvard Longwood Campus Institutional Review Board
Approval Number: IRB23-0403
Board Affiliation: Harvard University
Status: Approved
Contact:
Phone: 866-606-0573
Email: orarc@hsph.harvard.edu
Address: 90 Smith Street, 3rd Floor, Boston, MA 02120

9.3 Informed Consent

Per IRB determination, this study involves minimal risk to participants because it entails clinician re-review of de-identified slides with confirmed diagnoses. Informed consent procedures follow institutional guidelines and include:

- Explanation of study purpose and procedures
- Description of the AI system being evaluated
- Voluntary nature of participation
- Right to withdraw without penalty
- Confidentiality protections
- Contact information of investigators for questions or concerns

9.4 Confidentiality and Privacy

9.4.1 Patient Privacy

All pathology slides used in the study are fully de-identified according to HIPAA Safe Harbor standards prior to use. No protected health information (PHI) is accessible to study participants or visible in the digital images.

9.4.2 Participant Privacy

Clinician participants are assigned unique de-identified codes. Individual performance data are kept confidential and reported only in aggregate. Institutional affiliations may be reported in publications, but individual identities will not be disclosed without explicit consent.

9.5 Risk-Benefit Assessment

9.5.1 Risks

This study poses minimal risks to participants:

- Time commitment for the study session (2-3 hours)
- Potential fatigue from prolonged slide review
- Minimal risk of breach of confidentiality (mitigated by security measures)

Importantly, this study does not involve direct patient care, and no clinical decisions are based on study diagnoses. All patients have established ground-truth diagnoses and completed clinical management.

9.5.2 Benefits

Potential benefits include:

- Advancement of knowledge regarding AI-assisted diagnosis in hematopathology
- Contribution to the development of tools that may improve diagnostic accuracy and efficiency
- Professional development through exposure to challenging cases and AI technology

9.6 Conflicts of Interest

All investigators and collaborators will disclose any financial or personal conflicts of interest related to the LEAP platform, AI technology, or study outcomes according to institutional policies and journal requirements.

10. STUDY TIMELINE AND COMPLETION

10.2 Study Completion Summary

Actual Study Start Date: September 9, 2025
Actual Completion Date of Primary Outcome Measures: October 1, 2025
Actual Study Completion Date: October 10, 2025
Actual Number of Participants Enrolled: 10
Actual Number of Slides Reviewed per Participant: 102 (34 per condition)

10.3 Protocol Amendments

No amendments to the protocol were made. The study was conducted according to the original protocol version 1.0.