Beyond One-Shot Interaction: Framing Pathology Human–AI Collaboration via Workflow Evidence

Shangqin Lyu[1] and Zeyu Liu[1]

PuzzleLogic Pte Ltd, Singapore
{shangqinglyu, zeyuliu}@puzzlelogic.com

Abstract. Pathology AI is often framed as one-shot interaction: a model highlights a region, assigns a label, or drafts text, and the pathologist accepts, edits, or discards it. We argue that this is too narrow. Clinical collaboration unfolds across a diagnostic workflow, and its value depends on whether evidence, corrections, and responsibility remain connected. We propose *workflow evidence* as the basic unit of pathology human–AI collaboration: a persistent, slide-linked record of model proposals, human inspection, granular corrections, confirmation state, provenance, and downstream use. A Select–Modify–Confirm loop operationalizes the concept, but our claim extends beyond interface design: interaction traces should become accountable clinical objects rather than transient clicks. This reframing shifts design from answer delivery to evidence continuity, and evaluation from accuracy alone to correction efficiency, traceability, calibrated reliance, error recovery, and longitudinal reuse. We illustrate the position through PuzzleVision and outline a workflow-level research agenda.

Keywords: Human–AI collaboration · Computational pathology · Workflow evidence · Interactive AI.

1 Beyond One-Shot Interaction

Pathology AI can detect suspicious tissue, rank regions of interest (ROIs), measure lesions, and generate report drafts [9]. Yet technical accuracy alone does not determine clinical value [6]. Once an algorithm enters a diagnostic workspace, its contribution depends on how findings are presented, whether whole-slide context is preserved, how disagreement is expressed, and whether the final decision remains traceable to image evidence [10]. Human–AI collaboration is therefore a property of the complete workflow, not merely of the model.

Reader studies show both the promise and limits of assistance. In prostate biopsy grading, AI support increased pathologists' agreement with an expert reference standard, demonstrating that the combined team can outperform unaided review [1]. However, assistance can also create new failure modes: pathology experiments report automation bias when erroneous advice overturns initially correct human judgments [8]. These findings shift the design question from whether AI is accurate to when, where, and in what form its advice should enter clinical reasoning.

Human–computer interaction studies make this workflow dependence explicit. xPath was designed around pathologists' native examination process after field observations identified poor workflow compatibility as a barrier to adoption [3]. NaviPath further showed that collaborative navigation could expose more relevant patterns per unit time while improving examination consistency [4]. Jarkman et al. compared interface concepts for AI-assisted lymph-node metastasis detection and identified a consequential trade-off: conventional whole-slide interaction preserves familiar navigation and contextual overview, whereas an AI-curated ROI workflow makes triage, review, and case summarization more explicit [5]. Complementing these controlled studies, CONFIDENT-B evaluated AI assistance inside routine sentinel-lymph-node assessment, emphasizing workflow integration, diagnostic safety, user behavior, and operational consequences rather than stand-alone retrospective accuracy [2].

Yet most systems still organize collaboration around isolated outputs: the AI proposes, the pathologist reacts once, and the interaction ends. Corrections may remain detached from slide context, disappear after reporting, or be stored without distinguishing model suggestion from clinician decision. This one-shot framing is misaligned with pathology, where diagnosis emerges through repeated navigation between whole-slide architecture and local morphology, comparison across candidate regions, refinement of measurements and language, and explicit commitment to a final interpretation [7].

Our position. The basic unit of pathology human–AI collaboration should be *workflow evidence*, not a prediction, prompt, or click. Workflow evidence is a persistent, slide-linked object recording: (i) what the AI proposed; (ii) what the pathologist inspected; (iii) what was corrected; (iv) what was confirmed; and (v) where confirmed content subsequently flowed. Clinical trust is not produced by a single accurate answer, but by continuity between source pixels, human judgment, and the final record. This goes beyond a generic "human-in-the-loop" label by making control, provenance, and responsibility explicit.

Relation to existing concepts. Workflow evidence builds on, but is not equivalent to, data provenance, audit trails, or versioned annotations. These mechanisms primarily record where data originated, what changed, or who performed an action. Workflow evidence instead treats the *evolving clinical evidence object itself* as the unit of collaboration, jointly linking source pixels, AI proposals, human inspection and correction, clinical commitment, and governed downstream use. Provenance and versioning are therefore enabling properties of workflow evidence rather than its defining endpoint.

2 Framing Collaboration via Workflow Evidence

Figure 1 contrasts one-shot interaction with the proposed workflow framing. Select–Modify–Confirm is one mechanism for constructing workflow evidence.

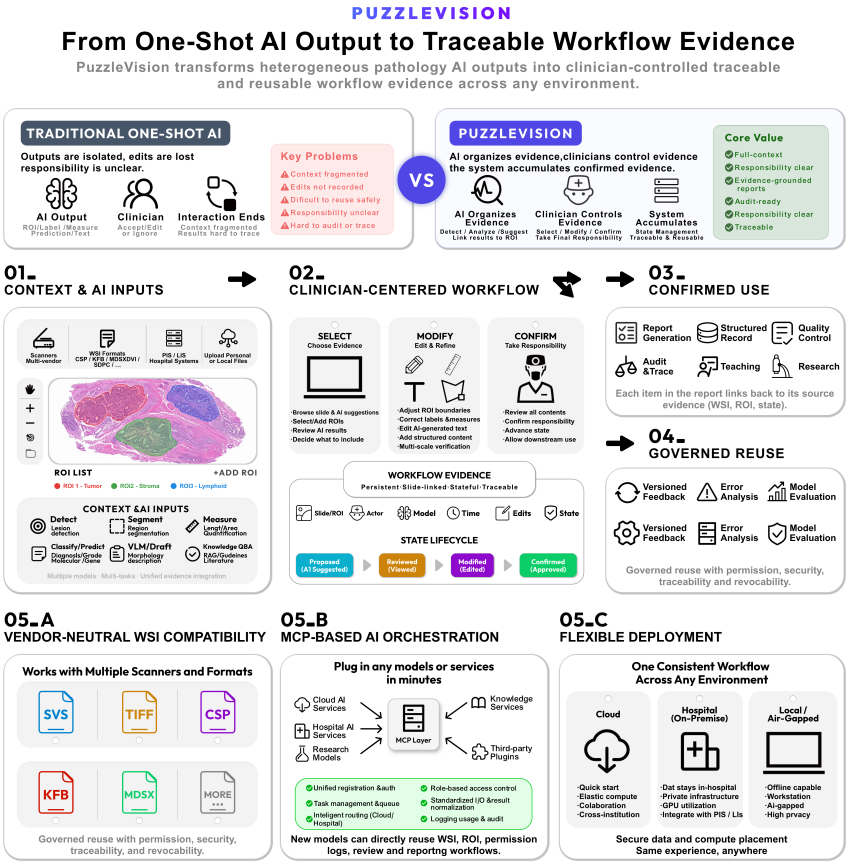

Fig. 1. From transient outputs to workflow evidence: inspection, correction, confirmation, provenance, and downstream use accumulate in a persistent slide-linked object.

2.1 Select: Preserve Diagnostic Freedom

AI can reduce search by ranking ROIs using morphology, confidence, risk, or case context, but selection should not become a forced route. Pathologists must retain unrestricted WSI navigation, inspect recommendations spatially, and add overlooked regions. The evidence object records both AI-proposed and clinician-selected regions.

2.2 Modify: Convert Output into Joint Work

Boundaries, labels, measurements, and generated text should remain independently editable. Modification reconciles local morphology, global architecture, case context, and expert judgment. Retaining the delta between proposal and correction captures evidence about both model failure and clinical reasoning.

Table 1. Evaluation dimensions induced by workflow evidence.

Dimension	Example measures
Evidence continuity	Traceable claims, decisive-region coverage, unsupported statements
Correction efficiency	Review time, edit magnitude, navigation burden
Reliance quality	Appropriate acceptance/rejection, calibration, automation bias
Responsibility clarity	Confirmation completeness, provenance, reversibility
Longitudinal value	Repeated-error reduction, audit and reuse utility

2.3 Confirm: Mark the Responsibility Boundary

Unconfirmed content remains a model suggestion. Confirmation marks clinical commitment at a particular point in the workflow and permits entry into structured data, reports, quality control, or retrospective analysis. The object records who confirmed what, when, and after which edits, preventing silent promotion of model output to clinical truth. Importantly, confirmation does not imply immutable ground truth: confirmed evidence may later be revised or withdrawn, independently validated, and separately approved or restricted for downstream reuse. Clinical state, validation status, and reuse permission should therefore remain distinct and versioned.

2.4 Five Properties

Workflow evidence requires **continuity** with source pixels and context; **correctability** of meaningful components; explicit **state** separating proposed, reviewed, modified, and confirmed content; auditable **provenance**; and governed **reusability** for reporting, quality review, education, and model improvement.

3 Illustration and Research Agenda

PuzzleVision illustrates the proposal. In one WSI workspace, pathologists browse the full slide, inspect AI-ranked ROIs, invoke heterogeneous models, edit outputs, and assemble confirmed findings into structured records or report drafts. The contribution is the interaction grammar: multiple models update one evolving evidence object rather than produce disconnected outputs.

This framing retains model metrics but adds evaluation of evidence transformation and team performance.

Four questions follow. **First**, when does AI navigation reduce search without narrowing attention? Studies should measure missed regions, not only speed. **Second**, which traces best represent meaningful correction: final edits, trajectories, rejected proposals, or combinations? **Third**, how should clinical confirmation, subsequent validation, and downstream reuse be separated? A confirmed finding is not necessarily ground truth or an approved training label; revision,

withdrawal, validation, and reuse permission should remain independently versioned. **Fourth**, how should interaction adapt to risk? Low-risk automation may need lightweight oversight, whereas consequential claims should retain visible evidence and explicit confirmation.

Granular provenance can increase workload, excessive prompts can disrupt reading, and some screening settings may favor near-automation. We therefore do not require manual approval for every action; we require a legible path from model proposal to clinical commitment whenever an output has diagnostic consequence.

4 Conclusion

Pathology human–AI collaboration should not be reduced to one-shot exchanges. Its clinically meaningful object is the evolving evidence produced across the workflow: what was proposed, inspected, corrected, confirmed, and reused. Workflow evidence turns interaction from a transient interface event into an accountable clinical artifact. Systems should therefore be evaluated not only by answer accuracy, but by whether human and machine can construct, repair, and preserve a trustworthy evidence chain together.

References

1. Bulten, W., Balkenhol, M., Belinga, J.J.A., Brilhante, A., Çakır, A., Egevad, L., Eklund, M., Farré, X., Geronatsiou, K., Molinié, V., et al.: Artificial intelligence assistance significantly improves gleason grading of prostate biopsies by pathologists. Modern Pathology **34**(3), 660–671 (2021)
2. van Dooijeweert, C., Flach, R., Ter Hoeve, N., Vreuls, C., Goldschmeding, R., Freund, J.E., Pham, P., Nguyen, T., Van der Wall, E., Frederix, G., et al.: Clinical implementation of artificial-intelligence-assisted detection of breast cancer metastases in sentinel lymph nodes: the confident-b single-center, non-randomized clinical trial. Nature cancer **5**(8), 1195–1205 (2024)
3. Gu, H., Liang, Y., Xu, Y., Williams, C.K., Magaki, S., Khanlou, N., Vinters, H., Chen, Z., Ni, S., Yang, C., et al.: Improving workflow integration with xpath: Design and evaluation of a human-ai diagnosis system in pathology. ACM Transactions on Computer-Human Interaction **30**(2), 1–37 (2023)
4. Gu, H., Yang, C., Haeri, M., Wang, J., Tang, S., Yan, W., He, S., Williams, C.K., Magaki, S., Chen, X.A.: Augmenting pathologists with NaviPath: Design and evaluation of a human–AI collaborative navigation system. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. pp. 1–19. Association for Computing Machinery (2023)
5. Jarkman, S., Lindvall, M., Lundström, C., Treanor, D., van der Laak, J.: Designing ai tools for pathology: A mixed-method study on user interface design for breast cancer lymph node metastases detection. Intelligence-Based Medicine **14**, 100396 (2026)
6. Liu, Z., Zhang, T., Chen, B.K., Feng, Y., Lyu, S., Lei, Y., Ying, N., Feng, Y., Zhao, Y., Zhang, P., et al.: Rethinking pathology image analysis through shuffling. npj Biomedical Innovations **3**(1), 42 (2026)

7. Molin, J., Fjeld, M., Mello-Thoms, C., Lundström, C.: Slide navigation patterns among pathologists with long experience of digital review. Histopathology **67**(2), 185–192 (2015)
8. Rosbach, E., Ganz, J., Ammeling, J., Riener, A., Aubreville, M.: Automation bias in ai-assisted medical decision-making under time pressure in computational pathology. In: BVM Workshop. pp. 129–134. Springer (2025)
9. Song, A.H., Jaume, G., Williamson, D.F., Lu, M.Y., Vaidya, A., Miller, T.R., Mahmood, F.: Artificial intelligence for digital and computational pathology. Nature Reviews Bioengineering **1**(12), 930–949 (2023)
10. You, Y., Li, X.: A scoping review of human-ai collaboration patterns and task divisions in healthcare applications. npj Digital Medicine (2026)