Journal of Clinical and Translational Pathology **2026** vol. 6(1) | 13–24
DOI: 10.14218/JCTP.2025.00056

Review Article

Artificial Intelligence in Genitourinary Pathology: A Translational Readiness Map

Ankush U. Patel[1*], Amanda Dy[2], Anil V. Parwani[1] and Swati Satturwar[1]

[1]The Ohio State University Wexner Medical Center, Department of Pathology, Columbus, OH, USA; [2]Electrical, Computer, and Biomedical Engineering, Toronto Metropolitan University, Toronto, ON, Canada

Received: December 31, 2025 | **Revised:** February 20, 2026 | **Accepted:** February 26, 2026 | **Published online:** March 13, 2026

Abstract

Background and objectives: Artificial intelligence (AI) translation in genitourinary (GU) pathology has progressed unevenly across organs and tasks. This review addresses a central clinical question: which GU pathology AI applications are deployment-ready, which require further validation, and what frameworks can guide safe implementation? We synthesize evidence across GU organs and introduce pragmatic translation frameworks to guide deployment and prioritize translational research. ***Methods:*** Narrative review integrating foundational literature with targeted 2023–2025 publications, emphasizing regulatory milestones, external validation, and prospective studies. Literature was identified through PubMed, Embase, and conference proceedings using structured search terms for AI, digital pathology, and GU organ-specific queries. For each organ/task, we mapped evidence strength, regulatory maturity, generalizability, workflow integration, safety, and feasibility to a Translational Readiness Index (TRI) rubric (0–30 scale). ***Results:*** Prostate biopsy AI demonstrates the strongest maturity (TRI 26/30), supported by U.S. Food and Drug Administration-cleared systems, multi-site validation, and prospective implementations showing efficiency gains and reduced ancillary testing. Bladder cytology shows moderate readiness (TRI 19/30), with commercial offerings supporting pilotable prescreening workflows aligned with the Paris System when paired with uncertainty-aware deferral. Bladder histology, renal neoplasia, and low-prevalence domains (testis, penis) remain emerging (TRI 6–15/30), constrained by label variability, rare subtype underrepresentation, and limited external validation. ***Conclusions:*** The TRI rubric, SURE-Path safety bundle, and VALIDATED/ORCHESTRATE implementation pathway provide a practical template for evidence-based deployment in GU pathology. Clinically defensible translation requires matching intended use to validation evidence, with explicit safeguards for emerging applications.

Citation of this article: Patel AU, Dy A, Parwani AV, Satturwar S. Artificial intelligence in Genitourinary Pathology: A Translational Readiness Map. J Clin Transl Pathol 2026;6(1): 13–24. doi: 10.14218/JCTP.2025.00056.

Keywords: Artificial intelligence; Genitourinary pathology; Digital pathology; Translational readiness; Clinical validation; Workflow integration; Prostate cancer; Bladder cytology; Explainable artificial intelligence.

***Correspondence to:** Ankush U. Patel, Department of Pathology, The Ohio State University, Wexner Medical Center, 410 W 10th Avenue, Columbus, OH 43210, USA. ORCID: https://orcid.org/0000-0003-3706-2320. Tel: +1-6142922064, E-mail: ankush@digitalpathomics.com

Introduction

Artificial intelligence (AI) has crossed the threshold from experimental technology to operational utility in pathology. Genitourinary (GU) pathology is an active frontier for clinical translation, driven by high disease burden, rich morphological complexity, and mature computational infrastructure capable of handling gigapixel whole slide images (WSIs) and multimodal inputs. As a result, GU cancers, especially prostate, bladder, and renal, have become leading use cases for AI systems that support detection, characterization, grading-adjacent tasks, and quantitative assessment.[1–4]

This review addresses a fundamental translational question: which AI applications in GU pathology are ready for clinical deployment, which require further validation before routine use, and what frameworks can guide safe implementation decisions? We focus specifically on histopathology and cytology applications for prostate, bladder, renal, testicular, and penile specimens, examining evidence from detection/triage through quantification and risk stratification. We exclude radiology-only AI, liquid biopsy molecular assays without morphologic correlation, and pure genomic prediction models that do not incorporate pathology inputs.

The practical value of AI in GU pathology concentrates in three workflow phases: pre-sign-out (specimen intake, quality control, and case prioritization), during sign-out (region-of-interest identification, assistive review, and reportable quantification), and post-sign-out (quality assurance (QA), monitoring, and tumor board preparation). Two success factors recur across implementations: (1) robust pre-analytic quality control to prevent downstream failure (e.g., focus and tissue-coverage checks for WSIs; adequacy/cellularity constraints for cytology), and (2) interpretable, evidence-linked outputs that allow pathologists to verify findings and retain diagnostic control.

Real-world deployments show meaningful efficiency gains in selected settings, including faster review, streamlined workflows, and reduced immunohistochemistry (IHC) utilization, alongside improvements in standardization and consistency for well-bounded tasks.[5–8] At the same time, translation remains uneven across organs and entities. Domain shift (scanner and stain variability), label variability (especially in borderline lesions), and the long tail of rare GU subtypes lim-

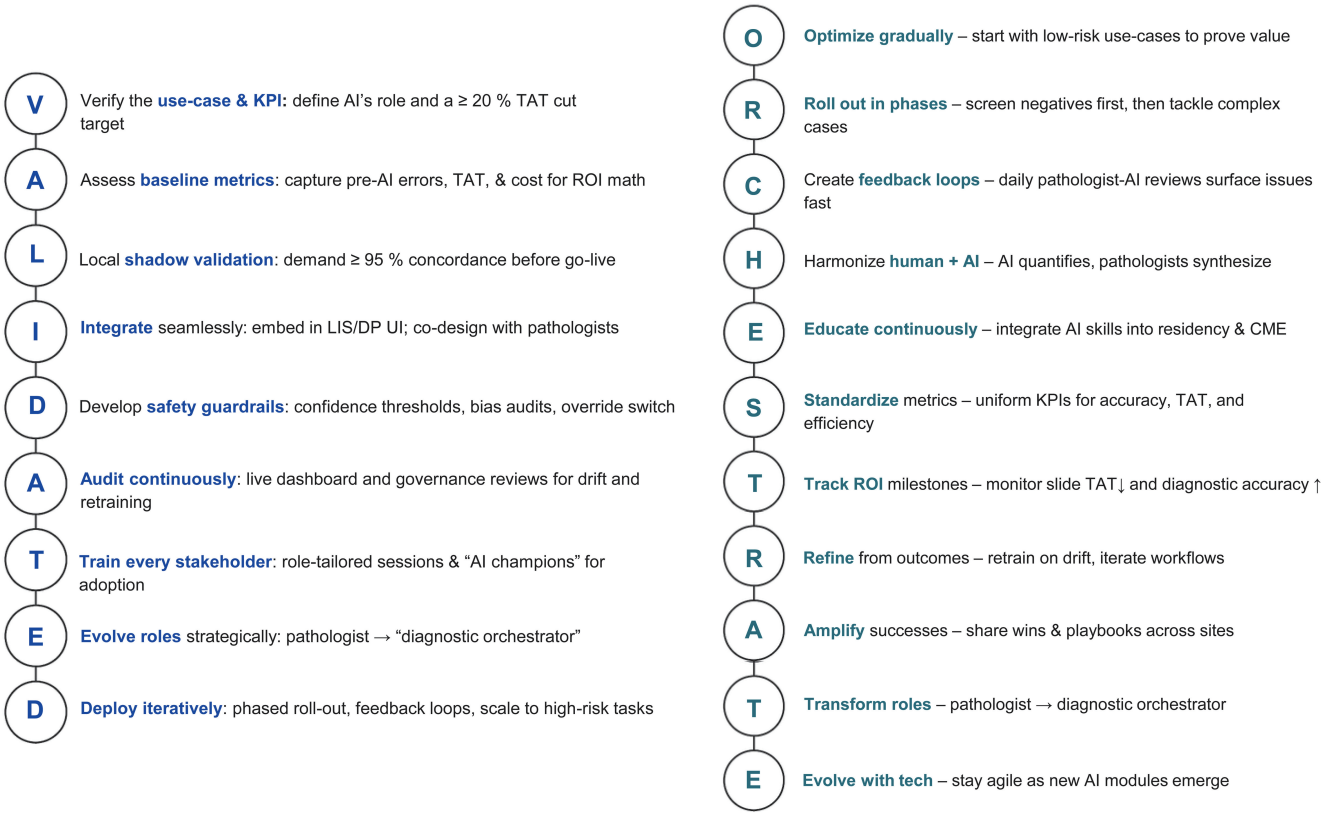

Fig. 1. The VALIDATED and ORCHESTRATE frameworks for pathology AI integration. This schematic illustrates the complementary relationship between the two implementation frameworks. VALIDATED (left panel) governs pre-deployment governance and safety oversight through nine sequential steps: Verify use case scope, Assess baseline metrics, Local shadow-mode validation, Integrate with systems, Develop guardrails, Audit continuously, Train stakeholders, Evolve roles, and Deploy with measured confidence. ORCHESTRATE (right panel) guides ongoing operational excellence through ten iterative components: Optimize workflow gradually, Roll out in phases, Create feedback loops, Harmonize human-AI collaboration, Educate continuously, Standardize metrics, Track return on investment, Refine based on outcomes, Amplify successes, and Transform with evolving technology. The frameworks function as a continuous cycle: VALIDATED ensures safe initial deployment, while ORCHESTRATE drives sustained operational improvement and adaptation. Together with TRI (assessment) and SURE-Path (safety requirements), these frameworks form a complete translational stack from evidence evaluation through implementation to long-term governance. AI, artificial intelligence; CME, continuing medical education; KPI, key performance indicator; LIS/DP, laboratory information system/digital pathology; ROI, return on investment; TAT, turnaround time; TRI, Translational Readiness Index; UI, user interface.

it generalizability and elevate safety risk, underscoring the need for clear intended use, conservative deployment guardrails, and rigorous multi-site validation.

To address these translation challenges, we propose a pragmatic evidence-to-operations stack: the Translational Readiness Index (TRI) to map maturity and prioritize next validation steps, the SURE-Path minimum safety bundle to define clinically defensible safeguards, and the VALIDATED/ORCHESTRATE pathway to operationalize deployment, monitoring, and governance. Together, these frameworks aim to strengthen clinical reliability and translational impact by linking scientific claims to auditable implementation decisions that can be reproduced across institutions.

Roadmap of this review: We first define the TRI and apply it to GU pathology tasks to create an organ-level readiness map, distinguishing what is deployment-ready from what remains validation-limited. We then examine cross-cutting constraints and enablers that shape generalizability, particularly for low-prevalence entities, including data limitations, opportunities offered by foundation and vision-language models (VLMs), and the evolving role of explainability. Finally, we translate these insights into operational guidance through workflow integration patterns, the SURE-Path safety bundle, and the VALIDATED/ORCHESTRATE implementation framework, providing a practical blueprint for safe adoption and continuous performance monitoring in real-world practice.

Core frameworks overview

Before examining organ-specific evidence, we introduce the three interconnected frameworks that structure this review. These frameworks operate in a progressive sequence: assessment, safety, and implementation (Fig. 1).

TRI: The TRI provides a structured rubric for assessing AI deployment readiness across six domains: evidence strength, regulatory maturity, external generalization, workflow integration, safety/explainability, and health economics. Each domain is scored 0–5, yielding a total score of 0–30. TRI scores stratify applications into four tiers: deployment-ready (≥22), pilotable with guardrails (16–21), emerging/validation-required (10–15), and preclinical concept (<10). The TRI serves as a diagnostic tool to identify which validation gaps must be addressed before clinical adoption.

SURE-Path minimum safety bundle: Once an application achieves sufficient TRI maturity for deployment consideration, SURE-Path defines the minimum safeguards required for clinically defensible use. The acronym captures five essential

Table 1. Translational Readiness Index across GU pathology tasks

Task (Setting)	Evid.	Reg.	Gen.	Flow.	Safety.	Econ.	Total	Category
Prostate biopsy: cancer detection/triage	5	5	4	4	4	4	26	Deployment-ready
Prostate biopsy: grading/quantification	4	4	4	4	4	3	23	Deployment-ready
Bladder cytology: AI prescreen	3	3	3	3	4	3	19	Pilotable
Bladder histology: invasion/grade assist	3	2	2	2	3	2	14	Emerging
Renal neoplasia: subtype/grade assist	3	1	2	2	3	2	13	Emerging
Lymph-node triage (GU): metastasis screen	3	1	3	3	4	2	16	Pilotable
Testicular tumors: TIL/LVI/GCNIS quant.	2	0	1	1	3	1	8	Preclinical/Emerging
Penile squamous lesions: diagnosis	1	0	1	1	2	1	6	Preclinical

AI, artificial intelligence; Econ., Health economics; Evid., Evidence strength; Flow., Workflow integration; GCNIS, germ cell neoplasia in situ; Gen., Generalization; GU, genitourinary; LVI, lymphovascular invasion; Reg., Regulatory maturity; Safety., Safety and explainability; TIL, tumor-infiltrating lymphocytes.

elements: Safety thresholds (pre-specified operating points and stop rules), Uncertainty and abstention (calibrated confidence with explicit deferral states), Reproducibility (external validation with site-stratified reporting), Evidence-linked explainability (region-grounded outputs with audit trails), and Path-of-use governance (workflow integration, training, documentation, and version control). SURE-Path translates TRI-identified readiness into operational safety requirements.

VALIDATED/ORCHESTRATE implementation pathway: For applications meeting both TRI thresholds and SURE-Path requirements, VALIDATED/ORCHESTRATE provides a structured implementation roadmap. VALIDATED governs pre-deployment activities: Verify use case scope, Assess baseline metrics, Local shadow-mode validation, Integrate with systems, Develop guardrails, Audit continuously, Train stakeholders, and Evolve roles strategically before final Deployment. ORCHESTRATE guides ongoing operations: Optimize workflow gradually, Roll out in phases, Create feedback loops, Harmonize human-AI collaboration, Educate continuously, Standardize metrics, Track return on investment, Refine based on outcomes, Amplify successes, and Transform roles with evolving Technology. Together, these frameworks form a coherent progression from evidence assessment through safe deployment to sustained operational excellence.

Synergistic value: The three frameworks address different but complementary questions. TRI asks "Is this application ready?" SURE-Path asks "What safeguards are required?" VALIDATED/ORCHESTRATE asks "How do we implement and sustain it?" An application with high TRI scores but missing SURE-Path elements should not be deployed; an application with all safety elements but without structured implementation governance risks operational failure. The frameworks function as a translational stack that laboratories can apply systematically to evaluate, deploy, and monitor AI tools in clinical practice.

Limitations of these frameworks: We acknowledge that these frameworks represent pragmatic synthesis rather than empirically validated instruments. TRI domain weights reflect expert judgment and published validation priorities rather than formal utility derivation. SURE-Path and VALIDATED/ORCHESTRATE have not been tested in controlled implementation studies. We present these frameworks as structured starting points for institutional adaptation rather than prescriptive standards and encourage prospective evaluation of their utility in guiding deployment decisions.

TRI

To provide a structured approach for evaluating AI deployment readiness across GU pathology applications, we introduce the TRI (Table 1). TRI scores each organ-task pair across six domains from 0 (nascent) to 5 (mature):

- *Evidence strength:* External, multi-site, multi-reader multi-case validation;
- *Regulatory maturity:* U.S. Food and Drug Administration (FDA) clearance, CE-IVDR certification, or equivalent;
- *External generalization:* Multi-scanner, multi-stain, out-of-distribution (OOD) resilience;
- *Workflow integration:* Laboratory information system (LIS) connectivity, prospective use, efficiency/IHC impact;
- *Safety and explainability:* Region-level grounding, calibrated uncertainty, abstention policies;
- *Health economics:* Return on investment evidence, cost-effectiveness analyses, or microsimulations;
- *TRI categories:*
  - ◊ ≥22: Ready for clinical deployment;
  - ◊ 16–21: Pilotable with guardrails;
  - ◊ 10–15: Emerging (requires significant validation);
  - ◊ <10: Preclinical concept;
- *Methodological limitations:* The TRI scoring rubric reflects synthesis of published validation frameworks, regulatory guidance, and implementation literature rather than empirical derivation from outcome data. Individual domain scores involve expert judgment and may vary across assessors. We did not perform formal inter-rater reliability testing. The equal weighting of domains (each 0–5) represents a pragmatic simplification; different clinical contexts may warrant differential weighting (e.g., prioritizing safety over health economics for high-risk applications). We encourage institutions to calibrate TRI assessments against their local validation experience and risk tolerance.

Prostate pathology

Domain-specific catalysts for AI development

Among GU subspecialties, prostate pathology is the most deployment-ready for AI because it combines high case volume, standardized pattern-based grading, and clinically consequential but visually focal findings (e.g., small malignant foci in core biopsies). In many Western laboratories, prostate specimens constitute a major fraction of routine GU histopathology workload, creating sustained operational pressure that favors tools designed for triage, localization, and standardized quantification.

Translation has been accelerated by the availability of large, multi-institutional annotated datasets that enable

Table 2. Commercial AI tools for prostate cancer detection/grading (approvals vary by region/version)

Product	Company	Detection	Grading	Quantification	Approvals
Paige Prostate Detect/Grade & Quantify	Paige AI (USA)	✓	✓	✓	FDA (Detect); CE-IVD
Galen Prostate	Ibex (Israel)	✓	✓	✓	CE-IVD
Aiforia Prostate	Aiforia (Finland)	✓	✓	✓	CE-IVD
DeepDx	Deep Bio (South Korea)	✓	✓	✓	CE-IVD; MFDS
Inify Prostate	Inify Laboratories (Sweden)	✓	✓	–	CE-IVD
HALO Prostate	Indica Labs (USA)	✓	✓	–	CE-IVD

CE-IVD, Conformité Européenne – In Vitro Diagnostic; FDA, Food and Drug Administration; MFDS, Ministry of Food and Drug Safety.

stress testing across sites, scanners, and staining variation. The Prostate cANcer graDe Assessment (hereinafter referred to as PANDA) challenge (10,616 WSIs spanning scanner platforms and staining protocols) exemplifies prostate AI development under real-world variability and has become a widely cited template for multi-site validation.[9] Foundational work predating 2023, including early deep learning grading studies and the development of multiple-instance learning approaches, established the computational foundations upon which current commercial systems are built.[10–12]

Currently available commercial solutions

Commercialization in prostate pathology is more mature than in other GU domains, with regulatory traction in both the United States and Europe. In the United States, FDA clearance of Paige Prostate Detect (2021) for prostate cancer detection in core biopsies catalyzed broader clinical adoption, while multiple CE-marked solutions are used or piloted across European networks (Table 2).[13,14]

Clinical performance and validation evidence

Across studies, deep learning models achieve high discrimination for benign versus malignant tissue, with many reports showing area under the curve (AUC) values exceeding 0.95 and multi-center validation cohorts reporting AUC ≥ 0.99 for detection tasks.[9,15–17] However, high headline performance metrics should be interpreted in the context of clinically meaningful discordance and task framing.

Even in benchmark settings demonstrating strong expert-level concordance (e.g., quadratically weighted Cohen's κ approximately 0.862–0.868), clinically meaningful disagreements remain non-trivial (reported in approximately 13–14% of PANDA cases).[9,18] In practice, many deployed systems are optimized for high-sensitivity workflows (e.g., negative-case exclusion, triage, region-of-interest highlighting), consistent with sensitivity values often reported in the 97–98% range alongside more modest specificity (approximately 75–84%). This operating point supports safe throughput gains when paired with explicit guardrails, but it also increases the likelihood of benign tissue being flagged as suspicious.

Workflow efficiency and economic impact

Prostate AI is increasingly integrated in structured, multi-site implementations (e.g., regional laboratory networks and consortium-style frameworks). Real-world reports suggest meaningful time savings when AI is used to automate or pre-assemble routine interpretive steps and quantitative outputs (tumor area/burden, Grade Group support, glandular architecture features, pattern quantitation, and related morphometric summaries). Single-center experience (e.g., Ohio State University Wexner Medical Center) has reported approximately 20–25% pathologist time savings in AI-assisted workflows through automation and synthesis of slide-derived parameters.[19]

Beyond single-center reports, studies using FDA-cleared and CE-marked systems describe efficiency gains as high as approximately 65% in AI-assisted pre-screening or concurrent reading paradigms.[5–8,20] Modeling studies (e.g., Swedish microsimulation) have projected substantial reductions in manual review burden (e.g., approximately 80% fewer cores requiring full review) without compromising detection of clinically significant cancer.[21,22] Reported downstream effects include reduced turnaround time, fewer ancillary immunohistochemical studies, and reduced secondary consultation burden.

Economic evidence limitations: We note that much of the published health-economic evidence for prostate AI derives from modeling studies and microsimulations rather than prospective cost-effectiveness trials. Real-world economic validation across diverse practice settings remains limited, and published efficiency estimates may not generalize to laboratories with different case volumes, staffing models, or reimbursement environments.

Limitations, edge cases, and regulatory considerations

Clinically important failure modes remain concentrated in borderline/atypical proliferations (e.g., atypical glands, limited foci); stain and pre-analytic variability; rare morphologic subtypes (e.g., foamy gland, pseudohyperplastic patterns) and unusual architecture; and artifacts that mimic tumor morphology and can produce false confidence.

Generalizability remains a central translational challenge: performance can drop substantially across institutions, scanners, or staining protocols without explicit domain-adaptation strategies and monitoring. Practical implementation also introduces medico-legal and operational issues, including documentation expectations for AI-assisted reads, reimbursement uncertainty, and liability concerns related to missed diagnoses, reinforcing the need for defined "assistive use" boundaries and auditable evidence outputs.

Post-market surveillance: While regulatory clearances provide important validation milestones, post-market performance monitoring remains essential. The published literature includes limited systematic reporting of post-deployment failures, algorithm drift, or performance degradation over time. Institutions implementing cleared systems should establish prospective monitoring for concordance patterns, abstention rates, and site-specific failure modes that may not have been captured in pivotal trials.

Clinical value proposition

The strongest near-term value proposition for prostate AI is

not autonomous diagnosis, but standardization of grading-adjacent decisions and support for throughput. In practice, these systems can improve reproducibility in grading-related tasks, reduce interobserver variability, and produce consistent quantitative outputs (e.g., pattern percentages, tumor burden) that can be reviewed and edited by the pathologist.

Prospective performance monitoring remains critical to verify that AI prioritization and highlighting do not preferentially miss rare patterns or small, high-grade foci, and that efficiency gains translate into clinically meaningful endpoints, including turnaround time, ancillary test utilization, and more uniform risk stratification.

TRI-aligned summary (prostate)

Evidence and validation status: Prostate biopsy assistance represents the most mature GU use case (TRI approximately 26/30), supported by multi-site validation, prospective reader-in-the-loop studies, and regulatory-cleared commercial solutions for bounded tasks.

Workflow integration and QA hooks: The most defensible deployments center on worklist triage, region-of-interest localization, and standardized quantitative outputs embedded in sign-out, with post-deployment monitoring of concordance/discordance patterns, abstention/deferral rates, and drift indicators tied to scanner/stain variation.

Key safety/abstention pitfalls: Silent failure risk concentrates in OOD slides (artifact, unusual stains/scanners), rare variants, and mimics (e.g., inflammation/atrophy), reinforcing the need for explicit intended use, calibrated uncertainty, and systematic deferral/audit pathways.

What would move TRI up: Broader prospective, multi-site deployments with standardized endpoints, transparent failure-mode reporting, and longitudinal monitoring linking workflow impact to clinically meaningful quality metrics.

Bladder pathology

Current landscape

Bladder AI is advancing rapidly, but routine histopathology deployment lags behind prostate. Development spans histopathology, urine cytology, and cystoscopy, targeting diagnostic consistency, efficiency, and risk stratification. Across reported cohorts, AI systems commonly achieve greater than 80% accuracy for tasks such as tumor detection, grade prediction, and compartment segmentation (urothelium, stroma, muscle).[23] Beyond morphology, some models aim to predict recurrence/progression, response to Bacillus Calmette-Guerin, and higher-risk trajectories.

Regulatory status and commercial platforms

Regulatory traction remains modest compared with prostate, but is increasing. The TOBY Test received FDA Breakthrough Device Designation (2025), with pivotal validation underway. Other tools (e.g., VisioCyt, Menarini/Nucleix, Techcyte/CytoBay, and URO17 collaborations) are CE-marked or in late-stage evaluation.[24,25] Meta-analyses commonly report that AI assistance can raise sensitivity and reduce subjective variability, but interpretability, domain-shift vulnerability, and evolving regulatory expectations continue to slow broad deployment.

Cytology applications and the Paris System

Urine cytology is currently the most translationally advanced bladder domain for AI, in part because it can be paired with structured quality gates and standardized pre-analytic controls. The Paris System for Reporting Urinary Cytology provides a standardized diagnostic framework that aligns well with AI development, offering reproducible category definitions (negative, atypical urothelial cells, suspicious, and positive for high-grade urothelial carcinoma) that can serve as training labels and evaluation endpoints.[26–28]

Practical workflow use cases: AI-assisted urine cytology is most defensibly positioned for prescreening and triage workflows. In a prescreening model, AI evaluates slides first and routes clearly negative specimens for expedited verification review while flagging atypical or suspicious cases for detailed pathologist assessment. This approach can reduce time spent on high-volume negative cases while concentrating expert attention on diagnostically challenging specimens. Multi-center evaluations including VISIOCYT support feasibility for such workflows.[24] Prospective studies emphasize the importance of cellularity optimization (e.g., reported thresholds such as ≥2,644 urothelial cells per slide) to ensure robust downstream AI performance.[29]

Foundational work by Vaickus *et al*.[30] demonstrated automated Paris System categorization, establishing proof of concept for AI-assisted urine cytology. Subsequent large-scale validation by Levy *et al*.[31] with the AutoParis-X system across multiple institutions provided evidence for clinical-grade performance. Additional work by Levy and colleagues on longitudinal recurrence markers suggests potential for AI to inform surveillance strategies beyond single-specimen diagnosis.[32]

Histopathology applications

For transurethral resection of bladder tumor (TURBT) specimens, AI development focuses on grade assessment and invasion detection, often implemented via compartment segmentation (urothelium, lamina propria, muscularis propria). Given the artifact-prone nature of TURBT, clinically defensible systems increasingly incorporate artifact awareness and explicit abstention rules to prevent false-positive invasion calls driven by cautery/crush artifact or dense inflammation.

Translation is limited by persistent interobserver variability in training labels and underrepresentation of rare variants (e.g., plasmacytoid, micropapillary), which can reduce sensitivity and increase bias risk.

Emerging capabilities

Emerging bladder applications include histology-based molecular surrogate prediction (e.g., FGFR3 status) to prioritize cases for confirmatory testing and risk stratification models for Bacillus Calmette-Guerin-treated non-muscle-invasive disease in prospective multi-center evaluation.[33–35] In tumor board settings, quantitative outputs can improve communication and decision-making. However, durable adoption will depend on transparent performance reporting, continuous validation, and clearly explainable evidence outputs aligned with clinical workflow.

TRI-aligned summary (bladder)

Evidence and validation status: Bladder cytology is the most pilotable non-prostate GU domain (TRI approximately 19/30), supported by multi-center efforts and commercial offerings for prescreening/triage aligned with Paris System categories. Bladder histopathology remains emerging (TRI approximately 12/30), with translation limited by label variability and heterogeneous specimen quality.

Workflow integration and QA hooks: The strongest near-term fit is cytology prescreening that triages clearly negative material while deferring uncertain or high-risk cases, coupled with explicit adequacy thresholds, audit logs, and periodic review of discordant cases and missed high-grade lesions.

Key safety/abstention pitfalls: Risk concentrates in low-prevalence/high-consequence findings, borderline/reactive atypia, specimen heterogeneity (preparation differences), and domain shift; conservative deferral rules and evidence-linked visualization are essential.

What would move TRI up: Larger prospective, multi-site reader-in-the-loop studies with standardized labeling for atypia categories, explicit failure-mode analysis, and harmonized adequacy/quality control (QC) criteria across preparation types.

Renal pathology

Current applications

In renal oncology, renal cell carcinoma (RCC) subtyping is among the most mature applications, with models distinguishing clear cell, papillary, and chromophobe RCC with reported AUC values exceeding 0.90 across validation cohorts.[36–39] Clear cell RCC grading performance has been reported in the AUC approximately 0.89–0.96 range, with some studies demonstrating added prognostic value when combined with clinical parameters.

Renal AI development is also expanding into quantification and compartment-level segmentation (e.g., fibrosis, tubular atrophy, glomerulosclerosis), supporting standardization in research reporting and mixed tumor-parenchyma resections. Multi-institutional studies using The Cancer Genome Atlas (TCGA) images have reported strong performance distinguishing tumor/normal/non-neoplastic tissue (e.g., F1 approximately 0.88; AUC approximately 0.97), despite notable histologic diversity.[40–42]

Workflow integration

Current renal tools are often positioned as adjuncts and, in some settings, remain research-use-oriented. Translationally relevant integration points include pre-sign-out triage (region-of-interest preview generation; subtle component flagging across blocks); during sign-out support (exportable subtype/grade evidence with region-level grounding); and post-sign-out monitoring (site-specific performance tracking, rare-variant sensitivity surveillance).

Emerging capabilities and limitations

Renal AI is moving toward pathology-genomic integration, including prediction of genomic alterations (e.g., VHL pathway features, tumor mutational burden) and radiology-pathology fusion models.[43,44] Early slide-based inference of VHL pathway alterations (AUC approximately 0.75–0.85) illustrates the potential direction, but clinical translation is constrained by limited multi-center datasets, required validation rigor, and the complexity of integrating multimodal inputs (brightfield/IHC/IF/EM and clinico-radiologic context). Renal disease heterogeneity creates a high bar for generalization: subtle, focal lesions require expert labeling across glomeruli, tubules, interstitium, and vessels, while major public repositories often lack granular labels. Rare RCC variants and mixed-histology specimens remain underrepresented, emphasizing the need for multi-institutional data sharing and rigorous external validation.

TRI-aligned summary (renal)

Evidence and validation status: Renal neoplasia remains emerging in translational maturity (TRI approximately 14/30), reflecting substantial morphologic heterogeneity, rare subtype frequency, and variability in grading/staging-adjacent labels across institutions.

Workflow integration and QA hooks: Near-term value is most defensible as assistive review and structured quantification in narrowly scoped tasks, paired with strong case-selection constraints, uncertainty-aware deferral, and site-specific shadow validation before clinical use.

Key safety/abstention pitfalls: Silent failure risk is elevated by rare variants, mixed patterns within tumors, limited representation of unusual preparations, and domain shift; robust abstention behavior and subtype-aware performance reporting are critical.

What would move TRI up: Consortium-level datasets with enriched rare subtypes, standardized annotation protocols, and multi-site external validation (ideally prospective) demonstrating stable performance across scanners/stains and diverse practice settings.

Low-prevalence GU domains (testis and penis pathology)

Testicular and penile pathology share fundamental translational constraints: low case volumes, morphologic heterogeneity, and minimal dedicated data resources. These shared challenges justify consolidated discussion while preserving attention to domain-specific considerations.

Shared challenges

Data scarcity: Both domains are constrained by limited sample counts relative to prostate and bladder pathology. Most reported datasets contain fewer than 200 digitized slides, insufficient for robust training and validation of deep learning models across the spectrum of clinically important entities. This scarcity is compounded by institutional heterogeneity, scanner diversity, and the rarity of key diagnostic subtypes within already small cohorts.

Morphologic heterogeneity: Both testicular germ cell tumors (TGCTs) and penile squamous neoplasms exhibit substantial morphologic diversity. Testicular tumors range from seminoma through embryonal carcinoma, yolk sac tumor, choriocarcinoma, and teratoma, with mixed patterns common. Penile lesions span the spectrum from condyloma through differentiated squamous cell carcinoma variants, with diagnostic overlap with other squamous lesions presenting additional challenges. This heterogeneity elevates the risk of spectrum bias in model training and silent failure for rare entities.

Label ambiguity: In both domains, interobserver variability for borderline lesions and grading-adjacent decisions creates label noise that can propagate through training datasets. For testicular tumors, distinction between subtypes and quantification of mixed components involves judgment that varies across pathologists. For penile lesions, distinguishing precursor lesions from invasive carcinoma and differentiating penile squamous lesions from those arising at other sites requires contextual information not always available from morphology alone.

Testicular pathology: Domain-specific considerations

Testicular AI remains early stage and is dominated by screening/quantification tasks in small cohorts. Initial models trained on limited datasets have reported high performance for TGCT versus benign classification (e.g., F1 approximately 0.92) and variable subtype true-positive rates (e.g., 75–95%).[45] Tumor-infiltrating lymphocyte mapping and germ cell neoplasia in situ quantification are emerging as structured quantification targets, while lymphovascular invasion (LVI) detection shows more modest precision and typically

requires conservative thresholds, mandatory region-level evidence, and liberal abstention.[46]

Multimodal diagnostic dependence: A unique challenge for testicular pathology is that TGCT workups often depend on clinical features, imaging, and serum tumor markers (alpha-fetoprotein, human chorionic gonadotropin, lactate dehydrogenase) in addition to histomorphology. Slide-only models risk misclassification when decisive context is multimodal, limiting the scope of purely histology-based AI.

Translation pathway: Near-term clinical roles will likely remain supportive (screening adjuncts, small-focus detection, standardized quantification). The most practical path to translation is transfer learning from high-volume GU tissues with fine-tuning on pooled TGCT datasets, supported by multi-institutional collaboration.

Penile pathology: Domain-specific considerations

At present, no established commercial or mature research-stage AI applications are widely reported for penile pathology. The primary challenge is distinguishing penile squamous neoplasms from other squamous lesions, including cutaneous and mucosal squamous cell carcinomas arising at other sites. Human papillomavirus status adds another dimension relevant to both pathogenesis and prognosis that is not directly visible on routine H&E sections.

Diagnostic complexity: Penile carcinoma includes multiple histologic variants (usual type, basaloid, warty, verrucous, mixed) with prognostic implications. Accurate subtyping and grading require experienced subspecialty review, and the lack of large annotated datasets makes supervised model development particularly challenging.

Translation pathway: Progress will likely require multi-institutional consortia and transfer learning from cutaneous and mucosal squamous neoplasia models, paired with careful validation for variant-rich, low-prevalence entities. Human papillomavirus-prediction models from histology represent a potential entry point, given analogous work in cervical and oropharyngeal squamous carcinomas.

Shared mitigation strategies

For both testicular and penile pathology, several strategies may accelerate translation despite data constraints:

- *Consortium-level data aggregation:* Multi-institutional collaboration to pool cases, standardize annotations, and create enriched validation sets for rare subtypes.
- *Transfer learning:* Leveraging pre-trained models from higher-volume tissues (prostate, bladder, or general pathology foundation models) with fine-tuning on domain-specific curated datasets.
- *Conservative intended use:* Framing AI applications as decision support or narrowly scoped assistive tasks (e.g., flagging potential LVI for pathologist review) rather than autonomous classification.
- *Strict deferral rules:* Implementing low thresholds for abstention and mandatory routing to expert review for any uncertain or low-confidence outputs.
- *Shadow-mode validation:* Running AI outputs in parallel with routine diagnosis before any clinical integration, with prospective tracking of concordance and failure patterns.

TRI-aligned summary (low-prevalence domains)

Evidence and validation status: Both testicular (TRI approximately 8/30) and penile (TRI approximately 6/30) pathology remain low-readiness, primarily due to low prevalence, limited curated datasets, and high diversity of clinically important but uncommon entities.

Workflow integration and QA hooks: Early translation is most appropriate as research-grade decision support (education, retrieval of similar cases, structured checklists) or narrowly scoped assistive tasks validated locally in shadow mode with strict deferral rules.

Key safety/abstention pitfalls: Disproportionate risk arises from rare/high-impact diagnoses, spectrum bias, and limited external generalizability; conservative intended use and mandatory deferral for low-confidence outputs are essential.

What would move TRI up: Multi-institutional aggregation with enriched rare diagnoses, standardized ground-truth adjudication, and external validation designed specifically to measure performance on the long tail rather than only common entities.

Cross-cutting data limitations and dataset biases

Having mapped organ-level readiness, we next consider the constraints and potential enablers that determine whether a model survives translation. The primary barrier to translation for low-prevalence entities is data scarcity, which presents as limited sample counts, narrow institutional and scanner diversity, under-representation of rare variants, and labels that are insufficiently granular for the clinical question being modeled.

Dataset imbalances and AI biases: A critical limitation across GU pathology AI is the systematic over-representation of common entities in training datasets. Prostate AI models are predominantly trained on cases with identifiable carcinoma, with less representation of mimics, borderline lesions, and rare variants. Bladder datasets skew toward high-grade urothelial carcinoma, with under-representation of low-grade tumors, unusual variants, and non-neoplastic mimics. Renal AI faces the greatest challenge, with clear cell RCC dominating datasets, while chromophobe, papillary type 2, translocation-associated, and other rare subtypes remain sparse.

These imbalances translate directly to performance disparities: models may achieve headline AUC values exceeding 0.95 for common entities while failing silently on rare but clinically important subtypes. For clinically consequential low-prevalence findings (e.g., micropapillary bladder carcinoma, collecting duct RCC, LVI in testicular tumors), published performance metrics often derive from samples too small for reliable estimation.

Mitigation strategies: Defensible approaches to address dataset bias include stratified performance reporting with explicit metrics for rare subtypes; enriched validation sets over-sampling low-prevalence entities; explicit intended-use boundaries acknowledging where model performance is uncertain; and abstention policies triggered by morphologic features associated with rare entities. Importantly, aggregate performance metrics should not be used to imply generalizability across the full diagnostic spectrum without subtype-specific evidence.

These limitations are compounded by routine archival practices, where slides are often categorized broadly at the case level (e.g., "tumor present") without region- or feature-level ground truth. For tasks requiring focal evidence (e.g., LVI, germ cell neoplasia in situ at margins), weak labels can introduce label leakage, inflate metrics, and bias training toward spurious correlates rather than the lesion of interest.

New frontiers: Foundation models and multimodal integration

Foundation models: From narrow AI to generalist AI

Most deployed pathology AI remains task-specific (single or-

gan, single endpoint). Foundation models represent a shift toward large-scale, broadly trained image or multimodal encoders that can be more readily adapted to multiple downstream GU tasks with less labeled data.[47,48] In early reports, models such as GigaPath (trained on 1.3 billion image tiles from more than 171,000 WSIs), UNI, and multimodal approaches such as BioMedCLIP demonstrate that scale and diversity can support strong performance across multiple benchmarks, sometimes approaching specialized systems without extensive task-specific training.[49,50]

The Virchow foundation model further illustrates the potential of large in-domain self-supervised pretraining. Trained on more than 1.5 million H&E WSIs, Virchow-derived embeddings supported pan-cancer detection and biomarker prediction, with reported robustness across external institution slides and improved performance on some rare histologic variants, suggesting that broad pretraining can reduce (but not eliminate) the dependency on large labeled datasets for every new task.[47]

Translational advantages and practical limitations

For GU pathology, the practical appeal of foundation models is clear: transfer learning for rare entities, improved tolerance to variation in staining and scanners, and accelerated development of niche classifiers (e.g., difficult renal oncocytic neoplasms) via fine-tuning on smaller curated cohorts.

However, clinical translation requires careful framing. High benchmark performance can obscure limitations that emerge under targeted scrutiny, especially for rare cancers, mixed histologies, and non-neoplastic pathology, unless explicitly represented and evaluated. In addition, many foundation models remain operationally limited in interpretability at the decision level. For clinical deployment, this shifts the focus from philosophical explainability to evidence-grounded outputs (region localization, quantification, calibrated uncertainty), rigorous external validation, and clear governance for model updates and drift monitoring. Regulatory pathways for rapidly evolving generalist models remain an active area of uncertainty, reinforcing the need for conservative deployment boundaries and auditable evidence trails.

Multimodal AI

GU oncology is inherently multimodal: pathology, imaging, molecular testing, and clinical context collectively drive management. Multimodal AI seeks to integrate these inputs to produce more reliable risk stratification than any single modality alone.[51,52] Early work in prostate cancer suggests improved recurrence or response prediction when combining digitized histology with MRI and genomic markers, highlighting the potential value of integrated models.

Conceptually, multimodal integration may support prostate cancer (coupling histologic grade and quantitative tumor burden with MRI risk scores and genomic features to refine prognostic estimates); bladder cancer (linking cystoscopic imaging with histopathology to support real-time assessment of grade and invasion risk); and renal lesions (improving risk stratification for imaging-ambiguous cystic lesions using biopsy histology plus clinical variables).

For translation, key requirements include robust handling of missing modalities, standardized data harmonization, and evaluation against clinically meaningful endpoints. Multimodal outputs are most useful when presented as decision support (with confidence and provenance), not as autonomous summaries.

VLMs

VLMs extend multimodal learning by pairing histologic images with text, enabling models to learn associations between microscopic patterns and the diagnostic lexicon (e.g., "cribriform," "hobnail," "Schiller-Duval bodies"). CONCH, trained on more than 1.17 million image-caption pairs, illustrates how text grounding can enable strong performance on certain zero-shot or low-shot tasks, including GU-relevant classifications.[53]

For GU pathology, where many entities are rare but richly described, VLMs may be particularly valuable for education and decision support (interactive morphology search, retrieval of prototypical regions, and literature-aligned pattern descriptions) and low-data settings (aiding differential diagnosis and triage when examples are sparse but textual descriptors are abundant).

Hallucination risk: A major limitation of VLMs is the risk of plausible but incorrect language output ("hallucinations"), including false statements about findings such as LVI, fabricated diagnostic criteria, or invented references.[54] In clinical settings, hallucinated outputs could lead to inappropriate management decisions if not recognized. Clinically safe use therefore requires strict guardrails: expert oversight for all VLM outputs, retrieval-augmented approaches that ground outputs in verifiable sources, evidence-linked region visualization, and explicit labeling that VLM-generated text requires pathologist verification before clinical action.

Emerging interactive assistants (e.g., PathChat) suggest a future in which pathologists can query foundation or vision-language systems conversationally during sign-out to refine differentials and ancillary testing strategies, provided outputs remain bounded, auditable, and anchored to slide evidence.

Explainability and clinical assurance in genitourinary pathology

Why explainability matters in clinical GU pathology

In diagnostic pathology, "explainability" should be defined pragmatically as the set of user-facing evidence and governance artifacts that allow a pathologist to verify, contextualize, and safely act on an algorithm's output. For WSI applications, explainability is not an interpretability philosophy; it is a clinical safety requirement that supports (i) verification of region-level evidence, (ii) reproducible quantification aligned with routine reporting, (iii) calibrated uncertainty with clear "do-not-trust" behavior, and (iv) traceability for QA, discordance review, and regulatory audit.

Across GU tasks, the most clinically meaningful explainability can be organized into three core "questions" a system must answer for the pathologist: Where did the model look? (region-level grounding); What did it measure, and in what units? (feature-level quantification); and How confident is it, and when will it abstain? (uncertainty and abstention).

Clinically relevant explainability outputs

In day-to-day GU sign-out, explainability is most useful when it produces reviewable evidence at the same granularity as the diagnostic act (e.g., small cribriform foci, intraductal carcinoma, carcinoma in situ, subtle LVI) and when outputs map to reportable quantities (e.g., tumor percentage, Gleason pattern percentage, linear extent, mitotic counts, tumor-infiltrating lymphocyte density).

Limitations of current methods: Importantly, "explanations" that are visually compelling but unstable or not demonstrably linked to model decision pathways should be treated as supplementary and not relied upon as sole justification for clinical decisions. Attention maps, in particular, have known limitations: they may highlight artifacts, back-

ground tissue, or regions unrelated to the diagnostic feature; they can be unstable across minor input perturbations; and their relationship to model predictions is often indirect. For artifact-heavy specimens (common in TURBT and biopsy material), attention-based explanations may be misleading rather than informative.

Foundation-model and vision-language era: Same safety requirements

Foundation models and VLMs introduce additional explanation modalities (e.g., prototype retrieval and concept scoring). These can be valuable for education, rare entity support, and second-pass verification in GU pathology, but they do not replace region-grounded evidence. For clinical use, any retrieved "similar cases" or named "concept scores" (e.g., "cribriform") should be treated as hypothesis generators that must be anchored to explicit slide regions and governed by the same abstention and audit requirements as conventional models.

How explainability changes daily practice

When deployed with clinical guardrails, explainability shifts AI from a "black box" to a reviewable assistant: faster initial triage and localization (region-grounded overlays and patch galleries accelerate identification of subtle or high-impact foci); more reproducible reporting (standardized quantification reduces interobserver variability and manual transcription error); safer adoption (calibrated abstention prevents over-reliance by routing uncertain or OOD cases to fully manual review); and improved QA and trust (audit trails and drift monitoring make discordance review feasible and support continuous improvement without compromising patient safety).

Workflow integration across the diagnostic cycle

Clinical value from AI in GU pathology depends less on any single model and more on how algorithm outputs are embedded into routine diagnostic operations. Across institutions, successful implementations typically align AI functions with three phases of the diagnostic cycle: pre-sign-out, during sign-out, and post-sign-out, with explicit safety gates and performance monitoring.

Pre-sign-out: Quality gates and worklist triage

Pre-sign-out AI functions are most defensible when framed as quality control and prioritization rather than diagnosis. Common pre-review capabilities include automated slide/image QC to identify focus defects, scanning artifacts, inadequate tissue coverage, and other conditions that can invalidate downstream inference; case prioritization using probabilistic ranking to support operational triage (e.g., routing "higher suspicion" cases for earlier review while batching low-suspicion cases for efficient verification); and region-of-interest pre-localization, generating thumbnails or candidate regions linked to the digital worklist to reduce manual navigation burden.

For clinical deployment, these pre-review functions should be coupled to hard-stop criteria (e.g., QC fail → no AI inference; OOD detected → manual review only) and continuously audited for site-specific drift related to scanner, stain, or pre-analytic variation.

During sign-out: Assistive review and reportable quantification

During sign-out, AI is most clinically useful when it provides reviewable, region-grounded evidence and report-aligned quantitative outputs, not when it attempts to replace diagnostic judgment. Typical assistive outputs include heatmap overlays and ranked patch galleries that enable rapid verification of why a case is flagged and where attention should be focused; structured quantification aligned with routine reporting (e.g., tumor extent metrics, pattern proportions, linear measurements) presented as editable suggestions under pathologist control; and calibrated confidence and abstention behavior, explicitly signaling when outputs should not be used (e.g., artifact-heavy regions, atypical morphologies, inadequate quality, or OOD inputs).

Interfaces should prevent automation bias by making uncertainty visible and requiring active verification for any AI-suggested change that could alter grade, stage-relevant features, or management.

Post-sign-out: Assurance, monitoring, and continuous improvement

Post-sign-out functions are central to converting assistance into assurance. The three most translationally meaningful capabilities include concordance and discordance tracking between AI outputs and finalized diagnoses, stratified by institution, scanner, stain batch, and specimen type to detect performance drift; operational QA dashboards capturing error patterns, abstention rates, and QC failure frequencies as leading indicators of changing performance; and structured data reuse (where appropriate) to streamline downstream tasks (e.g., tumor board preparation and synoptic summaries) while maintaining clear provenance and audit trails.

Continuous improvement should be governed by prespecified triggers (e.g., drift thresholds, rising abstention rates, recurring failure modes) that mandate revalidation, retraining, or temporary suspension of AI outputs.

Ethical considerations

The deployment of AI in GU pathology raises important ethical considerations that extend beyond technical performance metrics. Responsible implementation requires attention to data governance, equity, and medico-legal frameworks.

Data privacy and security: AI development requires large annotated datasets, raising questions about patient consent, data de-identification, and cross-institutional data sharing. While pathology images are generally considered low risk for re-identification, the aggregation of morphologic data with clinical outcomes creates potential privacy concerns. Institutions deploying AI tools should ensure compliance with applicable regulations (HIPAA in the United States, GDPR in Europe) and establish clear data governance frameworks for both model development and ongoing performance monitoring.

Equity in AI access: Current AI development is concentrated in well-resourced academic centers and high-income countries, potentially widening disparities in diagnostic quality. Laboratories without digital pathology infrastructure, high-performance computing, or informatics expertise may be unable to benefit from AI advances. Equitable translation requires attention to implementation costs, training requirements, and interoperability standards that enable broader adoption. Foundation models and cloud-based deployment may partially address infrastructure barriers, but connectivity, cost, and data sovereignty concerns remain.

Medico-legal implications: AI-assisted diagnosis creates new questions regarding professional responsibility and liability. When AI contributes to a diagnostic error, the allocation of responsibility among the pathologist, the AI devel-

oper, and the deploying institution remains legally unsettled in most jurisdictions. Defensive practices may include explicit documentation of AI as "assistive" rather than autonomous; clear audit trails showing pathologist review of AI outputs; defined protocols for AI disagreement with pathologist interpretation; and informed consent considerations when AI plays a substantive role in diagnosis.

Algorithmic bias and fairness: AI models trained on non-representative populations may perform differently across demographic groups. While pathology AI is less directly affected by skin tone biases that impact dermatologic and radiologic AI, institutional and geographic biases in training data can affect generalizability. Performance should be monitored across available demographic strata, and systematic disparities should trigger investigation and mitigation.

From assistance to assurance: The SURE-Path minimum safety bundle

To support clinically defensible adoption, we propose a five-element "minimum safety bundle" (SURE-Path) that operationalizes trustworthy AI behavior in routine GU sign-out. The goal is not maximal interpretability but pre-specified performance boundaries, auditable evidence, and monitored reliability. We emphasize that SURE-Path represents a pragmatic synthesis of existing guidance rather than an empirically validated standard; prospective evaluation in diverse practice settings is needed.

S - Safety thresholds: Define intended use (triage vs. assistive diagnosis vs. quantification) and pre-specify operating points aligned to local risk tolerance (e.g., sensitivity-prioritized triage). Include explicit stop rules for conditions in which AI outputs must not be used (e.g., QC failure, OOD detection).

U - Uncertainty and abstention: Implement calibrated confidence (including conformal or other uncertainty approaches when feasible) with explicit abstain states that automatically route cases to full manual review. Track abstention as a monitored metric rather than a nuisance variable.

R - Reproducibility: Require external validation and site-stratified performance reporting (scanner, stain, specimen type, case mix). Establish periodic revalidation and "stress testing" with artifacts and rare histologies by design rather than relying on convenience cohorts.

E - Evidence-linked explainability: Provide region-grounded evidence (e.g., patch ranking with click-through WSI context) and report-aligned quantification with clear units and provenance. Explanatory outputs should be treated as clinical evidence only when they are stable, reviewable, and logged.

Path - Path-of-use governance: Operationalize integration: LIS/digital workflow embedding, user training, documentation practices for AI-assisted review, incident logging, and versioned audit trails (model version, thresholds, QC state) for every AI-assisted case. Governance should specify ownership (clinical, operational, informatics) and change control for updates.

Together, these elements shift AI from a "helpful output" to a monitored, auditable clinical tool, positioning performance claims within an implementation framework that is testable, reviewable, and aligned with routine pathology QA.

Implementation strategy: The VALIDATED and ORCHESTRATE frameworks

Implementing AI in pathology demands structured, iterative planning, validation, and governance, guided by regulatory bodies (CAP, WHO, FDA) and validated through systematic evaluation and stakeholder engagement. To ensure both technical success and cultural adoption, positioning pathology departments for sustainable AI integration, we propose two complementary frameworks: VALIDATED for governance and safety oversight, and ORCHESTRATE for day-to-day operations (Fig. 1).[55]

Framework limitations: We acknowledge that VALIDATED and ORCHESTRATE have not been tested in controlled implementation studies. These frameworks synthesize published implementation guidance, regulatory recommendations, and expert consensus rather than empirical validation from deployment outcomes. We present them as structured starting points that institutions should adapt to local contexts, workflows, and governance structures.

The VALIDATED framework

V - Verify use case and define scope: Clearly delineate AI's intended role (primary diagnostics, secondary quality control, decision support). Establish measurable success criteria upfront.

A - Assess baseline performance metrics: Document current performance metrics, including turnaround times, error rates, and resource utilization.

L - Local validation in shadow mode: Perform rigorous, local parallel validation comparing AI outputs directly against pathologist interpretations.

I - Integrate with existing systems: Integrate AI seamlessly into LIS and digital pathology platforms, emphasizing user-friendly interfaces.

D - Develop safety guardrails: Develop clear safety mechanisms: confidence thresholds, abstention guidelines, and transparent explanatory outputs.

A - Audit continuously: Implement ongoing quality monitoring of AI performance, systematically tracking discordances and systematic issues.

T - Train all stakeholders: Provide comprehensive training for pathologists, technologists, and clinicians, clarifying AI capabilities and limitations.

E - Evolve roles strategically: Recognize that AI implementation transforms professional roles from pure diagnosticians to diagnostic orchestrators.

D - Deploy with measured confidence: Phased rollouts, continuous feedback loops, and responsive adjustments promote smoother transitions.

The ORCHESTRATE framework

O - Optimize workflow gradually: Begin with lower-risk applications before expanding to critical diagnostic tasks.

R - Roll out in phases: Implement AI capabilities incrementally, starting with screening negative cases.

C - Create feedback loops: Establish robust communication channels between pathologists, technologists, and AI systems.

H - Harmonize human-AI collaboration: Design workflows that leverage the strengths of both human expertise and AI efficiency.

E - Educate continuously: Embed AI competencies into daily practice through ongoing training.

S - Standardize quality metrics: Develop consistent measures for AI performance and workflow efficiency.

T - Track return on investment milestones: Monitor key performance indicators aligned with institutional goals.

R - Refine based on outcomes: Use performance data to continuously improve AI algorithms and workflow integration.

A - Amplify successful practices: Share successes across departments and institutions to accelerate adoption.

T - Transform roles strategically: Support the evolution of professional roles through targeted education.

E - Evolve with technology: Remain adaptable as AI capabilities advance.

VALIDATED and ORCHESTRATE together provide a blueprint for AI transformation: VALIDATED ensures safe, systematic implementation, while ORCHESTRATE drives daily operational excellence.

Conclusions

The center of gravity in GU pathology AI is moving from isolated task assistance to clinical assurance. Translation is most reliable when case volumes are high, labels are stable, datasets are multi-institutional, and outputs are embedded into workflows with clear intended use, quality gates, and continuous performance monitoring.

Prostate pathology remains the most mature example of clinical integration, with multiple commercial solutions supported by regulatory clearances, multi-site validation, and prospective reader-in-the-loop implementations demonstrating meaningful efficiency gains and increased standardization of quantification and grading-adjacent tasks. The most defensible near-term value is a structured human-AI partnership: region-level localization, consistent quantitative summaries, and reproducibility improvements under pathologist control, backed by auditable QA.

Bladder cytology is the most pilotable non-prostate GU domain for prescreening workflows when paired with explicit adequacy thresholds, uncertainty-aware deferral, and traceable evidence artifacts aligned with Paris System categories. By contrast, bladder histology and renal neoplasia remain emerging domains where interobserver variability, artifact susceptibility, and under-representation of rare variants elevate the risk of silent failure unless systems incorporate robust abstention behavior, external validation across heterogeneous sites, and continuous drift surveillance.

For low-prevalence domains (testis, penis), the limiting factor is less algorithmic promise than data reality. Translation will require consortium-level aggregation, carefully defined endpoints, and transfer-learning strategies evaluated with enriched external test sets and subtype-aware reporting. Foundation and VLMs may reduce labeling burden and improve adaptation in low-data settings, but they do not replace the translational requirements of evidence grounding, uncertainty calibration, auditability, and governance.

The TRI rubric, SURE-Path minimum safety bundle, and the VALIDATED-ORCHESTRATE implementation pathway convert a rapidly expanding literature into operational decisions for clinical and translational pathology audiences. Used together, these frameworks help laboratories define intended use, validate locally in shadow mode, deploy with measurable safeguards, and monitor performance over time using workflow and quality outcomes. Ultimately, the institutions best positioned to benefit are those that adopt available tools with disciplined validation, realistic scope, and rigorous governance, treating AI as a monitored clinical instrument.

Acknowledgments

The authors express gratitude to the anonymous peer reviewers who volunteered their time and perspective for this manuscript.

Funding

The authors received no financial support from any public, commercial, or not-for-profit funding agency for the preparation of this manuscript.

Conflict of interest

The authors declare no conflicts of interest related to the content of this manuscript.

Author contributions

Conceptualization (AUP, AVP, SS), methodology (AUP), investigation and data curation of literature synthesis (AUP, AD, SS), visualization (AUP), writing - original draft (AUP), writing - review and editing (AUP, AD, SS, AVP), supervision (SS, AVP), and project administration (SS, AVP). All authors read and approved the final manuscript.

References

[1] Verghese G, Lennerz JK, Ruta D, Ng W, Thavaraj S, Siziopikou KP, *et al*. Computational pathology in cancer diagnosis, prognosis, and prediction - present day and prospects. J Pathol 2023;260(5):551–563. doi:10.1002/path.6163, PMID:37580849.

[2] Parwani AV, Patel A, Zhou M, Cheville JC, Tizhoosh H, Humphrey P, *et al*. An update on computational pathology tools for genitourinary pathology practice: A review paper from the Genitourinary Pathology Society (GUPS). J Pathol Inform 2023;14:100177. doi:10.1016/j.jpi.2022.100177, PMID:36654741.

[3] Patel AU, Mohanty SK, Parwani AV. Applications of Digital and Computational Pathology and Artificial Intelligence in Genitourinary Pathology Diagnostics. Surg Pathol Clin 2022;15(4):759–785. doi:10.1016/j.path.2022.08.001, PMID:36344188.

[4] Patel AU, Shaker N, Mohanty S, Sharma S, Gangal S, Eloy C, *et al*. Cultivating Clinical Clarity through Computer Vision: A Current Perspective on Whole Slide Imaging and Artificial Intelligence. Diagnostics (Basel) 2022;12(8):1778. doi:10.3390/diagnostics12081778, PMID:35892487.

[5] Eloy C, Marques A, Pinto J, Pinheiro J, Campelos S, Curado M, *et al*. Artificial intelligence-assisted cancer diagnosis improves the efficiency of pathologists in prostatic biopsies. Virchows Arch 2023;482(3):595–604. doi:10.1007/s00428-023-03518-5, PMID:36809483.

[6] Flach RN, van Dooijeweert C, Nguyen TQ, Lynch M, Jonges TN, Meijer RP, *et al*. Prospective Clinical Implementation of Paige Prostate Detect Artificial Intelligence Assistance in the Detection of Prostate Cancer in Prostate Biopsies: CONFIDENT P Trial Implementation of Artificial Intelligence Assistance in Prostate Cancer Detection. JCO Clin Cancer Inform 2025;9:e2400193. doi:10.1200/CCI-24-00193, PMID:40036728.

[7] Chatrian A, Colling RT, Browning L, Alham NK, Sirinukunwattana K, Malacrino S, *et al*. Artificial intelligence for advance requesting of immunohistochemistry in diagnostically uncertain prostate biopsies. Mod Pathol 2021;34(9):1780–1794. doi:10.1038/s41379-021-00826-6, PMID:34017063.

[8] Raciti P, Sue J, Retamero JA, Ceballos R, Godrich R, Kunz JD, *et al*. Clinical Validation of Artificial Intelligence-Augmented Pathology Diagnosis Demonstrates Significant Gains in Diagnostic Accuracy in Prostate Cancer Detection. Arch Pathol Lab Med 2023;147(10):1178–1185. doi:10.5858/arpa.2022-0066-OA, PMID:36538386.

[9] Bulten W, Kartasalo K, Chen PC, Ström P, Pinckaers H, Nagpal K, *et al*. Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge. Nat Med 2022;28(1):154–163. doi:10.1038/s41591-021-01620-2, PMID:35027755.

[10] Ström P, Kartasalo K, Olsson H, Solorzano L, Delahunt B, Berney DM, *et al*. Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study. Lancet Oncol 2020;21(2):222–232. doi:10.1016/S1470-2045(19)30738-7, PMID:31926806.

[11] Lu MY, Williamson DFK, Chen TY, Chen RJ, Barbieri M, Mahmood F. Data-efficient and weakly supervised computational pathology on whole-slide images. Nat Biomed Eng 2021;5(6):555–570. doi:10.1038/s41551-020-00682-w, PMID:33649564.

[12] Bulten W, Pinckaers H, van Boven H, Vink R, de Bel T, van Ginneken B, *et al*. Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. Lancet Oncol 2020;21(2):233–241. doi:10.1016/S1470-2045(19)30739-9, PMID:31926805.

[13] U.S. Food and Drug Administration. DEN200080: Paige prostate de novo classification orde. 2021.

[14] The Paige Prostate Suite: Assistive Artificial Intelligence for Prostate Cancer Diagnosis: Emerging Health Technologies. Ottawa (ON): Canadian Agency for Drugs and Technologies in Health; 2024. Report No.: EH0123. PMID:39466926.

[15] Steiner DF, Nagpal K, Sayres R, Foote DJ, Wedin BD, Pearce A, *et al*. Evaluation of the Use of Combined Artificial Intelligence and Pathologist Assessment to Review and Grade Prostate Biopsies. JAMA Netw Open 2020;3(11):e2023267. doi:10.1001/jamanetworkopen.2020.23267, PMID:33180129.

[16] Tolkach Y, Ovtcharov V, Pryalukhin A, Eich ML, Gaisa NT, Braun M, *et al*. An

international multi-institutional validation study of the algorithm for prostate cancer detection and Gleason grading. NPJ Precis Oncol 2023;7(1):77. doi:10.1038/s41698-023-00424-6, PMID:37582946.
[17] Marletta S, Eccher A, Martelli FM, Santonicco N, Girolami I, Scarpa A, *et al*. Artificial intelligence-based algorithms for the diagnosis of prostate cancer: A systematic review. Am J Clin Pathol 2024;161(6):526–534. doi:10.1093/ajcp/aqad182, PMID:38381582.
[18] Faryna K, Tessier L, Retamero J, Bonthu S, Samanta P, Singhal N, *et al*. Evaluation of Artificial Intelligence-Based Gleason Grading Algorithms "in the Wild". Mod Pathol 2024;37(11):100563. doi:10.1016/j.modpat.2024.100563, PMID:39025402.
[19] Hagland M. At Ohio State, a breakthrough leveraging AI in pathology. Healthcare Innovation. 2024. Available from: https://www.hcinnovationgroup.com/analytics-ai/article/53096393/at-ohio-state-a-breakthrough-leveraging-ai-in-pathology.
[20] Rienda I, Vale J, Pinto J, Polónia A, Eloy C. Using artificial intelligence to prioritize pathology samples: report of a test drive. Virchows Arch 2025;487(1):203–208. doi:10.1007/s00428-024-03988-1, PMID:39627613.
[21] Du X, Hao S, Olsson H, Kartasalo K, Mulliqi N, Rai B, *et al*. Effectiveness and Cost-effectiveness of Artificial Intelligence-assisted Pathology for Prostate Cancer Diagnosis in Sweden: A Microsimulation Study. Eur Urol Oncol 2025;8(1):80–86. doi:10.1016/j.euo.2024.05.004, PMID:38789385.
[22] Olsson H, Kartasalo K, Mulliqi N, Capuccini M, Ruusuvuori P, Samaratunga H, *et al*. Estimating diagnostic uncertainty in artificial intelligence assisted pathology using conformal prediction. Nat Commun 2022;13(1):7761. doi:10.1038/s41467-022-34945-8, PMID:36522311.
[23] Khoraminia F, Fuster S, Kanwal N, Olislagers M, Engan K, van Leenders GJLH, *et al*. Artificial Intelligence in Digital Pathology for Bladder Cancer: Hype or Hope? A Systematic Review. Cancers (Basel) 2023;15(18):4518. doi:10.3390/cancers15184518, PMID:37760487.
[24] Lebret T, Paoletti X, Pignot G, Roumiguié M, Colombel M, Savareux L, *et al*. Artificial intelligence to improve cytology performance in urothelial carcinoma diagnosis: results from validation phase of the French, multicenter, prospective VISIOCYT1 trial. World J Urol 2023;41(9):2381–2388. doi:10.1007/s00345-023-04519-4, PMID:37480491.
[25] Ibrahim M, Rabinowitz J, Hilbert R, Ghose A, Agarwal S, Swamy R, *et al*. The role of URO17® in diagnosis and follow up of bladder cancer patients. BMC Urol 2024;24(1):34. doi:10.1186/s12894-024-01426-7, PMID:38336681.
[26] Rosenthal DL, Wojcik EM, Kurtycz DFI, editors. The Paris System for Reporting Urinary Cytology. 1st ed. Cham: Springer; 2016. doi:10.1007/978-3-319-22864-8.
[27] Barkan GA, Wojcik EM, Nayar R, Savic-Prince S, Quek ML, Kurtycz DF, *et al*. The Paris System for Reporting Urinary Cytology: The Quest to Develop a Standardized Terminology. Acta Cytol 2016;60(3):185–197. doi:10.1159/000446270, PMID:27318895.
[28] Wojcik EM, Kurtycz DFI, Rosenthal DL. We'll always have Paris: the Paris system for reporting urinary cytology 2022. J Am Soc Cytopathol 2022;11(2):62–66. doi:10.1016/j.jasc.2021.12.003.
[29] Patel AU, Atiya S, Song Y, Chu W, Parwani AV. Novel liquid immunocytochemistry with machine learning analysis for bladder cancer detection. J Histotechnol 2025:1–9. doi:10.1080/01478885.2025.2546655, PMID:40815554.
[30] Vaickus LJ, Suriawinata AA, Wei JW, Liu X. Automating the Paris System for urine cytopathology-A hybrid deep-learning and morphometric approach. Cancer Cytopathol 2019;127(2):98–115. doi:10.1002/cncy.22099, PMID:30702803.
[31] Levy JJ, Chan N, Marotti JD, Kerr DA, Gutmann EJ, Glass RE, *et al*. Large-scale validation study of an improved semiautonomous urine cytology assessment tool: AutoParis-X. Cancer Cytopathol 2023;131(10):637–654. doi:10.1002/cncy.22732, PMID:37377320.
[32] Levy JJ, Chan N, Marotti JD, Rodrigues NJ, Ismail AAO, Kerr DA, *et al*. Examining longitudinal markers of bladder cancer recurrence through a semiautonomous machine learning system for quantifying specimen atypia from urine cytology. Cancer Cytopathol 2023;131(9):561–573. doi:10.1002/cncy.22725, PMID:37358142.
[33] Bannier PA, Saillard C, Mann P, Touzot M, Maussion C, Matek C, *et al*. AI allows pre-screening of FGFR3 mutational status using routine histology slides of muscle-invasive bladder cancer. Nat Commun 2024;15(1):10914. doi:10.1038/s41467-024-55331-6, PMID:39738108.
[34] Loeffler CML, Ortiz Bruechle N, Jung M, Seillier L, Rose M, Laleh NG, *et al*. Artificial Intelligence-based Detection of FGFR3 Mutational Status Directly from Routine Histology in Bladder Cancer: A Possible Preselection for Molecular Testing? Eur Urol Focus 2022;8(2):472–479. doi:10.1016/j.euf.2021.04.007, PMID:33895087.
[35] Lotan Y, Krishna V, Abuzeid WM, Launer B, Chang SS, Krishna V, *et al*. Predicting Response to Intravesical Bacillus Calmette-Guérin in High-Risk Nonmuscle-Invasive Bladder Cancer Using an Artificial Intelligence-Powered Pathology Assay: Development and Validation in an International 12-Center Cohort. J Urol 2025;213(2):192–204. doi:10.1097/JU.0000000000004278, PMID:39383345.
[36] Chandramohan D, Garapati HN, Nangia U, Simhadri PK, Lapsiwala B, Jena NK, *et al*. Diagnostic accuracy of deep learning in detection and prognostication of renal cell carcinoma: a systematic review and meta-analysis. Front Med (Lausanne) 2024;11:1447057. doi:10.3389/fmed.2024.1447057, PMID:39301494.
[37] Hermsen M, de Bel T, den Boer M, Steenbergen EJ, Kers J, Florquin S, *et al*. Deep Learning-Based Histopathologic Assessment of Kidney Tissue. J Am Soc Nephrol 2019;30(10):1968–1979. doi:10.1681/ASN.2019020144, PMID:31488607.
[38] Zheng Q, Mei H, Weng X, Yang R, Jiao P, Ni X, *et al*. Artificial intelligence-based multimodal prediction for nuclear grading status and prognosis of clear cell renal cell carcinoma: a multicenter cohort study. Int J Surg 2025;111(6):3722–3730. doi:10.1097/JS9.0000000000002368, PMID:40146270.
[39] Tian K, Rubadue CA, Lin DI, Veta M, Pyle ME, Irshad H, *et al*. Automated clear cell renal carcinoma grade classification with prognostic significance. PLoS One 2019;14(10):e0222641. doi:10.1371/journal.pone.0222641, PMID:31581201.
[40] Tabibu S, Vinod PK, Jawahar CV. Pan-Renal Cell Carcinoma classification and survival prediction from histopathology images using deep learning. Sci Rep 2019;9(1):10509. doi:10.1038/s41598-019-46718-3, PMID:31324828.
[41] Ricketts CJ, De Cubas AA, Fan H, Smith CC, Lang M, Reznik E, *et al*. The Cancer Genome Atlas Comprehensive Molecular Characterization of Renal Cell Carcinoma. Cell Rep 2018;23(1):313–326.e5. doi:10.1016/j.celrep.2018.03.075, PMID:29617669.
[42] Liu J, Lichtenberg T, Hoadley KA, Poisson LM, Lazar AJ, Cherniack AD, *et al*. An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. Cell 2018;173(2):400–416.e11. doi:10.1016/j.cell.2018.02.052, PMID:29625055.
[43] Zheng Q, Wang X, Yang R, Fan J, Yuan J, Liu X, *et al*. Predicting tumor mutation burden and VHL mutation from renal cancer pathology slides with self-supervised deep learning. Cancer Med 2024;13(16):e70112. doi:10.1002/cam4.70112, PMID:39166457.
[44] Xiong Y, Yao L, Lin J, Yao J, Bai Q, Huang Y, *et al*. Artificial intelligence links CT images to pathologic features and survival outcomes of renal masses. Nat Commun 2025;16(1):1425. doi:10.1038/s41467-025-56784-z, PMID:39915478.
[45] Chuluunbaatar Y, Bansal S, Brodie A, Sharma A, Vasdev N. The current use of artificial intelligence in testicular cancer: a systematic review. Art Int Surg 2023;3:195–206. doi:10.20517/ais.2023.26.
[46] Ghosh A, Sirinukunwattana K, Khalid Alham N, Browning L, Colling R, Protheroe A, *et al*. The Potential of Artificial Intelligence to Detect Lymphovascular Invasion in Testicular Cancer. Cancers (Basel) 2021;13(6):1325. doi:10.3390/cancers13061325, PMID:33809521.
[47] Vorontsov E, Bozkurt A, Casson A, Shaikovski G, Zelechowski M, Severson K, *et al*. A foundation model for clinical-grade computational pathology and rare cancers detection. Nat Med 2024;30(10):2924–2935. doi:10.1038/s41591-024-03141-0, PMID:39039250.
[48] Chen RJ, Ding T, Lu MY, Williamson DFK, Jaume G, Song AH, *et al*. Towards a general-purpose foundation model for computational pathology. Nat Med 2024;30(3):850–862. doi:10.1038/s41591-024-02857-3, PMID:38504018.
[49] Xu H, Usuyama N, Bagga J, Zhang S, Rao R, Naumann T, *et al*. A whole-slide foundation model for digital pathology from real-world data. Nature 2024;630(8015):181–188. doi:10.1038/s41586-024-07441-w, PMID:38778098.
[50] Zhang S, Xu Y, Usuyama N, Bagga J, Tinn R, Preston S, *et al*. Large-scale domain-specific pretraining for biomedical vision-language processing. arXiv 2023. Available from: https://doi.org/10.48550/arXiv.2303.00915.
[51] Zhou C, Zhang YF, Guo S, Huang YQ, Qiao XN, Wang R, *et al*. Multimodal data integration for predicting progression risk in castration-resistant prostate cancer using deep learning: a multicenter retrospective study. Front Oncol 2024;14:1287995. doi:10.3389/fonc.2024.1287995, PMID:38549937.
[52] Hu C, Qiao X, Huang R, Hu C, Bao J, Wang X. Development and Validation of a Multimodality Model Based on Whole-Slide Imaging and Biparametric MRI for Predicting Postoperative Biochemical Recurrence in Prostate Cancer. Radiol Imaging Cancer 2024;6(3):e230143. doi:10.1148/rycan.230143, PMID:38758079.
[53] Lu MY, Chen B, Williamson DFK, Chen RJ, Liang I, Ding T, *et al*. A visual-language foundation model for computational pathology. Nat Med 2024;30(3):863–874. doi:10.1038/s41591-024-02856-4, PMID:38504017.
[54] Liu H, Xue W, Chen Y, Chen D, Zhao X, Wang K, *et al*. A survey on hallucination in large vision-language models. arXiv 2024. Available from: https://doi.org/10.48550/arXiv.2402.00253.
[55] Patel AU, Parwani AV, Satturwar S. Artificial intelligence in genitourinary pathology. Histopathology 2026;88(1):353–373. doi:10.1111/his.70020, PMID:41384695.