Article

Not peer-reviewed version

A Guide for the Deployment, Validation, and Accreditation of Clinical Digital Pathology Tools

Andrew Janowczyk [†] , Johan Ferrari [†] , Nicolas Brandt , Laura Padayachy , Aurélie Bornand , Suzanne Chartier , Thomas Mallet , Nicolas Roduit , Paul Seed , Medya Tekes Mizrakli , Julien Massonnet , Petros Liakopoulos , Mario Kreutzfeldt , Laura Rubbia-Brandt [‡] , Doron Merkler [*,‡]

Posted Date: 12 September 2025

doi: 10.20944/preprints202509.1090.v1

Keywords: digital pathology; lab developed test; clinical validation; accreditation; algorithms; computer aided diagnostics

Article

A Guide for the Deployment, Validation, and Accreditation of Clinical Digital Pathology Tools

Andrew Janowczyk [1,2,3,†], Johan Ferrari [3,†], Nicolas Brandt [3], Laura Padayachy [3], Aurélie Bornand [3], Suzanne Chartier [3], Thomas Mallet [3], Nicolas Roduit [4], Paul Seed [4], Medya Tekes Mizrakli [3], Julien Massonnet [3,5], Petros Liakopoulos [2], Mario Kreutzfeldt [3,5], Laura Rubbia-Brandt [3,‡] and Doron Merkler [3,5,*,‡]

[1] Department of Biomedical Engineering, Emory University and Georgia Institute of Technology, Atlanta, USA.

[2] Division of Precision Oncology, Geneva University Hospitals, Geneva, Switzerland.

[3] Division of Clinical Pathology, Geneva University Hospitals, Geneva, Switzerland.

[4] Digital Transformation and Augmented Intelligence Department, Geneva University Hospitals, Geneva, Switzerland.

[5] Department of Pathology and Immunology, University of Geneva, Geneva, Switzerland.

[*] Correspondence: doron.merkler@hug.ch

[†] Denotes shared first authorship.

[‡] Denotes shared senior authorship.

Abstract

Digital pathology (DP) has become an integral part of modern diagnostic practice, enabling efficient workflows, reproducible assessments, and the integration of computational tools. While many institutions are transitioning toward digital diagnostics, deployment, validation, and routine usage of DP tools in clinical settings remains a major challenge. Existing recommendations often remain high-level, leaving laboratories without detailed, practical guidance on how to structure and document validation processes in real-world settings. To address this gap, we present a framework derived from the experience of Geneva University Hospitals (HUG), which since 2019 has operated a fully digital primary diagnostic workflow processing more than 230,000 slides annually. Drawing on this implementation, we outline best practices and lessons learned for the development, deployment, validation, and accreditation of DP tools. Document templates are also provided, which may serve as adaptable blueprints for other laboratories. This practical guide is intended to support reproducible, transparent, and audit-ready validation processes, helping institutions bridge the gap between general recommendations and operational deployment of DP tools in accredited diagnostic environments.

Keywords: digital pathology; lab developed test; clinical validation; accreditation; algorithms; computer aided diagnostics

1. Introduction

Digital pathology (DP) represents a transformative shift in the field of pathology, where traditional methods of analyzing tissue samples using glass slides and microscopes are replaced with digital high-resolution images and potentially augmented by computational analysis [1]. The digitization of histopathological specimens facilitates their retrieval, evaluation and sharing while enabling advanced computational tools to enhance the efficiency and reproducibility of diagnostic processes [2,3]. These tools support pathologists by quantitative and objective quality control [4,5], automating routine tasks like quantification [6–8], and prioritizing diagnostically relevant slides for review [9,10]. Furthermore, DP opens new avenues for personalized medicine by enabling the reproducible extraction of quantitative data for image-based biomarkers, thus improving diagnostic accuracy, guiding treatment decisions, and monitoring of treatment response [11–15].

However, before using internally developed DP tools in a routine clinical workflow, their validation and accreditation remains paramount in ensuring the reliability, accuracy, and consistency of diagnostic processes. To begin, it is crucial to differentiate between *qualification*, *validation*, and *accreditation* processes [16]. **Qualification** pertains to verifying that equipment and systems are correctly installed and functioning according to specified requirements, encompassing stages such as Design Qualification (DQ), Installation Qualification (IQ), Operational Qualification (OQ), and Performance Qualification (PQ) [17]. Once the system has been qualified, **validation** focuses on ensuring that the entire process in question (e.g., creating diagnostic grade WSIs) consistently meets its intended purpose and predetermined specifications [18]. This involves confirming that e.g., tools produce reliable, accurate, and consistent results in their application. Lastly, **accreditation** broadly implies evaluation and certification of laboratory practices, personnel competence, equipment performance, and quality management systems against established standards and guidelines, which adhere to stringent quality control measures, ensuring the integrity and traceability of data from specimen collection to reporting [19]. It is important to note that a lab *cannot accredit itself;* only an external auditor may provide formal approval of the completed validation and confirm it complies with current regulations (e.g., in Switzerland the Swiss Accreditation Service (SAS)[1], in Germany the Deutsche Akkreditierungsstelle (DAkkS)[2], and in the US the internationally recognized College of American Pathologists (CAP) Accreditation program[3]). Thus, in summary, *qualification* is specific to the operational readiness of equipment, *validation* encompasses broader assessment of the process's effectiveness, and an external accreditor certifies the validation process.

Of note, while an accredited laboratory workflow encompasses all validated steps from specimen preparation to diagnostic reporting, individual components such as DP tools may not themselves need to be independently accredited. In practice, accreditation applies to the controlled use of these tools within the laboratory process rather than to the tool's internal logic or algorithmic design. For instance, a laboratory may develop and validate a DP tool as a Laboratory Developed Test (LDT) [20–22], demonstrating its reliability, accuracy, and reproducibility within the intended accredited workflow. Nevertheless, the tool in isolation is itself not formally accredited, as accrediting bodies typically do not certify proprietary software; they instead evaluate the laboratory's implementation, documentation, quality control measures, and adherence to regulatory standards. Consequently, a tool can be safely employed in clinical practice under an accredited workflow provided that its validation is rigorously documented, but its standalone accreditation would require separate regulatory approval, such as through medical device clearance or compliance with software-specific standards [23]. In this manuscript, care has been taken to precisely use these terms, with the phrase "validation/accreditation" used when the choice of terminal state is left to the discretion of the institution.

That said, tools often begin as LDTs with the intent of being developed and validated for use *within a specific laboratory* adhering to local or national standards (e.g., ISO 15189:2022 [24]) where they undergo regular audits and evaluations to ensure accuracy and reliability. Once mature, institutions and companies often want to commercialize them for *broader multi-site use*, and thus must comply with the European Directive 98/79/EC[4] or Regulation (EU) 2017/746[5], [25] or FDA regulations in the United States [26,27]. As such, the main difference between an LDT and a CE+IVDR (Conformité Européenne - In Vitro Diagnostic) or FDA-approved technique lies primarily in their regulatory pathways and oversight requirements. Beyond LDT requirements, CE-marked IVDs

[1] https://www.sas.admin.ch/sas/en/home.html

[2] https://www.dakks.de/en/home-en.html

[3] https://www.cap.org/laboratory-improvement/accreditation/laboratory-accreditation-program

[4] https://eur-lex.europa.eu/eli/dir/1998/79/oj/eng

[5] https://eur-lex.europa.eu/eli/reg/2017/746/oj/eng

undergo conformity assessment procedures under the IVDR, which typically involve a notified body for higher-risk devices (Classes B, C, and D) that conducts quality management system audits and technical documentation reviews. FDA-cleared IVDs follow either the 510(k) pathway [28,29] requiring demonstration of substantial equivalence to predicate devices through analytical performance studies, or the premarket approval pathway (PMA) [30] for high-risk devices requiring clinical trial data. Importantly, regardless of regulatory approval status, all diagnostic tests—including CE-marked IVDs and FDA-cleared devices—must undergo local verification or validation before clinical implementation to confirm they perform reliably within the specific context of local laboratory conditions, sample types, and clinical workflows [31].

Despite growing adoption of DP [32], standardized implementation of validation protocols varies markedly across institutions, an issue that may in part stem from the fact that most available guidance remains at a high-level rather than providing detailed, prescriptive procedures [33]. This guide aims to address that gap by presenting practical real-world guide based on insights from Geneva University Hospitals (HUG), which completed its transition to a fully digital primary diagnostic workflow begun in 2019, and now develops and deploys in-house tools for routine use. Drawing on HUG's experience processing over 1,500 slides daily (230,000+ annually), this paper outlines key considerations, challenges, and best practices for tool validation and accreditation. Document templates are provided in the Appendix that can serve as blueprints to further facilitate implementation of the processes outlined.

1.1. Roles and Responsibilities

Personnel engaged in the tool creation, deployment, validation, and accreditation process can be defined by their roles. Depending on the size of the institution, these roles may be assigned several people, or conversely, one person may be assigned to several roles. At HUG we defined following roles:

Technical Roles

Algorithm Researcher (AR): The **AR** is responsible for developing specific technologies, models, or algorithms. This role may be undertaken, for example, by a (a) student working on a terminal project (e.g., a Master's or PhD thesis), (b) in-house/consulting professional software developer, or (c) an external third-party institution. Depending on the context, **AR** may thus be temporary, permanent, or externally contracted.

Deployment Engineer (**DE**): The **DE**'s role involves taking an *algorithm* developed, by e.g., **AR** , and hardening it into a *tool* for practical, real-world usage. These responsibilities include: (a) strengthening the algorithm to ensure it is robust and reliable for production use, (b) creating the necessary connectors to enable interaction with the hospital's infrastructure e.g., via standardizing software development kits (SDKs), and (c) deploying the tool within the IT infrastructure. Typically, a DE is a hospital employee, thus unlike **AR**, holds a long-term responsibility for maintenance and sustainability of the tool.

IT Staff (**IT**): **IT** staff concentrate on the broader technical environment, ensuring necessary hardware and software systems (e.g., high-performance computing solutions) support the tool's deployment and ongoing operation. They need to ensure that underlying technical infrastructure is robust, secure, and compliant with regulatory standards. This includes maintaining network stability, managing storage solutions, and ensuring cybersecurity measures protecting patient data.

Medical Roles

Experimental Pathologist (**EP**): **EP**s collaborate closely with **AR**s and **DE**s to define scope and requirements of the tool, ensuring it addresses relevant clinical challenges. As domain experts, they help identify relevant cases/slides, and provide feedback/annotations as needed. Acting as "beta testers", they rigorously evaluate the algorithm during its development phase, providing essential feedback to refine and enhance its performance. Their clinical expertise and hands-on involvement help bridge the gap between theoretical development and practical application, ensuring the final tool is clinically viable, user-friendly (ergonomic), and relevant.

Routine Pathologist (**RP**): **RP**s are the intended end-users of the tool who will use it in daily practice. Unlike **EP**s, they were not actively involved in the tool's development. Instead, they are exposed to it only after it (a) has been fully integrated in their routine clinical workflows and validated, and (b) they have received training on how to use it. As such, effectiveness and user-friendliness of the tool in real-world scenarios are paramount to them.

Laboratory Roles

Laboratory Technician (**LT**): **LT**s assure the day-to-day operations of the laboratory, executing tests, maintaining equipment, troubleshooting issues, and ensuring accurate results. When new tools are introduced, e.g., via hand-off from **DE**, they are responsible for supporting (e.g., diagnostic tools) and/or applying (e.g., quality control tools) them in daily workflow, following protocols and guidelines established by **DE** and Laboratory Managers. By adhering to quality control standards and providing feedback on performance and usability they ensure that integration of new tools is in line with the laboratory's needed efficiency and reliability.

Laboratory Manager (**LM**): **LM**s oversee overall operations of the laboratory, ensuring that all processes run smoothly and efficiently. In the context of implementing new tools, they are responsible for coordinating between all the various stakeholders listed here. They ensure that the laboratory is equipped with the necessary resources and infrastructure to support the new technology. Additionally, they oversee compliance with regulatory standards, manage quality control, and ensure that staff are adequately trained to use new tools effectively.

Departmental Leadership (**DL**): They provide strategic direction and oversight for the laboratory's operations. In the context of implementing new tools, they play a pivotal role in road-mapping, decision-making, resource allocation, and driving organizational change. **DL** ensures that the development and adoption of new technologies aligns with institutional goals, guidelines, and objectives. They facilitate communication and coordination among various departments, including research, clinical, and administrative teams, to ensure a smooth integration process. Additionally, they champion innovation and quality improvement initiatives, encouraging a culture of continuous learning and adaptation.

Regulatory Roles

Quality Officer (**QO**): A hospital employee ensuring compliance with internal standards and external accreditation requirements. They serve as liaison between the hospital and accreditation bodies, maintaining transparency and upholding the integrity of the hospital's quality management system.

Accreditors (**AC**): Representatives of the external body responsible (e.g., SAS, CAP) for evaluating and certifying that the laboratory, including its processes and tools meet industry standards and regulatory requirements. Their role involves conducting thorough audits of the laboratory's processes, equipment, personnel and documentation.

1.2. High Level Overview

The high-level overview of the validation/accreditation process (see **Figure 1**) provides a structured framework for understanding the key stages involved. Each process box corresponds to a numbered section below for easy reference and comprehension. The workflow is delineated into four distinct phases, described below, represented by different colors.

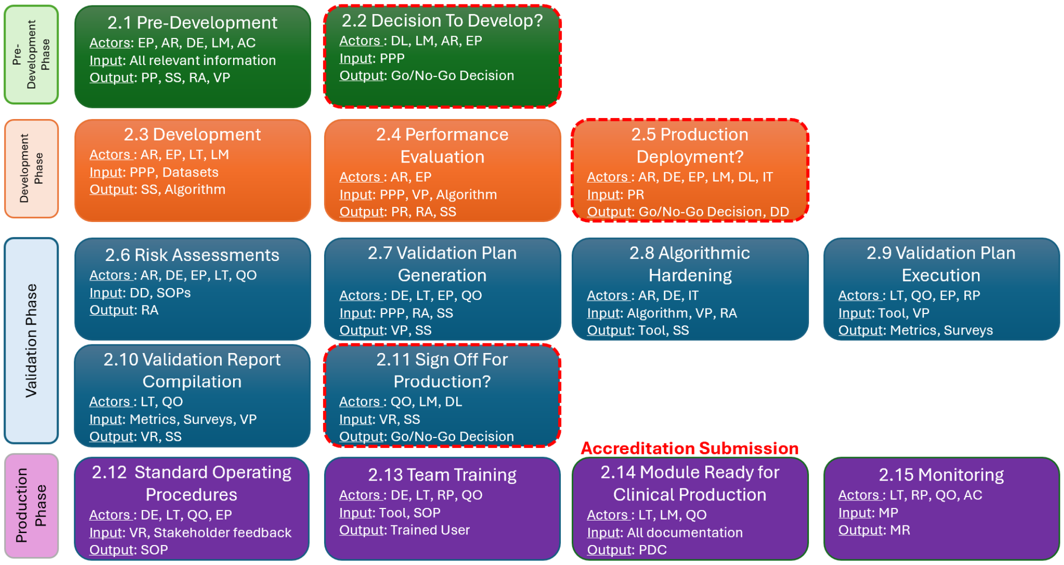
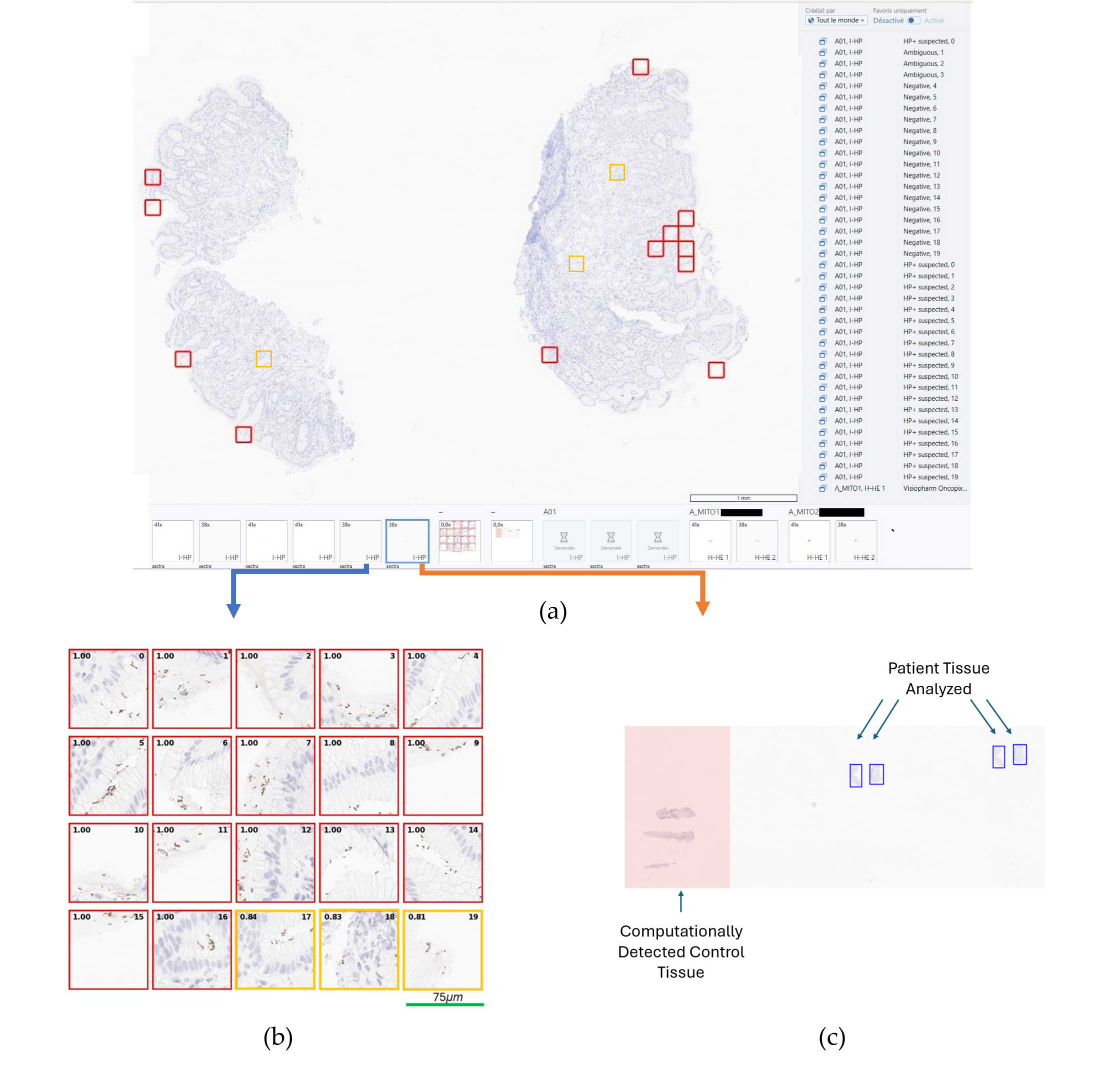


Figure 1. High level overview with each stage coinciding with a numbered sub-section in the main text below for easy reference. The workflow is broken into 4 phases, as indicated by the different colorings. Pre-development Phase and Development Phase, occur pre-validation/accreditation, wherein the algorithms are designed and developed. The Validation Phase sees the now hardened tool validated and deployed for routine clinical usage. Lastly, the Production Phase is a process of continual assessment and monitoring, as defined by the validation/accreditation needs, to ensure long-term stability and effectiveness of the deployed tool. Stages surrounded by a red dashed border indicate decision points. Additional acronyms: Decision Document (DD), Monitoring Plan (MP), Monitoring Report (MR), Production Deployment Checklist (PDC), Project Plan (PP), Project Plan Package (PPP), Performance Report (PR), Risk Assessment (RA), Standard Operating Procedures (SOPs), Software Specification (SS), Validation Plan (VP), Validation Report (VR).

The Pre-Developmental Phase determines if criteria such as feasibility and expected added value are sufficient to justify moving to the developmental phase, and yields a precise **Project Plan Package**. In the Developmental Phase, the algorithm is designed, developed, and evaluated according to that **Project Plan Package**. Establishing a realistic timeline—with clear milestones, deadlines, and checkpoints—is essential for an effective developmental phase, and can be guided by best practices from software and research projects, such as Agile and Scrum methodologies [34,35] that emphasize phased progress, collaboration, and continuous feedback and improvement. Next, the Validation Stage ensues, wherein algorithms are forged into hardened tools which undergo rigorous validation and deployment for routine clinical usage. The Production Phase signifies an on-going process of assessment and monitoring post-deployment. This phase is essential for maintaining long-term stability and effectiveness of deployed algorithms, incorporating quality control measures, regular maintenance and audits, along with feedback mechanisms to address evolving validation/accreditation needs.

1.3. Introduction to Worked HPyloriDet Example

To provide real-world context for the various stages, our *Helicobacter pylori* (HP) detection tool (HPyloriDet) is used as an example. Briefly, HPyloriDet is a deep learning based tool which analyses HP stained slides (IHC, DAB), and integrates outputs into Sectra for pathologist review (see **Figure 2**). This tool was deployed as a beta version in April 2024, and then fully clinically validated since March 2025. During that time, it has processed approximately 6,000 and 1,700 WSIs respectively (*manuscript under preparation*).



A01, I-HP	HP+ suspected, 0
A01, I-HP	Ambiguous, 1
A01, I-HP	Ambiguous, 2
A01, I-HP	Ambiguous, 3
A01, I-HP	Negative, 4
A01, I-HP	Negative, 5
A01, I-HP	Negative, 6
A01, I-HP	Negative, 7
A01, I-HP	Negative, 8
A01, I-HP	Negative, 9
A01, I-HP	Negative, 10
A01, I-HP	Negative, 11
A01, I-HP	Negative, 12
A01, I-HP	Negative, 13
A01, I-HP	Negative, 14
A01, I-HP	Negative, 15
A01, I-HP	Negative, 16
A01, I-HP	Negative, 17
A01, I-HP	Negative, 18
A01, I-HP	Negative, 19
A01, I-HP	HP+ suspected, 0
A01, I-HP	HP+ suspected, 1
A01, I-HP	HP+ suspected, 2
A01, I-HP	HP+ suspected, 3
A01, I-HP	HP+ suspected, 4
A01, I-HP	HP+ suspected, 5
A01, I-HP	HP+ suspected, 6
A01, I-HP	HP+ suspected, 7
A01, I-HP	HP+ suspected, 8
A01, I-HP	HP+ suspected, 9
A01, I-HP	HP+ suspected, 10
A01, I-HP	HP+ suspected, 11
A01, I-HP	HP+ suspected, 12
A01, I-HP	HP+ suspected, 13
A01, I-HP	HP+ suspected, 14
A01, I-HP	HP+ suspected, 15
A01, I-HP	HP+ suspected, 16
A01, I-HP	HP+ suspected, 17
A01, I-HP	HP+ suspected, 18
A01, I-HP	HP+ suspected, 19
A_MITO1, H-HE 1	Visiopharm Oncopix...

Figure 2. Example of integrated HPyloriDet output into Sectra. (a) Sectra image viewer showing computationally added annotations, colored by confidence (red=high, orange=medium, blue=low), for potential HP positive regions. They are clickable on the right-hand side, enabling immediate navigation to those regions. (b) Zoomed view of the contact sheet corresponding to the blue arrow in (a), displaying the top 20 highest-probability HP positive patches. (c) Zoomed view of the item indicated by the orange arrow in (a), the slide overview image, with blue bounding boxes showing where the analysis was performed. The red overlay marks computationally detected control tissue, which was subsequently excluded from analysis.

2. The Workflow

2.1. Pre-Development Stage

Scope: The scope of the pre-development stage encompasses defining objectives, constraints, and requirements of the project. This involves clearly outlining clinical needs, desired outcomes, and target users for the algorithm, but excludes any actual tool development until after 2.2 Decision to Develop Stage.

Participants: **EP**, **AR**, **DE**, **LM**, **AC** AC is optional but suggested on first LDT accreditation.

Input: These include gathering relevant data, literature reviews, stakeholder input, and existing algorithms or methodologies that may inform the development process, helping to serve as a foundation for defining scope and project requirements.

Output: This **Project Plan** includes an overarching summary document containing sections: Background, Motivation, Aims, Timeline, Expected Deliverables (both quantitively and qualitatively in retrospective and prospective validations), and Initial Feasibility Assessments.

Preliminary versions of (a) **Appendix 1 - Software Specification, Responsibility, and Confirmation Plan** (referred to as **"Software Specification"** going forward) should be completed, documenting in more detail the expectations of such a tool, its technical requirements, software stack, and assign personnel ownership/responsibility of the various components of its life-cycle to competent personnel, (b) **Appendix 2 - Analysis Of Risks Impacting Work Processes During Major Modifications** (referred to as "**Risk Assessment**"), and (c) **Appendix 3 - Validation Plan and Report** (referred to going forward as "**Validation Plan**" or "**Validation Report**", as appropriate) with its Section 1 completed. These documents should be updated throughout the validation process until 2.11 Sign Off For Production Stage is completed.

Importantly – these documents are often similar across tools of the same class (e.g., all detection tools, all segmentation tools), and thus benefit from standardization and templating.

Together, the Project Plan, Software Specification, Risk Assessment, and Validation Report make up what is referred to as the Project Plan Package.

Timeline: Algorithms with easily interpretable output and clear performance requirements that can be met with current technology, like HPyloriDet, typically require 15-20 hours for the pre-development phase. More sophisticated algorithms, where existing technology may not be readily applicable, or the output less obvious how to interpret (e.g., an image-based biomarker risk score), will likely require substantially more of a time investment.

Context and Considerations: This stage is potentially the most critical as it has an extensive impact on the success and completion efficiency of all subsequent stages, and therefore warrants appropriate investment. For example, when the acceptance criteria outlined in the **Validation Plan** are known in advance, relevant elements can be incorporated during the algorithm design and data collection phases, thereby greatly facilitating these processes and avoiding rework.

Appreciating that many tools will be aimed at efficiency and time-gain, it is advisable to record a set of unbiased time measurements—both unaided and aided (e.g., task completion time per minute). Two complementary types of experiments can be considered: (i) a baseline measurement of unaided performance within the Image Management System (IMS), ideally performed on a larger number of slides, to establish how much room for improvement exists in practice and to provide a benchmark for later aided comparisons; and (ii) a preliminary time-gain experiment, which may be estimated outside the IMS (e.g., using tools such as QuPath [8]), to provide early evidence that tool usage could yield efficiency benefits. It may also be useful to simulate the expected outputs of the tool—for example, by having a second pathologist manually generate a minimal set of results representative of those the tool would ideally produce for the first pathologist to complete their evaluation with. This allows for a controlled estimation of the time or effort differential, even before an algorithm is created. A point of anecdotal evidence suggests that algorithm development may bias the pathologists' interpretative processes; therefore, these measurements should ideally be obtained independently or prior to the development phase.

Additionally, estimating the annual volume of cases to which the tool will be applicable is crucial for balancing development effort against anticipated benefits (see **Software Specification 4**). If, in practice, the task is completed efficiently without assistance—leaving little room for meaningful improvements at scale—it may be worth reconsidering the development of the tool unless other advantages, such as increased precision/sensitivity or financial benefits, justify its implementation.

To enhance the efficiency of the potential validation process, all collected whole slide images (WSIs)—particularly those used during the time measurements—should be retained. Since tool development and validation typically span several months, this period aligns well with the ideal pathology reading washout interval [36,37]. By preserving WSIs for later use, the same slides can be reused for validation, enabling direct comparisons of evaluation speed and accuracy, thereby increasing statistical power and reliability while minimizing human effort.

Lastly, the scope of the 2.1 Pre-Development Stage should likely be reviewed by the Institutional Review Board (IRB) to ensure ethical compliance, assess potential risks to patients, and oversee proper use of patient data, as responsibility for these aspects ultimately resides with them. It also remains important to emphasize that when the pre-development phase is performed wholly or partially by individuals external to the institution—such as students, contractors, or commercial partners—all data released to them remains subject to institutional regulations (e.g., established data usage agreements), and should be thoroughly anonymized to prevent any linkage between patients, pathology samples, and these pre-development activities.

Worked HPyloriDet Example: The team was as follows: (i) **EP**: the senior pathologist that proposed the project, (ii) **AR** and **DE** roles were assigned to the same hospital employee, as this was an internal project, (iii) a senior **AR** to supervise the project, with oversight from the (iv) **LM** and **DL**.

The **Project Plan** included a concise background section and literature review on automatic HP detection algorithms, revealing most previous studies focused on non-IHC staining and were not clinically deployed. After a cost-benefit analysis, the primary goal was identified as a clinically deployable screening tool for detecting HP on IHC-stained slides with nearly 100% sensitivity. The key metric in demonstrating value-add was determined to be time-to-diagnosis gain. The user-facing usability/ergonomic requirement (see **Software Specification 7e**) stated that results were to be integrated within the same IMS used during routine diagnostic sign-out for optimal accessibility.

The cohort size needed for developing the tool was estimated to be approximately 100 HP-positive IHC slides scanned at 40x magnification, with 20–30 of them to be annotated by **EP**. The planned baseline approach involved the following steps: (1) tissue detection, (2) identification of regions of interest (ROIs), i.e., those containing putative HP-positive objects, using stain deconvolution, (3) their associated patch extraction, and (4) their classification as positive or negative via deep learning.

The project was divided into three components with the following 3 deliverables: (1) Tissue detection, ROI identification and patch generation, (2) Training the deep learning model focusing on high sensitivity, and (3) Hardening and deploying the tool within the existing workflow, ensuring DICOM compatibility and IMS integration, with clear, interpretable outputs presented as either heatmaps highlighting HP-rich areas or direct slide annotations. The latter decision was left open for iterative refinement and consensus-building during development.

The expected duration for these 3 deliverables was approximately 3 weeks, 2 months, and 5 weeks, respectively. The overall 2.1 Pre-Development Stage was completed over the course of 2 weeks.

2.2. Decision to Develop Stage

Scope: Review material collected in the 2.1 Pre-Development Stage and decide if development is warranted.

Participants: DL, LM, AR, EP

Input: The **Project Plan Package** created in the previous stage.

Output: A go/no-go decision to advance to the 2.3 Development Stage.

Timeline: The decision-making process can likely happen during a single discussion when all necessary documentation is available and reviewed beforehand.

Context and Considerations: Upon reviewing the **Project Plan Package**, it may become clear that proceeding with tool development is no longer desirable or feasible. Several factors could contribute to this assessment, including: (a) the anticipated improvement over existing clinical methods may be less significant than initially expected; (b) the development plan may prove overly complex, due to unforeseen technical challenges, a broader-than-expected scope, or need for specialized expertise not readily available; (c) the overall available data may be insufficient to support robust training and validation due to a low projected caseload; (d) even if the dataset is large enough for training, rare subtypes (i.e., those part of the long tail [38]) intended for inclusion may be inadequately represented, limiting validation potential; (e) inadequate compute or IT infrastructure to support needed computational throughput or turnaround time; or (f) lack of functionality needed to ergonomically integrate results within the IMS.

Relatedly, once a decision is made to proceed, revisiting the data collection strategy is beneficial. Since the validation set will likely need to include challenging or rare cases—often identifiable only through retrospective and labor-intensive review—prospective/concurrent identification and retention can be facilitated by enabling **EP** and **RP** proactive flagging within the IMS during their routine clinical workflows.

Worked HPyloriDet Example: **DL** approved the decision to develop upon reviewing the **Project Plan Package**.

2.3. Development Stage

Scope: Focuses on the collaborative design, coding, and iterative refinement of the algorithm aimed at instantiating the tool for eventual clinical usage.

Participants: AR, EP, LT, LM

Input: Includes clinical data sets, domain knowledge from **EP**, existing algorithms/tools/models that can be adapted, and the **Project Plan Package** (e.g., **Project Plan**, **Software Specification, Risk Assessment, Validation Plan**)

Output: A containerized (e.g., docker/apptainer) version of the algorithm along with a fully drafted **Software Specification**, which includes items like well documented source code, hardware/software requirements for running the algorithm, and tests to be completed during 2.9 Validation Plan Execution Stage. The tool's user facing output/result format should be approved by **EP**, and in line with prior specifications.

Timeline: While the complexity of the algorithm itself naturally influences the timeline—for instance, a simple IHC detection algorithm may take only a few weeks to develop, whereas a comprehensive image-based biomarker system could require month to years—this component is often the most *predictable*. Development efforts generally follow patterns from similar "research-use-only" or software development projects. By contrast, the factors discussed below in **Context and Considerations** are less constrained and more variable, often driving uncertainty and delays. Clarifying them early help create a realistic, reliable project timeline.

Context and Considerations: A key consideration is assessing data availability and allocating adequate time for data collection. Studies may require (a) ethics approval, (b) retrospective curation of WSIs and associated clinical metadata, and (c) translation of this data into anonymized machine-readable formats suitable for tool development. As data volumes increase, the time needed—particularly for tasks requiring manual effort—is often underestimated.

Another key consideration is the choice of technology used to develop the algorithms. Some platforms may be less suitable for integration into high-throughput clinical workflows. For example, while MATLAB is well-suited for prototyping, Python is more widely accepted for production-grade development and integration. Selecting technologies that support scalability and compatibility with existing clinical systems and **IT** infrastructure from the outset is essential.

It is strongly recommended to avoid changing technologies between development and deployment, as both 2.4 Performance Evaluation Stage and 2.9 Validation Plan Execution Stage benefit significantly from using a consistent technology stack. Any changes at this point could invalidate the results of the 2.4 Performance Evaluation Stage and necessitate repeating the 2.8 Algorithm Hardening Stage.

Worked HPyloriDet Example: Development began with **EP** annotating 20 HP positive WSIs (SVS or MRXS format), selected for their abundance of stained objects affording exposure to both clear true positives and potential false positives. To improve tool sensitivity, regions with ambiguous staining—where HP positivity was uncertain but would still warrant pathologist review—were labeled as positive. The model was developed in Python using the PyTorch deep-learning framework. After tissue detection was performed using thresholding on a 2.5x magnification black-and-white version of the WSI, ROI detection was built using stain deconvolution and applying a minimum DAB-stain intensity threshold. During training sensitivity, specificity, PPV and NPV were tracked, and errors manually reviewed, paying particular attention to false negatives. The pathologist-facing results display reached its final design, as shown in **Figure 2**. For integration, DICOM support was implemented to ensure IMS compatibility: slides are downloaded from the IMS in DICOM format, annotations are transmitted through the IMS API, and image results are transferred via DICOM C-Store.

2.4. Performance Evaluation Stage

Scope: To evaluate the algorithm's performance on retrospective data in a way that informs whether it merits deployment and prospective validation.

Participants: AR, EP

Input: An algorithm, as developed in 2.3 Development Stage, along with its associated **Project Plan Package** and **Validation Plan**, outlining key metrics and the corresponding datasets to compute them on.

Output: A scientific manuscript style document, termed a **Performance Report**, including sections for: Methods, Experimental Design, Experimental Results, Discussion, and Conclusion. Results should be presented both as aggregate summary statistics, and at the level of individual patients and/or items (e.g., patches), as inspecting metric distributions may offer deeper insights into potential failure modes. Evident failures of the algorithm should be noted in detail, as these will provide critical information for either (a) refining the algorithm (return to 2.3 Development Stage, or address during 2.8 Algorithm Hardening Stage), (b) amending the **Risk Assessment** or **Software Specification**, and/or (c) inclusion into the 2.12 Standard Operating Procedure (SOP) Stage as documented edge cases for awareness of downstream tool users.

Timeline: For fully automatic algorithms with known available ground truths, the evaluation process can take only a few minutes, as the computational comparison of the algorithm output to the ground truth can be measured in an unaided way. On the other hand, it is often too expensive in terms of human labor to collect ground truth for use cases which have large numbers of objects (e.g., cell phenotyping), so a visual verification approach may be warranted, wherein the **EP** grades results on a semi-qualitative scale indicating {poor, modest, acceptable, ideal}, with requirements defined as a percentage of these categories. The challenge in this context is notably that this involves additional **EP** time which can add delays to the overall process. Documentation timeline is similar to that of an equivalent peer-reviewed scientific manuscript.

Context and Considerations: From our experience, there is often a large gap between the public research cohorts which algorithms are typically built with (e.g., Grand Challenge datasets [39], see **Figure 3**) versus those with which the tool must operate on daily. Clinical data often involves greater complexity and heterogeneity, but typically maintains consistently high quality. As a result, to be more efficient in the evaluation, it is not uncommon to break the evaluation into increasingly larger steps. For example, a pilot study on a small subset of slides can and should be conducted to ensure that the results are as expected before beginning evaluation on the totality of the cohort. This enables a more rapid iteration, while also not exposing the algorithm to all the testing data, so that a confident assessment of the algorithm on held-out test data can still take place. Naturally, the closer this evaluation is to the real-world assessment (see 2.7 Validation Plan Generation Stage and 2.9 Validation Plan Execution Stage), particularly if the development and this performance evaluation only employ real-world data, the more efficient the overall validation process will be, as there will be fewer opportunities for unexpected situations to arise in later stages.

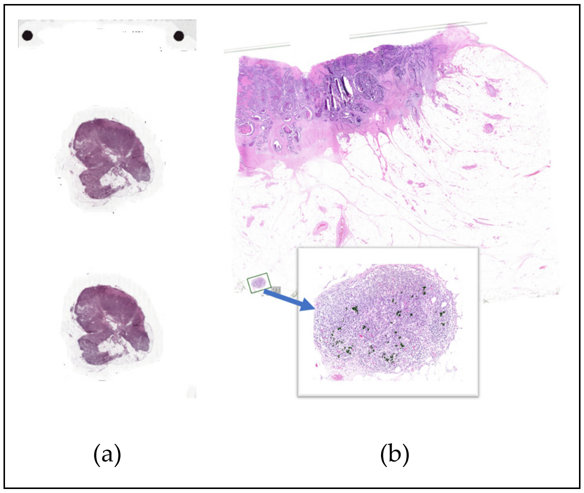


Figure 3. An example demonstrating the difference between (a) a public research dataset for metastasis detection in lymph nodes, versus (b) a routine clinical slide, with the equivalent lymph node region highlighted in the bottom left box. When transitioning an algorithm developed using cohorts on the left to the real-world data on the right, it is likely that a performance gap will be exposed which will need to be addressed before the algorithm can be integrated into a tool meeting the needed performance specifications.

Worked HPyloriDet Example: A **Performance Report** was generated for HPyloriDet, beginning with a summary of its functionality and representative examples of output within the IMS. The evaluation proceeded in three parts: (i) patch-level performance metrics (sensitivity, specificity, positive predictive value, and negative predictive value) were computed on the validation set of the 20 annotated slides (see 2.3 Development Stage), supplemented by visual inspection of false negatives and illustrative patch examples; (ii) a broader assessment was conducted on 132 routine histopathology slides collected over a two-week period, with the primary aim of confirming correct identification of HP-positive cases, noting minor output issues (e.g., scattered false-positive patches) versus major issues (e.g., widespread false positives or predictions off tissue), and illustrating representative contact sheet outputs; and (iii) a preliminary estimate of time-to-diagnosis gain was derived using QuPath outside the IMS environment. The QuPath analysis suggested a ~20% reduction in review time for negative slides; however, absolute review times differed substantially (~2 minutes in QuPath vs. 20–30 seconds in IMS), likely due to (a) the pathologist's greater familiarity with IMS and smoother navigation (e.g., 3Dconnexion "Spacemouse Pro" support), and (b) the QuPath-aided read being performed before inception of the contact sheet output format.

Several limitations were identified at this stage, including positive predictions occurring outside tissue regions (e.g., on artifacts), inability to ignore juxtapositioned control tissue when present on slides, and a relatively high rate of false positives—likely attributable to an initial training strategy focused on minimizing false negatives. Future development priorities included integrating quality control mechanisms, accommodating control tissue, and retraining the model to reduce false positive rates. In hindsight, pre-collection of the two-week slide cohort would have streamlined the assessment process.

2.5. Production Deployment Decision Stage

Scope: Evaluate the algorithm's current performance, scope, and intended use case to determine its suitability for deployment.

Participants: AR, DE, EP, LM, DL, IT (optional)

Input: The **Performance Report** generated in the previous step which summarizes the intended use of the algorithm, what validation was performed, on what data, the results of that evaluation in contrast to those expected and identified during the 2.1 Pre-Development Stage.

doi:10.20944/preprints202509.1090.v1

Output: A go/no-go decision by **DL** to proceed with the deployment and validation process. For traceability a **Decision Document** is created, being the signed version of the final **Performance Report** with any additional comments or notes which need to be propagated to subsequent steps.

Timeline: With appropriate documentation in place, a decision can often be reached in one or two meetings. That said, however, it is not uncommon that during the review process of **Input**, additional specificity or refinement is needed to inform the final decisions, and thus iterations of **Input** between the **Participants** to generate an approved **Decision Document** are common.

Context and Considerations: The validation/accreditation process requires a dedicated effort to be completed, which stretches across many different departments, domains, and disciplines. As such, it is not "free" to validate/accredit algorithms, and only those that are likely to (a) meet the required standards, (b) provide the expected added value, and (c) justify long-term support and maintenance should proceed with this process. It is also important to factor in that as one validates more algorithms of the same class, and better workflows and document templates are created, the easier the accreditation is likely to become, thus further increasing the number of algorithms which pass from the "research" into the "clinical" sphere.

Worked HPyloriDet Example: HPyloriDet was approved to move into the Deployment Phase by **DL** upon review of the **Performance Report**.

2.6. Risk Assessment Stage

Scope: Analyze and assess all risks associated with deploying the tool into routine operation.

Participants: AR, DE, EP, LT, QO

Input: The **Decision Document**, as well as established laboratory SOPs that connect to this tool. For example, most DP tools will be inserted into accredited DP workflows, which themselves contain a set of accreditation documents. Reviewing these to identify any interconnectivity that may pose a risk for the tool under consideration is suggested [40–42].

Output: A formalized **Risk Assessment** document which aims to highlight the frequency, severity, and detectability of risks both before and after their respective mitigation strategies are employed.

Timeline: It is reasonable to expect this stage to take approximately one week, appreciating that the first time one completes it may require substantial time and effort. However, tools within the same class (e.g., diagnostic, prognostic, detection, counting) tend to share a core set of risk profiles, supporting reuse and refinement of templated risk tables. As a result, subsequent revisions and iterations are expected to proceed with increasing efficiency in a matter of hours, as familiar risks are systematically identified, scoped, and addressed.

Context and Considerations: Assessing risk is one of the major components of the validation process, which as a reminder is focused on guaranteeing the quality, reliability, and safety of its processes and outcomes.

As **RP** and **LT** increasingly rely on deployed tools, it is crucial to consider the implications of tool failures. Any malfunction or unexpected behavior can significantly disrupt clinical workflows and potentially compromise patient care. Understanding the potential impacts of tool downtime or errors is therefore essential. Risk assessment must include scenarios where the tool stops working as expected, examining both the consequences for clinical operations and patient outcomes, as well as how such failures can be rapidly detected through continuous monitoring and quality measurements. Contingency plans should be developed to address these scenarios, ensuring that alternative processes or manual interventions are in place to maintain continuity of care. Additionally, the risks extend beyond patient care to include workflow disruptions, cybersecurity threats, data leakages, and the ongoing training of pathologists to ensure they do not experience a loss of competence in interpreting results independently due to reliance on the tool [7].

The insights gained from risk assessment directly inform the 2.7 Validation Plan Generation Stage. By identifying potential risks, the validation plan can incorporate specific tests and procedures to ensure that the mitigation strategies are effective in practice. This implies designing validation protocols that simulate real-world conditions and potential failure scenarios to verify the tool's reliability and robustness. For instance, if a risk is identified that the tool might misinterpret certain types of data, the validation plan should include extensive testing on these types to evaluate and improve the tool's performance under those conditions. Ensuring that risk mitigation approaches are thoroughly tested helps build confidence in the tool's readiness for routine use.

Worked HPyloriDet Example: The primary risks associated with HPyloriDet were identified as: incorrect algorithmic HP detection, failure of the tool to execute, misinterpretation of results by the user, and potential impacts on overall system performance and workflow. Regarding risk of incorrect detection—where the tool may miss true positive regions—a low frequency and detection risk was anticipated, with medium clinical gravity (see **Risk Assessment**). Mitigation strategies included a formal validation plan to assess detection performance, clearly communicated guidance to **RPs** regarding scenarios where the output may not be reliable, and *explicit retention of the* ***RP****'s final diagnostic authority*. The latter clearly signifying that HPyloriDet is not an autonomous tool, but a computer-aided diagnostic assay. HPyloriDet execution failure, although unlikely and of low impact, was mitigated by fallback to standard unaided slide review. To prevent misinterpretation—such as control tissue being mistaken for a patient's positive sample, or assuming high-confidence patches are always correct—training sessions and SOP documentation were defined, and further algorithmic improvements were planned to explicitly exclude control regions. Finally, to assess the potential impact on computational resources and workflow, HPyloriDet's resource requirements were benchmarked in a test environment prior to deployment. A dedicated high-performance server for DP tools was purchased, providing a scalable infrastructure for production. Testing and execution were carried out on the same system, enabling a more accurate risk assessment than would be possible using the previous desktop environment. Monitoring metrics were defined, aiming to detect systematic deviations from validated performance to ensure ongoing reliability (see 2.15 Monitoring Stage).

2.7. Validation Plan Generation Stage

Scope: Define performance goals, expected criteria, and the methods to be used including number of samples, execution, and metrics.

Participants: DE, LT, EP, QO

Input: In practice, all previous inputs should be present to ensure that the **Validation Plan** covers the entire scope needed for validation. Of particular importance is the **Risk Assessment** plan and **Software Specification**, as these will have identified concerns of varying levels of severity, that should be particularly addressed and evaluated during validation.

Output: A revised **Software Specification** and a **Validation Plan** (see **Appendix 3**, **1. Planning Section**) that outlines the specific steps, methods, and metrics to be used in evaluating the tool. This includes defining the number and type of samples to be tested, execution plan, and performance metrics to be measured. The **Validation Plan** should specify objective acceptance criteria, such as accuracy, sensitivity, specificity, reliability, or time-gain that the tool must meet to be considered successful. Additionally, it should include procedures for analyzing data, recording results, and reporting findings. An ideal, well-structured, **Validation Plan** ensures that the validation process is systematic, transparent, and reproducible.

Timeline: Generation of the first **Validation Plan** is likely on the order of 100 hours, with subsequent ones on the order of 10 hours. Similar to the **Risk Assessment**, the first **Validation Plan** may take longer as the requirements should be iteratively fleshed out with the **QO** and potentially the **AC**, but once in place, generation for subsequent tools of the same class should be increasingly efficient as core requirements are not likely to change significantly.

Context and Considerations: This stage involves understanding the purpose and scope of the validation process and ensuring it aligns with clinical and regulatory requirements. The primary purpose is to define a **Validation Plan** which enables the verification of the tool performing as intended and meets predefined performance goals. This includes confirming that improvements in any subsequent versions of an established tool enhance its functionality as compared to the previous version.

Important considerations include the clinical relevance of validation metrics, representativeness of the sample data, and feasibility of validation procedures within clinical settings. It is important to anticipate potential challenges, such as variability in sample quality or integration issues with existing systems, and include strategies in the **Validation Plan** to ensure they are effectively addressed. For example, the 2.6 Risk Assessment Stage helps determine the types and characteristics of cases, especially complex or edge cases, needed during this validation. The pathologist/end-user's experiences using the tool in their routine practice remains crucial, thus planned dedicated surveys documenting their acceptance and approval is warranted.

Time spent during earlier stages reflecting on validation requirements should now yield benefits, as the **Validation Plan** may already be largely in place due to those efforts.

Worked HPyloriDet Example: Building on the successful performance evaluation (see 2.4 Performance Evaluation Stage) the following acceptance criteria were established for promotion of HPyloriDet to production:

1. Detection and removal of positive control samples placed adjacent to patient tissue: When not properly identified, these highly positive regions overwhelm patient-specific results, causing a 5–10% slide-level false positive rate. A key task identified for 2.8 Algorithm Hardening Stage, was to accurately exclude this control tissue computationally. The acceptance threshold was set so that HPyloriDet must correctly handle more slides than simply ignoring all slides containing control tissue (which would be the default fallback). This functionality was evaluated in 2.9 Validation Plan Execution Stage.

2. Patch-level sensitivity and specificity: For acceptance, patch-level sensitivity and specificity was set to at least match those from the 2.4 Performance Evaluation Stage. While these metrics are critical for rapidly computationally assessing HPyloriDet, the primary focus remained on slide-level accuracy, i.e., correctly identifying slide HP status using the top 20 predicted regions. As such, patch-level metrics acted as a computational surrogate appreciating that they do not fully capture clinical relevance. Manual review of false negatives patches was also required, and false negatives were considered acceptable if they occurred in patches with ambiguous staining, but were not acceptable in any patch showing clearly visible HP. Validation used the original evaluation patches along with any additional patches collected during the hardening phase.

3. False negative rate: As a sensitive screening tool, HPyloriDet was required to identify all positive slides The acceptance criterion mandated that at least one of the top 20 predicted regions on every positive slide must contain HP. This requirement was evaluated retrospectively using the 132 sequential slides from the 2.4 Performance Evaluation Stage, and prospectively during the survey period (see **List Item 5,** below). Slides exhibiting known failure modes—identifiable by pathologists within the IMS—were excluded. These included errors in tissue control handling, major quality control failures such as all top 20 predictions highlighting artifacts, incorrectly identified tissue regions, predictions falsely positive on red blood cells, blurry slides, and slides containing unusually large tissue samples such as surgical specimens. These situations were appropriately documented in the 2.6 Risk Assessment Stage and 2.12 Standard Operating Procedure (SOP) Stage.

4. Time saving: A reduction in time to diagnosis serves as the primary measure of clinical benefit. For acceptance, HPyloriDet had to demonstrate a statistically significant improvement within the IMS, as used by pathologists, in average time to diagnosis ($p < 0.05$) compared to routine slide examination without it. For rigor, the validation targeted matched numbers of positive and negative slides, with all slides containing comparable amounts of tissue. Slides meeting exclusion criteria outlined in **Item 3** were omitted from analysis.

5. Pathologist survey: A mandatory survey would be conducted to assess **RPs** confidence in the tool's outputs. For acceptance, the average confidence rating must be at least 2.7 out of 3 [43,44], with high=3 points, medium=2 points, or low=1 point. Slides excluded under the criteria outlined in **Item 3** were not considered in this evaluation.

6. Average analysis time: HPyloriDet must process slides quickly enough to ensure result availability when **RPs** review cases, particularly since IHC-stained slides are batch scanned in the late afternoon. Acceptance requires that average total analysis times remain below 5 minutes, including downloading, analyzing, and uploading results for each slide.

7. Repeatability and reproducibility: HPyloriDet had to demonstrate consistent performance, with reanalysis of the same slide yielding identical top 20 patch scores to 2 decimals. Furthermore, analyses of the same slide performed on the two different scanner models used in our clinical workflow must yield the same diagnostic conclusions.

Based on the **Software Specification**, the following criteria were identified for validation:

1. Unit tests: Passing unit tests for key functions.

2. Integration tests: Prior to deployment, HPyloriDet must complete successfully in the IMS test environment on one positive and one negative slide per scanner, with results displayed as expected.

3. Resource usage: The maximum RAM and VRAM usage must remain below 10GB and 5GB, respectively.

4. Concurrent usage / reliability: Using two concurrent HPyloriDet instances, all WSIs must be processed so that results are available for pathologists to begin case review according to their regular rhythm of clinical workflow.

2.8. Algorithm Hardening Stage

Scope: Transforming a research-grade *algorithm* into a robust *tool* suitable for clinical use.

Participants: AR, DE, IT

Input: A research grade algorithm along with the **Validation Plan** and **Risk Assessment**.

Output: A fully hardened tool integrated and working in the clinical environment, for execution in 2.9 Validation Plan Execution Stage. A revised **Software Specification** incorporating results from execution of the **Software Verification Plan**. The code should be well documented and have undergone code review to ensure (a) adherence to best practices and internal coding guidelines (e.g., SDK usage), and (b) understandability by **LT.**

Timeline: Here again, the advantage of having precise tool specifications becomes evident. If the tool was developed with the required standards in mind, this phase can be completed quickly and with minimal effort. However, if those standards were not adequately addressed early on, substantial rework may be needed to achieve the necessary performance, throughput, and integration. Consequently, this phase can range from just a few hours to several months, depending on the alignment between the prior 2.3 Development Stage and 2.7 Validation Plan Generation Stage stages.

Context and considerations: The transition from a *research algorithm* to a *clinically deployable tool* involves (a) refining the approach to meet strict performance and regulatory standards, (b) integrating it seamlessly into the clinical ecosystem, and (c) preparing it for rigorous validation according to the **Validation Plan**. During this hardening process, **DE** and **AR** collaborate closely: **DE** is responsible for adapting the algorithm to the healthcare facility's technical infrastructure, ensuring compatibility with existing electronic health records (EHR) and associated systems such as laboratory information systems (LIS), and IMS. Meanwhile, **AR** provides ongoing support and adjustments based on all prior feedback. Together, they optimize and test the algorithm's performance to handle real-world clinical data with high accuracy and reliability, while respecting any IRB requirements (e.g., AR may potentially not have access to clinical data). This may include fine-tuning parameters by specimen type (e.g., biopsies versus resections), improving computational efficiency, and ensuring the tool can process large data volumes rapidly without compromising accuracy.

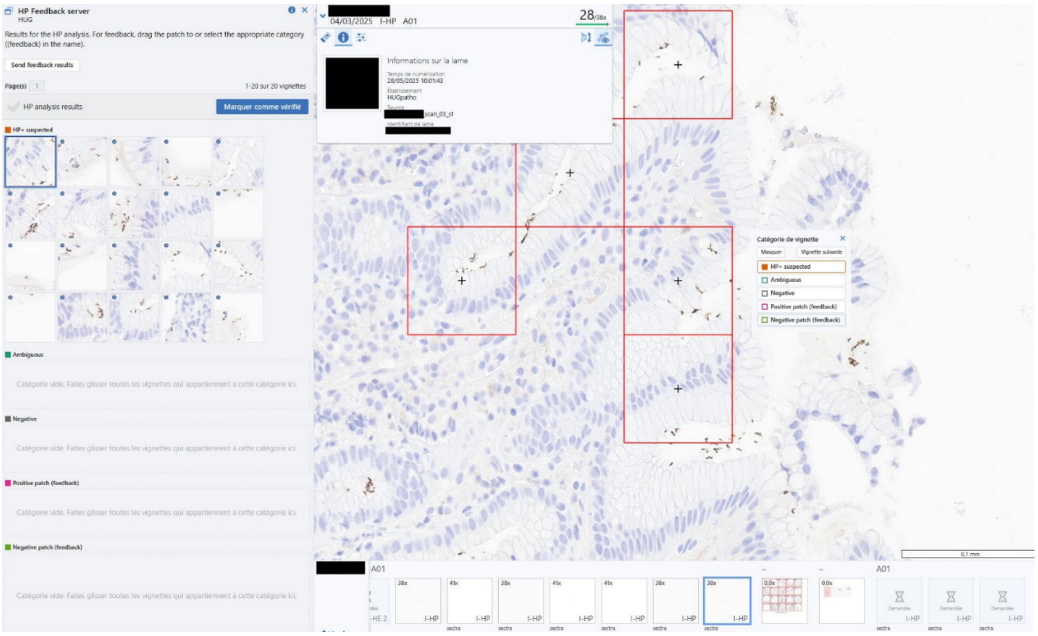
Security and compliance are critical aspects of algorithm hardening. **DE** must ensure the algorithm complies with data privacy regulations and cybersecurity best practices. This often involves collaborating with **IT** and data security officers to implement measures such as data encryption, secure transmission protocols, and safeguards that prevent unauthorized access to sensitive patient information.

Lastly, robust error handling and monitoring is essential, in particular to mitigate risks identified in **Risk Assessment**. The system should include mechanisms for detecting and managing crashes or failures, with comprehensive error logging that captures sufficiently detailed diagnostic information to enable debugging. Logs should be securely stored and accessible for troubleshooting and auditing purposes. Error reporting should be automated, with alerts sent to relevant **LT**/**DE** for timely intervention. Integrating the tool into a centralized monitoring dashboard enables real-time tracking of errors, performance metrics, and key performance indicators (KPIs) across all deployed tools. This centralized approach facilitates rapid identification of issues, trend analysis, and supports continuous improvement. Additionally, automated fail-safe procedures should be implemented to ensure that, in the event of a critical failure, disruption is contained within the failing component. These mechanisms should prevent cascading failures by isolating faults, enabling a controlled and limited shutdown that minimizes impact on other algorithms and preserves overall clinical operations.

Importantly, this is the last time that the tool should be refined and adjusted, as once it is validated, any subsequent modifications will require re-testing, reporting, and documentation. As a result, carefully reviewing the **Validation Plan** to ensure that the algorithm is in line with expectations is critical.

Worked HPyloriDet Example: During the algorithm hardening phase, a beta version of HPyloriDet was deployed in routine clinical practice. It was evaluated through multiple rounds of **RP** surveys with clear indication that it was not yet validated for use as a clinical aid (i.e., Research Use Only). To facilitate deployment, HPyloriDet was containerized (docker) and hosted on a dedicated high-performance computing server, with two additional containers for (a) logging and managing daily slide monitoring, and (b) supporting the HP feedback mechanism (see **Figure 4**).

A significant improvement involved enhanced control handling: when control tissue was detected, it was automatically excluded from the analysis, and the omitted regions were clearly highlighted on the "Regions analyzed Mask" for pathologists' reference (see **Figure 2**). Additionally, an in-IMS feedback mechanism was implemented (see **Figure 4**), allowing pathologists to flag patches as correctly or incorrectly classified directly within the viewer, enabling real-time feedback during routine use.y



HP Feedback server
HUG

Results for the HP analysis. For feedback, drag the patch to or select the appropriate category ((feedback) in the name).

Page(s) 1 1-20 sur 20 vignettes

HP analysis results

Ambiguous
Catégorie vide. Faites glisser toutes les vignettes qui appartiennent à cette catégorie ici.

Negative
Catégorie vide. Faites glisser toutes les vignettes qui appartiennent à cette catégorie ici.

Positive patch (feedback)
Catégorie vide. Faites glisser toutes les vignettes qui appartiennent à cette catégorie ici.

Negative patch (feedback)
Catégorie vide. Faites glisser toutes les vignettes qui appartiennent à cette catégorie ici.

[redacted]
04/03/2025 I-HP A01

[redacted]
Informations sur la lame
Temps de numérisation
28/05/2025 10:01:43
Établissement
HUGpatho
Source
[redacted]scan_03_st
Identifiant de lame
[redacted]

[redacted]

Figure 4. Gallery View within the Sectra viewer, which enables users to review algorithm predictions and provide corrective feedback. Predicted patches are organized into categories: "HP+ suspected" (high-confidence positives), "Ambiguous" (low-confidence positives), and "Negative." Two additional categories, "Positive patch (feedback)" and "Negative patch (feedback)," allow users to reassign labels either by dragging patches into the

correct group or by selecting the appropriate category. Once feedback has been applied, updated labels can be stored by clicking the "Send Feedback Results" button.

Quality control was another focus area. After the initial survey yielded a confidence score of 2.5—below the target of 2.7—largely due to numerous false positives, HistoQC [4] was integrated to filter out artifacts outside of tissue areas. This integration helped reduce slide level artifacts and improve overall tool accuracy.

Despite improvements, the second survey round only raised pathologist confidence score to 2.6, still short of the target. In response, the model was retrained using patches flagged by pathologists during routine practice, as well as selected HP analysis output patches saved during regular use.

2.9. Validation Plan Execution Stage

Scope: Conduct validation tests using both retrospective (previously analyzed cases) and prospective (ongoing case analyses alongside the validated method) approaches.

Participants: LT, QO, EP, RP

Input: Completed **Validation Plan** Planning Section.

Output: Metrics and values, including **RP** surveys, needed for the completion of the Report Section of the **Validation Plan**.

Timeline: The duration may vary significantly based on the associated **Validation Plan** and ability to meet the acceptance criteria. A primary factor contributing to this variability is the minimum number of slides required for evaluation, which directly ties the completion time to the case load managed by the specific tool.

Context and considerations: During the execution of the **Validation Plan**, it is crucial to closely monitor and document any deviations from expected outcomes, such as unexpected false positives or false negatives that may not have been anticipated during the earlier 2.6 Risk Assessment Stage or 2.8 Algorithm Hardening Stage phases.

Identifying these deviations promptly allows for this process to evolve into an iterative cycle, where new findings—especially rare or edge cases that were too infrequent to be randomly sampled in retrospective evaluations—may prompt a return to earlier stages, including 2.6 Risk Assessment Stage, 2.7 Validation Plan Generation Stage, and 2.8 Algorithm Hardening Stage. Rather than viewing these discoveries as failures, it is important to recognize them as valuable opportunities for continuous improvement. Such iterative refinement ensures that the tool ultimately deployed maintains safety, reliability, and robustness, which support the delivery of the highest quality patient care.

Recording analyzed cases in this stage facilitates long-term traceability and enables re-evaluation as the tool evolves.

Worked HPyloriDet Example: During evaluation, insufficient confidence levels on the pathologist survey were addressed by reverting back to the 2.8 Algorithm Hardening Stage to: (a) integrate HistoQC, and (b) perform additional model retraining based on annotations provided during routine clinical usage. Once both were implemented, this stage was re-executed, and results curated.

2.10. Validation Plan Compilation Stage

Scope: Completion of the **Software Specification** and **Validation Plan**, incorporating final statistics and conclusion.

Participants: LT, QO

Input: Test results from the previous stage and the associated **Validation Plan**

Output: A completed (a) **Software Specification**, and (b) "Report" Section of the **Validation Plan**, converting it from a **Validation Plan** into a **Validation Report**

Timeline: If all data is present, computing the final statistics and writing the report typically takes less than 10 hours.

Context and considerations: Summarizing the results within the institutionally approved templates typically requires noting any discrepancies identified during the testing phase. These should be meticulously documented, analyzed, and addressed to both refine the tool, if needed, but as well be taken into consideration when developing 2.12 Standard Operating Procedure (SOP) Stage to help mitigate any risks associated with its clinical application.

As seen in **Validation Report (Section 2 Item 7)**, the Conclusion Section essentially states if the tool met the required operating characteristics and should be considered for deployment. If this process fails, i.e., specific performance criteria are not met, one needs to determine if it is a catastrophic failure, requiring reverting back to 2.3 Development Stage, or if the failure is due to relatively minor points that can be addressed by proceeding with additional 2.8 Algorithm Hardening Stage**;** in either case, a revision of the **Validation Plan** and/or the **Software Specification** may be needed.

Notably, at this advanced stage, either outcome is considered costly and therefore should aim to be avoided whenever possible. Such situations may reflect insufficient clarification of the project scope or an incomplete assessment of anticipated risks. In particular, when completing this phase using an internally developed tool, unforeseen issues should be rare—any potential challenges should have been identified and addressed during earlier stages of development and planning.

Worked HPyloriDet Example: The results for the validation were the following:

1. Control Detection Rate**:** Of 103 validation slides (9 with controls), HPyloriDet correctly identified 5 controls, retaining 99 slides for analysis versus 94 without control handling, exceeding the stated baseline tolerance (see 2.7 Validation Plan Generation Stage List item 1). Notably, during the survey period all 172 slides had the control handling performed correctly, indicating that the effective rate in practice was higher than observed in the validation step. The primary source of residual error was attributable to coverslip artifacts.

2. Patch-Level Sensitivity and Specificity: An additional 630 patches were added to the original validation dataset, bringing the total to 5,316 patches with 549 being positive. The retrained model retained a sensitivity of 93%, while improving specificity from 88% to 94%. Importantly, all patch-level false negatives occurred in regions with ambiguous staining—no patches with clearly visible HP were missed.

3. False Negative Rate (Slide-Level): Across both original validation slides and cases evaluated during the survey period, no positive slides were missed, yielding a 0% false negative rate. One slide was excluded from analysis during the survey due to the predefined documented criteria.

4. Time to Diagnosis: An experiment comparing aided versus unaided diagnosis was conducted on 25 slides per condition (with 9 positive cases in each). The tissue burden was matched across groups to ensure comparability (3 low, 9 medium, and 4 high tissue volume slides in the negative set). Tool-assisted review led to an average 26% reduction in time to diagnosis (26.68 seconds vs. 36.21 seconds, $p = 0.02$), meeting acceptance criteria.

5. Pathologist Confidence Survey: Over a two-week survey period, 80 cases and 172 slides (33 positive) were reviewed. One case was excluded due to slide debris exceeding QC tolerances. The average confidence score reported by **RPs** was 2.93 out of 3, exceeding the required threshold of 2.7. Only one case had a low confidence rating due to black ink generating excessive false positives. Medium confidence was reported in four additional cases: two involving predictions outside tissue, one due to over-prediction on red blood cells, and one positive case with only a small amount of HP, where HP regions were correctly displayed in the top 20 patches but the pathologist noted that a relevant patch was not included.

6. Average Analysis Time: HPyloriDet processed slides in an average of 1 minute and 50 seconds, well below the 5-minute threshold required for clinical viability.

7. Repeatability and Reproducibility: The tool produced identical results when run multiple times on the same slide and generated comparable outputs when analyzing the same slide scanned on different scanners, confirming its consistency across runs and hardware variations.

All **Software Specification** requirements were also met, including (a) unit tests, (b) integration tests, (c) resource usage limits, and (d) concurrent usage /reliability tests as described in 2.7 Validation Plan Generation Stage.

2.11. Sign Off for Production Stage

Scope: The final validation report is reviewed and approved
Participants: QO, LM, DL
Input: **Completed Validation Report** and **Software Specification**.
Output: Go/no-go decision after reviewing **Validation Report** and **Software Specification**.
Timeline: This stage can be completed in as little as a few hours, but will depend highly on the time required for each signing official to review the documentation, provide feedback, and resolve any outstanding questions or concerns

Context and considerations: The main considerations at this step are a thorough review by the responsible parties to ensure the documents and tests have been completed as needed for promotion of the tool to routine clinical usage.

Worked HPyloriDet Example: The tool was approved for deployment in production without incident.

2.12. Standard Operating Procedure (SOP) Stage

Scope: Drafting the SOP describing the usage of this new tool
Participants: DE, LT, QO, EP
Input: **Validation Report** along with feedback from the various stakeholders who have experience in using the tool.
Output: A versioned SOP uploaded to the institutional document management system.
Timeline: Producing a complete first version of the SOP typically takes around one week, and may be faster for subsequent tools as some sections are reusable. However, multiple iterations are often part of the broader revision process, as feedback from stakeholders is gradually integrated to ensure the SOP is clear, comprehensive, and aligned with clinical expectations. Certain sections—such as troubleshooting—may continue to expand over time, with subsequent revisions issued as new issues arise during routine use. In this sense, this is a "living" document which implies this stage does not terminate, but instead becomes less frequent overtime as the tool matures.

Context and considerations: The development of SOPs is critical to the correct usage of the tool. Items that are typically included are listed in **Appendix 5 - Standard Operating Procedure Template**. These SOPs cover a few core concepts (a) when and what the tool should be applied to, (b) who should use the tool, (c) how they should use it, (d) installation, (e) maintenance and backup, and (f) troubleshooting. When drafting the SOP, it is worthwhile to think about who will be reading it and when. For example, most likely, experienced users or developers are not regularly reviewing the documentation. It is more likely that the tool has unexpectedly broken, and **IT**/**LT** are looking for ways to repair/restart it, or new people (e.g., **RP**) are onboarded and aiming to learn how to use the tool safely. Considering these points will yield an SOP which is not only needed for validation but improves usage by its users.

Worked HPyloriDet Example: SOPs were divided into two main components, tailored respectively for **RPs** and **LTs**:

For **RPs:** the SOP outlines both access and use of HPyloriDet within their clinical workflow. This includes stating that (a) to view results, the **RP** must have access to the IMS and be working on a case involving an HP-IHC slide, and (b) results are integrated directly into the IMS environment, appearing as DICOM images positioned next to the analyzed slide and as overlaid annotations on the slide itself. Importantly, **RP** is informed that the analysis is launched automatically in the background when a slide is received by the IMS, and results are typically available within five minutes after completion of the slide scanning process.

doi:10.20944/preprints202509.1090.v1

The SOP also includes detailed guidance on result interpretation. This section explains the presentation format of the outputs, highlights illustrative examples for clarity, and outlines known tool limitations (e.g., exclusion criteria described in 2.7 Validation Plan Generation Stage). Finally, it provides instructions for submitting feedback via the integrated viewer, in line with the tool validation protocol described in 2.8 Algorithm Hardening Stage.

For **LTs**: the SOP focuses on the installation, operation, and ongoing maintenance of the system. It specifies the location of the codebase and documentation, lists required supplementary files with corresponding storage paths, and provides instructions for containerization—detailing how to build and deploy the analysis container. To prevent data duplication or configuration drift, certain technical elements may instead opt to be cross-referenced from other systems, such as README files in version control.

Maintenance procedures include a description of potential errors, how to identify and address them, and how to restart the container when necessary. To ensure traceability, the SOP also defines the storage structure and content of log files, outlining what is retained and where it can be accessed for audit or debugging purposes.

2.13. Team Training Stage

Scope: Provide training for the relevant individuals identified in the SOP

Participants: DE, LT, RP, QO

Input: Access to a running instance of the tool, along with the SOP.

Output: Record training as completed for the individuals in the training register, which marks them as competent to use it.

Timeline: Depending on the sophistication of the tool, the training can typically be completed in a few hours.

Context and considerations: As stated by current standards, the laboratory needs to identify the competence requirements for each tool to ensure that users are able to use the tool correctly [45]. This training is typically done via interactive sessions where users are introduced to the tool, its capabilities, and its integration into their workflow. These sessions can be led by **DE** or **LT** who can provide in-depth explanations and answer questions. Practical sessions then follow where users can interact with the tool in a controlled environment, using intentionally curated sample data to simulate real-world scenarios. This helps users gain confidence and familiarity with the tool.

Regularly scheduled sessions to update users on newly validated version of the tool may be required to ensure their skills remain current. Other approaches, such as online courses that cover various aspects of the algorithm, including theoretical background, practical use, and case studies, or live/recorded webinars may be used to provide supplemental information about the tool to increase familiarity if needed or desired.

The discretion to mark a user as competent is given to the individual conducting the training session, who may implement quizzes or assessments to evaluate understanding. In our experience, the SOP is explained, questions addressed, and the trainee instructed to complete the steps mentioned in the SOP on their own once, with the output verified as correct.

Worked HPyloriDet Example: Similar to the SOP above, two dedicated training sessions were developed to ensure proper understanding by **RPs** and **LTs**.

For **RPs**, training included a comprehensive walkthrough of the SOPs and live demonstrations within the IMS. These sessions covered all essential components of tool usage, including how to access and interpret results, recommended practices, known limitations (such as exclusion cases outlined in 2.7 Validation Plan Generation Stage), and how to submit patch-level feedback directly through the system. Upon completion, **RPs** signed a confirmation form to document their participation. In addition, HPyloriDet training was incorporated into the standard IMS onboarding protocol for all new **RPs**.

For **LTs**, the session covered an SOP walkthrough and additional technical aspects of deployment and maintenance. These included detailed instructions on installation and

doi:10.20944/preprints202509.1090.v1

configuration—covering code location, necessary files, documentation access, and container build and run procedures. The training also addressed maintenance and troubleshooting, offering guidance on handling common errors and restarting the container when needed. Finally, participants were taught how to manage records and maintain traceability through the automated structured logging systems. Like the **RP** training, completion is confirmed with a signed form.

2.14. Module Ready for Clinical Production Stage

Scope: Deployment to production checklist is filled, reviewed, and approved

Participants: LT, LM, QO

Input: All previous documentation.

Output: A signed **Appendix 4- Production Deployment Checklist**.

Timeline: Execution time lies in the range of a few hours, as items should have in fact been addressed in previous phases. It is suggested to review this document from the onset of **Validation Phase** to ensure needed milestones are successfully reached along the way.

Context and Considerations: A final check to ensure that all the needed documentation, training, and tests have been completed and put in place. Once signed, the tool is officially available for usage by all those competent to use it. As such, the checklist should contain all steps that should be verified, such as (a) placement of the SOP into an accessible document repository, (b) availability of the computing resources, and (c) tool now enabled for daily use.

It is important to note that at this stage, although the tool has been validated as an LDT, it is *not yet accredited*. Upon DL's request, necessary documentation can be submitted by QO to initiate a request to extend existing accreditation to include this new tool. During the next scheduled audit visit—which typically occurs every two years—the tool will be reviewed on-site by the auditors (e.g., SAS). Based on their evaluation, accredited status may be granted. If accreditation is not awarded, necessary adjustments must be made to ensure that all regulatory requirements are fully met before resubmitting for approval.

Worked HPyloriDet Example: All documents were approved, and HPyloriDet was promoted to production.

2.15. Monitoring Stage

Scope: Monitoring of tool performance over time to ensure constancy in performance and functioning.

Participants: LT, RP, QO, AC

Input: At a specified interval, a subset of cases reasonably covering the expected gamut of presentations, re-read unaided by a qualified user and compared against the tool's output.

Output: A **Monitoring Report** detailing the tests performed, slides used, and noting any variations from expected performance metrics identified in the **Validation Report**. This can essentially be viewed as a subset of the **Validation Report**, which focuses on repeating and confirming performance metrics. Additionally, the report identifies the personnel who conducted the tests, along with any anomalies observed, and any recommendations for algorithm refinement based on the findings.

Timeline: Monitoring occurs as long as the tool is in use, with reporting requirements at identified time intervals.

Context and Considerations: The tool should have an established procedure for monitoring the validity of the results it produces. This implies recording results in such a way that trends and shifts are detectable and, where practicable, statistical techniques can be applied to review the results and identify sub-optimal behavior or outliers.

There are two types of monitoring which can take place, Internal Quality Control (IQC) or External Quality Assessment (EQA), with some tools taking advantage of both approaches. A more complete discussion is available in "Ensuring the validity of examination results" of the ISO standard [24], but the concepts are briefly elucidated here.

For prospective IQC, a subset of slides is re-evaluated at a fixed time (e.g., quarterly) to ensure concordance with the performance previously demonstrated during the **Validation Plan**. Any discrepancies between the original and current performance mechanisms trigger a review of the tool or related processes to maintain its accuracy and reliability. EQA sees the results of the tool compared with results obtained by other laboratories via e.g., the exchange of WSI, or 3rd parties provided references standard samples for analysis. For accredited tools, when EQA is not available, period IQC is required.

To ensure ongoing reliability and safety, a formal mechanism for user feedback should be established. This enables early detection of unexpected shifts in tool performance, which may arise from sources such as stain variability, batch effects, or other latent variables (e.g., subtle changes in tissue thickness [46]) that can evolve during routine laboratory operation. User-reported concerns—particularly when aggregated—can serve as valuable early warning signals of tool degradation or emerging edge cases not previously identified.

In parallel, due to the computational nature of these tools, automated monitoring systems can be implemented to perform real-time statistical surveillance of performance metrics. These systems can apply well-established quality control rules—such as Westgard [47,48] or Nelson [49,50] rules—to detect statistically significant deviations from expected behavior. When triggered, such alerts can prompt immediate review or intervention, minimizing the risk of performance drift going unnoticed.

Together, these user-driven and algorithm-driven monitoring approaches form a robust feedback and quality assurance loop, ensuring the tool remains safe, accurate, and clinically valuable throughout its operational lifetime. Ultimately, however, it is important to recognize that a properly deployed algorithm—being deterministic and pinned to fixed software versions—should not drift on its own. Exiting from "in-control" performance typically indicates that an external factor in the surrounding system (e.g., staining, scanning, or data handling) has shifted, rather than the tool itself. In such cases, the priority should be to bring the input or upstream process back under control. Any algorithmic adjustments should be undertaken with great caution, as they may cross a threshold that effectively creates a new version of the tool, thereby invalidating prior validation and necessitating revalidation, starting from 2.8 Algorithm Hardening Stage. This process can be further complicated if significant time has passed since the original development, as the necessary expertise, development context, or training data may no longer be readily available.

Worked HPyloriDet Example: Globally, the DP workflow at HUG undergoes periodic NEQAS quality evaluation (an EQA), which ensures the quality of slide cutting, staining, and digitization [51,52]. NEQAS requires both physical and digital slides for comparison; however, the DP module is currently available only for HE slides. To ensure the continued reliability of the HPyloriDet tool, in particular, both manual and automated review processes were implemented during deployment. Manual review was and will be conducted quarterly by selecting four representative slides per scanner model to verify that the tool produced consistent outputs across devices and maintained diagnostic concordance. This periodic check helped validate that hardware variability did not impact tool performance, and also occurs when e.g., a scanner software update occurs.

In parallel, automated monitoring was implemented to continuously detect performance drift that might not be identified through manual checks alone. On the input side, HistoQC outputs are continuously monitored to detect shifts in slide quality, such as scanning artifacts or staining anomalies, which could impact tool reliability. On the output side, the distribution of prediction scores is tracked separately for slides predicted as positive and negative, along with the overall ratio of predicted positive to negative cases. This allowed the identification of gradual shifts in scoring behavior, which could suggest emerging issues with the model or upstream processing.

3. Discussion

Based on practical experiences in implementing a full digital pathology workflow at HUG, this manuscript outlines essential steps for developing, validating and optionally accredit a clinically deployable digital pathology tool under the framework of a laboratory-developed test (LDT). In addition to describing a generalizable pathway from research to clinical integration, we provide supplemental materials and templates for the core documentation needed throughout this process. Our discussion is further grounded in practical experience, drawing on the end-to-end development and validation of HPyloriDet, an in-house computer-aided detection (CAD) tool for identifying HP in gastric biopsies.

A clear understanding of the full workflow—from design through deployment—is essential for the successful development and validation/accreditation of clinical algorithms. By anticipating regulatory and operational requirements early, developers can integrate necessary elements organically, reducing the risk of major rework later in the process. This proactive, holistic approach also supports better coordination and feedback across multidisciplinary teams, enabling clinical, technical, and regulatory stakeholders to contribute meaningfully from the outset. In turn, this fosters transparency, alignment, and efficient decision-making. Ultimately, embedding validation considerations throughout the development cycle helps ensure a smoother, more robust pathway to clinical integration.

One important aspect not explored in depth here is the associated financial cost—particularly when partnering with third-party tool providers. It is essential to ensure that their business models align with the budgetary constraints of pathology services. While a tool may demonstrate clinical utility, its implementation can lead to significant ongoing expenses, both anticipated and unforeseen. In addition to the considerable effort required for validation, integration often demands substantial technical resources, which can lead to hidden costs (e.g., custom integration fees, in particular with/for/to legacy systems). Institutions may also experience sticker shock post-integration when faced with recurring licensing fees or support costs. As such, it is critical to fully understand contractual obligations, billing structures, and the potential for reimbursement before committing to the development, deployment, and validation of any tool.

Regardless of the origin of the tool, be it in-house or via a third-party, both development and validation remain inherently iterative processes—particularly in the still-maturing field of clinical digital pathology. These efforts should be viewed as scientific endeavors aimed at discovering and applying knowledge to real-world clinical systems in a rigorous and reproducible manner. Given the inherent complexity of biology and the variability encountered in routine practice, it is important to approach this work with humility and an understanding that the first iteration is unlikely to be perfect. Our own experience, shared here through the lens of a relatively simple diagnostic algorithm, reflects the culmination of more than three years of work by a mid-sized multidisciplinary team. However, the lessons learned through this effort have already significantly enhanced our capacity to develop and validate subsequent tools, several of which are currently under way within our institution.

One of the key lessons learned was the value of early technical prototyping to reduce uncertainty and manage risk. While the initial time estimates for developing and validating the algorithm were fairly accurate—taking approximately 3 to 4 months in total—the integration phase, particularly with the IMS, proved significantly more time-consuming than anticipated. Much of this delay stemmed from the fact that it was our first attempt at transmitting DICOM images in this manner, revealing unexpected technical hurdles. In retrospect, having an operational prototype of that integration earlier in the process would likely have accelerated development. Since model training and platform integration can proceed in parallel to some extent, earlier prototyping could have allowed for more efficient allocation of resources and a reduction in overall development time. Relatedly, we recognized the value of developing a site-specific software development kit (SDK) to centralize and standardize common functions such as slide downloading, result uploading, and logging. Without such a framework, even minor updates become increasingly burdensome, as having a multitude of differentially implemented tools creates an unsustainable maintenance load for **DEs** supporting a growing collection of tools.

Although not addressed in detail here, tool generalizability across institutions, scanners, and use cases remains an ongoing and significant challenge. Variability in staining protocols, scanner hardware, and slide preparation can substantially impact algorithm performance, often necessitating site-specific adaptations or fine-tuning. However, an even greater challenge may lie in the wide-ranging differences between hospitals themselves—including their infrastructure stacks, IMS and LMS systems, compute resources, clinical coding standards (e.g., stain code, diagnostic code), and reimbursement frameworks. Each hospital's setup, having evolved largely in isolation over many decades, often requires substantial rework of existing tools that may already be validated at a single site. This lack of standardization limits the potential for seamless "plug-and-play" deployment as tools transition between institutions. At each site they typically begin in a Research Use Only mode until locally validated, since Laboratory Developed Test (LDT) status cannot be transferred across laboratories. Therefore, establishing robust frameworks for assessing and adapting tools to diverse clinical environments will be essential to ensure safe, reliable, and effective broader deployment across healthcare systems. To that end, the Swiss Digital Pathology Consortium (SDiPath), has formed a Swiss Digital Pathology Algorithm Alliance (SDiPAA) working group, to foster inter-site communication and standardization of best practices[6].

While technical validation is essential, fostering user trust through transparency, education, and robust feedback mechanisms is equally critical. It is important to remember that the end users—primarily pathologists and, in some cases, laboratory technicians (e.g., during quality control)—must feel confident and comfortable when using these tools. This becomes especially vital as DP tools grow more sophisticated, moving beyond simple detection to complex tasks such as prognostic assessments or therapy response predictions based on image-based biomarkers. Ultimately, these users are responsible for signing off on reports and integrating the tool's outputs into clinical decision-making in collaboration with their clinical partners, whether they be oncologists, surgeons, or other specialists. Ensuring their acceptance and confidence requires more than just delivering superior technology; it demands clear communication, comprehensive training, and ongoing opportunities for users to provide feedback. Only through building trust and transparency can these tools be successfully adopted in routine clinical practice.

The need for strong institutional support and clear governance structures cannot be overstated when deploying such tools in clinical environments. Successful implementation relies not only on technical excellence but also on well-defined policies, oversight committees, and cross-disciplinary collaboration to manage risks, ensure compliance, and oversee ongoing performance monitoring. Institutional leadership must prioritize resource allocation, establish accountability frameworks, and foster a culture that embraces innovation while maintaining patient safety as the paramount concern. Without such comprehensive governance, even the most advanced tools risk underutilization, misuse, or unintended consequences, ultimately limiting their clinical impact and value.

4. Conclusions

We are at a pivotal moment in the evolution of digital pathology, where research-driven tools are beginning to enter routine clinical workflows. This transition brings tremendous opportunities for innovation, learning, and improved patient outcomes. However, it also underscores the critical responsibility to ensure these tools are safe, reliable, and clinically sound. This manuscript, with its associated supplemental templates, has proposed a general framework for validating digital pathology tools as laboratory-developed tests (LDTs), emphasizing patient safety at every stage. As the field matures, open sharing of methodologies, challenges, and best practices will be essential to accelerate development, streamline deployment, and maximize clinical impact. We encourage the broader community to contribute their insights so that, together, we can advance the safe and effective integration of these transformative technologies into everyday care.

[6] https://sdipath.ch/swiss-digital-pathology-algorithm-alliance/

References

1. Zia S, Yildiz-Aktas IZ, Zia F, Parwani AV. An update on applications of digital pathology: primary diagnosis; telepathology, education and research. *Diagnostic Pathology*. 2025;20(1):17. doi:10.1186/s13000-025-01610-9
2. Dawson H. Digital pathology – Rising to the challenge. *Frontiers in Medicine*. 2022;9. Accessed September 5, 2023. https://www.frontiersin.org/articles/10.3389/fmed.2022.888896
3. Nam S, Chong Y, Jung CK, et al. Introduction to digital pathology and computer-aided pathology. *J Pathol Transl Med*. 2020;54(2):125-134. doi:10.4132/jptm.2019.12.31
4. Janowczyk A, Zuo R, Gilmore H, Feldman M, Madabhushi A. HistoQC: An Open-Source Quality Control Tool for Digital Pathology Slides. *JCO Clin Cancer Inform*. 2019;3:1-7. doi:10.1200/CCI.18.00157
5. Chen Y, Zee J, Smith A, et al. Assessment of a computerized quantitative quality control tool for whole slide images of kidney biopsies. *The Journal of Pathology*. 2021;253(3):268-278. doi:10.1002/path.5590
6. Frei AL, Oberson R, Baumann E, et al. Pathologist Computer-Aided Diagnostic Scoring of Tumor Cell Fraction: A Swiss National Study. *Mod Pathol*. 2023;36(12):100335. doi:10.1016/j.modpat.2023.100335
7. Frei AL, Khan A, Oberson R, et al. Computer-aided tumor cell fraction (TCF) estimation by medical students, residents, and pathologists improves inter-observer agreement while highlighting the risk of automation bias. *Virchows Arch*. Published online July 4, 2025. doi:10.1007/s00428-025-04163-w
8. Bankhead P, Loughrey MB, Fernández JA, et al. QuPath: Open source software for digital pathology image analysis. *Sci Rep*. 2017;7(1):16878. doi:10.1038/s41598-017-17204-5
9. Jarkman S, Karlberg M, Pocevičiūtė M, et al. Generalization of Deep Learning in Digital Pathology: Experience in Breast Cancer Metastasis Detection. *Cancers*. 2022;14(21):5424. doi:10.3390/cancers14215424
10. Challa B, Tahir M, Hu Y, et al. Artificial Intelligence–Aided Diagnosis of Breast Cancer Lymph Node Metastasis on Histologic Slides in a Digital Workflow. *Modern Pathology*. 2023;36(8):100216. doi:10.1016/j.modpat.2023.100216
11. Leo P, Janowczyk A, Elliott R, et al. Computer extracted gland features from H&E predicts prostate cancer recurrence comparably to a genomic companion diagnostic test: a large multi-site study. *npj Precis Onc*. 2021;5(1):1-11. doi:10.1038/s41698-021-00174-3
12. Chen Y, Li H, Janowczyk A, et al. Computational pathology improves risk stratification of a multi-gene assay for early stage ER+ breast cancer. *NPJ Breast Cancer*. 2023;9(1):40.
13. Barisoni L, Lafata KJ, Hewitt SM, Madabhushi A, Balis UGJ. Digital pathology and computational image analysis in nephropathology. *Nat Rev Nephrol*. 2020;16(11):669-685. doi:10.1038/s41581-020-0321-6
14. Ni HM, Kouzy R, Sabbagh A, et al. The state of the art in artificial intelligence and digital pathology in prostate cancer. *Nat Rev Urol*. Published online August 4, 2025:1-16. doi:10.1038/s41585-025-01070-2
15. Leo P, Chandramouli S, Farré X, et al. Computationally Derived Cribriform Area Index from Prostate Cancer Hematoxylin and Eosin Images Is Associated with Biochemical Recurrence Following Radical Prostatectomy and Is Most Prognostic in Gleason Grade Group 2. *Eur Urol Focus*. Published online April 30, 2021:S2405-4569(21)00122-X. doi:10.1016/j.euf.2021.04.016
16. Williams BJ, Knowles C, Treanor D. Maintaining quality diagnosis with digital pathology: a practical guide to ISO 15189 accreditation. *J Clin Pathol*. 2019;72(10):663-668. doi:10.1136/jclinpath-2019-205944
17. Shukla SS, Pandey RK, Gidwani B, Kalyani G. Objectives of Qualification. In: Shukla SS, Pandey RK, Gidwani B, Kalyani G, eds. *Pharmaceutical Calibration, Validation and Qualification: A Comprehensive Approach*. Springer Nature; 2023:49-61. doi:10.1007/978-981-19-9002-1_5
18. Long RE, Smith A, Machotka SV, et al. Scientific and Regulatory Policy Committee (SRPC) paper: validation of digital pathology systems in the regulated nonclinical environment. *Toxicol Pathol*. 2013;41(1):115-124. doi:10.1177/0192623312451162
19. García-Rojo M. International Clinical Guidelines for the Adoption of Digital Pathology: A Review of Technical Aspects. *Pathobiology*. 2016;83(2-3):99-109. doi:10.1159/000441192
20. Genzen JR. Regulation of Laboratory-Developed Tests: A Clinical Laboratory Perspective. *Am J Clin Pathol*. 2019;152(2):122-131. doi:10.1093/ajcp/aqz096
21. Genzen JR, Mohlman JS, Lynch JL, Squires MW, Weiss RL. Laboratory-Developed Tests: A Legislative and Regulatory Review. *Clin Chem*. 2017;63(10):1575-1584. doi:10.1373/clinchem.2017.275164

22. Genzen JR, Miller LJ, Rets AV, Affolter KE. Laboratory-developed tests and in vitro diagnostics: A regulatory overview for anatomic pathology. *Am J Clin Pathol*. 2025;163(5):730-743. doi:10.1093/ajcp/aqae181
23. Bogdanoski G, Lucas F, Kern W, Czechowska K. Translating the regulatory landscape of medical devices to create fit-for-purpose artificial intelligence (AI) cytometry solutions. *Cytometry Part B: Clinical Cytometry*. 2024;106(4):294-307. doi:10.1002/cyto.b.22167
24. International Organization for Standardization. ISO 15189:2022 - Medical laboratories — Requirements for quality and competence. ISO. Accessed March 9, 2023. https://www.iso.org/standard/76677.html
25. Kahles A, Goldschmid H, Volckmar AL, et al. Regulation (EU) 2017/746 (IVDR): practical implementation of annex I in pathology. *Pathologie (Heidelb)*. 2023;44(Suppl 2):86-95. doi:10.1007/s00292-023-01274-6
26. Benjamens S, Dhunnoo P, Meskó B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. *NPJ Digit Med*. 2020;3:118. doi:10.1038/s41746-020-00324-0
27. Van Norman GA. Drugs, Devices, and the FDA: Part 2. *JACC Basic Transl Sci*. 2016;1(4):277-287. doi:10.1016/j.jacbts.2016.03.009
28. Advancements in Clinical Evaluation and Regulatory Frameworks for AI-Driven Software as a Medical Device (SaMD). *IEEE Open J Eng Med Biol*. 2024;6:147-151. doi:10.1109/OJEMB.2024.3485534
29. Aboy M, Crespo C, Stern A. Beyond the 510(k): The regulation of novel moderate-risk medical devices, intellectual property considerations, and innovation incentives in the FDA's De Novo pathway. *NPJ Digit Med*. 2024;7(1):29. doi:10.1038/s41746-024-01021-y
30. Liotta LA, Petricoin EF. Regulatory approval pathways for molecular diagnostic technology. *Methods Mol Biol*. 2012;823:409-420. doi:10.1007/978-1-60327-216-2_27
31. Vanstapel FJLA, Orth M, Streichert T, et al. ISO 15189 is a sufficient instrument to guarantee high-quality manufacture of laboratory developed tests for in-house-use conform requirements of the European In-Vitro-Diagnostics Regulation: Joint opinion of task force on European regulatory affairs and working group accreditation and ISO/CEN standards of the European Federation of Clinical Chemistry and Laboratory Medicine. *Clinical Chemistry and Laboratory Medicine (CCLM)*. 2023;61(4):608-626. doi:10.1515/cclm-2023-0045
32. Koelzer VH, Grobholz R, Zlobec I, Janowczyk A. Update on the current opinion, status and future development of digital pathology in Switzerland in light of COVID-19. *Journal of Clinical Pathology*. 2022;75(10):687-689. doi:10.1136/jclinpath-2021-207768
33. Janowczyk A, Zlobec I, Walker C, et al. Swiss digital pathology recommendations: results from a Delphi process conducted by the Swiss Digital Pathology Consortium of the Swiss Society of Pathology. *Virchows Arch*. 2024;485(1):13-30. doi:10.1007/s00428-023-03712-5
34. Srivastava A, Bhardwaj S, Saraswat S. SCRUM model for agile methodology. In: *2017 International Conference on Computing, Communication and Automation (ICCCA)*. 2017:864-869. doi:10.1109/CCAA.2017.8229928
35. Schwaber K, Beedle M. *Agile Software Development with Scrum*. 1st ed. Prentice Hall PTR; 2001.
36. Baidoshvili A, Khacheishvili M, van der Laak JAWM, van Diest PJ. A whole-slide imaging based workflow reduces the reading time of pathologists. *Pathology International*. 2023;73(3):127-134. doi:10.1111/pin.13309
37. Fraggetta F, L'Imperio V, Ameisen D, et al. Best Practice Recommendations for the Implementation of a Digital Pathology Workflow in the Anatomic Pathology Laboratory by the European Society of Digital and Integrative Pathology (ESDIP). *Diagnostics*. 2021;11(11):2167. doi:10.3390/diagnostics11112167
38. Wu Z, Guo K, Luo E, et al. Medical long-tailed learning for imbalanced data: Bibliometric analysis. *Computer Methods and Programs in Biomedicine*. 2024;247:108106. doi:10.1016/j.cmpb.2024.108106
39. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, et al. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. *JAMA*. 2017;318(22):2199-2210. doi:10.1001/jama.2017.14585
40. International Organization for Standardization and International Electrotechnical Commission. ISO/IEC 31010: Risk Assessment Techniques. Published online 2019. https://www.iso.org/standard/72140.html
41. International Organization for Standardization. ISO 14971: Medical Devices – Application of Risk Management to Medical Devices. Published online 2019. https://www.iso.org/standard/72704.html

42. International Organization for Standardization. ISO 31000: Risk Management – Guidelines. Published online 2018. https://www.iso.org/standard/65694.html
43. Fónyad L, Krenács T, Nagy P, et al. Validation of diagnostic accuracy using digital slides in routine histopathology. *Diagn Pathol*. 2012;7:35. doi:10.1186/1746-1596-7-35
44. Båth M, Månsson LG. Visual grading characteristics (VGC) analysis: a non-parametric rank-invariant statistical method for image quality evaluation. *Br J Radiol*. 2007;80(951):169-176. doi:10.1259/bjr/35012658
45. International Organization for Standardization. ISO/IEC 17025:2017 - General requirements for the competence of testing and calibration laboratories. ISO. January 26, 2021. Accessed March 9, 2023. https://www.iso.org/standard/66912.html
46. Shah M, Polónia A, Curado M, Vale J, Janowczyk A, Eloy C. Impact of Tissue Thickness on Computational Quantification of Features in Whole Slide Images for Diagnostic Pathology. *Endocr Pathol*. 2025;36(1):10. doi:10.1007/s12022-025-09855-2
47. Westgard JO, Westgard SA. Quality control review: implementing a scientifically based quality control system. *Ann Clin Biochem*. 2016;53(1):32-50. doi:10.1177/0004563215597248
48. Westgard JO, Barry PL, Hunt MR, Groth T. A multi-rule Shewhart chart for quality control in clinical chemistry. *Clin Chem*. 1981;27(3):493-501.
49. Nelson LS. Interpreting Shewhart $\bar{X}$ Control Charts. *Journal of Quality Technology*. Published online April 1, 1985. Accessed August 21, 2025. https://www.tandfonline.com/doi/abs/10.1080/00224065.1985.11978945
50. Nelson LS. The Shewhart Control Chart—Tests for Special Causes. *Journal of Quality Technology*. Published online October 1, 1984. Accessed August 21, 2025. https://www.tandfonline.com/doi/abs/10.1080/00224065.1984.11978921
51. Jennings I, Kitchen DP, Woods TAL, Kitchen S, Walker ID. Emerging Technologies and Quality Assurance: The United Kingdom National External Quality Assessment Scheme Perspective. *Seminars in Thrombosis and Hemostasis*. 2007;33:243-249. doi:10.1055/s-2007-971810
52. Dunn C, Brettle D, Hodgson C, Hughes R, Treanor D. An international study of stain variability in histopathology using qualitative and quantitative analysis. *Journal of Pathology Informatics*. 2025;17:100423. doi:10.1016/j.jpi.2025.100423

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.