FRIENDS of CANCER RESEARCH

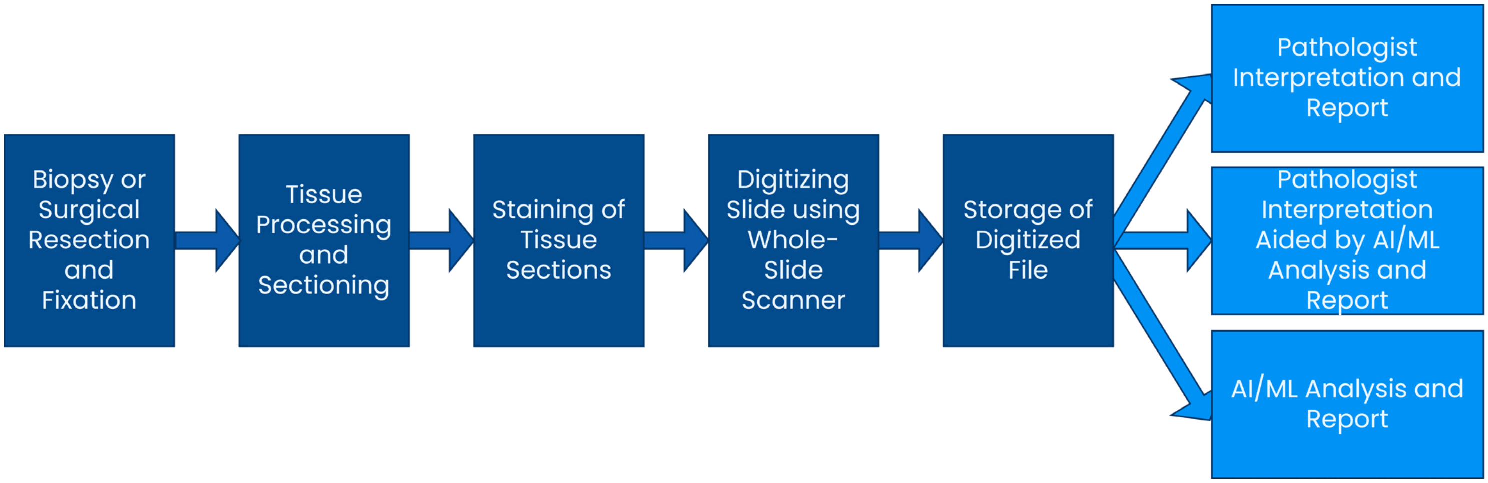
A FRIENDS OF CANCER RESEARCH WHITE PAPER

Supporting the Application of Computational Pathology in Oncology

Introduction

Biological heterogeneity of cancers causes tumors to respond differently to the same treatments. Thus, there is a compelling need to appropriately diagnose patients and identify relevant biomarkers for oncology treatments in both clinical practice and trials. Digital pathology is an emerging application in oncology drug development and clinical care, which allows for whole-slide image creation for storage, viewing, analyses, and interpretation. Digitized images are used directly by pathologists for biomarker interpretation, cellular annotation, and diagnosis. These images can also be used to support development of computational pathology platforms that utilize techniques such as artificial intelligence (AI) and machine learning (ML) to analyze and measure specific image elements, such as subvisual morphological patterns and phenotypes, identify features, and generate reproducible and structured data. These AI and ML platforms referred to in aggregate as computational pathology, may establish novel biomarkers, aid in quantifying prognostic and predictive biomarkers currently assessed or categorized by a pathologist, and expedite diagnosis or pathological scoring, all of which may go towards identifying and selecting patients for oncology treatments. Digital and computational pathology encompass several linked workflow components including both the digitization of the whole slides as well as the platforms for analysis (**Figure 1**).

Figure 1: Workflow Components of Digital and Computational Pathology


Biopsy or Surgical Resection and Fixation
Tissue Processing and Sectioning
Staining of Tissue Sections
Digitizing Slide using Whole-Slide Scanner
Storage of Digitized File
Pathologist Interpretation and Report
Pathologist Interpretation Aided by AI/ML Analysis and Report
AI/ML Analysis and Report

Thank You to Our Contributors

Nick Anderson, PathAI

J. Carl Barrett, formerly AstraZeneca, University of North Carolina at Chapel Hill

Anne-Laure Bauchet, Sanofi

Vipul Baxi, Bristol Myers Squibb

Nike Beaubier, Tempus Labs, Inc.

Daniel N. Cohen, GlaxoSmithKline

Suzana Couto, Neomorph, Inc.

Tathagata Dasgupta, 4D Path Inc.

Megan Doyle, Amgen

Dorothy French, Amgen

Brandon D. Gallas, U.S. FDA Center for Devices and Radiological Health

Emre Gulturk, Paige AI

Mark Gustavson, AstraZeneca

Sarah Hersey, Bristol Myers Squibb

Kimary Kulig, Kulig Consulting and My Biomarker Navigator

Alexander Lazar, MD Anderson Cancer Center

Jochen K. Lennerz, Massachusetts General Hospital

Mike Montalto, PathAI

Thomas Mrowiec, The Healthcare business of Merck KGaA, Darmstadt, Germany operates as EMD Serono in the U.S. and Canada

Satabhisa Mukhopadhyay, 4D Path Inc.

E. Tom Richardson, Merck & Co., Inc.

Roberto Salgado, University Hospital of Antwerp

Alain Silk, Tempus Labs, Inc.

Nino Sireci, Loxo@Lilly

Martin Stumpe, Tempus Labs, Inc.

Jian Wang, AstraZeneca

This paper reflects discussions that occurred among stakeholder groups, including the U.S. FDA. The topics covered in the paper, including recommendations, therefore, are intended to capture key discussion points. The recommendations provided should not be used in lieu of FDA published guidance or direct conversations with the Agency about a specific development program. This paper should not be construed to represent FDA's views or policies.

Objectives

Computational pathology has the potential to generate novel insights and biomarkers, and provide greater accuracy, reproducibility, and standardization of pathology-based features to aid in oncology drug development. Friends of Cancer Research (*Friends*) assembled a multi-stakeholder group of experts including government officials, computational pathology platform developers, academic pathologists and researchers, and biopharmaceutical industry members to outline proposals that facilitate robust development of computational pathology platforms for oncology drug development. The objectives of this group were to:

- Characterize current and future uses of computational pathology in oncology drug development and how they can facilitate clinical research.
- Identify the challenges in current drug and diagnostic co-development and articulate lessons learned to circumvent these in computational pathology.
- Provide proposals to facilitate robust development of computational pathology platforms for oncology drug development, including:
    1. Outline input and platform performance characteristics to report for optimized transparency.
    2. Establish a risk classification framework to inform evidentiary needs and performance criteria.
    3. Establish common reference standards and repositories of reference materials to support future platform development and cross-validation of platforms.

Uses of Digital and Computational Pathology in Oncology Drug Development

Digital pathology currently aids oncology drug development in operational and logistical tasks by supporting remote sharing of slides, storage of data for future analyses, and promoting efficient training of pathologists (**Table 1**). However, this white paper will focus on the use of AI/ML and other (image-based) computational pathology methods into a digital pathology workflow. Computational pathology can identify and quantify features from image data beyond human analytic capability. As such, computational pathology can establish novel biomarkers and improve current assessment of pathological features that would not otherwise be produced through conventional pathological evaluation. While this white paper focuses on the use of computational pathology in oncology, there is promise in other applications such as in non-alcoholic steatohepatitis (NASH)[1], inflammatory bowel disease (IBD)[2], and other diseases, and the proposals described herein may be relevant to these other applications.

Computational Pathology Applications in Oncology Drug Development

There is a spectrum of applications for digital and computational pathology throughout oncology drug development, including early discovery, pre-clinical and translational research, early phase trials, registrational trials, post-market/clinical use (**Table 1**). While some applications are currently in use in oncology drug development (e.g., digitization of tumor slides for future biomarker correlation to outcomes), others are currently in various stages of development (e.g., prediction of biomarker status) or are not yet ready for trials or clinical use (e.g., exploratory endpoints). Further, while each phase of development is depicted as distinct, the long-term goal for an integrated computational pathology workflow should be considered as it will determine the types of evidence and validation necessary for the platform. For example, a computational pathology platform used in exploratory translational research or early phase trials may not be intended for use in later phase trials or clinical care, while the goal for a platform used in a late phase trial may be to develop a companion diagnostic (CDx) for use in the post-market setting. Therefore, as some platforms may be used in several phases of drug development, developers should consider the various validation needs of these uses early in the platform development process.

Table 1. Examples of Potential Uses of Digital Pathology Workflows in Oncology Drug Development

	Drug Development Phase				
Use of Digital Pathology	**Pre-Clinical/Research**	**Exploratory/ Translational**	**Early Phase Prospective Trials (I/IIa)**	**Late Phase Prospective Trials (IIb/III)**	**Post-Launch/Clinical Use**
Digital Pathology Workflow	• Reading/ interpretation of pharmacology models and toxicology studies from digital images • Peer review of toxicology digital slides from non-GLP or GLP studies • Qualitative assessment of multiplex biomarker assays • Evaluation of drug distribution (PD)	• Pathologist manual annotation and semi-quantitative scoring of images for biomarker prevalence, discovery and validation • Exploration of manual/ human interpretable features as biomarkers (i.e., mitosis count, IHC scores, % immune cell infiltrate)	• Archiving slides and retrospective qualitative or semi-quantitative biomarker analysis • Review of biomarker status for trial enrollment • Pharmacodynamic biomarker measurements using digital images • Promote pathologist peer review, document decision-making process, and improve quality	• Visual review for trial enrollment: biomarker measurement as inclusion criteria • Foster pathologist peer review, document decision process, and improve quality • Measure clinical trial endpoint (e.g., pCR) using digital images	• Pathologist remote disease diagnosis, second opinion consult, tumor boards, pathologist education/board certification • Store images for future analysis • Pre-screen for selection of treatment • Promote efficient training of pathologists
Computational Pathology Analyses Based on Digital Pathology Workflow	• Toxicology/ veterinary pathology read/ count assistance on digital images • Investigation of novel biomarkers, spatial characterization • Evaluation and quantitation of drug distribution (PD) • Quantitative assessment and interpretation of multiplexed biomarkers	• Retrospective analysis of clinical trial data to discover new biomarker spatial correlations or image features with prognostic or predictive value • Enables quantification of histologic feature to create a unique biomarker(s) • Data driven biomarker scoring	• PD biomarker quantification • Tool for clinical trial enrollment • Exploration of predictive biomarkers with more precise and continuous cutoffs • Support, guide, and monitor pathologist scoring	• Trial enrollment: Biomarker measurement as inclusion criteria • Evaluating exploratory endpoints • Clinical trial outcomes assessment (e.g., pCR, RCB) • Discovery of biomarkers in TME that may correlate to efficacy and/or safety • Support, guide, and monitor pathologist scoring	• Biomarker assessment for targeted treatment identification • Pre-screen followed by confirmatory testing for treatment selection • Support, guide, and monitor pathologist scoring

GLP: Good Laboratory Practice, IHC: immunohistochemistry, PD: pharmacodynamics, pCR: pathologic complete response, RCB: residual cancer burden, TME: tumor microenvironment.

Platform Description and Use

For each computational pathology application, it will be useful to have a clear description of what it does and how it will be used, including the level of reliance on the output. This description will impact the evidentiary needs for validation. Some platforms may improve existing manual processes and assist the pathologist by enhancing or providing efficiencies (e.g., image quality control and low-level tasks like object or feature recognition, counting, and segmentation). Results generated from platforms that assist the pathologist in routine tasks or workflow support rely on the pathologist's final judgment and "sign-off."

However, computational pathology platforms are likely to provide novel insights that go beyond traditional histopathology assessments of pathologists, such as novel quantitative biomarker discovery or detection of spatial relationships between multiple biomarkers. These platforms may be further divided into those that produce an output that can be independently validated by a pathologist or other orthogonal method (e.g., DNA/ RNA sequencing) and those with an output that cannot be independently generated by a pathologist or other mechanism (i.e., "black box"). The ability to verify a platform's output by an alternate method may impact the level of evidence necessary to support its use. For example, in a clinical setting, a platform used as a pre-screen for a biomarker followed by confirmatory testing with a gold standard methodology (e.g., sequencing) may have different evidentiary needs for validation than if the output is the sole determinant for a patient receiving treatment.

Challenges in the Current Diagnostic and Drug Development Landscape

Currently, oncology diagnostic development for a predictive biomarker generally follows the paradigm where a single test or assay defines a single biomarker for a specific drug in a drug-diagnostic co-development model. This paradigm usually results in the U.S. Food and Drug Administration (FDA) approval of a CDx, which provides information that is essential for the safe and effective use of the corresponding drug or biological product.[3] However, in clinical practice, additional assays, including laboratory developed tests, are often independently developed for the same biomarker and may be used in lieu of the approved CDx. As a result, a diverse set of assays with varying performance and predictive ability will be in use to detect the same biomarker to assist with treatment selection. Without robust data about performance and comparability across assays, this may result in confusion and lack of confidence in the diagnostics. This concern is reflected in FDA's recently released final guidance: Oncology Drug Products Used with Certain In Vitro Diagnostic Tests: Pilot Program.[4] The pilot aims to increase transparency regarding performance characteristics for tests used to identify biomarkers for selection of oncology drug products.

Previous biomarker alignment and concordance demonstration projects on programmed death ligand 1 (PD-L1) immunohistochemistry (IHC)[5,6] and tumor mutational burden (TMB),[7] highlight disparate methodologies in biomarker assessment across available assays, with various clinical cutoffs used for reporting results and supporting treatment decision-making, possibly leading to disparate care for patients. The discordance seen in these projects provides lessons learned for improved prospective harmonization and transparency in the pre-market stage for computational pathology.

Disparate platforms and methodologies for biomarker assessment may make comparing computational pathology platforms challenging unless harmonization efforts exist. Currently, there is not a simple mechanism for comparing the performance of the multiple available computational pathology platforms assessing the same biomarker. However, addressing this gap could support broader clinical use of computational pathology derived biomarkers, in addition to supporting broader regulatory authorizations outside of the single-platform, single-drug paradigm. Outlining best practices for validation studies, identifying and reporting key input and platform performance characteristics, and establishing standards to support the consistent performance of different computational pathology platforms can address concerns around test accuracy, reliability, and comparability.

Proposals for Robust Use of Computational Pathology in Drug Development

The following proposals for computational pathology development and use in oncology drug development will help to ensure the development of robust and well characterized platforms while enabling innovation.

Proposal 1: Input and Platform Performance Characteristics Reported for Optimized Transparency

Transparent methodology, input requirements, output scale and units, and performance characteristics will aid drug developers in identifying platforms that are appropriate for a given use case and aid platform developers and regulatory agencies in validating and evaluating robustness of platforms.

To increase transparency of the platform's methodology, the design and testing of the algorithm should be described, as well as the types of data used as training and validation sets, how the datasets were used, and how the datasets are related to the distribution of outputs. This information can support critical evaluation of the algorithm development and validation process, ensuring that datasets capture real-world parameters and are representative of the heterogeneity of treatment settings, patients, and tumor characteristics. Transparency in the baseline performance characteristics of a computational pathology platform for specific use cases can also help harmonize future development efforts resulting in high quality performance irrespective of the platform and developer.

Input Parameters to Consider in Development and Reporting

Given the multiple workflow components involved in computational pathology (**Figure 1**), it is important to clearly state and define the multiple input parameters that can influence the platform's robustness and performance. Defining the input parameters encourages more robust and transparent platform development and use and can be used to develop quality metrics, which can be applied across platforms. In turn, this can aid in the development of pathology practice standards to ensure consistent practice irrespective of where tissue is collected and processed, scanning devices used, and what platform is used. The two relevant categories of input parameters to define for computational pathology platforms are tissue processing (slide preparation) and image acquisition (scanning). Within these categories, key parameters to consider when evaluating input quality and the robustness of a platform for a given intended use are listed in **Table 2** and are informed by FDA guidance on the technical performance assessment of digital pathology whole slide imaging devices.[8] Each input parameter can be described or measured, and the appropriate specifications and quality metrics required will depend on the platform's application. Certain input parameters may be easier to control for quality (e.g., slide age) than others (e.g., tissue artifacts) and future work is needed to define quality metrics. The input parameters described in **Table 2** are intended to help computational pathology developers directly by describing the specifications of their platform with regards to variation in the input parameters. This can also help drug developers understand and evaluate the capabilities and limitations of algorithms when considering their potential use in supporting drug development.

Table 2: Input Parameters to Define and Evaluate*

Parameter	Definition	Considerations
Tissue and Slide Processing		
Means of Tissue Acquisition	The type of tissue sample (e.g., excisional, core needle, fine needle aspirate, cytology, etc.)	Relevant to sampling bias and potentially algorithm performance; Relevant to valid use of platform per sample type
Tissue Sample Origin	The origin of the tissue sample (e.g., primary tumor vs. metastatic lesion vs. lymph node) as well as organ site	Relevant to valid use of platform per tissue origin
Tissue Processing	Specific steps for processing tissue (e.g., freezing, type of fixative, fixation time, etc.)	May impact tissue quality or usability with the platform; Some artifacts are specific to tissue processing and may affect the quality of the stain applied; Platforms may work differently on fresh frozen vs. FFPE tissue, etc.
Glass Slide Type	A description of the slide including thickness and slide material	Slide type may impact coloring and depth of the tissue that is measured and may affect opacity
Tissue Thickness	Acceptable range of tissue thickness in microns	May affect image quality and characteristics such as color and optic density of features as well as the number of cells analyzed
Tissue Area	Minimum and maximum tissue area recommended for reliable and reproducible analysis, including tissue area alone as well as tumor content (as a percentage of total area)	There may be a minimum amount of tumor tissue/ tumor cells required for the analysis
Tissue Folds/ Tears	Description of any tissue folds or tears in the tissue, and how these are handled	Presence of tissue folds/tears which may cause out-of-focus digitization, in addition to the reduced usability of areas that are directly affected
Surgical Ink/ Pigments	Presence of surgical ink or other markings, and how these are handled	Markings may impact the software, and may result in false counts and misidentification of features
Other Tissue Artifacts	Other relevant artifacts (e.g., tissue lifting, incomplete decalcification, dust or surgical glove powder, bubbles, over fixation, improper dehydration, tissue bloating, etc.[9]) and how they are handled	Various artifacts may impact analyses when present
Tissue Age	The recommended duration between when slides are cut and stained	May impact the stability of some features and affect stain characteristics such as intensity and color
Slide Age	The recommended duration between tissue staining and scanning	The time post-staining may impact intensity and quality of the slide (e.g., chromogen stability, diffusion of chromogenic dyes, fading of fluorescent dyes, etc.)

*Concepts in this table may be specific to currently existing technologies (e.g., IHC). As emerging technologies evolve (e.g., multiplex immunofluorescence, RNA mass spectrometry, etc.) the input parameters may also evolve depending on the technology.

Parameter	Definition	Considerations
Antibody Used	The antibody used for staining including clone, company, and catalog number	The type, batch, and age of antibody used may impact staining results
Staining Conditions	The staining conditions, such as incubation, blocking, etc.	Staining conditions may alter the staining intensity and results[10]
Slide Storage	The manner and environment in which the physical slides are stored	Storage conditions (e.g., oxygen, humidity, sun or heat exposure) may impact staining results and/or tissue
Image Acquisition		
Scanner Hardware and Software Versions	Description of scanner hardware and software versions	Differences in hardware (e.g., optics) as well as software (e.g., pre/post-processing, color normalization, or application programming interface) can impact the algorithm performance
Scanner Software Configurable Parameters	Description of configurable parameters in the scanner software and the actual values, or acceptable ranges, which should be used during the scanning operation	Differences in scanner software configurable parameters (e.g., exposure and saturation) can impact the algorithm performance
Slide Viewer Used	Software and version used for slide viewing	Relevant to ability to use platform with different slide viewers and screens
Type of Image Files	Description of acceptable file formats and compression, and use of single plane images or image stacks	Relevant to whether image files can be appropriately processed by the algorithm
Region of Interest Selection	Information on whether the whole tissue, whole tumor area, or specific fields of view (including size) are used by the algorithm	The type of region may affect how algorithms are trained and their applicability to different tissue types
Magnification	The acceptable range of magnification of the digitized slide	Relevant to the use of the platform at different magnifications
Resolution	Specified magnification for image acquisition (e.g., 100x, 200x, 400x) and any requirements related to pixel resolution (expressed as micrometers per pixel)	Algorithms may require specific magnification during image acquisition and specific pixel density/resolution to identify features
Color	Details of the color processing, such as white-balance or contrast settings, which result in hue, saturation, brightness of the image; metrics for acceptable color settings and characteristics should be reported (with ranges of acceptability or a description of the color normalization procedure if used)	Algorithms can be sensitive to variations in color and contrast
Focus Quality	The focus quality required by the algorithm and a metric for acceptable focus quality	Focus quality can impact algorithms and should be quantified globally or locally as appropriate

Further, specifying performance/operating boundaries for the preprocess components of the workflow (e.g., scanners) will support use within the validated workflow. An appropriate description of the performance/operating boundaries may enable evaluation of the extent and conditions under which two different platforms used for the same purpose might produce similar results. The scanner model(s) and specific scanner configuration and acquisition protocol used for the training and testing of the computational pathology platform should be explicitly stated.

The specific parameters and acceptable ranges and values will depend on the computational pathology application. This includes the interaction of a human operator with the platform's output. For example, acceptable ranges may be wider when a human operator can independently check the output of the software or if it is being used to help direct a pathologist to examine certain slide areas, and narrower if the results cannot be independently verified by a human user.

Appendix 1 applies the reporting of input parameters to hypothetical use cases of computational pathology platforms. Some parameters, such as slide age, may be common across different use cases, whereas other input parameters may vary depending on the use case. Understanding the commonality or variability across use cases can also inform prioritization, by identifying parameters that may be relevant for model development and performance assessment for all studies.

Performance Characteristics and Assessment

Identifying and reporting key performance characteristics for computational pathology platforms will increase transparency, provide study designs and assessment methods for others to follow, and inform performance expectations for other quality and robust platforms. This may also increase confidence in using independently developed and validated platforms for a common purpose. Alignment is needed on standardized methods to report these characteristics to aid in transparency and the comparison.

Guidelines for establishing performance of AI or image analysis methods in computational pathology are limited. The FDA Center for Devices and Radiological Health (CDRH) has cleared one computational pathology device under a regulation that defines a broad intended use: "A software algorithm device to assist users in digital pathology […] to provide information to the user about presence, location, and characteristics of areas of the image with clinical implications."[11] The special controls provided outline what information should be included in a Class II marketing submission for performance assessment, and the decision summary of the FDA-authorized device includes a summary of the scientific evidence that served as the basis for FDA's decision.[12] Other relevant FDA resources to understand key performance characteristics include regulations, reclassification orders, decision summaries, guidance documents, and other written works on the regulation of software as a medical device (SaMD) in areas other than pathology.[13]

The platform description, what it does and how it will be used, will impact its key performance characteristics. The College of American Pathologists published recommendations for the validation of whole slide imaging systems in clinical practice[14] and further provides resources related to the validation of image analysis platforms in clinical practice.[15] The Digital Pathology Association also broadly noted both hurdles and solutions for implementing computational pathology and validating these platforms.[16] **Appendix 2** provides considerations for validation study designs, as well as examples of how the design elements were met in a few computational pathology validation studies.

Performance should be assessed on a dataset not used in the platform's development or training and is representative of the clinical population the platform is intended to evaluate to offer an unbiased assessment of performance.[17] Performance characteristics may be influenced by details such as true biomarker prevalence in the study population, as well as training and testing data sources and sampling. These details and their impact on performance should be described. The input parameters highlighted in **Table 2** will also impact performance and should be considered. Key performance characteristics should be evaluated in a manner consistent with what the platform does and how it will be used. This may include evaluation by standalone performance, a measure of the platform performance with little to no input or interpretation from the clinical end user, multi-reader multi-case study performance, and/or a measure of performance with interaction from the clinical end user or multiple end users. The end user involved in validation should be different than the user(s) involved in training.

Further, focusing on "explainable AI" (i.e., methods allowing for a representation of the input parameters used by the algorithm such as overlays of high attention areas or cell segmentation), may aid in the interpretability of "black box" algorithms. This interpretability could have two functions: allow for review of the impact of preanalytical variables, such as those detailed in **Table 2**, on the quality of the results, and bring additional confidence in the results to the end user.

Establishing Performance Comparisons

When performance comparisons to a "ground truth" or reference standard are possible and desirable, various study designs can be employed and careful consideration should be given to the method for establishing ground truth. Several methods exist for using pathologist interpretations as the reference standard, including using the original sign-out diagnosis, single readers, or consensus panels. Additionally, the concordance within pathologists should be considered when comparing concordance between a pathologist's interpretation and the platform's output, as there is also heterogeneity within pathologists' readings. Poor concordance within pathologists may indicate that multiple pathologists are needed to determine the reference standard. Also, the within-pathologist concordance may provide a performance criterion for model-pathologist concordance, assuming they are measured the same way.

In cases where comparison to a pathologist score/interpretation is not desired or possible, orthogonal methods that generate biological outputs such as gene or protein expression may be an acceptable comparison. For novel biomarkers, or in other cases where no orthogonal methods exist, native or contrived reference materials with a known or well-characterized status may be used as a comparison. Ultimately, establishing performance in relation to clinical characteristics or outcomes may be highly desirable, but is not always practical for certain use cases.

To compare the performance of several different computational platforms that report the same output, establishing a reference dataset with defined ground truth and pre-defined analysis methods is recommended. There is precedent for such approaches, such as the CAMELYON16 grand challenge,[18] in which several model developers created models to detect lymph node metastatic disease and then tested the performance of their models on a single validation dataset.

Reporting Quantitative Measurements

Platforms may measure or define the biomarker of interest differently and direct cross comparisons may be challenging, especially with binary outputs. Although dichotomization of continuous biomarkers to a binary reading (e.g., high vs. low) by establishing a cutoff correlated to a clinical feature or outcome is frequently used in registrational trials for drug-CDx approval, the quantitative biomarker value (i.e., continuous scale) is often provided by computational pathology platforms and should be retained. Binary readings are often clinically desirable for ease of interpretation. How cutoffs are defined and derived should be encouraged. As part of the effort to establish an adequate cut off, there should be clear understanding of the variability in measurement surrounding the cutoff and reporting of the relevant range of quantitative measurements, their use within a final platform, and their relationship (if any) to outcomes in clinical trial data.

Proposal 2: Establish a Risk Classification Framework to Inform Evidentiary Needs and Performance Criteria

Adequate evidence generation, in the form of analytical and clinical validation, is needed to support the use of computational pathology platforms in oncology drug development. Further, a risk-based framework can support and inform this evidence generation and establishment of performance criteria across platforms and intended uses. Regulatory flexibilities are critical to encourage innovation and applying a risk-based approach will build an understanding of when flexibility is appropriate, what types of evidence are needed for computational pathology use in clinical trials and supporting regulatory approval, and regulatory pathways associated with a given platform.

Current Regulatory Classification and Pathways for Marketing

Regulatory agencies have applied existing risk classification systems for medical devices and diagnostics to digital pathology platforms. This paper focuses on the U.S. regulatory pathways, but depending on the intended use outside of the U.S., additional regulatory requirements should be considered in development (e.g., IVDR regulations). Diagnostic tests and digital pathology platforms are regulated based on their risk classification (i.e., Class I-III FDA designations), which helps inform the performance and reporting requirements.

Certain digital pathology platforms have been regulated as "Whole Slide Imaging" systems. In the U.S., these have largely been regulated as moderate-risk, class II devices requiring clearance of a 510(k) to be marketed.[19] FDA issued recommendations regarding technical performance testing that should be completed to support a marketing submission for a whole slide imaging system.[8] FDA has also regulated some AI/ML platforms as moderate-risk, class II devices, and issued special controls for these.[12] Further, the FDA, Health Canada, and United Kingdom's Medicines and Healthcare products Regulatory Agency (MHRA) have put forth 10 guiding principles to inform Good Machine Learning Practices (GMLP) for medical devices using AI/ML, which could be applicable to computational pathology devices.[20] Further, FDA's Drug Development Tool program[21] and Medical

Device Development Tool program[22] offers opportunities for public health stakeholders to pursue FDA qualification of digital and computational pathology tools.

Table 3 summarizes the current U.S. regulatory pathways and some applicable regulatory controls for specific defined use cases. Of note, this is not an exhaustive list of all regulatory controls that apply to developing and marketing a computational pathology platform. In addition to existing use cases and regulatory controls, a risk-based approach should be applied to future, not yet established use cases. **Table 4** suggests example future use cases and a potential risk-based approach to regulating them. However, it is important to note there are currently no cleared or approved devices for these uses, and the FDA may not agree with the relationships between use cases and regulatory controls. Readers are encouraged to engage the FDA early and often, including with a Q-submission or a pre-IND to inquire about use cases and regulatory pathways.[23]

Table 3: Potential Regulatory Pathways and Regulatory Controls for Marketing Digital Pathology Platforms by Intended Use[24]

Device Name, Risk Classification, Regulatory Pathway	Intended Use Summary	Potential Development and Evidence Generation Expectations
Software Algorithm Device To Assist Users In Digital Pathology[25] Class II, 510(k)	Intended to aid a healthcare provider in determining a pathology diagnosis, provide information to the user about presence, location, and characteristics of areas of the image with clinical implications	Design Controls 21CFR820 Quality Management System ISO13485 Demonstrate substantial equivalence to a predicate Good ML Practices See Special Controls for Evidence Generation Expectations
Digital Pathology Image Viewing And Management Software[26] Class II, 510(k)	Intended for viewing and management of digital images of scanned surgical pathology slides, as an aid to the pathologist to review and interpret these digital images for the purposes of primary diagnosis	Design Controls 21CFR820 Quality Management System ISO13485 Demonstrate substantial equivalence to a predicate Bench testing[8] Clinical Validation Study comparing to reference standard or manual read
Digital Pathology Display[27] Class II, 510(k)	Intended for in vitro diagnostic use to display digital images of histopathology slides acquired by whole-slide imaging scanners that are used for review and interpretation by trained pathologists	Design Controls 21CFR820 Quality Management System ISO13485 Demonstrate substantial equivalence to a predicate Good ML Practices Bench testing[8] Display Equivalency Study
Whole Slide Imaging System[28] Class II, 510(k)	Intended to aid the pathologist in review and interpretation of digital images of surgical pathology slides by automating digital slide creation, viewing, and management	Design Controls 21CFR820 Quality Management System ISO13485 Demonstrate substantial equivalence to a predicate Good ML Practices Bench testing[8] Clinical Validation Study Human factors study

Table 4: Example Future Use Cases with Potential Regulatory Pathways and Controls*

Potential Intended Use	Anticipated Risk Classification & Regulatory Pathway	Possible Regulatory Controls	Potential Development and Evidence Generation Expectations
Intended for use as a companion diagnostic	De Novo (Class II) or PMA (Class III) based on risk	Premarket Approval	Design Controls 21CFR820 Quality Management System ISO13485 Good ML Practices Demonstrate reasonable assurance of safety and effectiveness (AV, CV) Analytical validation studies (e.g., sensitivity, specificity, precision, accuracy, limit of detection, etc.) Clinical validation studies
Prescreening (with confirmation by another central test or CDx)	De Novo (Class I or II), due to lack of existing product code or classification regulation	Special Controls	Design Controls 21CFR820 Quality Management System ISO13485 Demonstrate substantial equivalence to a predicate Bench testing[8] Other data and controls, as requested by regulators, e.g.: Good ML practices Clinical Validation or Concordance Study
Automated computational digital pathology system for scanning, converting, reading, and detecting/measuring a biomarker on a pathology slide, with oversight and confirmation of output by a physician.	De Novo (Class I or II), due to lack of existing product code or classification regulation	Special Controls	Design Controls 21CFR820 Quality Management System ISO13485 Good ML Practices Demonstrate reasonable assurance of safety and effectiveness (AV, CV) Analytical validation studies (e.g., sensitivity, specificity, precision, accuracy, limit of detection, etc.) Clinical validation studies Usability study

*These are suggestions for a risk-based approach but have not been formally established via FDA classification decisions, clearances, or approvals to date.

Use of Computational Pathology Platforms in Clinical Trials

Currently, FDA guidance has not specifically addressed the use of computational pathology methods in clinical trials. Although others have published information and recommendations that may be helpful,[29] regulatory expectations for use of computational pathology platforms in oncology trials are still nascent. Recent publications have highlighted regulatory considerations for medical imaging AI/ML devices, including the existing regulatory pathways.[30,31] Gaps in knowledge and test methods, and the novelty, pose a challenge for identifying regulatory expectations.[32] To this end, this proposal seeks to build on prior regulatory resources by providing suggestions for a risk-based approach to these items, to advance the use of computational pathology platforms in oncology drug development.

While FDA has not opined specifically on the use of computational pathology in clinical trials, the agency has issued guidance on use of diagnostics and CDx in drug trials, as well as the use of digital health technologies (DHTs) for remote data acquisition in clinical trials.[33–35] Depending on the intended use, computational pathology platforms could be considered a diagnostic device as well as a type of DHT. Similar to when using a diagnostic device, or when using a DHT, trial sponsors should demonstrate that the platforms are fit-for-purpose (i.e., that the level of validation and performance characteristics are sufficient to support its use and interpretability) prior to use in the trial. Of note, evidence needed to demonstrate the platform is fit-for-purpose may not be commensurate with what would be expected to support regulatory authorization. Verification and validation would be expected, although the extent is not clearly defined. Additionally, there is an open question as to which quality and design principles to apply when developing a computational pathology platform for clinical trial use. With uncertainty in the regulatory pathway, the best course of action is to engage the FDA with a pre-IND submission in which one describes the computational pathology platform, the verification and validation results and plans, and how it will be used in the clinical trial.

Considerations that may be relevant to determining the level of evidence and design principles needed to demonstrate a computational pathology platform is fit-for-purpose could include:

1. The intended use of the platform;
2. Risk to patient safety;
3. Intent to support a marketing application for the platform or a drug; and
4. Business and trial operational risks.

For example, regarding intended use and risk to patients, computational pathology platforms used for pre-screening and confirmation with another medically established method, or to enrich for biomarker positive patient enrollment, may not require testing that is as robust as a platform used as the sole method for selecting participants for a trial or treatment arm, given these use cases pose less risk to patient safety. However, it is important to understand the concordance between the computational pathology platform and the confirmatory method, to avoid biases. Similarly, platforms used in an early phase study for biomarker discovery or exploration of disease biology likely require less stringent levels of validation and technical performance testing than a platform being used in a registrational trial where the data will inform patient management and support marketing authorization of the platform. Good software engineering practices and state-of-the-art software validation practices may be sufficient, from a quality and design perspective, for these lower risk use cases. Meanwhile, platforms being developed as a CDx and with an intent to market should be developed in accordance with design controls, AI/ML GMLPs, and would likely need to generate technical performance results, as well as robust evidence of analytical and clinical validity, among other data, to support a marketing submission. Further, it is imperative that the algorithm used in the clinical trial is predefined and locked in prior to use, including establishment of a cutoff.

The International Medical Device Regulators Forum (IMDRF) has published a SaMD risk categorization framework with four risk categories (I-IV) based on significance of the information to the healthcare decision (e.g. whether output from a SaMD is used to treat or diagnose, drive or inform clinical management) and the severity of the health condition.[36] Given the serious nature of cancer, using this risk categorization to inform evidence generation would be largely influenced by the intended use (e.g., inform management vs. treat or diagnose) and sponsors may find value in applying this approach to the use of DHTs in clinical trials.

In addition to assessing the level of evidence needed to demonstrate a computational pathology platform is fit-for-purpose, sponsors should ensure compliance with other applicable regulatory requirements for the clinical trial. For example, for deployment into a clinical trial in the U.S., sponsors must follow 21 CFR part 812 to assess whether the platforms are considered to pose a significant risk to participants and/or seek an investigational device exemption (IDE) as needed.[37]

In addition to patient safety risk, and unrelated to regulatory expectations, the operational risks to a clinical trial (e.g., logistics of incorporating new technology and costs) are also important considerations when determining the required level of performance testing of computational pathology platforms that will be used in a clinical trial. For example, a platform may present very little, if any, risk to patient safety, but may have an impact on important business drug development decisions such as a go/no go decision to proceed from an early safety/dose escalation trial to a registrational trial. Additionally, there are various operational models for implementing computational pathology in a clinical trial and commercial use, which may raise different risks for trial operations/business decisions. For example, implementation could use a centralized model (similar to central lab testing for a trial or a single-site PMA for a marketed diagnostic) or a distributed model (similar to a distributed IVD kit). Therefore, sponsors may want to assess the risks to trial operations/business decisions, when deciding whether the level of evidence is sufficient to use a computational pathology platform in a clinical trial.

Below are sample questions and considerations when determining fit-for-purpose performance testing of computational pathology platforms in oncology drug development. If the answers to these questions indicate a high risk to patient safety, then an organization should employ a high level of testing and quality oversight during development (e.g., strong engineering practices and/or design controls). Alternatively, if the answers to these questions suggest less impact to patient safety, then a less stringent level of performance testing or quality oversight may be acceptable.

Questions to Consider When Determining Fit-for-Purpose Performance Testing

1. How will the platform be used?
   - Will it be used prospectively to select patients for a trial or a treatment?
   - Will it be used retrospectively for biomarker discovery, disease biology, or other exploratory purposes?
   - Will it be used for assessment of a primary or secondary endpoint?
   - Will it be used for futility analyses or other analyses for decision-making on the trial?
   - Will it be used in conjunction with one or more confirmatory tests?
2. What is the risk to patients of an inaccurate result?
   - Will patient management change?
   - Could patients be exposed to treatment toxicities?
   - Will the dosing of patients be modified inappropriately?
   - Could a patient forgo the standard of care or be enrolled when little benefit is to be expected?
   - Could a patient be falsely excluded from receiving care with expected benefit?
3. Will the platform be the subject of a marketing authorization application?

    - Will the platform be used to generate data in support of a marketing application for a drug?
    - Will the platform itself be the subject of a medical device marketing application?
    - Are both drug and device marketing applications intended?
4. What are the business risks of an inaccurate result?
    - Will implementation of the platform be using a centralized model?
    - Will implementation of the platform be using a distributed model?

Proposal 3: Establish Common Reference Standards

Establishing common reference standards and repositories will support future platform development and cross-validation. As multiple platforms are developed for the same biomarker, utilizing common datasets to validate and develop these platforms can support 1) wider access to biomarker testing across multiple platforms showing similar performance characteristics that may already be in place in testing labs, 2) platform developers producing concordant or comparable platforms, and 3) clinician end-users making informed decisions because they will understand the comparability of different platforms. This may help prevent future situations such as that observed with the various PD-L1 follow-on tests, in which multiple PD-L1 IHC assays were independently developed as follow-ons for different therapies without an understanding of how these different assays and scoring methodologies were related.[38]

While a single computational pathology platform may be used in a registrational trial for biomarker identification, additional, “follow-on” platforms measuring the same biomarker may be developed. Where available, the original slides could be used to ensure new platforms developed have high concordance with the originally approved platform, in addition to the other datasets used for validation of the follow-on platforms. However, institutional definitions of images as biospecimens versus de-identified data will impact the ease with which the images may be stored, shared, or used. Further, there are existing country-specific requirements and regulations regarding maintaining control of patient-level data that may impact the feasibility of sharing trial images. If the images cannot be shared, the platforms could be made available to the sponsor to evaluate performance across platforms using digital images from registrational studies, assessing the comparability of the performance of multiple platforms on its own dataset without sharing the slides. Although it would benefit drug developers to assess performance across platforms to identify a biomarker of interest, the scalability and management of such research is uncertain. The burden would be on drug developers to ensure proper consent for this future use and to conduct this work, as well as add potential regulatory or commercial risk to be involved with validation of third-party platforms outside of the CDx, which may limit the viability of this approach.

Unlike the banking of tissue and/or blood samples in which there is limited supply, banking slide images with proper informed consent for future use may be more attainable. However, criteria to define the appropriate number of images, or size of the training dataset will vary according to the platform being developed and the intent of use. Additionally, the storage, back-up, and auditing of the images are not negligible undertakings. The memory storage size and cost of databases needed to hold the images and associated metadata are substantial and should be considered when developing datasets. Furthermore, the workflow for digitization and interpretation of the images involves many different people, roles, as well as potentially different locations (e.g., where the slide is cut, digitized, and image analysis conducted). Therefore, developing robust reference datasets must encompass the relevant stakeholders (e.g., sponsors, pathology labs, platform developers).

There are additional opportunities to develop reference datasets outside of a single sponsor in a pre-competitive manner. Commercially acquired digitized images, or those collected through a consortium, could provide access to images that could be analyzed using the same platform and algorithm deployed for the registrational trial of interest as a comparator and reference for other platforms. Consortia have previously used a commissioned third-party to securely hold and analyze data from drug and/or diagnostic developers and share results with the community. Alternatively, a federated model for a reference dataset could be implemented, with those in control of the images maintaining control over their critical datasets (either a sponsor or a source institution) but allowing a model to run on the images without the images themselves leaving the virtual workspace. This federated model would allow for concordance testing both between different datasets as well as different algorithms. Depending on the intended use of the reference dataset, linked outcomes data may not be necessary, which may increase the comfort level of sponsors to share data. Lastly, existing infrastructure may be leveraged to share digital pathology images, including the National Cancer Institute's (NCI) Imaging Data Commons,[39] a cloud-based repository of publicly available cancer imaging data, as well as the precisionFDA[40] platform, a secure, cloud-based environment permitting collaborative research and data sharing on a secure platform.

A common reference set of slides are needed to support generating robust data repositories. Recommendations for establishing a reference dataset (also see these references[41,42]):

- Slides are digitized shortly after staining to minimize the impact of storage on the quality of slides. A timing threshold could be established and reported.
  - If slides are not digitized shortly after, such as when archived samples are imaged, detailed reporting of the slide age is needed.
- Images are stored appropriately and in the same file format to ensure the greatest amount of interoperability.
  - There are current initiatives to expand the DICOM standard to pathology imaging and could be one mechanism to enable alignment.
- Access to stored documents is secure and controlled, but not cumbersome.
- Relevant preprocess metadata including input parameters (**Table 2**) are linked to the images.
- Clinical metadata is ideally included, containing orthogonal information such as genomic and proteomic data, treatment regimens, and outcomes.
- Relevant characteristics of the intended patient population and measurement inputs are sufficiently represented in a sample of adequate size.
- All metadata are reported in a standardized format and of a given quality.
- Dataset represents the heterogeneity of real-world clinical/laboratory practices and patient populations, including slide preparation, scanning, patient characteristics, and tumor characteristics.

When platform developers leverage reference standards to perform comparisons and assess performance, it is important to consider what the platform does and how it will be used, as well as the purpose of the reference dataset to ensure the intentions are aligned and the reference dataset has the appropriate data. This includes considerations on the types of tissue and slide processing, diseases, digitization methodology, and relevant metadata (**Table 2**). Reference datasets should also be diverse in the relevant patient and tumor characteristics, preferably from multiple centers to be more generalizable to real-world patient and clinical practice populations. It is imperative that reference datasets have data reported in a standardized format, including reporting the input parameters for digitization, patient and tumor characteristics, treatment and outcome data, and platform performance metrics and output. As noted in **Proposal 2**, computational pathology biomarker measurements should be reported as continuous variables in addition to binary results even if performance metrics dichotomize the data.

Conclusions

This white paper highlights the promise of computational pathology to aid oncology drug development, as well as the possible future challenges to evaluating the robustness of these platforms to support their validation and use in drug development. As such, the proposals outlined support identification and reporting of key input and platform performance characteristics, a framework to inform evidentiary needs and performance criteria, and opportunities for establishing standards and common reference datasets. Computational pathology can be used across the spectrum of oncology drug development, from early discovery to registrational trials, and the intended use for each computational pathology application will impact the evidentiary needs to validate the platform. Computational pathology is an evolving field with evolving technologies, and as such, the possible applications and validation of these platforms will grow.

In addition to this working group, there are many ongoing consortia and efforts surrounding the use of digital and computational pathology platforms and their validation, and collaboration is needed to tackle outstanding questions.[43–45] Future efforts are needed to align on recommendations and benchmarks for quality metrics of preprocess input parameters to support transparency in platform development. Further, to support aligned data deposition into reference datasets, the development of standardized methodologies and data dictionaries is also needed. Alignment regarding data storage (e.g., on premises versus cloud solutions, ensuring data integrity and security, data transfer, redundancy/backups) is critical to ensure robust datasets for future use.

Formal guidance from regulatory bodies and relevant interest groups is needed to set regulatory expectations and establish performance metrics for computational pathology in drug development. FDA has signaled[46] their consideration of AI/ML in aiding drug development, with discussion ongoing. Clarity in the regulatory expectations for use of computational pathology in clinical trials would be valuable, including the evidence to demonstrate a platform is fit-for-purpose and the quality and design principles to apply when developing these platforms.

While this white paper demonstrates the potential promise of use of these platforms, there are currently regional differences in capabilities for using this technology. Many laboratories do not have digitization capabilities, due to lack of infrastructure, training, adequate funding, or other barriers. Additionally, if digitization capabilities are available, most have only access to one scanner type, which may impact the ability to use various platforms if they are not developed in an agnostic way to the digitization workflow. Significant uptake of robust digital pathology is needed to realize the promise of these platforms and future work should address these barriers to enable broader uptake.

Lastly, there is an opportunity to leverage existing data (e.g., pathology slides, metadata) from various stakeholders to generate an accessible digital pathology dataset to cross-evaluate different computational pathology platforms measuring the same biomarker to support the concepts in this white paper. There is a precedent in the AI development industry to conduct "Challenges" to evaluate the variability of AI models using standard datasets for training and testing, and precisionFDA also hosts challenges.[47] Further, *Friends* has conducted previous harmonization efforts to support aligned biomarker measurement and use, including the Tumor Mutational Burden (TMB) Harmonization and Homologous Recombination Deficiency (HRD) Harmonization Projects, and is poised to support a harmonization effort in computational pathology. Future work will focus on building out an appropriate use case to test the proposals herein, clarify workflows, and provide concrete data to support guidance efforts.

References

1. Naoumov N V., Brees D, Loeffler J, et al. Digital pathology with artificial intelligence analyses provides greater insights into treatment-induced fibrosis regression in NASH. J Hepatol. 2022;77(5):1399-1409. doi:10.1016/J.JHEP.2022.06.018
2. Najdawi F, Sucipto K, Mistry P, et al. Artificial Intelligence Enables Quantitative Assessment of Ulcerative Colitis Histology. Modern Pathology. 2023;36(6):100124. doi:10.1016/j.modpat.2023.100124
3. Companion Diagnostics | FDA. Accessed August 30, 2023. https://www.fda.gov/medical-devices/in-vitro-diagnostics/companion-diagnostics
4. Oncology Drug Products Used with Certain In Vitro Diagnostics Pilot Program | FDA. Accessed August 30, 2023. https://www.fda.gov/medical-devices/in-vitro-diagnostics/oncology-drug-products-used-certain-in-vitro-diagnostics-pilot-program
5. Tsao MS, Kerr KM, Kockx M, et al. PD-L1 Immunohistochemistry Comparability Study in Real-Life Clinical Samples: Results of Blueprint Phase 2 Project. Journal of Thoracic Oncology. 2018;13(9):1302-1311. doi:10.1016/J.JTHO.2018.05.013
6. Rimm DL, Han G, Taube JM, et al. A Prospective, Multi-Institutional Assessment of Four Assays for PD-L1 Expression in NSCLC by Immunohistochemistry. JAMA Oncol. 2017;3(8):1051. doi:10.1001/JAMAONCOL.2017.0013
7. Vega DM, Yee LM, McShane LM, et al. Aligning tumor mutational burden (TMB) quantification across diagnostic platforms: phase II of the Friends of Cancer Research TMB Harmonization Project. Annals of Oncology. 2021;32(12):1626-1636. doi:10.1016/j.annonc.2021.09.016
8. Technical Performance Assessment of Digital Pathology Whole Slide Imaging Devices Guidance for Industry and Food and Drug Administration Staff Preface Public Comment. Published online 2016. Accessed August 30, 2023. http://www.regulations.gov
9. Taqi SA, Sami SA, Sami LB, Zaki SA. A review of artifacts in histopathology. J Oral Maxillofac Pathol. 2018;22(2):279. doi:10.4103/JOMFP.JOMFP_125_15
10. Bagchi A, Madaj Z, Engel KB, et al. Impact of Preanalytical Factors on the Measurement of Tumor Tissue Biomarkers Using Immunohistochemistry. J Histochem Cytochem. 2021;69(5):297-320. doi:10.1369/0022155421995600
11. CFR - Code of Federal Regulations Title 21. Accessed August 30, 2023. https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/cfrsearch.cfm?fr=864.3750
12. Class II De Novo Paige Prostate FDA Letter. https://www.accessdata.fda.gov/cdrh_docs/pdf20/DEN200080.pdf.
13. Petrick NA, Chen W, Delfino JG, et al. Regulatory considerations for medical imaging AI/ML devices in the United States: concepts and challenges. https://doi.org/101117/1JMI105051804. 2023;10(5):051804. doi:10.1117/1.JMI.10.5.051804
14. Evans AJ, Brown RW, Bui MM, et al. Validating Whole Slide Imaging Systems for Diagnostic Purposes in Pathology. Arch Pathol Lab Med. 2022;146(4):440-450. doi:10.5858/ARPA.2020-0723-CP
15. How to Validate AI Algorithms in Anatomic Pathology | College of American Pathologists. Accessed August 30, 2023. https://www.cap.org/member-resources/clinical-informatics-resources/how-to-validate-ai-algorithms-in-anatomic-pathology
16. Abels E, Pantanowitz L, Aeffner F, et al. Computational pathology definitions, best practices,

and recommendations for regulatory guidance: a white paper from the Digital Pathology Association. J Pathol. 2019;249(3):286. doi:10.1002/PATH.5331

17. Moons KGM, Kengne AP, Grobbee DE, et al. Risk prediction models: II. External validation, model updating, and impact assessment. Heart. 2012;98(9):691-698. doi:10.1136/HEARTJNL-2011-301247
18. Bejnordi BE, Veta M, Van Diest PJ, 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
19. Evaluation of Automatic Class III Designation- De Novo Request . https://www.accessdata.fda.gov/cdrh_docs/pdf16/DEN160056.pdf.
20. Good Machine Learning Practice for Medical Device Development: Guiding Principles | FDA. Accessed August 30, 2023. https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles
21. Drug Development Tool (DDT) Qualification Programs | FDA. Accessed September 4, 2023. https://www.fda.gov/drugs/development-approval-process-drugs/drug-development-tool-ddt-qualification-programs
22. Medical Device Development Tools (MDDT) | FDA. Accessed September 4, 2023. https://www.fda.gov/medical-devices/medical-device-development-tools-mddt
23. CDRH. Requests for Feedback and Meetings for Medical Device Submissions: The Q-Submission Program Guidance for Industry and Food and Drug Administration Staff Preface Public Comment. Accessed August 30, 2023. https://www.reginfo.gov.
24. Lennerz JK, Green U, Williamson DFK, Mahmood F. A unifying force for the realization of medical AI. npj Digital Medicine 2022 5:1. 2022;5(1):1-3. doi:10.1038/s41746-022-00721-7
25. Product Classification. Accessed August 30, 2023. https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpcd/classification.cfm?id=QPN
26. Product Classification. Accessed August 30, 2023. https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfPCD/classification.cfm?ID=QKQ
27. Product Classification. Accessed August 30, 2023. https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfPCD/classification.cfm?id=5328
28. Product Classification. Accessed August 30, 2023. https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfPCD/classification.cfm?ID=PSY
29. Pell R, Oien K, Robinson M, et al. The use of digital pathology and image analysis in clinical trials. J Pathol Clin Res. 2019;5(2):81-90. doi:10.1002/CJP2.127
30. Petrick N, Chen W, Delfino JG, et al. Regulatory considerations for medical imaging AI/ML devices in the United States: concepts and challenges. J Med Imaging (Bellingham). 2023;10(5). doi:10.1117/1.JMI.10.5.051804
31. Chen W, Krainak D, Sahiner B, Petrick N. A Regulatory Science Perspective on Performance Assessment of Machine Learning Algorithms in Imaging. Published online 2023:705-752. doi:10.1007/978-1-0716-3195-9_23
32. Digital Pathology Program: Research on Digital Pathology Medical Devices | FDA. Accessed August 30, 2023. https://www.fda.gov/medical-devices/medical-device-regulatory-science-research-programs-conducted-osel/digital-pathology-program-research-digital-pathology-medical-devices
33. Principles for Codevelopment of an. Accessed August 30, 2023. http://www.fda.gov/Drugs/

GuidanceComplianceRegulatoryInformation/Guidances/de

34. Investigational IVDs Used in Clinical Investigations of Therapeutic Products Draft Guidance for Industry, Food and Drug Administration Staff, Sponsors, and Institutional Review Boards DRAFT GUIDANCE. Published online 2017. Accessed August 30, 2023. http://www.fda.gov/BiologicsBloodVaccines/GuidanceComplianceRegulatoryInformation/G
35. Digital Health Technologies for Remote Data Acquisition in Clinical Investigations. Accessed August 30, 2023. https://www.fda.gov/vaccines-blood-biologics/guidance-compliance-regulatory-information-biologics/biologics-guidances
36. Global Approach to Software as a Medical Device | FDA. Accessed August 30, 2023. https://www.fda.gov/medical-devices/software-medical-device-samd/global-approach-software-medical-device
37. CFR - Code of Federal Regulations Title 21. Accessed August 30, 2023. https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfCFR/CFRSearch.cfm?CFRPart=812
38. Salgado R, Bellizzi AM, Rimm D, et al. How current assay approval policies are leading to unintended imprecision medicine. Lancet Oncol. 2020;21(11):1399-1401. doi:10.1016/S1470-2045(20)30592-1
39. Fedorov A, Longabaugh WJR, Pot D, et al. NCI imaging data commons. Cancer Res. 2021;81(16):4188. doi:10.1158/0008-5472.CAN-21-0950/674291/AM/NCI-IMAGING-DATA-COMMONSNCI-IMAGING-DATA-COMMONS
40. precisionFDA - Overview. Accessed August 30, 2023. https://precision.fda.gov/
41. Homeyer A, Geißler C, Schwen LO, et al. Recommendations on compiling test datasets for evaluating artificial intelligence solutions in pathology. Modern Pathology. 2022;35(12):1759-1769. doi:10.1038/S41379-022-01147-Y
42. Wahab N, Miligy IM, Dodd K, et al. Semantic annotation for computational pathology: multidisciplinary experience and best practice recommendations. J Pathol Clin Res. 2022;8(2):116-128. doi:10.1002/CJP2.256
43. Digital and Computational Pathology Committee | College of American Pathologists. Accessed August 30, 2023. https://www.cap.org/member-resources/councils-committees/digital-pathology-committee/
44. Pathology Innovation Collaborative Community - MDIC. Accessed August 30, 2023. https://mdic.org/program/picc/
45. DPA: Digital Pathology Association. Accessed August 30, 2023. https://digitalpathologyassociation.org/
46. Cder Cber F. Using Artificial Intelligence & Machine Learning in the Development of Drug and Biological Products. Accessed August 30, 2023. https://www.imdrf.org/documents/machine-learning-enabled-medical-
47. Challenges - precisionFDA. Accessed August 30, 2023. https://precision.fda.gov/challenges
48. Wolff AC, Somerfield MR, Dowsett M, et al. Human Epidermal Growth Factor Receptor 2 Testing in Breast Cancer: American Society of Clinical Oncology–College of American Pathologists Guideline Update. Arch Pathol Lab Med. Published online June 7, 2023. doi:10.5858/ARPA.2023-0950-SA/493531/HUMAN-EPIDERMAL-GROWTH-FACTOR-RECEPTOR-2-TESTING
49. Dawson H. Digital pathology – Rising to the challenge. Front Med (Lausanne). 2022;9:888896. doi:10.3389/FMED.2022.888896/BIBTEX
50. Farahmand S, Fernandez AI, Ahmed FS, et al. Deep learning trained on hematoxylin and eosin

tumor region of Interest predicts HER2 status and trastuzumab treatment response in HER2+ breast cancer. Mod Pathol. 2022;35(1):44-51. doi:10.1038/S41379-021-00911-W

51. Shamai G, Livne A, Polónia A, et al. Deep learning-based image analysis predicts PD-L1 status from H&E-stained histopathology images in breast cancer. Nat Commun. 2022;13(1). doi:10.1038/S41467-022-34275-9
52. Baxi V, Lee G, Duan C, et al. Association of artificial intelligence-powered and manual quantification of programmed death-ligand 1 (PD-L1) expression with outcomes in patients treated with nivolumab ± ipilimumab. Mod Pathol. 2022;35(11):1529-1539. doi:10.1038/S41379-022-01119-2
53. Luchini C, Pantanowitz L, Adsay V, et al. Ki-67 assessment of pancreatic neuroendocrine neoplasms: Systematic review and meta-analysis of manual vs. digital pathology scoring. Mod Pathol. 2022;35(6):712-720. doi:10.1038/S41379-022-01055-1

Appendix 1: Hypothetical Use Cases for Considering and Reporting Input Parameters

Below are examples of hypothetical use cases to aid in how one may consider what parameters/conditions should be evaluated and characterized for a specific use case:

- Mitosis counting
  - Counts the number of mitoses/mm^2 in a sample using an algorithmic method for identifying the region of interest (ROI) or allows pathologists to select the ROI to be analyzed.
  - Requires a minimum area of sufficient quality for analysis.
  - Can tolerate slides with large regions that are inadequate for analysis.
- Prostate cancer Gleason grading
  - Algorithm to assign a Gleason score to prostate cancer samples.
  - Provides primary, secondary, and tertiary grades, and overall Gleason score by analyzing large scale histological patterns within a specimen.
  - Requires a minimum, representative total area of sufficient quality and with accurate location information for different prostatic regions; high magnification not required.
  - Less tolerant to slides with large regions that are inadequate for analysis or have artifacts.
- Metastases detection
  - Algorithm that detects the presence of metastatic cells within a biopsy.
  - High sensitivity task requiring a minimum total area of high-quality tissue and images.
  - Intolerant of slides with large regions that are inadequate for analysis or have artifacts.

Appendix Table 1: Reporting Input Parameters for Hypothetical Use Cases.
This table highlights how certain input parameters listed in Table 2 may be common considerations across use cases or may be different dependent on use case.

Parameter	Mitosis Counting	Metastasis Detection for Diagnostic Aid	HER2 Status Prediction for CDx
Tissue Sample Origin	fine needle aspirate not appropriate, other biopsy types acceptable	no specific requirements regarding biopsy type	Breast biopsies, breast resections, or specimens from metastatic sites (if applicable)
Tissue Processing	standard FFPE preparation, standard H&E staining		No frozen tissue Cold ischemic and fixation times within range as stipulated by interpretive guidelines[48]
Slide Type	standard glass slide, 1mm thick		
Tissue Thickness	3-5um sections		
Tissue Area	2mm^2 of tissue is analyzed, no tumor percentage minimum; region of interest based on algorithmic methods (details of method specified; i.e., random selection of X fields)	>1mm^2 of sufficient quality tissue area required, no tumor percentage minimum; entire tissue area evaluated	>1mm^2 of sufficient quality tissue area required and 20% minimum tumor content
Tissue Folds/ Tears	no tissue folds or tears in any analyzed region; algorithmic selection of regions of interest will prevent the presence of tissues folds/tears in analyzed regions	folds or tears and adjacent 5um distance will be excluded from analysis, excluded area must be less than 10% total analysis area	According to specific manufacturer QC procedure developed for intended use of the CDx, including detailed methodology and criteria to identify and exclude slides with tissue folds/tears.
Slide Age	scanned within X years/months of staining		
Slide Storage	slides protected from light, stored at room temperature		
Magnification	400X	200X or 400X	400x only

Appendix 2: Examples of Computational Pathology Platform Validation Study Designs

Below are examples of study designs to validate digital pathology platforms, provided based on the considerations for validation study design put forth in Frontiers in Medicine.[49]

	Ground Truth Definition	Case Selection	Acceptable Range of Output Values	Possible Confounding Effects	Identify Discrepant Cases
	Algorithm compared to a ground truth to establish precision and recall	Case mix should reflect real-world setting in terms of morphological heterogeneity and complexity	Define acceptable range of deviation from the ground truth. This may depend on clinically relevant cutoffs that determine therapy	Consider any variables in image preparation. For example, compare effect of different scanners.	Output values outside the defined acceptable range are discrepant to the ground truth (can systematic reasons be identified).
Her2[50]	Herceptest IHC scored based on 2018 ASCO/ CAP guidelines: Intense circumferential 3+ membrane staining in > 10% neoplastic cells are positive. The ground truth for the IHC results were defined as the consensus score reached by 3 pathologists for each case.	Trained a HER2 status predictor model on 188 HER2± H&E slides (93+/95-) and a test set of 187 HER2± H&E slides from The Cancer Genomic Atlas (TCGA) BRCA cohort.	The fully trained CNN model performance predicted the HER2 status with slide-level AUC of 0.90 (95% CI, 0.79–0.97). Model validation with an independent test set achieved an AUC of 0.80 (95% CI: 0.69–0.88) at the slide-level. Algorithm prediction of Trastuzumab clinical response is weak (sensitivity .56, specificity .58).	Utilized a deep learning based color normalization to remove batch effects and improve generalizability to independent datasets.	Borderline case confusion minimized by using only IHC 3+ cases training. Pathologist annotation improved model prediction.
PD-L1[51]	Breast Cancer PD-L1 expression was determined using the Ventana PD-L1 (SP142) assay as the proportion of tumor area occupied by PD-L1 expressing tumor-infiltrating immune cells (IC), and an expression in ≥1% IC was defined as PD-L1 positive status.	Tissue Microarrays: training set 2,516 (74.5%) cases; test set 860 (25.5%); and external test set 275.	CNN algorithm AUC performance with respect to the pathologist's binary PD-L1 status was 0.911 (95% CI: 0.891–0.925). Test set AUC performance was 0.915 (95% CI: 0.883–0.937). Independent test set AUC performance for PD-L1 prediction was 0.854 (95% CI: 0.771–0.908).	Data augmentation was performed to help the model deal with variability in staining methods and other differences between the cohorts.	Ground truth was not perfect; therefore, confirmed scores between 3 pathologists. Tissue microarrays used to train and test the algorithm may not have sufficient representation.

Appendix 2: Examples of Computational Pathology Platform Validation Study Designs

Below are examples of study designs to validate digital pathology platforms, provided based on the considerations for validation study design put forth in Frontiers in Medicine.[49]

PD-L1[52]	Dako PD-L1 28-8 IHC with cutoffs of TPS ≥1% and ≥5%.	217 samples from patients with NSCLC, 600 from MEL, 400 from SCCHN, and 293 from patients with UC.	AI-based assessment was highly correlated with the median score from manual assessment of PD-L1–expressing TCs by 5 pathologists (*r* ranging from 0.73 to 0.85).	Slides were scanned by two separate Aperio AT2 scanners across 5 days, two times per day (morning [AM] and afternoon [PM]).	A lower prevalence of PD-L1–positive patients was seen with AI-powered scoring (42.5% and 28.8%) compared with manual scoring (54.9% and 34.0%) at cutoffs of ≥1% and ≥5%, respectively, though the difference was not significant. This could be due to the presence of artifacts or low PD-L1 membrane staining with cytoplasmic positivity (blush).
Ki67[53]	Current guidelines for assessing Ki-67 recommended manual counting from a printed image that includes at least 500 neoplastic cells from tumor hotspots.	Review including 752 Pancreatic Neuroendocrine Neoplasms: G1 (55.3%), G2 (40.6%) and G3 (4.1%).	The pooled correlation estimate was 0.94 (95%CI: 0.83–0.98; I2 = 24.15%).	Risk of counting dividing non-neoplastic "contaminating" cells (endothelial cells, lymphocytes) and other brown pigment (hemosiderin).	Higher tumor grade generated due to overcounting "contaminating" cells or artifact.