FRIENDS of CANCER RESEARCH

Considerations for Developing Reference Data Sets for Digital Pathology Biomarkers

Friends of Cancer Research Discussion Document | 2025

Considerations for Developing Reference Data Sets for Digital Pathology Biomarkers

Discussion Document

Digital pathology enables innovative approaches for biomarker interpretation, cellular evaluation, and diagnosis. These approaches can leverage computational pathology models developed using artificial intelligence (AI) (including machine learning [ML] models) to aid in image analysis. While this holds the promise of enhancing accuracy, reproducibility, and standardization of pathology-based features to measure prognostic and predictive biomarkers, expedite diagnosis or pathological scoring, and identify novel biomarkers, there is currently a lack of robust publicly available data sets to support development and validation, and ensure consistent performance of different computational pathology models. Developing reference data sets of images and associated metadata can be challenging, requiring substantial time and money; however, adequately built data sets can support future platform development and validation and address concerns around model accuracy, reproducibility, reliability, and comparability.

Friends of Cancer Research (*Friends*) leveraged expertise from the ongoing Digital PATH Project working group, which included representatives from industry, the U.S. Food and Drug Administration (FDA), the National Cancer Institute (NCI), patient advocates, and academia, to discuss the promise of reference data sets. The Digital PATH Project evaluated the variability in HER2 assessments in a single breast cancer data set across multiple computational pathology models.[1] This document reflects a series of discussions on considerations for developing a reference data set, intended to spark ideas and facilitate further exploration of these critical topics to support future model development and validation.

Independent Reference Data Sets Provide Value

Multiple computational pathology models are often under development (by different developers) to assess the same biomarker and variability across models can cause challenges. The potential challenges are the same as when different immunohistochemistry (IHC) assays (i.e., without the application of computational pathology) are developed for the same biomarker. For example, multiple PD-L1 IHC assays were independently developed for various anti-PD-(L)1 therapies, each using different antibodies, scoring methods, and cut-offs. Analytical validation was performed on independent commercially acquired sample sets and clinical validation established using each developer's individual clinical trial data sets, resulting in inconsistency in how these IHC assays and scoring methodologies compare.[2, 3] To prevent similar challenges for future computational pathology models developed to assess the same biomarker, publicly available (non-proprietary) reference data sets can support an understanding of performance characteristics across multiple

models. Publicly available reference data sets also have the potential to enable more efficient regulatory evaluation of computational pathology models.

Many organizations have identified a need for reference data sets for assay validation,[4–6] including specifically for digital and computational pathology and AI-based models.[7–9] Unlike the development of reference data sets for assays requiring blood or tissue, data sets for digital pathology-based biomarkers are not limited by the constraints of obtaining and storing biological material. The banking of digitized slides is more feasible and provides the opportunity to develop reference data sets. Under the condition that a reference set is robustly built with relevant metadata and samples representative of the intended use of the assay, the data set can provide an objective measurement of model performance on data independent from any training data or data set unique to a specific model.

Ongoing efforts to develop digital pathology data sets largely source samples from individual academic sites.[10] These existing digital and computational pathology data sets are composed of various types of data (e.g., tissue and slide processing characteristics, and image acquisition characteristics, metadata, pathologist annotations, and clinical outcomes) with each uniquely contributing to their intended uses. However, single-source data sets may lack demographic and/or clinical characteristic representativeness of the larger patient population.

Reference data sets provide value to various groups who develop and use these models. Developers can have easy access to a rich data set to validate their model and assess performance. Drug developers can elucidate performance across multiple models to inform use in drug development and potential labeling. Clinician end-users can make informed decisions on model use as they understand the comparability of different models with a reference standard. Lastly, FDA's review process can evaluate validation with robust reference data sets as part of the body of evidence supporting regulatory decision-making.

Considerations for Developing Reference Data Sets

Possible Intended Uses of Reference Data Sets

Reference data sets may be developed for a variety of purposes and a single reference set may be leveraged to assess multiple aspects of model performance and multiple intended use populations, allowing for pre-specified analysis of data subsets in accordance with the specific use of a particular model. Alternatively, a reference data set may be designed to focus on one aspect of performance in specimens with particular characteristics. Table 1 provides an overview of possible intended uses for a reference data set. Analytical and clinical validation may require different types of data.

Each reference data set should be accompanied by a statement of its intended use(s), and the intended use(s) of any model evaluated on such a data set should be described. Any models assessed against the reference data set for validation purposes should be locked at the time of assessment, and the reference data set should be used for fully independent external validation of the model and not for training. To ensure the data set's utility is not limited to a single use, considerations for developer blinding and traceability between model versions should be explored. These approaches can help mitigate risks of overfitting to the reference data set and enable its reuse for testing modified versions of models.

Table 1. Possible Intended Uses of Reference Data Sets for Model Validation.

Intended Use	Performance Assessment	Considerations
Analytical Validation		
Accuracy	Demonstrate the extent to which the test model scores agree with the reference standard widely accepted as producing "truth"	• Accuracy refers to the assessment of the test compared to a reference standard, rather than the average of multiple values used as a proxy for the reference standard; therefore, assessment of accuracy using a consensus of multiple values is technically not a true measure of accuracy but may be necessary given the challenge to have a true "gold standard"
Precision	Demonstrate that the test model provides consistent scores when presented with similar or related inputs or under different conditions	• Reference data sets should include scenarios that capture known sources of variability, such as rescans of the same slide, scans of sequential sections, or scans of different biopsies from the same patient • The study design including number of replicates and samples, and factors considered, should be clearly defined[11]

Intended Use	Performance Assessment	Considerations
Interchangeability	Demonstrate that the test model scores are within the range of scores from multiple pathologists and/or current models (i.e., how well multiple raters agree when assessing the same sample)	● Recognizes that there may not be a singular, cost-effective, and independent reference standard, given the variability in models' and pathologists' scoring ● The number and the set of readers providing the reference scores will impact the assessment of interchangeability, and careful consideration is needed to ensure the readers are appropriately selected and trained ● Determine how inter-rater reliability and agreement will be assessed depending on the measurement scale ● The study design should be clearly defined to ascertain whether any factors other than raters (models) are changing
Clinical Validation		
Clinical Specificity	Assess the proportion of patients who are "negative" for the clinical outcome (denominator) who are correctly identified as "negative" by the digital pathology biomarker (numerator), e.g., assessment of tumor response by biomarker status	● For non-binary clinical outcomes, such as time-to-event outcomes, other metrics of clinical performance may be needed ● The definition of biomarker-positive, as well as the clinical outcome, will greatly impact the clinical specificity ● Findings may be specific to the clinical setting, including disease type and stage,

Intended Use	Performance Assessment	Considerations
Clinical Sensitivity	Assess the proportion of patients who are "positive" for the clinical outcome (denominator) who are correctly identified as "positive" by the digital pathology biomarker (numerator)	and particular treatment or drug mechanism of action • For predictive biomarkers, clinical sensitivity and specificity, as defined by patients' treatment benefits, cannot generally be estimated without restrictive assumptions. Therefore, a direct assessment of treatment effect (e.g., average probability of difference in "negative" outcomes versus biomarker values) could be employed[12, 13]

Intended Use of the Models

When leveraging reference data sets to assess performance, it is important to consider the intended use of the model, as well as the purpose of the reference data set, to ensure the intentions are aligned and the reference data set has the appropriate applicable data for the model's intended use. In practice, the intended use of models may be very diverse, and therefore, it may be difficult to develop a reference data set that is relevant to all models. Considering the utility of the reference set, given the current state of the intended uses of the models being developed, is important to ensure the reference data set is as broadly usable as possible.

One important consideration for intended use of the model is whether it will be used as a standalone test or to aid or assist the pathologist in interpretation or scoring. For standalone use, a more comprehensive and detailed understanding of performance compared to the reference standard might be important, and the reference data set would need to cover a broader range of potential cases/samples. As there is no need to involve a reader end user with standalone models, assessing performance on a reference set can be completed quickly. For pathologist-aided models, the reference data set might focus on borderline cases or a larger proportion of cases where there is known to be a higher degree of variability in pathologists' scoring. These models will likely require a reader study (e.g., comparing the reader with and without the model) which takes time and may encounter feasibility challenges depending on the size of the data set. However, even if the model is intended to be used as an aid, assessment of standalone performance is usually desired.

Infrastructure to Support Housing the Reference Data Set

Determining how to store, back up, and audit digitized slides will be critical, as the reference data set will likely require considerable memory storage space and cost to host the images and associated metadata. In addition to data storage, a platform to interface with model developers and allow for bidirectional data transfer will be necessary. The design of the infrastructure should align with the intended use(s) of the reference data set and the models it supports, ensuring these priorities guide subsequent decisions. The whole-slide images (WSIs) could either be transferred to the model developers, without the metadata and reference standard assessments, or the WSIs could remain sequestered on a platform within a federated framework that allows models to be executed or evaluated without requiring the WSIs to be transferred. There also needs to be a mechanism for analyzing the model output compared to the reference standard, which could be conducted by a third party. It is important to consider whether the model results remain blinded and what data would be made available to the model developer after conducting the analyses. Key governance considerations, such as contributions, quality control, accessibility, versioning, and validation, will be necessary for ensuring the data set's integrity and alignment with its intended use.

Considerations for Defining a Reference Standard

A single reference standard is necessary to establish accuracy. The current reference standard for many pathology-based biomarkers is generally considered to be the pathologist rendering an interpretation using a light microscope, which differs from reading a digital image. Given the variability in pathologists' manual biomarker readings, there may not be a single reference standard (i.e., "gold standard") for the biomarker for analytical validation. As such, a single reference standard can be based on a consensus across multiple pathologists. As biomarker development continues, including the development of novel biomarkers assessed by AI-models, pathologists' scores may not be feasible as a reference standard (e.g., HER2-ultra low may not be amenable to reproducible determination by pathologists). When considering the definition and measurement of the reference standard, there are strengths and limitations to various approaches, highlighted in Table 2. Additionally, the reference standard should align with the intended use of the data set to ensure relevance and applicability.

The target performance of a model in the assessment of accuracy compared to a given reference standard will depend on multiple factors. Performance targets for a biomarker assay will be context-dependent influenced by the level of risk due to inaccurate biomarker identification and its impact on clinical predictions and outcomes from clinical management decisions (e.g., treatment selection) guided by those predictions. Guidance on target performance goals would be helpful.

Rather than defining a single reference standard and performance goal for a model, it may be necessary to assess performance based on interchangeability with pathologists' scoring of the reference data set. This would require evaluating the level of agreement among multiple pathologists' scores on the WSIs comprising the reference data set. Following the determination of the agreement among pathologists, the level of agreement between the model and all the pathologists can be assessed to determine if the model performs within the distribution of pathologists' performance. However, this approach is only applicable when pathologist scoring is possible and does not support the development or validation of novel biomarkers or more quantitative scoring approaches that are independent of pathologists.

Considerations for Annotation of Region of Interest

Pathologists inherently have a different workflow for assessing a WSI compared to an AI-derived model, including their understanding of the overarching morphology depicted in the slide. The Digital PATH project anecdotally found that for many WSIs with discordance in HER2 scores across models, scoring variability stemmed from differences in how the models identified the area of invasive carcinoma. To promote alignment in biomarker assessment, a reference dataset could include a consensus-based reference standard for the identified invasive tumor area on a WSI, derived from a consensus of pathologist annotations. Performance could be assessed based on a model's ability to identify the area of invasive carcinoma, providing additional insight into its performance. Guidance is needed to understand how to set targets for performance specific to the task of tumor area identification. It is also important to note that similarity in identifying regions of interest may or may not support clinical validation of a computational pathology model, especially for models that identify or integrate signals not visually discernible or typically analyzed by pathologists.

Table 2. Strengths and Limitations of Reference Standards for Digital Pathology-Based Biomarkers.

Reference Standard	Considerations	Strengths	Limitations
Consensus Pathologists' Scores	• Assess the inter-rater variability of multiple raters to give context to variability observed between the model and reference standard • Capture the recruitment methods, applicable qualifications and requirements (training, board certification, specialization, etc.), inclusion/exclusion criteria for pathologists	• Established scoring guidelines and proficiency training for the biomarker (e.g., ASCO/CAP HER2) provide consistency in assessment • Current practice for ascertaining biomarker status	• Guidelines may become outdated or irrelevant to future use cases, limiting the utility of the reference data set (e.g., HER2-low/ultra-low designations) or requiring augmentation with new information • It is challenging to recruit and train experts, and the make-up of the pathologists included can impact the consensus derived and the assessment of inter-rater variability
Clinical Treatment Outcomes	• The reference standard for clinical validation will depend on clinical outcomes • Full assessment of the predictive ability of a model requires data from both biomarker-positive and biomarker-negative patients, with some in each group receiving biomarker-directed therapy versus non-biomarker-directed	• Variability in scoring might not always translate to major differences in predicted patient outcomes and variability should be viewed in the context of impact of performance on outcome prediction • Likely needed for novel biomarkers without standardized guidelines for assessment	• May be difficult to find biomarker-negative cases treated with biomarker-directed therapy when many trials use biomarker-based eligibility criteria; might only be possible to establish whether patients identified as biomarker-positive by a model benefit from a targeted therapy relative to an alternative • The reference standard will be narrowly applicable to a specific

Reference Standard	Considerations	Strengths	Limitations
			drug/mechanism of action that may not be relevant for other drugs that utilize the same biomarker, and the data set will also be biased towards any inclusion/exclusion criteria inherent to treatment selection
Other Biological Correlates	● Orthogonal assays to measure the biomarker, such as mRNA, in situ hybridization, or mass spectrometry, could provide an additional assessment of the biomarker	● More quantitative, objective measure of biomarker that is not reliant on human interpretation	● Requires additional tissue, slides to run analyses ● As new technology is developed, would need additional biological material to run new assays for the reference set to stay relevant ● Orthogonal methods themselves may not be standardized and introduce additional variability

Considerations for Reporting Metadata and Ensuring Representativeness in the Reference Set

Relevant clinical, sample, and patient data should be connected to the WSIs. Table 3 highlights relevant metadata to capture. A data dictionary should accompany the reference data set, including metadata definitions for demographic and clinical information, as well as how the data were identified/defined (e.g., chart review, central testing for biomarkers, consensus or single scoring for histological grade, etc.). Additionally, the data dictionary should also specify the expected format for each data field.

While de-identified data is likely to be used, patients should be properly consented for use of their samples in a publicly available reference data set. For example, certain institutions consider digital pathology images to be biospecimens, which may require additional patient consent for inclusion in a repository.

Table 3. Metadata to Include in a Reference Data Set.

Patient Characteristics	Clinical Characteristics	Tissue and Slide Processing Characteristics*	Image Acquisition Characteristics*
Age (at sample collection)	Diagnosis History (e.g., de novo or recurrence)	Glass Slide Type	Scanner Hardware and Software Versions
Sex	Histological Grade	Tissue Thickness	Scanner Software Configurable Parameters
Race	Histology	Tissue Area; Tumor area/size	Slide Viewer
Ethnicity	Clinical Stage	Tissue Artifacts	Image File Type
Geographic Location	Biomarker Status (relevant to disease of interest)	Tissue Age	Magnification
Relevant prognostic factors	Prior Treatments Received	Slide Age	Resolution

	Sample Type (e.g., core biopsy, FNA, cytology)	Antibody Used (Lot #)	
	Tumor Site/anatomic location	Staining Conditions/method of antigen retrieval	
	Associated molecular findings	Slide Storage	

*For more detail, see Supporting_the_Application_of_Computational_Pathology_in_Oncology.pdf

As relevant to the intended use of the reference data set, these characteristics should vary to represent the entire diagnostic spectrum of a diverse target population. The sampling strategy should be detailed in accompanying literature to the reference data set (e.g., data set includes all cases from one site within a specific time frame), as well as any relevant inclusion or exclusion criteria that impact the samples. Generally, reference data sets should include samples from multiple clinical sites to ensure diversity in patient populations and clinical practice, which allows for the potential to conduct subgroup analyses assessing the association of the model's performance with specific clinical, patient, or pre-analytical characteristics.

Several considerations are specific to ensuring representativeness in the reference data set of the biomarker. Within a biomarker category, a reference set should include a spectrum of staining positivity. For example, HER2 staining levels could include weak-to-moderate complete membrane staining (e.g., 11% or 50% of tumor cells) as well as intense membrane staining (e.g., 1% or 9% of cells). Therefore, data sets should not only consider representation across each broad biomarker category, but also within each category.

Characteristic categories (e.g., patient, clinical, etc.), including the biomarker category, may be representative of the intended use population in the reference data set as a general principle. However, if there are no indications of an association between a specific characteristic and the outcome, and no subgroup analyses are planned, strict balancing may not be necessary. This approach allows for flexibility while ensuring that the reference data set reflects the intended purpose and avoids unnecessary complexity. If risk profiles are different, subgroups should be sized for individual subgroup analyses for a more definitive understanding of performance. There is a concern about oversampling (or undersampling) as some agreement measures, such as kappa coefficients, are highly dependent on the distribution of scores in patient subgroups. It is important to understand the distribution of characteristic categories in the clinical population to ensure that subgroups can be weighted appropriately when estimating overall performance results.

Lastly, it is important to consider the sample quality included and represented in the reference data set, including artifacts, inadequate tumor cellularity, and edge cases to ensure robust and meaningful validation. Whether and how many lower-quality samples are included in the reference data set will depend on its intended use. For example, one may prioritize inclusion of more "pristine" samples to establish baseline concordance of the model, but "edge" or challenging cases should likely be included to assess robustness of the models in a data set more reflective of clinical practice. Challenging cases may include those with artifacts, complex architecture or morphology, rare or mixed histologies, etc. In biomarkers studied extensively, such as HER2 in breast cancer, defining challenging cases may be easier than in other biomarker contexts. Proactively identifying challenging cases may be difficult but could be informed by conducting interviews with pathologists to understand difficult cases.

Conclusions and Next Steps

This document provides an overview of discussions aimed at catalyzing further dialogue in the field on developing robust reference data sets. Reference data sets and models can have many intended uses, which require careful consideration during the development process. When creating a reference data set, it is important to narrow focus to a single intended use. For example, analytical validation for regulatory purposes could serve as a use case to propose specific criteria for developing a robust reference data set.

Developing a reference data set will require collaboration across multiple contributors and may emerge from community efforts, patient groups, federal initiatives, or professional societies. Recent opportunities, such as the Advanced Research Projects Agency for Health (ARPA-H) ImagiNg Data EXchange (INDEX) program, provide possible platforms to develop such reference data sets. Those interested in developing a reference data set for regulatory purposes should consult the FDA early in the planning process. An ideal opportunity for interaction with the FDA is through development of a medical device development tool (MDDT).[14] Additionally, the FDA provides various other pathways for engagement, depending on the intended use of the AI model or reference data set.[15] These options, outlined in a recent guidance document, include opportunities to discuss innovative trial designs, digital health technologies, and real-world evidence generation, among others. A voluntary pre-submission with the FDA would allow for early discussions on the scope, protocol, statistical approach, and patient population for the data set.

References

1. McKelvey B, Torres-Saavedra P, Li J, et al: Agreement Across 10 Artificial Intelligence Models in Assessing HER2 in Breast Cancer Whole Slide Images: Findings from the Friends of Cancer Research Digital PATH Project. Available from: https://friendsofcancerresearch.org/wp-content/uploads/SABCS-Poster_Digital-PATH-Project-Findings.pdf

2. September 26, 2024: Meeting of the Oncologic Drugs Advisory Committee Meeting Announcement - 09/26/2024 | FDA [cited 2025 Jan 15] Available from: https://www.fda.gov/advisory-committees/advisory-committee-calendar/september-26-2024-meeting-oncologic-drugs-advisory-committee-meeting-announcement-09262024

3. Tsao MS, Kerr KM, Kockx M, et al: PD-L1 Immunohistochemistry Comparability Study in Real-Life Clinical Samples: Results of Blueprint Phase 2 Project. J Thorac Oncol 13:1302, 2018. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC8386299/

4. Feng Z, Kagan J, Pepe M, et al: The Early Detection Research Network's Specimen Reference Sets: Paving the Way for Rapid Evaluation of Potential Biomarkers. Clin Chem 59:68–74, 2013. Available from: https://dx.doi.org/10.1373/clinchem.2012.185140

5. Khleif SN, Doroshow JH, Hait WN: AACR-FDA-NCI cancer biomarkers collaborative consensus report: Advancing the use of biomarkers in cancer drug development. Clinical Cancer Research 16:3299–3318, 2010. Available from: /clincancerres/article/16/13/3299/11147/AACR-FDA-NCI-Cancer-Biomarkers-Collaborative

6. Cancer Biomarker Measurements and Collaborations | NIST [cited 2025 Jan 15] Available from: https://www.nist.gov/programs-projects/cancer-biomarker-measurements-and-collaborations

7. Petrick N, Chen W, Delfino JG, et al: Regulatory considerations for medical imaging AI/ML devices in the United States: concepts and challenges. Journal of Medical Imaging 10:051804, 2023. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC10289177/

8. 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 249:286, 2019. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC6852275/

9. November 20-21, 2024: Digital Health Advisory Committee Meeting Announcement - 11/20/2024 | FDA [cited 2025 Jan 15] Available from: https://www.fda.gov/advisory-committees/advisory-committee-calendar/november-20-21-2024-digital-health-advisory-committee-meeting-announcement-11202024

10. GitHub - maduc7/Histopathology-Datasets: Ressources of histopathology datasets [cited 2025 Jan 15] Available from: https://github.com/maduc7/Histopathology-Datasets?tab=readme-ov-file

11. Yee LM, Lively TG, McShane LM: Biomarkers in early-phase trials: fundamental issues. Bioanalysis 10:933, 2018. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC6123886/

12. Janes H, Pepe MS, McShane LM, et al: The Fundamental Difficulty With Evaluating the Accuracy of Biomarkers for Guiding Treatment. J Natl Cancer Inst 107, 2015. Available from: https://pubmed.ncbi.nlm.nih.gov/26109106/

13. Polley MYC, Freidlin B, Korn EL, et al: Statistical and practical considerations for clinical evaluation of predictive biomarkers. J Natl Cancer Inst 105:1677–1683, 2013. Available from: https://pubmed.ncbi.nlm.nih.gov/24136891/

14. Medical Device Development Tools (MDDT) | FDA. Available from: https://www.fda.gov/medical-devices/medical-device-development-tools-mddt

15. Considerations for the Use of Artificial Intelligence To Support Regulatory Decision-Making for Drug and Biological Products | FDA [cited 2025 Jan 15] Available from: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/considerations-use-artificial-intelligence-support-regulatory-decision-making-drug-and-biological