Submitted July 17, 2025

Friends of Cancer Research Response to National Cancer Institute's Request for Information: Benchmarks for Artificial Intelligence in Cancer Research and Care

To Whom it May Concern:

Friends of Cancer Research (*Friends*) appreciates the opportunity to respond to the National Cancer Institute's (NCI) Request for Information (RFI) on benchmarks for artificial intelligence (AI) in cancer research and care. *Friends* is committed to accelerating the development of cutting-edge cancer care by advancing regulatory science, enabling collaboration across stakeholders, and promoting rigorous, data-driven policy solutions. We believe that standardized, high-quality benchmarks are essential to advancing the development, validation, and adoption of AI tools in oncology and ensuring they are clinically meaningful, trustworthy, and representative of real-world patient populations.

Through research partnerships and policy initiatives, *Friends* emphasizes the need for rigorous, transparent, and reproducible validation of AI tools. A core component of this work includes the use of common datasets to evaluate performance, variability across models, and alignment with clinical expectations – supporting the development and use of independent reference datasets and benchmarks that can enable consistent, efficient, and representative assessment of AI tools. These principles reflect considerations outlined in a white paper developed by *Friends* in collaboration with key stakeholders, which explored best practices for designing and applying reference datasets in digital pathology[1].

Below we provide input on specific use cases that would benefit from high-quality benchmarks, characteristics of ideal reference datasets, relevant existing resources, barriers to benchmark development, and actionable recommendations for advancing this work.

Prioritize Benchmark Development for High-Impact, Under-Validated AI Use Cases in Cancer Care

The development and widespread availability of benchmarks is essential to ensure the consistent, reproducible, and reliable performance of AI tools in oncology. Establishing benchmarks is a foundational step toward building trust in AI applications and supporting an efficient path to validation, helping ensure tools are fit for clinical use and can inform regulatory decision-making. Use cases where AI directly influences patient diagnosis, treatment decisions, or trial eligibility should be prioritized for benchmark development.

One such high-potential application is quantitative biomarker assessment in digital pathology. AI tools designed to assess expression levels of biomarkers such as HER2 are increasingly being explored to support treatment selection in cancer care. At the same time, emerging applications,

[1] Considerations for Developing Reference Data Sets for Digital Pathology Biomarkers.pdf

such as using AI to predict mutation status or molecular subtypes directly from histopathology images, could support more efficient patient enrichment in clinical trials, particularly when biomarkers are rare or tissue availability is limited[2]. These tools have the potential to improve reproducibility, reduce diagnostic burden, and expand access to precision oncology. However, they are also highly sensitive to variability in model development, dataset composition, and validation approaches, reinforcing the need for well-curated, representative benchmarks.

In the Digital and Computational Pathology Tool Harmonization (PATH) Project[3,4], *Friends* convened multiple stakeholders to evaluate the performance of 10 independently developed AI tools compared to human readers in assessing HER2 expression in breast cancer specimens. The project used a shared reference dataset of over 1,000 digitized tissue images. Each participating group applied their computational pathology models to this common dataset to evaluate concordance and variability.

This project showed that, when analyzing the same set of images, AI models produced varying HER2 scores in some instances, particularly for samples with low expression or borderline staining intensity (Figure 1). This reflected similar variability observed among pathologists. These differences likely stemmed from variation in training and validation data, implementation of scoring criteria, and technical aspects of model design. These findings underscore the value of applying a common dataset to multiple tools, which can reveal important sources of variation and inform future development. **Without a shared reference benchmark, it is difficult to interpret or compare model outputs in a consistent and clinically meaningful way.**

This case study exemplifies the value of creating well-curated, independent reference datasets that can serve as benchmarks for AI tools in digital pathology. Such benchmarks would allow developers and regulators to evaluate model performance under consistent conditions, identify sources of disagreement or bias, and build trust in AI tools used in cancer care decision-making.

While digital pathology is a clear priority area, several other high-impact use cases would also benefit from the development of robust benchmarks, including:

- Segmentation of solid tumors in radiographic imaging: Tasks such as identifying tumor boundaries in brain, lung, or liver imaging are critical for assessing disease progression or treatment response. Benchmarks can improve reproducibility of AI-based contours, which are often subject to inter-reader variability in manual workflows. *Friends* is currently leading the AI-Based Response Evaluation Criteria in Solid Tumors (ai.RECIST) project[5], a multi-stakeholder effort to compare AI-based tumor measurements with human assessments of

[2] Supporting_the_Application_of_Computational_Pathology_in_Oncology.pdf
[3] Digital PATH Project | Friends of Cancer Research
[4] Agreement Across 10 Artificial Intelligence Models in Assessing HER2 in Breast Cancer (Manuscript under review)
[5] ai.RECIST Project | Friends of Cancer Research

RECIST response using a shared, annotated imaging dataset. This work underscores the importance of well-curated benchmarks to evaluate concordance, reproducibility, and clinical applicability of AI tools in imaging analysis.

- Use of AI in Real-World Data (RWD) applications: As AI models are increasingly used to extract or infer key clinical variables from RWD sources such as electronic health records (EHRs), pathology reports, or unstructured clinical notes, benchmarks are needed to evaluate model accuracy and generalizability. For example, AI tools may be used to derive tumor response, progression, or biomarker status from text or imaging in the absence of structured data. Reference datasets that include expert-annotated ground truth and diverse source systems could enable objective assessment of these tools' performance and fitness-for-use in regulatory or research settings.

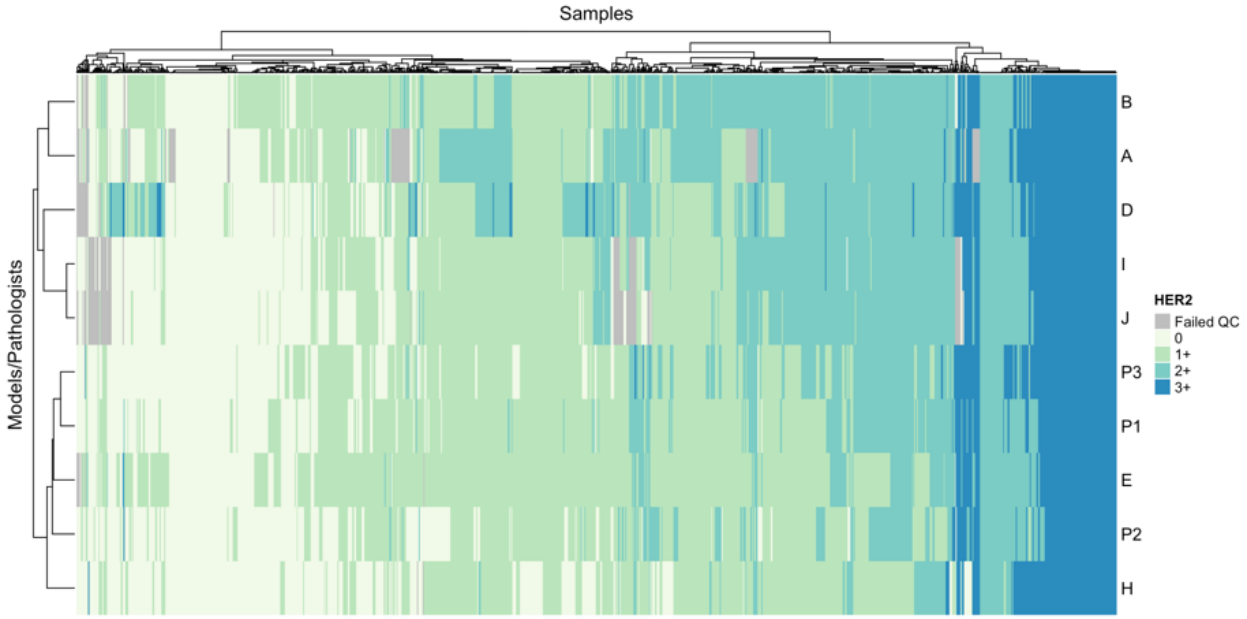

Figure 1: HER2 outputs show variability across AI models and pathologists, with the highest variability observed in 1+ and 2+ scores, while 3+ scores demonstrate greater consistency

Design Benchmarks That Reflect Real-World Complexity and Regulatory Relevance

To be effective and scalable, benchmarks must be thoughtfully designed to reflect both scientific rigor and real-world applicability. *Friends* recommends the following considerations:

- Standardized and representative: Datasets should reflect real-world heterogeneity in patient demographics, disease subtypes, imaging/scanning protocols, and clinical practice patterns.

- Well-annotated and transparent: Ground truth labels should be based on consensus reference standards (e.g., pathologist panels, molecular confirmation, clinical outcomes), and the provenance of data and annotations should be clearly documented.
- Publicly accessible and sustainably maintained: Benchmarks should be open-access or low-barrier to access, governed by transparent data use agreements, and supported by long-term infrastructure for hosting, updating, and curation. When benchmarks are used for formal performance evaluation, sequestering of test data or use of blinded evaluation processes may be necessary to avoid model overfitting and preserve the integrity of the assessment.
- Fit-for-purpose and linked to outcomes: Benchmarks should align with clinical use cases and regulatory expectations, and where possible, linked to treatment outcomes.

Leverage Existing Datasets to Accelerate Benchmark Creation

While few datasets have been developed specifically for benchmarking, several existing resources offer valuable content that could be adapted for this purpose with additional curation and standardization:

- The Cancer Imaging Archive (TCIA)[6]: A repository of over 160 de-identified radiographic and pathology image datasets across multiple tumor types. Many include expert-derived segmentations, clinical annotations, and structured metadata, which supports AI use cases such as tumor detection, segmentation, response assessment, and classification.
- The Cancer Genome Atlas (TCGA)[7]: A large-scale multi-modal dataset across 33 cancer types, including imaging, genomic, and clinical variables, which could enable integrated benchmarking use cases such as survival modeling and multi-omic classification.

Transforming these resources into benchmark datasets would require dedicated efforts to ensure representativeness, define shared intended-use contexts, and establish consistent standards. However, using these existing high-quality datasets could significantly accelerate the availability of benchmarks. In addition, they could inform the development of transparent, reproducible processes for dataset qualification to ensure the quality and credibility of future benchmarks.

Address Systemic Barriers to Benchmark Adoption and Use

Despite the recognized value of benchmarks, several challenges continue to limit their development and adoption:

- Data access and sharing constraints: Institutional policies, data silos, and privacy concerns often impede sharing of high-quality datasets, even for research or regulatory validation.

[6] Welcome to The Cancer Imaging Archive - The Cancer Imaging Archive (TCIA)
[7] The Cancer Genome Atlas Program (TCGA) - NCI

- Lack of centralized infrastructure: Without a long-term, well-funded infrastructure to host, curate, and govern benchmarks across the cancer research ecosystem, fragmented datasets may lose their value and consistency.
- Fragmentation in annotation and clinical standards: Disparities in clinical scoring thresholds, image quality, or labeling can limit interoperability and consistency in benchmarks.
- Resource-intensive curation processes: Creating well-annotated, diverse, and validated datasets requires significant investment in expert labor, technology infrastructure, and stakeholder coordination.
- Unclear regulatory alignment: Developers may hesitate to invest in benchmark development and use if their relevance to regulatory submissions or approval pathways is unclear.

Advance Benchmarking Through Collaboration and Regulatory Integration

To accelerate the development and impact of benchmarks in AI-driven cancer research and care, we recommend that NCI:

- Promote multi-sector collaborations to curate and govern disease-specific benchmarks, leveraging existing initiatives and guidance documents, such as the U.S. Food and Drug Administration's (FDA) AI framework[8] or the National Institutes of Health's Bridge2AI program[9].
- Align benchmarks with regulatory science, ensuring they are useful for demonstrating AI model reliability, reproducibility, and applicability in contexts relevant to FDA submissions or clinical decision-making.
- Incentivize contributions to open benchmarking efforts, including funding opportunities, publication tracks, or regulatory qualification programs that incentivize transparency and innovation.

Friends appreciates NCI's leadership in advancing safe and effective AI in oncology. We strongly support the development of high-quality benchmarks, particularly independent reference datasets, as a foundation for AI tools that serve all patients, foster trust, and meet the highest scientific and regulatory standards.

Sincerely,

Mark Stewart
Vice President, Science Policy
Friends of Cancer Research
mstewart@focr.org

[8] Considerations for the Use of Artificial Intelligence To Support Regulatory Decision-Making for Drug and Biological Products
[9] Bridge to Artificial Intelligence (Bridge2AI) | NIH Common Fund