Endocrine Society comments in response to NOT-OD-25-037, “Request for Information (RFI): Benchmarks for Artificial Intelligence in Cancer Research and Care”.

Response was informed by members of the Research Affairs Core Committee (RACC).

Comments were submitted electronically via online submission form on August 28, 2025.

Question 1: What are AI-relevant use cases or tasks in cancer research and care that could be advanced through the availability of high-quality benchmarks?

Artificial intelligence (AI) is a broad term used to encompass a wide array of technologies that can be used to advance basic science, clinical research, and patient care. To harness these capabilities, NIH should develop programs, such as instructional workshops and consortia funding, that encourage scientists and AI architects to collaborate and support the work needed to establish high-quality benchmarks.

We suggest several ways to advance cancer research through AI. Individuals who undergo life-saving chemotherapy and radiation treatments for their cancer will have negative effects on other organs, and the ability to have children is affected in approximately 60 percent of pediatric cancer patients. Research to investigate how oocytes, or the future eggs in girls and women, are affected by these necessary treatments to innovate ways to protect them from damage utilizes histology images to analyze oocytes within entire ovaries of model organisms. These animal models that specifically modulate genes within pathways important for DNA break repair and cell death, and small molecules to support viability and health of oocytes that mature into viable eggs. AI or algorithms designed with machine learning could facilitate analyses of oocyte counts and health within tissue sections, identify which transgenic animal models that selectively modulate genes may be redundant with other lines modulating other genes from the same pathways, and combined with enough foundational information, would support predictive models for identifying small molecules and translation to human tissue models. We emphasize that the use of AI would reduce but not eliminate use of animal models in this and other work as it would be impossible to predict off target effects that occur in whole body systems. Models do not replicate how a human would respond to an experimental variable and thus, we must still use animals until full body systems are replicated and tested by a different means. AI would likely increase efficiency in predicting fewer genes to target and small molecules to test, but would only be as good as the data that exists to create the AI pipelines and by the expertise in biology used to test these systems.

Additionally, there are several endocrine-related conditions and cancers in which patient care could be advanced using AI and given high-quality benchmarks. For example, nodules and tumors found in organs such as the thyroid, pituitary gland, adrenal gland, and in breast tissue are often difficult to evaluate for abnormalities based on pathology images alone. Information from spatial transcriptomic data and longitudinal electronic medical records can be used in conjunction with imaging data to train AI models through pattern recognition and predictive modeling to potentially help identify and diagnose malignant nodules that require further medical evaluation. Standard pathology specimens analyzed through AI-generated pipelines that are established with this information described above may accelerate the rate at which patients are diagnosed and receive care, especially in vulnerable patient populations.

Question 2: What are the desired characteristics of benchmarks for these use cases, including but not limited to considerations of quality, utility, and availability?

AI tools are only as good as the data used to develop underlying algorithms. Benchmarks used to develop AI models must consider the following points:

1. Quality of data: Datasets used for AI must include details on the technical preparation of specimens and incorporate quality control across batches of specimens. AI resources that use imaging data must consider the quality and resolution of images and be consistently matched with baseline, expert analyses to establish quality control metrics. Image capture and processing can significantly affect data consistency. Therefore, the completeness of patient tracking within datasets, with diagnostic imaging and labs, throughout progression of cancer or disease development and including outcomes would contribute to data quality.
2. Utilization: AI utilization must be planned for before the development and acquisition stages of data collection to ensure consistency in the quality and collection of data through the development and utilization of standardized protocols, specimen processing, and image acquisition.
3. Availability: Publicly available datasets with access to raw data would be an efficient and economical use of existing resources. Open-source software like QuPath can develop imaging analysis pipelines that are shareable. Publicly available datasets and software should allow researchers to access different patient cohorts and contribute to larger and more diverse datasets. The resulting software must be accessible to research and care providers who must test and utilize this information.
4. Biological heterogeneity: Any -omics analyses used in cancer research must consider the ages and strains of model organisms. The heterogeneity of the human population must also be accounted for with statistical tools that provide confidence intervals and appropriate statistical power. The human population included must be representative of individuals that contract these diseases.
5. Adaptability: To ensure that benchmarks are up-to-date and reflect new information and datasets, processes should be established to review and update benchmarks as datasets grow and expand their collection of representative data.
6. Transparency: AI tools developed for cancer research and care must disclose how the results are drawn. This will be important for periodically checking the quality of the AI tool and how results may be skewed because of previously unknown compounding factors. It may reveal new benchmarks that need to be set as more information is gathered and provide more context for projected results or confidence variables. Importantly, the staff, methods, funding, and development of each benchmark should be publicly disclosed. This

transparency will help researchers and healthcare providers determine how much weight should be given to the results obtained using AI technologies.
7. Utility: AI tools must be user friendly. A cancer biology researcher with minimal AI experience may not adopt an AI tool requiring additional staff support. A physician in cancer care needs to be able to utilize the AI tool within the regulatory framework of the hospital, including considerations for HIPPA compliance.

Question 3: What datasets currently exist that could contribute to or be adapted for benchmarking? Please include information about their size, annotation, availability, as well as AI use cases they could support.

The Cancer Genome Atlas (TCGA) is a program within the National Cancer Institute at NIH that contains a wealth of publicly available data from over 20,000 cancer and healthy patients representing 33 types of cancers. Genomic, epigenomic, transcriptomic, and proteomic data from this data set combined with imaging data can help train powerful, accurate AI models to identify and diagnose cancer.

The All of Us Research Program is another database maintained by NIH which contains patient and geographic information from over 865,000 participants and has collected over 609,000 biological samples. The wide breadth of information available through this program is especially important for incorporating health disparities into AI training.

ClinVar is another publicly available database maintained by NIH that contains information about human genetic variants with respect to their impact on human health and responses to drugs. The incorporation of genetic information into benchmark development is important for disease prediction.

The Centers for Disease Control and Prevention (CDC) maintains domestic and global epidemiological data sets that can be integrated into AI models for the development of benchmarks as well.

AI technology is being used for predictive modeling and developers can refer to existing tools as a guide for the development of benchmarks for cancer care and research. AlphaMissense is an example of one of these technologies used in oncoendocrinology as it is used to predict the pathogenicity of missense variants and ultimately help inform clinical decisions.

Question 4: What are the biggest barriers to creating and/or using benchmarks in cancer research and care?

Cost is a significant barrier to creating benchmarks for cancer research and care. Building and maintaining high-quality data sets will require significant financial investment. Clinical hours will be needed to obtain patient samples and additional medical staff hours will be required to generate images for AI model development. To ensure consistent and reliable information is collected, whether it be imaging data or other types of information, funding will be needed to make sure software is available to researchers and medical professionals to support benchmark development.

Funding cuts to NIH research could create significant barriers to creating and/or using benchmarks. Large initiatives like the All of Us Research Program will require sustained funding to enable utility now and in the future for the creation of AI benchmarks. Furthermore, funding cuts may disproportionately impact research that would create more effective benchmarks, such as research involving diverse patient populations. We also note that training programs to generate a diverse scientific workforce in cancer research and care will be necessary to generate sufficiently representative datasets. A representative and international workforce introduces new and different perspectives that would result in different approaches to collecting data, enhancing the inclusivity of datasets collected and strengthening the utility of the models and tools developed. Without funding programs that train scientists and generate data involving sufficiently representative populations, gaps in data and expertise will decrease the accuracy and reliability of benchmarks for cancer care and research.

Patient accessibility to proper health care and clinical research is another barrier to creating quality benchmarks. Some populations may not have the same level of access to care as others, leading to fewer opportunities to collect patient information and ability to contribute to representative datasets. Improving and expanding access to cancer care and research opportunities for patients who would not normally have access will expand image and data collection for the development of benchmarks to train AI models and to improve their accuracy. Additionally, lower-resourced communities may not have funding, technologies, or other resources needed for data collection. Such communities may not be equipped to collect data or other information that can be incorporated into databases for benchmark development. Thus, investments must be made in these communities to ensure that the quality of any data gathered is of comparable quality with existing data used for benchmark development.

Patient buy-in, trust in science and medicine, and the willingness to participate in clinical research are large barriers to overcome. Collaborations between cancer researchers and care providers with patient advocates should be considered early in this process to hear patient concerns and promote participation in data collection. Transparency at each stage of development will also foster trust and support by the public who could benefit from these tools.

Last, maintaining patient privacy and providing appropriate Institutional Review Board (IRB) oversight for the use of patient information in the development of new benchmarks is another potential barrier. Use of the single-site IRB framework would alleviate some of this barrier.