**December 2024**

AACI Commentary

# The Role of Data and Data Science in Cancer Research

*By W. Kimryn Rathmell, MD, PhD, MHHC, and Warren A. Kibbe, PhD*

***W. Kimryn Rathmell, MD, PhD, MHHC**, is director of the National Cancer Institute.*

***Warren A. Kibbe, PhD**, is deputy director for data science and strategy at the National Cancer Institute.*

## *Commentary* Overview

- 2024 has been the year of artificial intelligence (AI). Therefore, it is timely to elevate the National Cancer Plan goal to "Maximize Data Utility" for our national network of cancer centers.
- Necessitating that all forms of data be "AI ready" has become a key consideration for investigators and cancer centers, highlighting the importance of high-quality data for validation and training.
- Having accessible, usable data and a cancer research workforce fluent in data science will allow us to maximize the value of cancer research data and accelerate the translation of new knowledge into patient benefit.

**2024 has been the year of artificial intelligence (AI). The 2024 Nobel prizes in Physics and in Chemistry recognized and emphasized the fundamental research that has led to modern AI and the application of this research to designing and predicting protein structure. Both achievements have fundamentally changed our work to understand, control, and eradicate cancer. Therefore, it is timely to elevate the National Cancer Plan goal to "Maximize Data Utility" for our national network of cancer centers.**

Beginning with large language models (LLMs), these tools have democratized the AI universe and demonstrated the raw power in data science – highlighting the need for high-quality, well-annotated, and well-understood data for validation, as well as the need for well-characterized data for training. Interestingly, necessitating that all forms of data be "AI ready" has become an important consideration for a diverse array of investigators and cancer centers. (See AI Readiness for Biomedical Data, AI-READI.org, Artificial Intelligence Data-Readiness

Challenge, and EDRN AI Readiness Guidelines.) Research areas and programs with accessible, consistent, well-defined datasets will be (and are!) the first to benefit from rapid advances in machine learning and AI.

Data science is poised to revolutionize cancer research, and we envision a future in which data and AI intersect seamlessly with everything from discovery to clinical care and administration. As cancer centers prepare to meet this future, let's consider the factors that enable an AI-ready organization.

## Data Science in the Cancer Centers

Having had the chance to visit multiple cancer centers during the past year, we were impressed to see the infusion of data science into every aspect of cancer research. Seeing centers that have incorporated geospatial analysis techniques into the assessment of catchment areas; incorporated cancer registry data; and overlaid social determinants of health (SDoH) measures coming from the U.S. Census, the American Community Survey, the Social Vulnerability Index, and other sources, demonstrates the power of shared activities and approaches in community outreach and evaluation across the cancer center community.

We were similarly impressed by the breadth of data science-intensive approaches and understanding from multi-omic analyses coupled with spatial -omics; multiple imaging modalities; and innovative microfluidic methods for culturing and interrogating 3D tissues to identify pathways, systems, and perturbations that provide novel insights into both normal and cancer biology.

In the clinical realm, the use of digital twins to understand optimal radiation therapy has moved rapidly from a thought experiment to reality. Targeted therapies, immunotherapies, and integration of detailed molecular characterization generate deeply informative datasets; inform patient care; and are analyzed using informatics, computational biology, AI, and more. The use of patient reported outcomes to model and inform treatment requires the development of data science approaches, as does the use of wearable data in clinical research. Equally promising is the emergence of AI tools for clinical trials matching (**Jin et al., *Nature Communications*, 2024**). Currently being piloted as TrialGPT, this partnership with the National Library of Medicine provides an exciting view to a digitally enabled future. In cancer survivorship and cancer surveillance, the use of advanced machine learning promises to make reporting on cancer incidence and outcomes much faster. The cancer data science community is incredibly diverse, dynamic, and active!

## Well-Defined Data

Foundationally, access to data, including its provenance (dates and times, personnel, protocols used for data generation, experimental design, information on instruments, quality control methods, reagents, calibration, etc.) is critical for the three Rs: robust scientific analysis, rigorous experimentation, and reproducibility of data (**Vasilevsky et al., *PeerJ*, 2013**). Access to provenance documentation and inclusion of user-defined elements is also critical for meta-analysis, to make inferences across more than one dataset. Creating standard definitions and rigorously applying them to data lays the groundwork for large scale analyses. The **NCI Thesaurus**, **Metathesaurus**, **Enterprise Vocabulary System (EVS)**, and the **cancer Data Standards Repository (caDSR)** represent just some of the resources that provide semantic, ontological, and common data element resources for cancer research.

## Large Data Collection Programs

The power in data often comes when the available dataset is large enough to address statistically meaningful questions. It is relatively straightforward to analyze and make inferences when all the data have been generated and collected in a similar way, using either the same or harmonized protocols, such as the sequencing data curated across the different types of cancers analyzed as a part of **The Cancer Genome Atlas (TCGA)** program. It is harder, and more susceptible to batch effects and other systematic sources of bias, when looking across data representing many methods for collecting samples, storing, and sequencing the tissues – but it is still possible and valuable to do.

In fact, it can be a strategy to integrate across platforms. In the sequencing space, consider the analysis of clinical sequencing available through **AACR Project GENIE**. Likewise, **The Cancer Imaging Archive** and the **Imaging Data Commons** are excellent repositories of clinical imaging data. These foundational data resources can be used to train and validate new analysis methods,

make discoveries, and validate discoveries made in the lab. Moreover, the heterogeneity that is infused by the various imaging sources provides a natural extent of variation, such that emerging signals are more readily identified.

## Data Sharing Across Cancer Research Fields

As highlighted above, harmonized experimental protocols using mature technologies simplify efforts to make data consistent and accessible. Much of cancer research lacks that consistency, making it harder to co-analyze and infer findings across projects from multiple research fields.

However, many of the newer deep learning methods, such as LLMs, can now work with complex datasets generated using different protocols and experimental designs, assuming the datasets themselves are well annotated with this information (**Li et al., arxiv, 2024; Truhn et al., *npj Precis. Onc.* 2023**). Likewise, new methods for analyzing multimodal data (imaging, DNA sequencing, RNA sequencing, proteomics, metabolomics) are under development (**Steyaert, et al., *Nat Mach Intell*, 2023**; **Wakas et al., *Front Artif Intell*, 2024**). Adding these datasets to repositories like the **NCI Cancer Research Data Commons** will further fuel these developments.

## Data and Data Science as Enablers

Bringing together opportunities in data analysis, cancer data, and tools like AI and LLMs enables new methods of discovery and innovation in cancer research. At NCI, we strongly believe it is critical to make data generated through government-supported research accessible and available to facilitate serendipitous cancer research. Similarly, data science, broadly defined as any computational or data visualization method, is critical to "transforming data into information into insight into knowledge" and fundamental to enabling cancer research. Having accessible, usable cancer research data and a cancer research workforce fluent in data science will allow us to maximize the value of cancer research data and accelerate the translation of new knowledge into patient benefit.

Here, we must consider that data science provides a unique opportunity to engage patients, providers, and researchers in foundational discovery. Creating the language, the opportunities, and the resources will transform the health care system. It goes without saying that training in data science by students and clinical trainees is a must to develop the workforce of the future.

## Existing and Future Examples of the Importance of Cancer Data Repositories

Finally, well-defined, accessible data is transforming today's cancer research and influencing precision oncology. **TCGA**, **Therapeutically Applicable Research to Generate Effective Treatments (TARGET)**, and now the **Childhood Cancer Data Initiative (CCDI) Molecular Characterization Initiative** are providing the reference data for cancer molecular tumor boards so they can make the best treatment assignments for individual patients. There are similar, exciting opportunities for dramatically improving the diagnostic and prognostic value of many types of imaging in cancer; the key to realizing that value will be making those images and their associated clinical annotations accessible for data scientists to test and refine their tools. The ability of researchers to glean insights from public data, while enabling a rich portfolio of open-source research and commercial tools, will be key to bringing academic and industry researchers and entrepreneurs together to continue to innovate and provide real value to patients, cancer survivors, and everyone at risk for cancer.

## Educating Cancer Researchers on the Capabilities and Limitations of Data Science Tools

Preparing novice or amateur investigators to apply AI tools is the cornerstone of a successful program. This preparation includes specific training in data input, analysis, and experimental prioritization. For example, while the tendency of ChatGPT to confabulate when given certain prompts has received significant attention, it is important to recognize that every analysis and visualization technique has appropriate uses and limitations. (Think **"p-hacking"** and other methods of "torturing data until it confesses.") Similarly, visualization techniques can be used to tell very different stories with the same data – and **can be misleading**. Image manipulation to improve clarity can be good. Manipulation designed to obfuscate or misrepresent is bad. We need to have a workforce that understands the power of these tools, how to use them properly, and how to discern when they have been used improperly. While we have focused on visualization techniques, many types of analyses can be misused, either accidentally or deliberately. It is very important to develop fluency in interpreting analysis results and data visualization.

## NCI Data Science Touchstones

Finally, an important part of NCI's role in data science is working with the entire cancer research community to train the next generation of data-savvy cancer researchers. Ensuring that tomorrow's data science workforce reflects the talent and diversity of our nation is vital to our goal of improving the health of every American. The **Genomic Data Commons** and the **Cancer Research Data Commons** are important parts of the cancer research data foundation, and help to define the ecosystem in which future generations of cancer scientists will drive discovery.

Together, we are preparing for a new era of cancer research – one with new and exciting patterns and features to discover, made possible by a radically new set of tools that gives us a view of cancer biology and disease processes that will lead to entirely new observations, discoveries, and understanding, and contribute to the end of cancer as we know it.

## Our Mission

The Association of American Cancer Institutes (AACI) represents over 100 premier academic and freestanding cancer centers in the United States and Canada. AACI is accelerating progress against cancer by enhancing the impact of academic cancer centers and promoting cancer health equity.

## About AACI *Commentary*

To promote the work of its members, AACI publishes *Commentary*, a monthly editorial series focusing on major issues of common interest to North American cancer centers, authored by cancer center leaders and subject matter experts.





Association of American Cancer Institutes (AACI) | Medical Arts Building 3708 Fifth Ave, Suite 503 | Pittsburgh, PA 15213 US