*Learning Health Systems*

COMMENTARY OPEN ACCESS

# Twenty-First Century Data Systems: Evolving Cancer Registries to a Learning Health System

Peter P. Yu[1] | W. Scott Campbell[2] | Eric B. Durbin[3] | Lawrence N. Shulman[4] | Jeremy L. Warner[5]

[1]Hartford HealthCare Cancer Institute, Hartford, Connecticut, USA | [2]University of Nebraska Medical Center, Omaha, Nebraska, USA | [3]Kentucky Cancer Registry, Markey Cancer Center, University of Kentucky, Lexington, Kentucky, USA | [4]Abramson Cancer Center, University of Pennsylvania, Philadelphia, Pennsylvania, USA | [5]Brown University Warren Alpert School of Medicine, Providence, Rhode Island, USA

**Correspondence:** Peter P. Yu (peter.yu@hhchealth.org)

**Received:** 11 August 2025 | **Revised:** 27 March 2026 | **Accepted:** 1 April 2026

**Keywords:** learning health system | neoplasms | registries

## ABSTRACT

**Introduction:** Oncology is a data-rich environment reflecting the increasing incidence of cancer in the US aging population, transformation of cancer into a chronic disease due to advances in treatment, and the emergence of new data categories such as genomics. Concomitantly, healthcare systems are challenged to meet regulatory and voluntary reporting of cancer data to support government, quality certification, research and strategic partnership needs.

**Materials and Methods:** The National Academies of Science, Engineering, and Medicine organized a workshop entitled "Enabling 21st-century applications for cancer surveillance through enhanced registries and beyond" with representation from NCI Comprehensive Cancer Centers, oncology and medical informatics professional societies, industry, CDC, NCI, and patient advocacy groups.

**Results:** The proliferation of cancer registries has resulted in heterogeneity in data vocabularies and data transport standards. Federal policy is complementing private initiatives to modernize cancer data architecture that would support a Learning Health System. However, business models are needed to provide sustained investments in data infrastructure.

**Conclusion:** A computational approach to cancer registries would set the stage for interoperability and data sharing within a learning health ecosystem. Healthcare systems need to invest in their data infrastructure to improve data quality and adaptation of new data sources such as genomics and wearables. The ecosystem must evolve business models to sustain these investments.

## 1 | The Learning Health System

The Learning Health System (LHS) imagines how digital health data obtained from and reflecting routine health care delivery can become a resource to drive discovery, innovation and improvement in the quality and sustainability of health. The LHS concept relies on building an ecosystem wherein all may benefit from shared or interoperable data resources, all are motivated to contribute, and business models support governance. Starting in 2007, The National Academies of Science, Engineering, and Medicine convened a series of workshops to advance a LHS in the United States [1]. Two early notable instances of an oncology focused LHS were initiated by the American Society of Clinical Oncology (CancerLinQ, subsequently acquired by ConcertAI) and Flatiron Health, a commercial entity now part of F-Hoffmann-LA Roche AG [2, 3]. These efforts have published insights into using real-world evidence but limitations on the quality of health data capture such as comprehensiveness, completeness, data structure and fidelity, metadata and interoperability have required heavy reliance on subject matter expert review of electronic health records (EHRs). Furthermore, both LHS examples require that medical data be exported to a central data repository, crossing institutional firewalls that protect health data security and privacy. The dependency on human

curation and the cost of redundant data hosting are limitations on scalability, sustainability and utility. Commercial based LHS rely on monetization of data and healthcare entities may have limited insight over how this is done or equitable sharing in revenue generation.

## 2 | Current State of the U.S. Cancer Registries

Health registries collect population level data aggregated from patients who share certain clinical characteristics. Such characteristics may be broadly inclusive (all cancer patients within a country, state or healthcare system) or more narrowly focused (patients with a specific disease undergoing a specific treatment intervention such as a certain hip prothesis model). For over 50 years the U.S. national cancer registry systems have collected comprehensive, real world cancer data compliant with data representation and transmission standards, and of quality sufficient for scientific and statistically valid analysis. The launch of NCI Surveillance, Epidemiology, and End Results (SEER) in 1973 and CDC National Program of Cancer Registries (NPCR) in 1992 created cancer data repositories to support research and public health. These federal programs in turn fund state central cancer registries who hold regulatory authority over healthcare delivery entities. This design has led to a series of stacking cancer registries where hospitals who must meet regulatory requirements to collect and report cancer data report to state level cancer registries maintained by their state Department of Health. State cancer registries are in turn responsible for collecting data coming from various hospitals into one patient level record. Subsequently those state cancer registries report data to the national registries. This data flow is largely unidirectional, so hospitals must maintain their own internal data registries if they wish to preserve data to support operational initiatives such as quality improvement. Hospitals must also deal with multiple sources of cancer data. Pathology data resides in Laboratory Information Systems, radiographic imaging data in Picture Archiving and Communication systems (PACS), radiation oncology in dedicated systems that support radiation field and dose calculations, cancer research data in Clinical Trials Management Systems, and patient demographic and treatment data in EHRs. Evolving data sources for patient reported outcomes (PROs) and wearable devices and new data categories such as genomic data add to this complexity, and each data source has its own set of terminologies contained within a vocabulary.

Over time, the growing complexity of cancer data coupled with rising volumes as cancer incidence increases among an aging population have challenged cancer registries to keep pace. Cancer registry data is commonly 2 years out of date at time of reporting, mirroring the situation for other types of health data such as the National Death Index. Given the rapid advances in cancer treatments, two-year-old data may no longer reflect current standards of clinical practice. Advancing the automation of data reporting is an opportunity to improve timeliness of data reporting while reducing the cost of manual data abstraction; it is common for cancer center budgets to include multi-million-dollar allocations for registry reporting. State central cancer registries have embarked upon efforts to modernize data reporting within their jurisdiction [4, 5]. In 2024 the National Academies of Science, Engineering, and Medicine National Cancer Policy Forum sponsored "Enabling 21st Century Applications for Cancer Surveillance through Enhanced Registries and Beyond", a workshop to explore what can be done to advance robust and sustainable data architecture that promotes data liquidity and sharing [6]. This perspective reflects on the design of data architecture systems that lead the evolution of our current cancer registry enterprise to a Learning Health System.

## 3 | Modernization of Cancer Registries

Data architecture supports three fundamental and co-dependent properties, data quality, data sharing, and data governance. The first of these, data quality, begins with precise data definitions to ensure that data is fit for the intended purpose and data analytics. Data dictionaries function as controlled vocabularies that specify the common use name of the data element and associated values. For example, blood glucose is the name of the data element and acceptable values are expressed as mg/dl. Data dictionaries are constructed, along with accompanying metadata that specify important features such as the source and provenance of data. In the latter half of the twentieth century, SEER and NPCR, working with the North American Association of Central Cancer Registries (NAACCR) created the foundations of a national cancer data dictionary to support public health reporting. The United States Core Data for Interoperability (USCDI) is a medical data dictionary overseen by the U.S. Department of Health Services that is updated annually by the Assistant Secretary for Technology Policy. Its purpose is to increase interoperability of data and support healthcare delivery and federal regulatory policy. Incorporation of USCDI data elements into EHRs is a requirement for vendors to obtain certified electronic health record technology (CEHRT) status and use of CEHRT EHRs is linked to provider reimbursement by Medicare and Medicaid. The Twenty-first Century Cures Act extended the USCDI model by establishing U.S. Core for Data Interoperability Plus Cancer (USCDI+Cancer) to define essential cancer data elements necessary to support cancer clinical trials, identification of adverse events of immunotherapy, Center for Medicare & Medicaid Services Enhancing Oncology Model, and cancer registries [7]. Although adoption of USCDI+Cancer data elements is not a requirement for CEHRT status, several EHRs vendors have signaled their intent to adopt. The multiple use cases supported by USCDI+Cancer highlight a basic tenet that cancer data repositories of the future be shared resources that can be used by a range of stakeholders. Oncology professional societies have through their programs created additional data dictionaries; College of American Pathologists with digital synoptic pathology reports (eCPs), American College of Surgeons through the National Cancer Database (NCDB), and American Society of Clinical Oncology (CancerLinQ).

The comprehensiveness of data capture in machine-readable format within EHRs is a critical component of data quality. Incomplete capture of data undermines the precision of downstream data analytics and introduces potential systemic bias in the data set by underrepresentation of sub-populations of patients with cancer. Data that can be extracted with minimal human effort and curation reduces the historic reliance on manual chart extraction by costly subject matter experts who

themselves may introduce subjective bias. Structured data capture (SDC) describes the collection of data within EHRs such that the data can be extracted by digital technology without transformation. SDC combines the semantic precision of a data dictionary with specified representation within the EHRs such that the data can be searchable, also known as machine-readable. College of American Pathologists maintains a library of eCPs digital synoptic reports specific for over 90 cancer types, each including detailed structured data fields with defined terminology and values that are integrated into laboratory information systems. The requirement to use SDC contributes to the completeness of data capture.

The implementation of SDC needs to be thoughtfully integrated into the clinical workflow with human factors engineering to avoid adding to work related stress and burnout. Data entry itself is a data quality concern due to errors in data entry or missingness of data entry. An emerging alternative to SDC is the use of large language models (LLMs), or pre-transformed generative transformer models for data extraction from narrative text documents. LLMs can also be used to create narrative text directly from patient-clinician verbal encounters using ambient AI. These models use advanced artificial intelligence tools such as neural networks (deep AI) to parse narrative texts and extract data that is semantically correct and fit for analytic purposes, while allowing clinicians to continue to use narrative text documents that might better express the nuances of complex oncologic cases and decision-making. To be clear, it is unlikely that LLMs will achieve the accuracy of SDC and subject matter oversight will be needed. However, the accuracy of LLMs can be improved by referencing data standard documents using a process called Retrieval- Augmented Generation (RAG). For example, cancer staging definitions periodically undergo revision by the American Joint Committee on Cancer (AJCC) to reflect changes in cancer survival brought on by advances in cancer therapies. Cancer staging is a fundamental descriptor of a patient's cancer that impacts treatment decisions in the clinic and public policy at the population health level. LLMs could support cancer staging reporting by analysis of narrative texts such as clinical notes and radiology reports to supplement the SDC of pathology reports, while utilizing RAG to reference the most recent staging criteria in the AJJC Cancer Staging Manual.

The selection of data elements that support multiple critical use cases, the implementation of SDC that improves clinical efficiency, and the introduction of LLMs for data extraction are important functions of data governance.

Assuming a trustworthy level of data quality has been acquired, that data needs to move among the LHS ecosystem of cancer registries, the different healthcare systems that a patient may receive care from, academic or industry research partners, professional societies that certify quality, and patient advocacy groups. More precisely, it involves both the transport of data as well as the mapping of that data into the respective data systems of the receiving entity. Data transport generally involves packaging data into standardized templates or forms. U.S. cancer registry reporting utilizes NAACCR data templates for the transport of cancer data to central cancer registries. These templates (Volume II for case reporting and Volume V for cancer pathology reporting) are continually updated as data elements are added or re-defined [8]. AJCC and NAACCR closely coordinate to harmonize data definitions and data transport standards.

Health Level 7 (HL7) has created several general use standards for data transport. Fast Healthcare Interoperability Resources (FHIR) is an emerging standard that is web-enabled and avoids the costly need for custom one-to-one interfaces to be built for each instance of data exchange. Its adoption is a specification of the Twenty-first Century Cures Act. HL7 is currently testing an implementation guide for transport of cancer pathology data using SDC on FHIR which could facilitate transport of eCPs data directly to cancer registries and reduce the current reliance on oncology data specialists (cancer registrars) for this work [9].

A LHS is an ecosystem of stakeholders whose interactions arise depending on alignment of mission, as illustrated by CAP, AJCC, and NAACCR for cancer pathology reporting, by research projects, or by contractual relationships in the case of industry. Governance of decisions on whom to share data with and for what purposes is a business decision of the organization separate but also dependent on governance of data architecture design that enables data sharing. As noted previously data dictionaries provide semantic precision and accompanying metadata strengthens data quality assessment, however multiple data dictionaries exist and are in use as individual dictionaries may be best suited for a certain purpose. Systemized Nomenclature of Medicine-Clinical Terms (SNOMED CT) is an international standard through which different dictionaries can be related, as has recently been applied for cancer pathology data [10].

The application of machine learning and large multi-modal models to solve pressing challenges in healthcare requires access to extraordinarily large data sets. Centralizing data to create databases of necessary volume requires considerable effort in data transport, raises concerns over data privacy, security, and governance of ethical use of data, and duplicates data storage costs at the data source and central repository. Federated models for data sharing wherein data does not leave the healthcare system but instead data analytics are performed within the systems firewall are an alternative approach within a LHS. However, a federated model requires that the data architecture achieve a higher order of data representation that relates the data into an ontology structure. Data Models organize cancer data into categories such as patient demographics, genomics, treatment, and outcomes, so that the relationship between data can be analyzed; for example, the interaction between genomics and health or economic outcomes. Data models that organize data around a given patient provide a more holistic representation of that patient's health, which in turn facilitates aggregation of patients for population health and health system operational needs. The power of data models is more fully realized when different stakeholders of a LHS use the same data model, known as a Common Data Model, which facilitates data queries to be shared rather than data itself. Common data models can harmonize and relate various data dictionaries into overarching vocabulary such as SNOMED-CT. Two such common data models are emerging in oncology, Observational Medical Outcomes Partnership (OMOP) and minimal Common Oncology Data Elements (mCODE) [11, 12]. OMOP is governed by Observational Health Data Sciences and Informatics (OHDSI) and is designed for capturing real-world data for general medical use, with a recent extension of OMOP

for oncology [11]. mCODE is purposely designed for oncology use cases, such as clinical trials research and value-based care, and was initiated by the American Society of Clinical and governed today by the CodeX community. The National Patient-Centered Clinical Research Network (PCORnet) has created a common data model to support comparative effectiveness research. As there is not a single Common Data Model, a functional blending of OMOP, mCODE, PCORnet, and other data models along with existing data collection tools such as NAACCR will likely evolve as best suited to an entity's data needs and the rapidly evolving capabilities of AI driven data science. Adoption and harmonization of common data models can enable organizations that generate and host health data to unlock the value of their data assets.

## 4 | Evolution Toward a Learning Health Ecosystem

Implementation of data architecture improves data quality through data dictionaries that codify the semantic meaning of data elements and support data analytics, and through SDC and LLMs that improve data accuracy, completeness, and consistency. Standard data sharing templates and transport standards are architectural features that enable data sharing. Common data models that underlie federated data sharing systems address important features of data governance concerning data movement and storage.

A LHS is as dependent on having high quality data and the technical capability to transport and receive data as it is on the governance of how a health care entity forges partnerships and business relationships with other organizations that comprise the Learning Health System. Figure 1 illustrates the central role of a cancer registry that incorporates a CDM and enables data science. Surrounding that core is a circle of data sources that feed the cancer registry and define data provenance. At the outer segments are consumers of cancer data that represent the ecosystem of LHS stakeholders.

The cost borne by a health system to construct and maintain its data architecture is significant, but in part could be offset by improving the health system's internal operational efficiency and positioning the system for value-based healthcare delivery through focused data analytics. A LHS creates the opportunity for a healthcare system to forge strategic partnerships with information technology, healthcare industry, and academic institutions. Those relationships could establish a business model that recognizes and rewards the creation of intellectual property within the LHS.

## 5 | Challenges to Transforming Cancer Registries Into a Federated Platform for Shared Learning

1. **Commitment of healthcare systems to LHS.** Healthcare systems are integrated delivery networks that encompass multiple patient entry points for receiving healthcare: hospitals, ambulatory clinics and surgical centers, urgent care centers, home care and long-term care facilities. Hospitals form the core of these systems and hold the regulatory responsibility for reporting to state and federal cancer registries. In addition, healthcare systems often voluntarily report to professional cancer registries for quality improvement and certification purposes (NCDB) or for research (COVID-19 and Cancer Consortium). Together, the integration of reporting to various cancer registries can provide a more complete picture of a patient's story. For example, reporting on cancer treatment to government registries is largely collected only for the first line of therapy whereas combining data across registries results in a more complete narrative of cancer as a chronic disease with

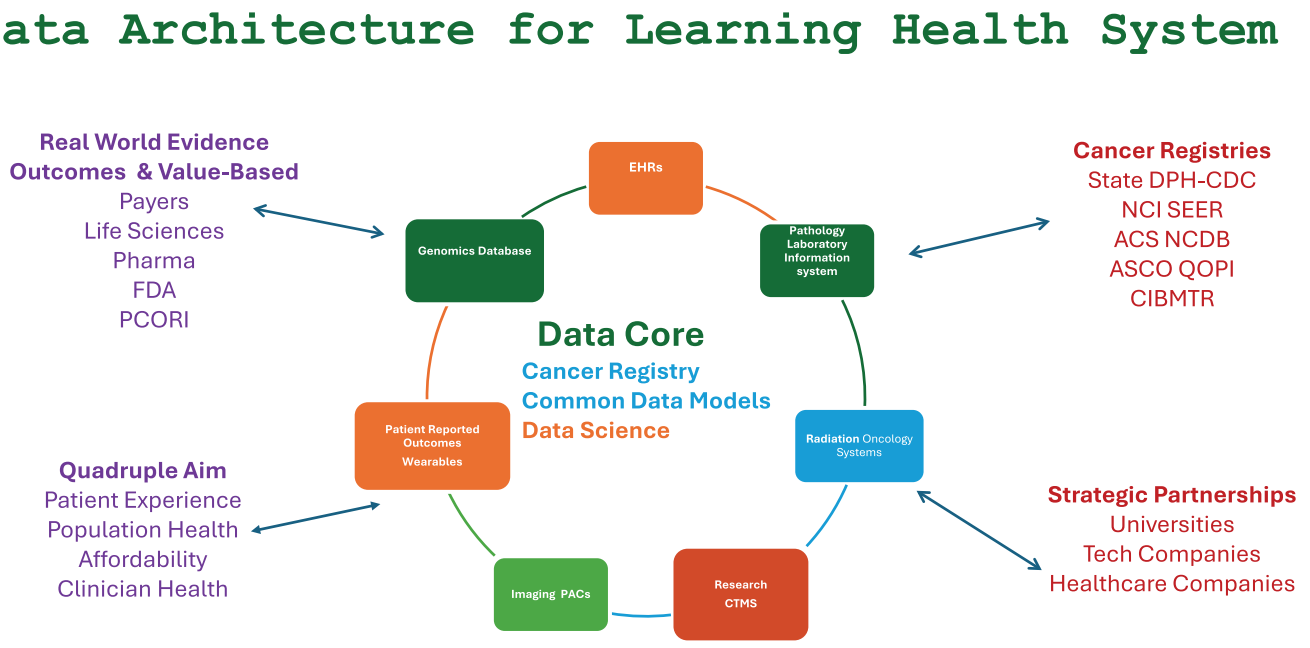


**FIGURE 1** | Data architecture built upon advances in data extraction, modeling, and liquidity creates a data resource that can be shared and consumed by a health ecosystem.

longitudinal patient trajectories. As the source of health data, healthcare systems hold responsibility for data quality, provenance and integration which requires a greater level of resource allocation in medical informatics, an investment that can be justified by gains in understanding and improving patient and population level health and economic outcomes.

2. **Improving the quality of health data**. Significant cancer data exists as structured data entered directly from EHRs (clinical lab values, chemotherapy and other administered medications) or by dedicated clinicians functioning at a high level of data integrity (pathologists entering eCPs digital synoptic reports). Data that capture cognitive thinking is less amenable to SDC, and here LLMs are introducing new solutions such as ambient AI and natural language processing of narrative medical documents. On the population health level, missing data can be partially overcome with data imputation, but another challenge is when a patient receives care across multiple healthcare systems and there is a need for cross-linking data. The state and federal cancer registries assume this responsibility for their registries, but this remains a challenge for healthcare systems that a federated data model does not solve. EHRs vendors such as Epic Systems provide the ability to share patient level data through Care Everywhere when healthcare systems elect to do so.

3. **Agile adaptation to emergent properties**. As medical science advances, molecular data on patient genomics, proteomics, and metabolomics is vital to the prediction of patient risk for cancer, the diagnosis and characterization of cancer, and subsequent precision medicine directed treatment. Molecular data have a high level of complexity in data representation. Patient reported outcomes capture data directly from the patient in a more complete manner without bias introduced by clinicians, and wearables provide real-time data at enormous scale. LHS will need to include new data forms and formats.

4. **Business models supporting investments in LHS**. The LHS is an ecosystem of producers of health data and consumers of that data, and ecosystem members can be both producers and consumers. Business models are constructed where there is convergence of interests. The Cancer AI Alliance is a collaboration of AWS, Microsoft, NVIDIA, Deloitte, and Dana-Farber Cancer Institute, Fred Hutchinson Cancer Center, Memorial Sloan Kettering Cancer Center, The Sidney Kimmel Comprehensive Cancer Center, and Whiting School of Engineering at Johns Hopkins. It is an ecosystem that unites industry and academic healthcare systems to support data science based on a federated data model.

## 6 | Conclusions

Cancer has a profound impact on patients, families, healthcare systems and the U.S. economy and that relevance is reflected in the heterogeneity of cancer data categories, the volume and rapidity of cancer data generation and number and types of cancer registries. Each cancer registry is designed for a specific purpose and contains data common across registries as well as data that is unique to that registry. At the center of this are healthcare systems that largely generate the data that feeds cancer registries and bear the burden of data reporting. That work can readily amount to annual costs measured in millions of dollars due to the heavy reliance on manual processes for data extraction, transformation and loading into registry data systems. Automation of this promises to improve the timeliness of data collection and quality while reducing costs. While a dollar saved is a dollar earned, unlocking the value of cancer data has the potential to generate return on investment in data architecture through strategic partnerships within the LHS ecosystem. The cost of standing up common data models and federated data sharing systems can be shared by industry or offset by shared revenue derived from intellectual property generated by the ecosystem. Cancer registries have laid a foundation for the systematic collection of cancer data. The opportunity is to build upon this through application of data architecture to improve the liquidity of data sharing and preparation of cancer data for data science led advances in achieving improved patient and economic outcomes.

### Author Contributions

All authors contributed to conceptualization, writing of original draft, review, and editing.


### Acknowledgments

The authors have nothing to report.


### Conflicts of Interest

W. Scott Campbell: Travel Expenses: GenomOncology. Lawrence N. Shulman: Research Funding: Gilead Sciences (Institutional). Jeremy L. Warner: Stock and Other Ownership Interests: HemeOnc.org. Honoraria: Henry Ford Health System, Prime Healthcare Services. Consulting or Advisory Role: Westat, Lewin Group, University of Texas Medical Branch, Nemesis Health. The authors declare no other conflicts of interest.

### Data Availability Statement

No data is presented in this commentary.

### References


1. National Academies of Science, Engineering, and Medicine, The Learning Health System Series, 2025, accessed April 25, 2025, https://nam.edu/our-work/programs/leadership-consortium/learning-health-system-series/.

2. D. Potter, R. Brothers, A. Kolacevski, et al., "Development of CancerLinQ, a Health Information Learning Platform From Multiple Electronic Health Record Systems to Support Improved Quality of Care," *JCO Clinical Cancer Informatics* 4 (2020): 929–937, https://doi.org/10.1200/CCI.20.00064.

3. E. H. Castellanos, B. K. Wittmershaus, and S. Chandwani, "Raising the Bar for Real-World Data in Oncology: Approaches to Quality Across Multiple Dimensions," *JCO Clinical Cancer Informatics* 8 (2024): 1–11, https://doi.org/10.1200/CCI.23.00046.

4. P. Wormeli, J. Mazreku, J. Pine, et al., "Next Generation of Central Cancer Registries," *JCO Clinical Cancer Informatics* 5 (2021): 288–294, https://doi.org/10.1200/CCI.20.00177.

5. T. C. Tucker, E. B. Durbin, J. K. McDowell, and B. Huang, "Unlocking the Potential of Population-Based Cancer Registries," *Cancer* 125 (2019): 3729–3737, https://doi.org/10.1002/cncr.32355.

6. National Academies of Science, Engineering, and Medicine, *Enabling 21st-Century Applications for Cancer Surveillance Through Enhanced Registries and Beyond: Proceedings of a Workshop* (National Academies Press, 2025), https://doi.org/10.17226/28676.TBD.

7. Accessed March 3, 2025, https://www.healthit.gov/topic/interoperability/uscdi-plus.

8. S. Jones, J. Mazuryk, L. Havener, et al., *Standards for Cancer Registries Pathology Laboratory Electronic Reporting. Version 5.1* (North American Association of Central Cancer Registries, Inc., 2024).

9. Home Page - Cancer Pathology Data Sharing v1.0.1, accessed March 3, 2025 SDC on FHIR.

10. W. S. Campbell, B. A. Rous, S. Dubois, et al., "Advancements in Interoperability: Achieving Anatomic Pathology Reports That Adhere to International Standards and Are Both Human-Readable and Readily Computable," *JCO Clinical Cancer Informatics* 9 (2025): 1–10, https://doi.org/10.1200/CCI-24-00180.

11. R. Belenkaya, M. J. Gurley, A. Golozar, et al., "Extending the OMOP Common Data Model and Standardized Vocabularies to Support Observational Cancer Research," *JCO Clinical Cancer Informatics* 5 (2021): 12–20, https://doi.org/10.1200/CCI.20.00079.

12. T. J. Osterman, M. Terry, and R. S. Miller, "Improving Cancer Data Interoperability: The Promise of the Minimal Common Oncology Data Elements (mCODE) Initiative," *JCO Clinical Cancer Informatics* 4 (2020): 993–1001, https://doi.org/10.1200/CC.20.00059.