Challenges to Using Big Data in Cancer

Shawn M. Sweeney[1], Hisham K. Hamadeh[2], Natalie Abrams[3], Stacey J. Adam[4], Sara Brenner[5], Dana E. Connors[4], Gerard J. Davis[6], Louis Fiore[7], Susan H. Gawel[6], Robert L. Grossman[8], Sean E. Hanlon[9], Karl Hsu[10], Gary J. Kelloff[11], Ilan R. Kirsch[12], Bill Louv[13], Deven McGraw[14], Frank Meng[15], Daniel Milgram[16], Robert S. Miller[17], Emily Morgan[4], Lata Mukundan[16], Thomas O'Brien[18], Paul Robbins[18], Eric H. Rubin[19], Wendy S. Rubinstein[5], Liz Salmi[20], Teilo Schaller[13], George Shi[6], Caroline C. Sigman[15], and Sudhir Srivastava[21]

ABSTRACT

Big data in healthcare can enable unprecedented understanding of diseases and their treatment, particularly in oncology. These data may include electronic health records, medical imaging, genomic sequencing, payor records, and data from pharmaceutical research, wearables, and medical devices. The ability to combine datasets and use data across many analyses is critical to the successful use of big data and is a concern for those who generate and use the data. Interoperability and data quality continue to be major challenges when working with different healthcare datasets. Mapping terminology across datasets, missing and incorrect data, and varying data structures make combining data an onerous and largely manual undertaking. Data privacy is another concern addressed by the Health Insurance Portability and Accountability Act, the Common Rule, and the General Data Protection Regulation. The use of big data is now included in the planning and activities of the FDA and the European Medicines Agency. The willingness of organizations to share data in a precompetitive fashion, agreements on data quality standards, and institution of universal and practical tenets on data privacy will be crucial to fully realizing the potential for big data in medicine.

Introduction

Precision medicine, wherein we learn from all patients to treat each patient, requires an end-to-end learning healthcare system (1, 2). In the absence of such a system, healthcare providers, health systems, and the global biomedical research community bring together available datasets when required and feasible. Globally, multitudes of patient-level data are generated daily; however, myriad factors prevent meaningful secondary use at the scale necessary to realize precision medicine fully (Box 1). This paper aims to enumerate some of the major impediments to this process and highlight good practices for future data generators. A companion paper ("Case studies for overcoming challenges in using big data in cancer") provides potential solutions through successful prior examples.

Healthcare big data refers to vast quantities of data arising from the digitization of individual patient healthcare journeys. The rise in use of such data in the medical setting and beyond promises to enable an unprecedented understanding of diseases and their treatment. These data may include electronic health records (EHR), medical imaging, genomic sequencing, payor records, pharmaceutical research, wearables, and medical devices. The landscape is continually complicated due to the increase in volume, types, and speed at which data are being generated. The ability to navigate this complexity and integrate seemingly disparate datasets to derive previously underappreciated insights could improve patient outcomes, increase efficiencies in healthcare systems, and drive discovery of new therapeutics with potential to dramatically improve the lives of people suffering from cancer and ultimately, other diseases. The road to leveraging big healthcare data is not without complexities and challenges, including awareness of existing datasets, data access, quality assurance (QA), lack of annotation, reconciliation and harmonization, storage, analysis, and derivation of insights.

While awareness, dissemination, and accessibility to the broader research community are basic requirements for publicly funded research studies, the authors of this paper strongly believe that all who generate data and develop major data commons should proactively embed these themes as part of their work from the outset to catalyze secondary data use and beyond. Interoperability and data quality continue to be major challenges when working with different healthcare datasets. Mapping terminology across

[1]American Association for Cancer Research, Philadelphia, Pennsylvania. [2]Genmab, Princeton, New Jersey. [3]Division of Cancer Prevention, Early Detection Research Network, National Cancer Institute, Rockville, Maryland. [4]Foundation for the National Institutes of Health, Bethesda, Maryland. [5]Office of In Vitro Diagnostics, Center for Devices and Radiological Health, U.S. Food and Drug Administration, Silver Spring, Maryland. [6]Abbott Diagnostics Division, Abbott Laboratories, Lake Forest, Illinois. [7]Boston University School of Medicine, Boston and New England Department of Veterans Affairs, Bedford, Massachusetts. [8]Center for Translational Data Science, The University of Chicago, Chicago, Illinois. [9]Center for Strategic Scientific Initiatives, National Cancer Institute, Bethesda, Maryland. [10]Sanofi, Bridgewater, New Jersey. [11]Division of Cancer Treatment and Diagnosis, National Cancer Institute, Bethesda, Maryland. [12]Adaptive Biotechnologies, Seattle, Washington. [13]Project Data Sphere, Morrisville, North Carolina. [14]Ciitizen Platform at Invitae, San Francisco, California. [15]Boston University and Veterans Administration Boston Healthcare System, Boston, Massachusetts. [16]CCS Associates, San Jose, California. [17]CancerLinQ, American Society of Clinical Oncology, Alexandria, Virginia. [18]Pfizer, Brooklyn, New York. [19]Merck, New York, New York. [20]Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, Massachusetts. [21]Cancer Biomarkers Research Group, Division of Cancer Prevention, National Cancer Institute, Rockville, Maryland.

Current address for T. Schaller: Kriya Therapeutics, Redwood, California.

Corresponding Authors: Shawn M. Sweeney, American Association for Cancer Research, 615 Chestnut Street, Floor 17, Philadelphia, PA 19106. Phone: 215-440-9300; E-mail: shawn.sweeney@aacr.org; and Hisham Hamadeh, Genmab US Inc., 777 Scudders Mill Road, Bldg. 2, 4th Floor, Plainsboro, NJ 08536. Phone: 609-455-7501; E-mail: hha@genmab.com

Cancer Res 2023;83:1175–82

doi: 10.1158/0008-5472.CAN-22-1274

Box 1: Patient Perspective by Liz Salmi

When we see health data at an individual level (aka "small data"), we better understand the value of data at a grand scale and may be more likely to appreciate science, support expanded research budgets, and maybe even enroll in a study ourselves. It was not long ago that the general public, including myself, thought of cancer as one big, homogenous disease. It was only after my own diagnosis nearly a decade ago that I learned cancer is actually hundreds of different diseases. Each person with cancer today is an n of 1, and a precision approach to treating these individuals is as complex as the disease itself.

The goals for big data guidance—digging into the sticky issues of standards, storage, and agreements, in an effort to glean insights and hopefully achieve a "superior outcome"—are necessary, honorable, and in many ways, exciting. But to those who see informed consent and data privacy as a derailment to your efforts, I suggest you look at it in another way: as an opportunity to send insights back to those you have pledged to respect and prevent unnecessary harm.

This is not a wild idea; it's already happening for the 1 million participants in the NIH All of Us Research Program (https://allofus.nih.gov). In exchange for enrolling in a longitudinal study, participants in All of Us are promised regular updates about how their data are used in research, including free access to academic publications and notifications about major findings.

Despite my enthusiasm for the potential and opportunities, I still have questions. What is the trade-off for loss of privacy? [Is it possible to re-identify that which has been de-identified (5)?] Will there be penalties for selling my information? Will my information be sold to for-profit companies, potentially leading to discrimination in other domains (such as housing or hiring decisions)? Might healthcare inequities be further perpetuated by big data? And will taxpayers who fund this research ever be notified of the results?

If public–private entities are successful in adhering to these guidelines and are able to design truly meaningful feedback mechanisms for the supporters and subjects of this work, we will have a real shot at advancing therapies, and maybe even some cures, for people facing cancer.

datasets, missing and incorrect data, and varying data structures make combining data an onerous and largely manual endeavor.

Issues of informed consent and data privacy, which have potential to derail data use and analysis when not appropriately addressed, further complicate matters. These complications can be exacerbated by continually evolving country-specific regulations that treat data with varying levels of conservatism, such as the General Data Protection Regulation (GDPR; https://ec.europa.eu/info/law/law-topic/data-protection_en) and Health Insurance Portability and Accountability Act (HIPAA; https://www.hhs.gov/hipaa/for-professionals/index.html); see the section, "Access to Data: the Role of Privacy Regulations," for further detail.

The authors represent organizations and foundations that recognize the potential promise of big healthcare data and the importance of collaboration in realizing the full potential of precision medicine. This document aims to offer a set of good practices as they apply to the collection of healthcare data for secondary use; however, these ideas are broadly applicable to all data. The authors recommend the "green paper" published by scientists from Sage Bionetworks and Microsoft, which offers a detailed framework and materials as a toolkit for legal and technical considerations regarding data sharing (3). Readers may also find the recent European Medicines Agency (EMA) draft guideline on registry-based studies insightful (4).

Access to Data: The Role of Privacy Regulations

There are multiple regulatory pathways to sharing data for research in the US. Much US cancer research is covered by HIPAA regulations (https://www.hhs.gov/hipaa/for-professionals/index.html) and the federal human subjects research rules governing identifiable data, also known as the Common Rule (https://www.hhs.gov/ohrp/regulations-and-policy/regulations/common-rule/index.html; refs. 6, 7). Although state laws (e.g., California Privacy Rights Act) can play a significant role in shaping data sharing solutions, a survey of state laws is beyond the scope of this section.

HIPAA and the common rule

Both HIPAA and the Common Rule require a HIPAA-compliant authorization or informed consent before identifiable information, such as names, locations, and procedure dates, can be used for research purposes; the Common Rule also requires Institutional Review Board (IRB) review of the research. However, both also provide paths forward for big data research that impose fewer preconditions on data access, such as de-identification, wherein researchers cannot readily ascertain participant identities. **Table 1** provides an overview of common elements of HIPAA and the Common Rule, and **Table 2** lists updates to both that help streamline research (8–10).

New Research Model: Participant-Contributed Data

Historically, data for research was obtained from clinical settings, or research organizations recruited participants to agree to research uses of their data. "Patient-driven" analysis is now becoming more common, where patient advocacy organizations establish and run research registries, or individuals agree to collect and donate their data directly for research purposes (11). The HIPAA Privacy Rule provides individuals with a right to copies of their medical and claims information. For example, the federal All of Us Research Program has a "direct volunteer" pathway that enables individuals to sign up for the program, obtain their clinical information and send it directly to the All of Us program (https://allofus.nih.gov/get-involved/participation). This will create other opportunities for individuals to gather relevant clinical information and donate it for research purposes (12). This approach enables patients to contribute to the types of research most important to them, potentially allowing them to feel more invested in the research. HIPAA or the Common Rule may still govern such patient-driven or patient-contributed research models, but with opportunities for greater patient buy-in and more meaningful participation.

Clinico-omic datasets highlight the risks between generating meaningful, large datasets and protecting patient privacy. Though de-identified, these datasets contain patient-unique data that could be exploited to re-identify patients, e.g., DNA in an attribute disclosure attack (13). Because there is no comprehensive data privacy law in

Table 1. Comparison of HIPAA and the Common Rule.

Topic	HIPAA	Common rule	Distinctions
Overview	The HIPAA Privacy Rule requires any entity that provides or processes healthcare data to obtain consent before disclosing a person's medical history to another entity. These covered entities may disclose medical information without consent if needed for obtaining treatment, payment, or healthcare operations, or if in the public interest. [45 CFR Part 160 and Part 164 Subparts A and E] The HIPAA Security Rule protects all individually identifiable health information a covered entity creates, receives, maintains, or transmits in electronic form. [45 CFR Subpart Part 160 and Part 164 Subparts A and C]	The Common Rule provides protection for human subjects in research conducted or supported by most federal departments and agencies. It is the baseline standard of ethics by which any government-funded research in the US is held; nearly all academic institutions hold their researchers to these statements of rights regardless of funding. Consideration is given to how various aspects of research projects (including privacy, confidentiality, data collection, data maintenance, and data retention) impact physical, emotional, financial, and informational harms. The policy established IRBs to help review and ensure compliance with the policy and the requirements for informed consent. [45 CFR 46]	HIPAA provides protection for all PHI accessed by entities who provide or process healthcare. The Common Rule provides protection for human subjects in research settings.
De-identified data	De-identified data (per HIPAA standards) not subject to further HIPAA requirements. [45 CFR 164.514 (a)]	Data that are "not identifiable" are not subject to the Common Rule. [45 CFR 46.102(e) (1) & (5)]	Although the Common Rule does not explicitly reference the HIPAA de-identified data standards, IRBs have been known to rely on it in determining if research is not subject to the Common Rule.
Limited data	A limited data set can be used or disclosed for research without the need for prior consent of participants if a data use agreement is executed, setting forth research purposes and prohibiting re-identification. [45 CFR 164.514(e)]	An IRB may determine that a study is not "human subjects research" or, in the case of secondary data research, consider the study to be "exempt" because the identity of participants is not "readily ascertainable" (researchers must agree not to re-identify or contact participants). [45 CFR 46.104(d) (4)(ii)]	Common Rule exemption is limited to secondary data (a HIPAA limited data set may involve data initially collected for research purposes). The Common Rule requires limited IRB review for determination of exemption.
Secondary use	Entities have the option to broadly consent individuals to future research uses of their identifiable PHI. [Federal Register vol. 78, no. 17, page 5612 (January 25, 2013)]	Entities may create secondary research databases with broad consent, subject to limited IRB review. [45 CFR 46.104(d) (7) & (8)]	Common Rule exemption is limited to secondary data (HIPAA broad consent provisions could involve data initially collected for research purposes). The Common Rule requires limited IRB review for determination of exemption. Broad consent under the Common Rule is subject to specific requirements.
Use of PHI	PHI used in research; when broad consent is not applicable, full HIPAA authorization is required (unless altered or waived; see below). [45 CFR 164.512(i)]	If governed by HIPAA, Common Rule exemption is available [45 CFR 46.104 (d) (4)(iii)]; otherwise, identifiable information used or disclosed in research—when not subject to any of the above exemptions—requires full IRB review and full informed consent (unless waived; see below). [45 CFR 46.106 & 46.116]	The Common Rule allows entities to rely on HIPAA where it applies; otherwise, the Common Rule requires IRB review.
Waiver of consent	Waiver of HIPAA authorization by IRB or Privacy Board. [45 CFR 164.512(i) (2) (ii)]	Waiver of informed consent by IRB. [45 CFR 46.116(f)]	Waiver criteria similar; the Common Rule also requires consideration of welfare and rights of participants and whether research is minimal risk.

Abbreviations: CFR, Code of Federal Regulations; PHI, protected health information.

Table 2. HIPAA and Common Rule Updates to Streamline Research.

HIPAA can be used solely to determine appropriate research access for HIPAA-governed research. [45 CFR 46.104(d) (4)(iii)] (https://www.hhs.gov/ohrp/regulations-and-policy/regulations/45-cfr-46/index.html)
Both HIPAA and the Common Rule allow participants to provide broad authorization (consent) for secondary research uses of data, without the need for study-specific consent.
Individuals can broadly authorize future research uses of their data, as long as the research is such that a "reasonable" individual would "expect that his or her health information could be used or disclosed for such future research" (10).
Research-ready databases of identifiable information can be created under a broad consent; secondary use needs IRB review but does not require re-consent of the participants. [45 CFR 46.104(d)(7&8)] (https://www.hhs.gov/ohrp/regulations-and-policy/regulations/45-cfr-46/index.html)
Researchers can request an IRB to waive or alter the elements of informed consent (or authorization). In response to the 21st Century Cures Act, FDA has indicated that it would honor IRB approvals of alterations or waivers of informed consent issued under the Common Rule (8).
The Common Rule now requires single IRBs for federally funded research involving more than 1 institution (9).

America, the rapid progress of analysis technologies increases the likelihood of re-identification without legal constraints or ramifications (14). Currently, there is no solution guaranteeing data privacy and it is outside the scope of this article to discuss mitigation technologies; however, there are combined system and process approaches such as access control, data anonymization, and cryptography that can serve as deterrents to re-identification. It is the best interest of the research community to continue working towards an unimpeachable technology solution and, meanwhile, adhere to data privacy best practices.

Data Quality

Perhaps the greatest challenge to optimizing the practical use of aggregated healthcare data is resolving the central tension between two fundamental data concepts: the paradigm of "collect once, use many" and the concept of data "fit for purpose." The volume, velocity, and variety that characterize big data require extraordinary technology and computational power to store, access, and analyze, which necessitate either a unique patient identifier to link different electronic sources or a parsimonious approach to data entry. However, this conflicts directly with the goals of secondary data reuse, because, in the healthcare context, that same data variety negatively impacts fitness by introducing bias and noise. Nevertheless, the concept of "data quality" is most commonly defined as showing the property of data fitness—the ability to be used for its intended purpose—and encompasses such characteristics as accuracy, completeness, consistency, reliability, lack of bias, and timeliness/currency. The attractiveness of real-world data (RWD), defined as data collected outside the context of a clinical trial from sources such as EHRs, disease registries, claims databases, and wearables, lies in its strong external validity and ability to capture characteristics and outcomes of patients commonly encountered in practice. However, reliance on RWD for clinical care and its use for discovery requires that data quality issues be addressed transparently.

Clinicians play a central role as the actors most responsible for data quality in EHRs, but these systems are optimized for patient care and billing documentation, not for population health research and downstream analytics. As a result, clinicians' use of EHRs is not fundamentally aligned with these secondary goals. In addition, most physicians are challenged daily by EHRs with poor usability, requiring time-consuming data entry detracting from patient contact and contributing greatly to professional dissatisfaction and burnout. To compensate, physicians often resort to dictating clinical notes rather than structured data entry, resulting in large volumes of unstructured text that is challenging to parse without advanced but imperfect techniques such as natural language processing (NLP). Consequently, enormous variability and heterogeneity exist across similar data elements from system to system, even for inherently structured data types such as drug or laboratory test names. Many of the most important concepts required to understand the cancer patient's journey, such as cancer stage, biomarkers, adverse events, and outcomes such as progression, are captured largely in unstructured notes. Manual data curation of unstructured documents, typically performed by trained human abstractors, has been effectively employed by some consortia and data aggregators. Further, some workflows can be enhanced by NLP and other forms of artificial intelligence (AI); however, this remains an enormously expensive, unscalable, and therefore unsustainable solution.

A potential solution to improving data quality is to enhance data interoperability. Standards-based solutions that address both syntactic interoperability (related to data formats and communication protocols) and semantic interoperability (related to the meaning of the data being exchanged) are needed. One such contemporary example is the Minimal Common Oncology Data Elements (mCODE) initiative (https://confluence.hl7.org/display/COD/mCODE/) led by the American Society of Clinical Oncology (ASCO), The MITRE Corp., and other collaborators (15). mCODE has established a foundational data specification in which clinical oncology data are subdivided into six domains (Patient, Disease, Lab/Vitals, Genomics, Treatment, and Outcomes) and individual data elements are defined by standard, nonproprietary terminologies such as Systematized Nomenclature of Medicine–Clinical Terms (SNOMED CT) and Logical Observation Identifiers Names and Codes (LOINC). Version 1.0 of mCODE was approved by Health Level Seven International (HL7) as a Standard for Trial Use in March 2020. Implementation pilots are currently underway.

Ensuring the fitness of RWD requires a systematic approach using data quality assessment best practices and tools to analyze the datasets themselves. A reasonable data quality assessment and remediation plan would include a formal data quality assessment program; validation of data against external comparator datasets; development and validation against clinical trials of RWD endpoints; and methods to address RWD bias. RWD bias has many sources, including nonrandom "missingness," where data dropout is related to the dependent variable; selection bias; performance bias; detection bias; and attrition bias (16). Common mitigation strategies include performing multivariate analyses to adjust for potential confounders, recognition of unmeasured confounders, and using propensity scoring techniques and sensitivity analyses (17), in addition to discarding incomplete cases, imputation, and performing Bayesian analyses.

Moreover, RWD continues to be redefined and may include multimodal integration of advanced molecular diagnostics, radiologic and histologic imaging, and codified clinical data. This presents real

opportunities to advance precision oncology beyond genomics and standard molecular techniques, but also requires increasingly sophisticated algorithms and approaches (18).

Storage, Annotation, Integration, and Analysis of Data

With the cost of sensors falling with Moore's law (19), it is easier each year to produce larger datasets; on the other hand, few if any advances make cleaning, annotation, integration, and analysis of large datasets any easier. Although cloud computing has created the underlying computing infrastructure to store and process large datasets, the ingestion, cleaning, integration, and analysis of these datasets is still largely limited by the availability of data "wranglers" and analysts.

The NCI has created several resources to assist in this regard. The NCI Cancer Research Data Commons (CRDC; https://datascience.cancer.gov/data-commons) is an example of the collection and integration of multiple types, sets, and sources of curated clinical data and tools for analysis. The CRDC provides access to repositories of genomic, proteomic, imaging, and other data types and data from NCI programs as well as analytic tools. The Imaging Data Commons component of CRDC includes images and associated clinical trial and metadata, and will include digital pathology images, and multispectral data from the Human Tumor Atlas Network (https://humantumoratlas.org/). The NCI Cloud Resources (The Broad Institute Firecloud, ISB Cancer Gateway in the Cloud, and Seven Bridges Cancer Genomics Commons) provide access to analysis tools. The Genomic Data Commons (GDC; ref. 20), one of the resources described in our companion review, is a component of the CRDC.

In 2000, the NCI Early Detection Research Network (EDRN) investigators envisioned the use of big data in biomarker discovery and planned for storing, curating, and disseminating biomarkers-related data using Common Data Models (CDM). Such data models include information about cellular and molecular phenotypes of cancer and stromal cells, clinical information, biospecimens, other varied biologic and epidemiologic information, and descriptions of experiments that will ensure that analytical results can be replicated.

The EDRN (https://edrn.cancer.gov/data-and-resources/informatics/) Informatics Center and Data Management and Coordinating Center developed more than 4000 common data elements (CDE) to enable search and retrieval of specific data from the data repository, as well as analysis. All CDEs and their interrelationships have been deposited in the cancer Data Standards Registry and Repository for community use. The EDRN also developed a data commons called Laboratory Catalog and Archive Service (LabCAS) that provides network researchers with a protected ecosystem for storing, searching, and analyzing data. For example, validated RNA sequencing (RNA-seq) pipelines have been implemented in the LabCAS architecture. Sites across the network use the pipeline to process sequencing data consistently and reproducibly. FASTQ files are generated in research laboratories, ingested into LabCAS with structured metadata (CDEs), and processed to produce results that are annotated with CDEs and stored in LabCAS. Investigators are then able to search using any combination of CDEs to identify sets of data for exploration and further analysis.

Biomarker research is intimately related to complex data arising from genome sequencing, gene expression profiling, proteomic and metabolomic analyses, and other point-of-care devices generating real-time, large amounts of data from special cohorts, populations, screening trials, etc. However, for these data to be useful as potential signatures for disease detection, the next step is to harmonize and integrate them to discover biomarker signatures for precision medicine. Successful integration of these data requires continued development and deployment of CDMs, NLP, and AI tools (including machine learning for predictive modeling), and health information exchange. As a simple example, initial harmonization of data in the GDC consisted of about a petabyte of data from a variety of experimental platforms, including DNA sequencing, RNA-seq, and methylation, and required hundreds of millions of core hours to process with a uniform set of bioinformatics pipelines (21).

As illustrated above, it is imperative that researchers have a deliberate plan regarding the significant resources associated with storage and computation of increasingly large amounts of data. The specific setup will depend on the size of datasets being transacted, access requirements including latency, frequency, and size of the user base, as well as the anticipated long or short-term nature of projects, to name a few. Options may include on premise or cloud resources, or a hybrid of both, and may also be a dedicated or a shared resource across multiple teams or departments in single or multiple institutions.

Regulatory Perspective: FDA

The FDA Center for Devices and Radiological Health (CDRH) has a long history of supporting and advancing use of real-world evidence (RWE) in regulatory decision-making. The Medical Device Epidemiology Network (MDEpiNet), established by CDRH in 2010, is a global RWE collaborative for health technologies. FDA also collaborates with the National Evaluation System for Health Technology (NEST) to foster evidence generation for regulatory decision-making and awarded a cooperative agreement to the Medical Device Innovation Consortium (MDIC) to establish a coordinating center for NEST. Among its objectives, NEST promotes the implementation of Unique Device Identifiers (22), without which the identity of devices and their use in routine healthcare settings cannot be discerned. Further, Systemic Harmonization and Interoperability Enhancement for Laboratory Data (SHIELD) is a public–private partnership working to solve the interoperability problem for laboratory data by providing an authoritative source for coding and supporting stakeholders in adoption of US Department of Health and Human Services—required reporting standards. SHIELD is the backbone for diagnostic test reporting for the COVID-19 pandemic; increasing adoption of SHIELD over time will help solve interoperability barriers for a plethora of laboratory data.

RWE has been used to support numerous marketing authorizations for Class II and III medical devices, and high-quality RWD sources have been leveraged to replace traditional post-approval studies. A small number of premarket authorizations for *in vitro* diagnostics (IVD) have used RWE, and the use of RWE for severe acute respiratory syndrome coronavirus 2 diagnostics is growing. The Illumina MiSeqDx Cystic Fibrosis Clinical Sequencing Assay (23) and the Illumina MiSeqDx Cystic Fibrosis 139-Variant Assay (24) illustrate the use of RWE to support premarket authorization of a genomics IVD. These assays detect genetic variations in the *CFTR* gene and are intended to be used as an aid in diagnosing individuals with suspected cystic fibrosis. Distinguishing pathogenic from benign variation is one of the greatest challenges in personalized medicine. Enormous variation exists in many human genes including *CFTR*, but evidence to help distinguish benign variation from disease-causing variation is often limited, especially for uncommon variants. Clearance of these IVDs was achieved by using a publicly maintained next-generation sequencing database called CFTR2 (Clinical and Functional

Translation of Cystic Fibrosis Transmembrane Conductance Regulator), which contains RWD from patients and families who have undergone *CFTR* sequencing and for whom phenotypic information is available (25). The CFTR2 database was used as a source of valid scientific evidence to establish which variants were disease-causing. Clinical sensitivity and specificity of the Illumina MiSeqDx Cystic Fibrosis Clinical Sequencing Assay was estimated on the basis of information from the CFTR2 database.

To support using public human genetic variant databases as sources of valid scientific evidence for clinical validity of genomic tests in premarket submissions, FDA has established a database recognition program and published a guidance (26). The first public genetic database to achieve FDA recognition was the ClinGen expert-curated human variant data, which pertains to assertions of clinical significance for germline-derived, highly penetrant variants (27). Recognition of additional databases, including somatic variant databases, along with increased concentration on pharmacogenetic databases, will help support future premarket submissions for genomics IVDs, as has been done with the OncoKB variant database (https://www.fda.gov/drugs/resources-information-approved-drugs/fda-recognizes-memorial-sloan-kettering-database-molecular-tumor-marker-information).

FDA has also published guidance on the use of RWE to support regulatory decision-making for medical devices (28), and in 2020, the MDIC IVD RWE Framework was established to build on the FDA guidance by providing additional contextual information for incorporating fit-for-purpose RWE into product development and regulatory decision-making, particularly in support of IVD authorizations (https://mdic.org/resource/ivd-rwe-framework/).

Regulatory Perspective: EMA

As part of evolving data-driven regulations, the EMA has identified a number of top priorities (https://www.hma.eu/about-hma/working-groups/hma/ema-joint-big-data-steering-group.html). These include: Delivering a sustainable platform to access and analyze healthcare data from across the EU; establishing an EU framework for data quality and representativeness; enabling data discoverability; developing EU Network skills in big data; strengthening EU Network processes for big data submissions; building EU Network capability to analyze big data; modernizing the delivery of expert advice; ensuring data are managed and analyzed within a secure and ethical governance framework; collaborate with international initiatives on big data; and creating an EU big data "stakeholder implementation forum."

The EMA has begun implementing their priorities through a number of activities. The Data Analysis and Real World Interrogation Network in Europe (DARWIN EU) project (https://www.ema.europa.eu/en/about-us/how-we-work/big-data/data-analysis-real-world-interrogation-network-darwin-eu), for example, will serve as a sustainable data platform connected to the European Health Data Space (EHDS; https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space_en). A major component of EHDS is a code of conduct for health data use, a governance framework including rules for primary and secondary use of health data, a regulatory framework for AI, principles such as free flow of data and free movement of digital health services, and digitization of healthcare systems. This code of conduct will be aligned with the European Data Governance Act adopted by the European Commission in November 2020 and the proposed EU Data Act. One of nine European data spaces, EHDS will be a space for exchanging and sharing different types of health data and will include storage, tools, data standards, as well as GDPR-compliant access and governance mechanisms.

Such initiatives complement existing EMA work in the big data space, including collaborations with the Clinical Data Interchange Standards Consortium on standards for RWD (https://www.cdisc.org/standards/real-world-data) that will readily be applicable to big data. EMA work is based on firsthand experience and white papers that explore the practical use of their CDM (29). Agency authors also published a call for unconventional analysis methods for validation by exercises independent from drug development, using the Agency's public qualification platform (30). Pathways have been described for RWD to be used for regulatory decision-making across therapeutic areas and across the life cycle of products (31).

In oncology, the European biopharmaceutical industry has mapped the initiatives and challenges for big data in a data landscape report (32), and the EMA held a workshop with cancer registries and cancer data collection initiatives to discuss data elements needed for regulatory purposes, and the QA and governance of registries and initiatives collecting data from patients with cancer (33). Key areas of need for regulators include data on patient function, detailed prognostic factors and biological features, as well as data on interventions, including reasons for discontinuation. The need to efficiently merge patient data independently collected by cancer registries and research initiatives was also identified as important. Other essential information that could assist regulators with cancer medicine development include data to support the safety and efficacy of medicines when these data address uncertainties about effects of a medicine or fill gaps in knowledge about cancers at the time of regulatory assessment or shortly after authorization (34).

Outside of the EMA, a number of European big data projects are already underway. The Innovative Medicines Initiative 2 public–private partnership: Big Data for Better Outcomes (which includes HARMONY for hematologic malignancies and PIONEER for prostate cancer), and the European Health Data & Evidence Network, plus several projects under the EU Horizon 2020 funding framework are but a few examples.

Summary and Recommendations

Cancer is both pervasive and increasingly complex. It is fascinating to observe the revolution in cancer treatment where the reliance on large amounts of data is becoming more mainstream. For the vast volumes of data to be useful, they need to be organized, shared, integrated, and readily accessible by teams that have decision-making responsibilities in cancer treatment. More importantly, the data need to reflect the intent-to-treat population. To date, most datasets largely reflect patients of European ancestry and conclusions drawn may not be broadly applicable to all. Efforts to be as inclusive as possible in future clinical studies will result in more representative datasets with time. This report highlighted some of the challenges associated with each of those steps and offered an array of existing efforts and opinions aimed at mitigating the respective roadblocks through real examples.

To maximize knowledge generation from data for patients, it is imperative that ecosystems of partnerships, standardization, and legislation continue to advance in concert to avoid weak links in the value chain of cancer treatment. New interventions will become more credible and adopted widely when data are used to inform treatment decisions. Therefore, willingness to share data by organizations in a precompetitive fashion, agreements on data quality standards, and

instituting universal and practical tenets on data privacy will be crucial to realize that aspiration.

Every stakeholder who touches patient data shares responsibility for delivering the vision of harnessing the totality of data that are available, to drive decision-making in favor of patients everywhere. The authors hope that by describing some of the emerging resources, we raise awareness and inspire generators, stewards, and consumers of health-care data to consider the secondary use of such data at the earliest possible step. This will ensure proper sharing of data to generate insights so that people suffering from cancer and their loved ones stand the best chance to benefit from collective knowledge of the cancer community.

Authors' Disclosures

S.M. Sweeney is an employee of the American Association for Cancer Research; however, this had no bearing on the review or acceptance of this manuscript by the journal. G.J. Davis reports other support from Abbott Laboratories outside the submitted work, being an employee of Abbott Laboratories, and also owner of Abbott stock. S.H. Gawel is an employee of Abbott Laboratories. I.R. Kirsch reports personal fees from Adaptive Biotechnologies outside the submitted work. D. McGraw reports other support from Invitae, Verily, and Datavant outside the submitted work. D. McGraw is on the Board of the CARIN Alliance, a nonprofit organization supporting the ability of patients to access all of their health information from clinical and claims records, including through FHIR APIs and is on the Board of Manifest Medex, a health information exchange in California. D. Milgram reports other support from NCI and from Foundation for the NIH outside the submitted work. R.S. Miller is an employee of ASCO. CancerLinQ is a wholly owned, nonprofit subsidiary of ASCO. T. O'Brien is a full-time employee of Pfizer. P. Robbins reports other support from Pfizer during the conduct of the study. E.H. Rubin reports other support from Merck & Co. outside the submitted work. C.C. Sigman reports other support from NCI and Foundation for the NIH outside the submitted work. No disclosures were reported by the other authors.

Acknowledgments

The authors would like to thank the Foundation for the NIH for their support and guidance, as well as Lukas Amler, Michelle Berny-Lang, Vladimir Popov, Elizabeth Pulte, Maggie Scully, Margaret Thompson, and Michael Taylor for their insight, review, and thoughtful contributions.

The publication costs of this article were defrayed in part by the payment of publication fees. Therefore, and solely to indicate this fact, this article is hereby marked "advertisement" in accordance with 18 USC section 1734.

Received April 15, 2022; revised July 29, 2022; accepted December 5, 2022; published first January 10, 2023.

References

1. Institute of Medicine. The learning healthcare system: workshop summary. Washington, DC: The National Academies Press; 2007.
2. National Research Council. Toward precision medicine: building a knowledge network for biomedical research and a new taxonomy of disease. Washington, DC: The National Academies Press; 2011.
3. Mangravite LM, Sen A, Wilbanks JT, Sage Bionetworks Governance Team. Mechanisms to govern responsible sharing of open data: a progress report. 2020.
4. European Medicines Agency (EMA). Draft guideline on registry-based studies. EMA/502388/2020. 2020.
5. Na L, Yang C, Lo CC, Zhao F, Fukuoka Y, Aswani A. Feasibility of reidentifying individuals in large national physical activity data sets from which protected health information has been removed with use of machine learning. JAMA Netw Open 2018;1:e186040.
6. Federal policy for the protection of human subjects. Final rule. Fed Regist 1991; 56:28003–18.
7. Federal policy for the protection of human subjects. Final rule. Fed Regist 2017; 82:7149–274.
8. U.S. Food and Drug Administration. IRB waiver or alteration of informed consent for clinical investigations involving no more than minimal risk to human subjects: guidance for sponsors, investigators, and institutional review boards. 2017.
9. Hahn C, Kaufmann P, Bang S, Calvert S. Resources to assist in the transition to a single IRB model for multisite clinical trials. Contemp Clin Trials Commun 2019;15:100423.
10. Modifications to the HIPAA privacy, security, enforcement, and breach notification rules under the health information technology for economic and clinical health act and the genetic information nondiscrimination act; other modifications to the HIPAA rules. Fed Regist 2013;78: 5565–702.
11. Woolf SH, Zimmerman E, Haley A, Krist AH. Authentic engagement of patients and communities can transform research, practice, and policy. Health Aff 2016; 35:590–4.
12. Savage LC, Neinstein AB, Savage M, Adler-Milstein J. ONC should not delay the release of its rule. Health Affairs Forefront 2020.
13. Erlich Y, Narayanan A. Routes for breaching and protecting genetic privacy. Nat Rev Genet 2014;15:409–21.
14. Lubarsky B. Re-identification of "anonymized data. Georgetown Law Tech Rev 2017;:202–12.
15. Osterman TJ, Terry M, Miller RS. Improving cancer data interoperability: the promise of the minimal common oncology data elements (mCODE) initiative. JCO Clin Cancer Inform 2020;4:993–1001.
16. Schilsky RL. Finding the evidence in real-world evidence: moving from data to information to knowledge. J Am Coll Surg 2017;224:1–7.
17. Booth CM, Karim S, Mackillop WJ. Real-world data: towards achieving the achievable in cancer care. Nat Rev Clin Oncol 2019;16:312–25.
18. Boehm KM, Khosravi P, Vanguri R, Gao J, Shah SP. Harnessing multimodal data integration to advance precision oncology. Nat Rev Cancer 2022;22: 114–26.
19. Moore GE. Cramming more components onto integrated circuits. Electronics 1965;38.
20. Wilson S, Fitzsimons M, Ferguson M, Heath A, Jensen M, Miller J, et al. Developing cancer informatics applications and tools using the NCI genomic data commons API. Cancer Res 2017;77:e15–e8.
21. Grossman RL. Data lakes, clouds, and commons: a review of platforms for analyzing and sharing genomic data. Trends Genet 2019;35: 223–34.
22. Daniel G, McClellan M, Gardina S, Deak D, Bryan J, Streit C. Unique Device Identifiers (UDIs): a roadmap for effective implementation. Washington DC: Engelberg Center for Health Care Reform at Brookings. 2014.
23. U.S. Food and Drug Administration. 510(k) Substantial equivalence determination decision summary. 510(k) Number: K132750. Illumina MiSeqDx Cystic Fibrosis Clinical Sequencing Assay. 2013.
24. U.S. Food and Drug Administration. 510(k) Substantial equivalence determination decision summary. 510(k) Number: K124006. Illumina MiSeqDx Cystic Fibrosis 139-Variant Assay. 2013.
25. Sosnay PR, Siklosi KR, Van Goor F, Kaniecki K, Yu H, Sharma N, et al. Defining the disease liability of variants in the cystic fibrosis transmembrane conductance regulator gene. Nat Genet 2013;45:1160–7.
26. U.S. Food and Drug Administration. Use of public human genetic variant databases to support clinical validity for genetic and genomic-based in vitro diagnostics. Guidance for Stakeholders and Food and Drug Administration Staff. 2018.
27. Koontz L., U.S. Food and Drug Administration. Genetic database recognition decision summary for ClinGen expert curated human variant data. Genetic Database Recognition Decision Summary (Q181150). 2018.
28. U.S. Food and Drug Administration. Use of real-world evidence to support regulatory decision-making for medical devices: guidance for industry and food and drug administration staff. 2017.
29. Candore G, Hedenmalm K, Slattery J, Cave A, Kurz X, Arlett P. Can we rely on results from IQVIA medical research data UK converted to the observational medical outcome partnership common data model? A validation

study based on prescribing codeine in children. Clin Pharmacol Ther 2020; 107:915–25.
30. Eichler HG, Koenig F, Arlett P, Enzmann H, Humphreys A, Pétavy F, et al. Are novel, nonrandomized analytic methods fit for decision-making? The need for prospective, controlled, and transparent validation. Clin Pharmacol Ther 2020; 107:773–9.
31. Cave A, Kurz X, Arlett P. Real-world data for regulatory decision-making: challenges and possible solutions for Europe. Clin Pharmacol Ther 2019;106:36–9.
32. Montouchet C, Thomas M, Anderson J, Foster S. The Oncology Data Landscape in Europe: Report [commissioned by EFPIA - European Federation of Pharmaceutical Industries and Associations]; 2018.
33. European Medicines Agency (EMA). Report of the workshop on the use of registries in the monitoring of cancer therapies based on tumors' genetic and molecular features. 2020.
34. Skovlund E, Leufkens HGM, Smyth JF. The use of real-world data in cancer drug development. Eur J Cancer 2018;101:69–76.