An overview of real-world data sources for oncology and considerations for research

Lynne T. Penberthy, MD, MPH [1]; Donna R. Rivera, PharmD, MSc [1]; Jennifer L. Lund, PhD, MSPH[2]; Melissa A. Bruno, MPH [1]; Anne-Marie Meyer, PhD, MSPH [2]

[1]Surveillance Research Program, Division of Cancer Control and Population Sciences, National Cancer Institute, Rockville, Maryland; [2]Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina.

Corresponding Author: Lynne T. Penberthy, MD, MPH, Surveillance Research Program, Division of Cancer Control and Population Sciences, 9609 Medical Center Drive, Room 4E456, Bethesda, MD 20892-2590 (lynnepenberthy.schumacher-penberthy@nih.gov).

DISCLOSURES: Jennifer L. Lund reports salary support to her institution from Westat, Inc, during the course of the study; and salary support to her institution from AbbVie to outside the submitted work; in addition, her spouse is an employee of GlaxoSmithKline and owns stock in the company. Anne-Marie Meyer is an employee of Roche and owns Roche stock, and is currently an employee of a healthtech start-up. Lynne T. Penberthy, Donna R. Rivera, and Melissa A. Bruno report no conflicts of interest.

CA Cancer J Clin 2022;72:287-300

doi: 10.3322/caac.21714. Available online at cacancerjournal.com

Abstract: Generating evidence on the use, effectiveness, and safety of new cancer therapies is a priority for researchers, health care providers, payers, and regulators given the rapid pace of change in cancer diagnosis and treatments. The use of real-world data (RWD) is integral to understanding the utilization patterns and outcomes of these new treatments among patients with cancer who are treated in clinical practice and community settings. An initial step in the use of RWD is careful study design to assess the suitability of an RWD source. This pivotal process can be guided by using a conceptual model that encourages predesign conceptualization. The primary types of RWD included are electronic health records, administrative claims data, cancer registries, and specialty data providers and networks. Careful consideration of each data type is necessary because they are collected for a specific purpose, capturing a set of data elements within a certain population for that purpose, and they vary by population coverage and longitudinality. In this review, the authors provide a high-level assessment of the strengths and limitations of each data category to inform data source selection appropriate to the study question. Overall, the development and accessibility of RWD sources for cancer research are rapidly increasing, and the use of these data requires careful consideration of composition and utility to assess important questions in understanding the use and effectiveness of new therapies.

Keywords: cancer registries, data, oncology, real-world data, research methods

Introduction

Generating accurate evidence on the patterns and effectiveness of preventing, diagnosing, and treating cancer in real-world settings is a priority for researchers, health care providers, payers, and regulators. Real-world data (RWD), or data relating to patient health and/or the delivery of health care from routinely collected sources as opposed to clinical trials,[1] can be an important component in addressing a range of important research questions across the cancer continuum. When combined with rigorous design and analytic methods, RWD can be used to generate real-world evidence about preventive and cancer-focused care delivered outside the selected trial populations in which they are often studied. Previous reviews have summarized different RWD sources for oncology research, their potential uses, and important biases for consideration.[2-4] In this review, we extend this prior work to: 1) introduce a conceptual model to help researchers with the process of RWD source selection for a given research question; 2) update and describe features of commonly used RWD source types, including their strengths and limitations; and 3) provide an example of RWD source selection using a case study from a recently published article.

Conceptual Model

We propose a conceptual model (Fig. 1) to assist researchers in assessing the suitability of an RWD source for answering a specific cancer-related research question. The model has 3 primary steps: 1) clearly define the research question, 2) understand the data source contents and target population coverage, and 3) assess the data source

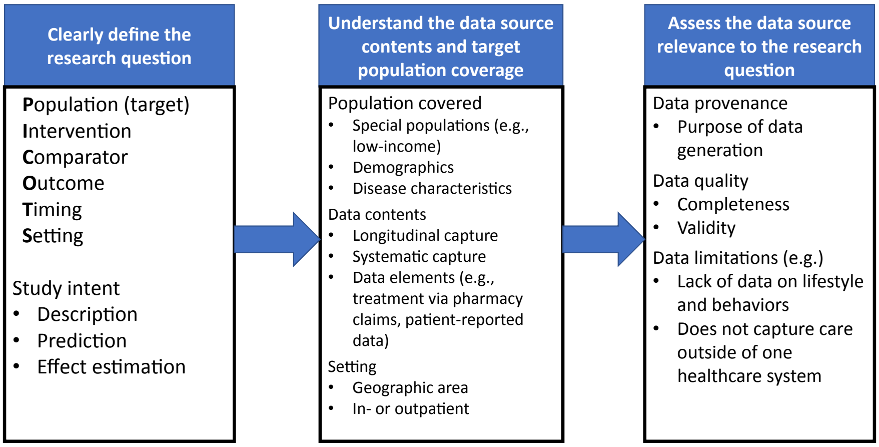
Clearly define the research question
Population (target)
Intervention
Comparator
Outcome
Timing
Setting
Study intent
• Description
• Prediction
• Effect estimation
Understand the data source contents and target population coverage
Population covered
• Special populations (e.g., low-income)
• Demographics
• Disease characteristics
Data contents
• Longitudinal capture
• Systematic capture
• Data elements (e.g., treatment via pharmacy claims, patient-reported data)
Setting
• Geographic area
• In- or outpatient
Assess the data source relevance to the research question
Data provenance
• Purpose of data generation
Data quality
• Completeness
• Validity
Data limitations (e.g.)
• Lack of data on lifestyle and behaviors
• Does not capture care outside of one healthcare system

FIGURE 1. Conceptual Model to Guide the Selection of a Real-World Data Source for a Specific Cancer-Related Research Question.

relevance to the research question. In step 1, we recommend applying the previously published PICOTS framework to clearly delineate the *population, intervention, comparator, outcome, timing, and setting*.[5,6] This framework is often used in evidence-based practice and thus can be adapted as a way of emulating a target trial using nonrandomized RWD.[7] It may also be useful for researchers to think through the study goal—description (eg, summarizing patterns), prediction (eg, identifying those likely at risk of an event), or effect estimation (eg, identifying effects of interventions or policies)—to clarify objectives and interpretation.[8] Step 2 highlights the importance of fully understanding representation and content of the data and coverage of the target population to which study the results will ultimately apply. The PICOTS framework and a clear specification of the target population outline the data requirements for a specific study question. In step 3, researchers must then assess the relevance of the data source for the proposed research question. This includes understanding the original purpose for which it was generated and key steps in data provenance and processing. Understanding the original data collection processes provides insight into the quality of specific data elements and whether the RWD source is suitable, or *fit-for-purpose*, for the intended use case.[9,10] Information about the availability of specific variables, their completeness, and their validity is another key component of this assessment. There is substantial variability across RWD sources in the type, breadth, completeness, and quality of data elements. Understanding the underlying differences in RWD sources is central to the appropriate selection and valid use of RWD for cancer research.

Real-World Data Types

The landscape of RWD is broad, expanding, and includes a variety of data source types that represent the complex and fragmented delivery of health care in the US system. Because of this fragmentation, RWD may be influenced by several key factors: who is paying for care (insurer), who is delivering the care (provider), where the care is delivered (geography or health system), or the specific population represented (disease or demographic). The main categories of RWD sources covered in this review, although not fully comprehensive, include those most commonly used by researchers. These include the following: 1) administrative claims, 2) electronic health records (EHRs), 3) registries, 4) health care data aggregators, and 5) specialty data providers and networks (Table 1).[11-26] Of note, these categories are somewhat subjective and data sources are dynamic, continually expanding their capture of information through data linkage and collation of other resources. As such, we acknowledge that others may consider specific datasets in different categories. Within each of these high-level categories are more detailed types and subtypes related to the network, organization, facility, setting, or modality of health care covered by the data source. Appropriate analysis of RWD requires an understanding of both the original purpose and current use cases of the data because the primary use case and subsequent changes made provide important context about the data elements captured and the data structure. Table 2 provides details about each RWD type, including the population and estimated coverage, strengths and limitations, and example studies from the literature.

Administrative Claims Data

Administrative claims data have been a longstanding source of RWD for cancer research. These data recorded for reimbursement purposes include information about coded diagnoses and services rendered during patient visits from claims for insurance providers. Longitudinal data from claims can be captured on individuals who are continuously enrolled in specific health insurance plans or pharmacy or other specific programs. Common sources of administrative data used by cancer researchers include enrollment and claims data generated from government insurers, including Medicare (federal level) and Medicaid (state level); commercial insurance providers; and health care claims data aggregators.

TABLE 1. Overview of Oncology Real-World Data Sources: Data Elements, Intended Purpose, and Examples

DATA TYPE	DATA SUBTYPE	DESCRIPTION	TYPES OF DATA AVAILABLE	INTENDED PURPOSE	EXAMPLE
Administrative claims	Private insurers	Administrative claims are generated to record health care transactions between a health care plan and health care providers for covered individuals; private health insurers may provide accessibility to these data for researchers through licensing and signing a data use agreement directly or through third-party vendors	Enrollment, demographics, dates of service, diagnosis codes, procedure codes, vital status, and pharmacy transactions	Data are collected for the purposes of billing and reimbursement for health care services (eg, medical, pharmacy)	Sharma 2020[11]
	Public/federal insurers (Medicare)	Federally sponsored health insurance coverage for adults aged 65 y and older and selected individuals with disabilities; administrative claims capture health care transactions between covered individuals and health care providers; researchers can access these data through a submission and approval process, which also requires a data use agreement	Enrollment, demographics, dates of service, diagnosis codes, procedure codes, vital status, and pharmacy transactions		Potosky 1992[12]
	Public/state insurers (Medicaid)	State-provided health insurance coverage for specific populations (eg, income-based, pregnant women, and children); administrative claims generated through the reimbursement of covered services are recorded; researchers can access these data through a submission and approval process, which also requires a data use agreement	Enrollment, demographics, dates of service, diagnosis codes, procedure codes, vital status, and pharmacy transactions		Maclean 2020[13]
Electronic health record (EHR)	Health maintenance organizations (HMOs)	Health system or catchment area provides patient care aggregated through an integrated model of health care delivery, including coordination of a health care plan, medical physician groups, and a health care facility system	Varies by HMO; typically includes data required for the provision of clinical care across the HMO settings as well as billing purposes, such as demographics, clinical variables, diagnosis, radiology, laboratory, diagnostics, and pharmacy	Data are collected for the documentation, assessment, and provision of clinical care and treatment pathways within health care systems or inpatient or outpatient settings	Bowles 2012[14]
	Ambulatory care	EHR systems developed to facilitate the provision of care in the outpatient setting, including physician office visits, radiology centers, laboratories, and other treatment centers, for the primary purposes of clinical care, documentation, and quality assessment	Demographics, diagnosis, clinical variables, medical oncology, radiation oncology, radiology, laboratory, diagnostics, and pharmacy		Lau 2011[15]
	Inpatient care	EHR systems developed to facilitate the provision of care in the inpatient setting, including hospitals, systems, and long-term care facilities, for the primary purposes of highly monitored clinical care, documentation, and quality assessment	Demographics, diagnosis, clinical variables, radiology, laboratory, diagnostics, and pharmacy		Callahan 2020[16]
Registry data	Federally sponsored (SEER, NPCR)	Data that are collected and curated systematically on a specific disease, condition, or population and entered into federally managed registry	Data variables are typically organized around variables to evaluate the etiology, diagnosis, treatment, and outcomes of patients within the registry	Provides epidemiology of disease incidence, prevalence, and trends for disease monitoring; data that are curated systematically according to data standards as part of public health reporting; data are HIPAA-exempt for maintaining PII and linking longitudinally	Cronin 2018[17]

(Continues)

TABLE 1. (Continued)

DATA TYPE	DATA SUBTYPE	DESCRIPTION	TYPES OF DATA AVAILABLE	INTENDED PURPOSE	EXAMPLE
	State or regional registries	Data that are collected and curated systematically on a specific disease, condition, or population that cover a specific geography, population, or is captured under state regulation (public health reporting)	Incidence data and trends, high-level data collection, survival trends	Contributes to epidemiology of disease, monitors trends in disease, and supports public health planning; data that are curated systematically according to data standards as part of public health reporting for the specific disease state; data are HIPAA-exempt for maintaining personal identities and linking longitudinally	Gearhart-Serna 2020[18]
	Industry-sponsored (drug specific)	Voluntarily developed or mandated for postmarketing to collect data specifically on patients receiving a specific drug or combination of drugs to allow longitudinal exposure monitoring for potential adverse events, safety, outcomes, and follow-up data	Demographics; pharmacy, including drug dosing and administration, laboratory, and adverse drug events (related to the drug of interest)	Collect data elements on patients receiving a specific agent	Brown 2013[19]
	Hospital-based registries (NCDB)	Registries developed to capture information for quality assurance at the facility level, with the focus on patients treated within the health care system	Demographics, clinical variables, limited treatment, incidence data and trends, survival outcomes; subset of practices include detailed data on cases, including quality measures	Monitoring quality of care at the facility level	Boffa 2017[20]
	Disease-specific registries (cancer site)	Registries developed or established to collect data on patients with a specific disease (eg, rare cancer)	Demographics, treatment data, pharmacy, diagnosis, laboratory, and clinical variables (related to the specific disease)	Data collected from patients with a disease for longitudinal monitoring and epidemiologic studies	Steele 2006[22]
Health care data aggregators	Nonprofit	Data aggregators combine data across varied sources using a specified data model (eg, federated or software system-based) to provide composite data for evaluation	Demographics, diagnosis, and clinical variables (can vary: radiology, laboratory, diagnostics, pharmacy)	Used to measure care delivery or improve quality of care (CancerLinQ); used to gather data on patient-centered outcomes (PCORnet); used to query signals to assess drug safety for marketed products (Sentinel)	Brown 2020[23]
	Commercial	Single-sourced, curated data	Demographics, diagnosis, clinical variables (can vary: radiology, laboratory, diagnostics, pharmacy) from one system or group of systems	Clinical data that are refined and cleaned for research purposes (eg, Flatiron, primarily Oncology EMR software)	Khozin 2019[24]
		Multiple unique sourced	Demographics, diagnosis (can vary: radiology, laboratory, diagnostics, pharmacy) from multiple sources across geographic areas	Data that are collected and curated from heterogeneous data sources to fit commercial research models (eg, COTA Inc, Symphony Health, HealthVerity, OptumLabs)	Kabadi 2019[25]
Specialty data providers and networks	Varied	Organizations capturing a specific individual data type, such as radiology images or reports, administrative pharmacy data, or genomic information	Demographics, diagnosis, clinical variables (can vary: radiology, laboratory, diagnostics, pharmacy), document type can vary by image or report (eg, DICOM or pdf)	Data exchange—typically to enhance clinical care—enabling providers across different entities to view patient results	Gajra & Feinberg 2020[26]

Abbreviations: DICOM, Digital Imaging and Communications in Medicine system; HIPAA, the Health Insurance Portability and Accountability Act of 1996; NCDB, National Cancer Data Base; NPCR, National Program of Cancer Registries; pdf, portable document format; PII, personally identifiable information; SEER, Surveillance, Epidemiology, and End Results program of the National Cancer Institute.

TABLE 2. Characteristics of Oncology Real-World Data Sources: Coverage, Strengths, and Limitations

DATA TYPE	DATA SUBTYPE	POPULATION	COVERAGE/ LONGITUDINALITY	COVERAGE ACROSS SETTINGS	STRENGTHS	LIMITATIONS
Administrative claims	Private insurers	Individuals with specific insurance coverage (eg, employer-based, self-insured, other)	Longitudinal capture of health care service encounters while enrolled in benefits; vital status is often available for state and federal insurance programs	Medium to high; coverage is based on benefits enrollment; can include capture of inpatient, outpatient, and pharmacy services	Clear population denominator; longitudinal data capture	Often short enrollment periods; lacks clinical details and laboratory results; no information on provider or patient intent/preference
	Public federal insurers (Medicare)	Adults aged 65 y and older or with qualifying disabilities			Clear population denominator; longitudinal data capture; often a more stable enrolled population; has been linked to other forms of data (eg, registry); vital status data available	Does not include individuals enrolled in Medicare Advantage; lacks clinical details and laboratory results; no information on provider or patient intent/ preference
	Public/state insurers (Medicaid)	Income-based eligibility or coverage for special populations (eg, pregnant women and children)			Clear population denominator; longitudinal data capture; several states' data can be accessed through centralized processes; vital status data available	Often population is unstable because of fluctuating eligibility requirements; lacks clinical details and laboratory results; no information on provider or patient intent/preference
Electronic health record (EHR)	Integrated delivery organizations	Individuals enrolled and receiving care in a health maintenance organization	Longitudinal capture of health care service encounters while enrolled in benefits	High	Clear population denominator; longitudinal data capture; high level of completeness across care settings; low rates of attrition within plan	Not representative of general population or patients in fee-for-service plans
	Ambulatory care	Patients receiving care within the specified outpatient setting captured through the source	Coverage may be sporadic, depending on sharing between centers based on the use specific of EHR software	Medium to low; only available through central linkage by EHR software or previously linked clinical centers	Contains data that may not be captured elsewhere	Care received outside of the system would not be documented; not longitudinal (question-dependent); population denominator often unclear
	Inpatient care	Patients receiving care within the specified inpatient setting captured through the source	Lacks longitudinality because the data are episodic and typically best used for short-term studies	Medium to low; only available within health systems with a common EHR software	Provides detailed data for episodic study	Care received outside of the system would not be documented; population denominator may not be clear
Registry data	Federally sponsored (SEER, NPCR)	Combined data from all patients who have cancer within a specific set of geographic catchment area (based on regional and/or state registries)	Longitudinal capture of health care for available data sources; may have gaps in knowledge (eg, treatment over time, recurrence)	Medium to high; consolidates data from across multiple health care settings and providers	Large sample size of population-based data; facilitates temporal trends assessment across different strata	Delays in reporting of data; limited detail currently
	State or regional registries	All patients who have cancer within a specific geographic catchment area	Longitudinal capture of health care for available data sources; may have gaps in knowledge (eg, treatment over time, recurrence)	Medium; consolidates data from across multiple health care settings and providers	Includes all cancers diagnosed within geographic area; followed longitudinally	Limited outcomes; limited detailed data on treatment, genomic characterization

(Continues)

TABLE 2. (Continued)

DATA TYPE	DATA SUBTYPE	POPULATION	COVERAGE/ LONGITUDINALITY	COVERAGE ACROSS SETTINGS	STRENGTHS	LIMITATIONS
	Industry sponsored (drug-specific)	Limited, drug-specific only	Coverage is limited; longitudinality is typically good	Medium to high; for a very specific population only	Very detailed information on specific data elements	Very narrow data collection
	Hospital-based registries (NCDB)	Patients receiving diagnosis or treatment in inpatient facilities or associated outpatient facilities	Limited capture of longitudinal follow-up of patient—dependent on access to information outside the institutional setting	Medium; consolidates data from across multiple health care settings and providers	More detailed data on each patient if within selected site; focus on facility quality of care	Not a population-based sample; limited information on care delivered outside the facility setting
	Disease-specific registries (voluntary)	Voluntary submission for a specific disease	Coverage is limited as focus on a particular disease; longitudinality is typically good; volunteer-based	Medium to high; for a very specific population only	Well defined cohort of interest; potential to target rare or unusual cancers	Limited data because of volunteer reporting
Health care data aggregators	Nonprofit	Variable, depends on the aggregator's purpose	Highly varied on data source; may be similar to EHR or claims-based sources	Medium to high; varies by source, although the objective is often longitudinal	If purpose is well defined, produces high-quality studies	Convenience, not population-based, sample
	Commercial	Patients receiving care within the specified setting captured through the source	Highly varied on data source; may be similar to EHR or claims-based sources; based on care received in the specific system	Medium to high; varies by source, although the objective is often longitudinal	Ability to curate data for specific purpose or extract variables (eg, EGFR)	Convenience, not population-based, sample
		Patients receiving care within the specified setting captured through the source	Coverage is complex and varies significantly by the intersection of linked sources; highly varied on data source; may be similar to EHR or claims-based sources	Medium to high; includes various settings	Includes multiple, heterogenous data sources to provide a detailed, longitudinal understanding of clinical interaction	Complete coverage for all data types may be sparse; convenience, not population-based, sample
Specialty data providers and networks	Varied	Variable, depending on the network size and mission	Highly varied based on data source purpose and structure	Typically crosses multiple health care settings	Variable by data source; may provide detailed clinical data elements from the specific source	Variable by data source; may have limited capture of complete clinical picture; may require linkage with other sources

Abbreviations: NCDB, National Cancer Data Base; NPCR, National Program of Cancer Registries; SEER, Surveillance, Epidemiology, and End Results program of the National Cancer Institute.

Approval of the Health Insurance Portability and Accountability Act (HIPAA) of 1996 led to requirements that resulted in claims data sources sharing many common data elements. Importantly, most administrative claims databases contain enrollment files, which track individual monthly enrollment in a covered health plan over the time span of the data source. This distinct longitudinal feature enables a clear description of a population over time that can be used to define a study denominator. In addition, many claims data sources contain patient health data across health care settings, including inpatient visits, outpatient visits, or other specialty health care providers. Increasingly, health plans provide additional pharmacy benefits and thus include prescription medication dispensing information from outpatient or community pharmacies. The latter data are increasingly important in understanding cancer outcomes in the context of treatment. In general, health care services that are not reimbursable by the health plan or program (eg, over-the-counter medicines or services paid out of pocket by the patient) are not captured. In addition, the type of insurance plan or program participation by the patient or provider may influence the sensitivity and specificity of care as recorded in the claim (eg, fee-for service vs managed care, such as health maintenance organizations or accountable care organizations). Claims data can also include valuable information on health care delivery that enables research on providers, care quality, access, hospital volume, and prescribing patterns.

Because administrative claims are generated for billing purposes rather than for patient care, the validity and completeness of costly procedures (eg, surgical resection) are likely to be high; however, the accuracy of specific diagnoses (eg, hypertension) is variable and depends on several factors. These include the specific patient population and the provider setting (eg, physician office vs inpatient care). On a specific claim, only the diagnoses and procedures that are needed to describe clinical care provided for reimbursement are likely to be included, which may lead to reduced sensitivity in the capture of certain outcomes. In addition, administrative data often lack important clinical, laboratory, or behavioral health information that may be important for cancer research, such as the cancer stage, genomic biomarker testing results, and smoking status.

Substantial efforts have been made to address some of these limitations in oncology by linking administrative claims with registry (eg, the National Cancer Institute's Surveillance Epidemiology and End Results [SEER] program) and survey (eg, the Health and Retirement Survey) resources. The National Cancer Institute has led several efforts to enhance cancer research data for the scientific community, resulting in widely used resources, including SEER-Medicare,[27] SEER-Consumer Assessment of Healthcare Providers and Systems,[28] and the SEER-Medicare Health Outcomes Survey.[29,30]

Electronic Health Records

EHRs are another increasingly prevalent RWD type. EHRs can provide rich information that may not be available from other types of RWD because they contain data from multiple sources within the health care system (eg, pathology reports, laboratory results, medication records, provider notes). However, the vast majority of information held within EHRs is maintained in unstructured text documents or is captured as a scanned, nonoptical character recognition portable document format, requiring curation and translation to extract structured data. Furthermore, EHRs do not include comprehensive information on health care provided outside the facility covered by the system. A patient with cancer may have data held independently within multiple EHRs across hospitals, community oncology practices, radiation oncology practices, or other settings, depending on the software used and its integration across these practices.[31,32] This is especially true for patients at different stages in their cancer journey. For example, a newly diagnosed patient may see a general practitioner or urologist, and early treatment phases may have a combination of surgical and systemic treatments provided in different practices. Patients undergoing passive surveillance or cancer survivors may also receive a large proportion of their care outside of an oncology practice.

There are now emerging opportunities to extract data across electronic health systems using *fast health care interoperability resources technologies*. There are also potential opportunities for manually assisted *natural language processing* or *deep-learning methods* to capture vast, unstructured data directly within EHRs. These tools may partially overcome the limitations of fragmented and unstructured data but are still early in implementation or systematic use. EHRs may include granular information, but the data are not adjudicated or quality checked as part of routine practice, which may result in inconsistencies in key data elements (eg, cancer stage).[33]

EHR data systems often are not interoperable, even across the same EHR system, which is a critical barrier to their use in research. However, new requirements issued by the US Department of Health and Human Services Office of the National Coordinator for Health Information Technology (ONC) mandate an increased ability to share data across these various systems to assure continuity of patient care.[34,35] As part of the 21st Century Cures Act,[36] 2 new laws colloquially known as *information (or data) blocking laws* are being enacted by the ONC and the Centers for Medicare and Medicaid Services. The ONC rule specifically requires health care providers to adopt or integrate standardized *application programming interfaces* into their electronic medical records. These requirements mean that all patients will have direct access to their electronic health information (structured and/or unstructured) using smart phones (or computers) at no cost. Similar to the application of HIPAA on claims data, this law will require a standardized set of data (referred to as data classes and data elements) outlined as the United States Core Data for Interoperability.[37] Although these data are still untested and their ability to capture specialty care like oncology is less clear, broad adoption of these application programming interfaces are likely to significantly improve data interoperability and the ability to share electronic data between and across health care systems.

Integrated care delivery conducted by health care maintenance organizations represents a different type of health care delivery in which comprehensive care is provided to patients for almost all health care services. The integrated care delivery model includes the coordination of a health care plan, medical physician groups, and a health care facility system.[38] From a data perspective within the EHR, the model includes all billed services for patients within the *closed* system—unlike fee-for-service insurance plans, in which patients can select care across multiple systems. Several integrated care organizations have consolidated their data into a virtual data warehouse to facilitate research.[38,39] The use of an EHR system across integrated care providers facilitates data access and, if patients do not receive care outside that system, potentially enables complete data on each patient.

This is in contrast to the data from fee-for-service care plans and systems that are fragmented across various practices and EHRs. An additional caution is that there is little assessment of quality of the data contained in the EHR system.

In summary, although EHRs can provide rich and deep data on a patient, the data may yield only a partial picture of the patient trajectory longitudinally because cancer care may be received in multiple facilities with different EHR systems.[31] Nevertheless, with the appropriate evaluation and study design, EHR data can be used effectively and appropriately to address many research questions.[40-43]

Cancer Registries

Registries are designed to collect uniform and systematic data on a population of patients based on exposure, disease, or outcome. Registries may or may not be independent of any one health system, payer, or EHR data vendor. Cancer registries compile records specifically on patients with cancer. These can be convenience-type samples (eg, volunteer registries for specific cancer types or drug-specific registries) based on a health care setting (eg, hospital registries) or population-based (state or central cancer registries). Most registries are designed for a specific purpose—for example, registries of patients representing rare tumors or familial syndromes. Many registries frequently collect information that may not be available from more traditional sources, such as detailed exposure data (eg, diet, physical activity) and patient-reported outcomes, but they typically represent a nonrandom sample of patients. Hospital-based registries capture detailed data on each patient and are useful for understanding the quality of care provided within a specific hospital setting. However, these health care system-based registries may not capture data on care provided outside the system in which the registry is based, similar to the limitations of EHR data. For example, if recurrence is diagnosed in the oncology office, the recurrence may be unidentified by the hospital-based registry. Facility-based registries may also have limitations for understanding the outcomes of tests ordered by the oncologist because the test results are sent directly to the oncologist and are not entered into the hospital-based information system.[44] These hospital-based registries are becoming increasingly agile, as new data items may be added readily, and data are often available in real time to enable analysis of quickly evolving clinical issues. The most important role for these hospital-based registries is in monitoring provider metrics and improving the quality of care for patients treated at that facility.[45]

Central cancer registries are unique in that they are legally mandated in each state and provide a census of all patients with cancer in a well defined geographic area. Central registries (usually state-based) collect data under state regulations that require the reporting of patient identifying health information (personally identifiable information [PII] and protected health information) from all health care providers. This data collection is HIPAA-exempt as part of public health reporting. Registries must maintain PII to comply with the requirement to consolidate data from multiple sources into a single record and to follow patients over time. This consolidation of multiple sources provides a more comprehensive picture of the cancer case, although currently the data collection is focused on the incident diagnosis and subsequent therapies. These registries do perform routine, and often active, follow-up of every patient from diagnosis until death. They contain detailed data on the characterization of each cancer case. By using new linkage methodologies, that characterization is being enhanced to include more clinically relevant information—such as genomic characterization of the cancer and detailed treatment received by each patient. National cancer registries include SEER and the National Program of Cancer Registries, which collate de-identified data from participating state central cancer registries that are then made accessible to researchers.

Limitations of registry data include a lack of information about longitudinal treatment and outcomes other than survival. Those deficiencies are being addressed through several new initiatives, including linkage of registry data with data collected by other organizations and external partners.[46,47] These new methods, along with the integration of real-time access to pathology reports, will also enable data to be reported in a more contemporary interval. With the addition of these new data, and because population-based cancer registries cover all patients within a defined geographic area, such registries provide an important opportunity to supplement understanding of therapeutic advances and their impact and effectiveness outside the clinical trial setting for population subgroups that may be underrepresented in clinical trials. An additional important component of population-based registries is that linked studies, even if not linked to the entire population, can provide information on characteristics of those individuals not included in the linkage to better understand bias.

Health Data Aggregators

The use of health data aggregators is increasingly common with the development of novel technology platforms, privacy-preserving linkages via encryption, and the need for more rapid and advanced data analytics. Health data aggregators, often called health technology data companies, enable health care data to be harnessed from across different clinical sources and sites in an integrated fashion. In the current review, we define data aggregators as entities that combine data across varied clinical sources and sites using a specified data model (eg, federated or software system-based) to provide multimodal composite data for evaluation. The resulting data sources may include patients from the general population and diverse clinical settings (ie, general practice, hospital, specialty clinic, pharmacy, etc) or may be restricted to certain diseased populations (eg, oncology clinics). The organization performing the aggregation may be gathering data for either nonprofit purposes (eg, quality improvement), or commercial purposes (eg, drug development), or both. It is critical to understand the diversity of sources being aggregated, the primary research intention, and the business model driving the data aggregation as well as to recognize that these data generally do not represent the entire population of patients.

Generally, the objective of data aggregators is to try to address the longitudinal and disparate challenges of data capture in the US health care delivery system. Therefore, they provide an infrastructure to capture patient care across the various health care facilities, physician practices, and laboratories that comprise the fragmented US health care system. Examples of data aggregators include, but are not limited to, HealthVerity,[48] IQVIA,[49] Symphony Health,[50] Flatiron,[51] and OptumLabs.[52] Although individual data sets may be limited to a single practice, health care system, or EHR software vendor, health data aggregators reduce those barriers by linking on a common protected identifier (usually encrypted) to provide aggregated, individual-level data across data sources. This approach provides a potentially more complete picture of health care utilization. The ability of aggregators to securely link patients may also result in an increased sample size, especially in rare diseases.

Limitations of data aggregation include the potential for selection bias (because patients with linkable data across clinical settings may differ from those without), and missing information is unlikely to occur at random. In designing a study, missingness across different clinical data types might be challenging to interpret or adequately understand. This can be particularly problematic for data analysts, who must be familiar with the underlying data structure and provenance of all data types. Moreover, the data pipelines and capture of elements often lack transparency and may not be systematic across all data sources. Often, these data are more challenging to use because privacy-preserving aggregation does not allow the source data to be reviewed for required comparisons when data quality issues or discrepancies arise. Although data aggregators may have an increased sample size and a large representation of heterogeneous data, close examination of the completeness, systematic capture, and longitudinality of the data requires close collaboration with the data vendor and an analyst who has appropriate training. The appropriate use of these large data sets requires familiarity with each data component and the potential impact of selection bias and/or data limitations (eg, missingness).

Specialty Data Providers and Networks

In addition to more *traditional* RWD sources, another category exists that is less well defined and more heterogeneous, but potentially important. It includes organizations that gather data or provide networks based on specialized data or for specific purposes. Examples include large electronic medical records interoperability networks, such as Carequality,[53] that coordinate information across multiple health care settings to support secure information exchange among disparate health care providers. Other examples of specialty networks include SureScripts,[54] which gathers electronic prescription data in a central repository for clinical decision making, and AmbraHealth, a specialty organization that networks radiology images and reports across different health care centers.[55] Although these may be less commonly used as RWD sources and are less easily categorized, they represent potentially important sources that may provide more comprehensive clinical data in specific areas, such as imaging.[56,57] Some of these data sources may be inclusive, whereas others serve as a convenience sample with the associated limitations.

Case Study: From Research Question to Selection of RWD Source

Here, we demonstrate how the conceptual model presented above can be applied to a given cancer-focused research question and guide the selection of an RWD source. A recently published article by Reeder-Hayes and colleagues[58] aimed to describe the uptake of ovarian suppression concurrent with endocrine therapy and its effects on endocrine therapy persistence among premenopausal women with invasive, nonmetastatic breast cancer. That RWD study question arose after a recent trial showing a clear benefit from adding ovarian suppression to endocrine therapy, particularly for premenopausal women, in prolonging disease-specific survival.[59,60] However, these benefits came at the cost of increased patient-reported side effects, including worsening hot flashes, loss of sexual interest, vaginal dryness, and sleep problems. In turn, the decision to use ovarian suppression concurrent with endocrine therapy among this population in clinical practice is complex. Therefore, the authors decided to conduct a study to examine questions surrounding the uptake and effects of concurrent variance suppression on endocrine therapy persistence using RWD.

In step 1 of the conceptual model, the authors had to clearly identify the key components of the research question using the PICOTS framework, as detailed in the column headed *STEP 1* in Table 3.[58] Specifying each component

TABLE 3. Application of the 3-Step Conceptual Model to the Case Study by Reeder-Hayes et al[a]

STEP 1: CLEARLY DEFINE THE RESEARCH QUESTION			STEP 2: UNDERSTAND THE DATA SOURCE CONTENTS AND TARGET POPULATION COVERAGE			STEP 3: ASSESS THE DATA SOURCE RELEVANCE TO THE RESEARCH QUESTION					
PICOTS FRAMEWORK			CONTENTS/COVERAGE			PROVENANCE		COMPLETENESS/QUALITY		DATA LIMITATIONS	
DESIGN COMPONENT	COMPONENT DESCRIPTION	RESEARCH QUESTION DESCRIPTION	IBM MS CCAE	SINGLE EHR	STATE CANCER REGISTRY	IBM MS CCAE	SINGLE EHR	IBM MS CCAE	SINGLE EHR	IBM MS CCAE	SINGLE EHR
Population	Define the patient population that will be studied	Premenopausal women with a diagnosis of hormone receptor-positive, early breast cancer	Proxy, all	Stage (unstructured), receptor (unstructured), menopause (proxy)	Stage, receptor, menopause (proxy)	Billing and reimbursement	Clinical care/ practice management	Proxy	X (unstructured)	Potential for measurement error of proxies	Potential for misclassification of patient exposure
Intervention	Define the intervention, including all components	Initiation of concurrent ovarian suppression administration plus endocrine therapy	X	X		Dispensing and Admin	Prescribing and Admin	X	X (potential for missing data)	Prescriptions paid out-of-pocket are missed (rare)	Prescribing data; may not be filled
Comparator	Define whether there is a placebo or active control comparator	Initiation of endocrine therapy alone	X	X		Dispensing	Prescribing	X	X (potential for missing data)	Prescriptions paid out-of-pocket are missed (rare)	Prescribing data; may not be filled
Outcome	Define the outcomes that matter to patients	Persistence and adherence to endocrine therapy	X	X		Dispensing	Prescribing		X (potential for missing data)	Prescriptions paid out-of-pocket are missed (rare)	Prescribing data; may not be filled
Timing	Define the duration of treatment and the follow-up schedule for outcome assessment	Treatment initiation as intervention (ie, intention-to-treat approach)	X	X		Dispensing	Prescribing	X	X (potential for missing data)	Prescriptions paid out-of-pocket are missed (rare)	Prescribing data; may not be filled
Setting	Define the setting where the study is implemented and relevance to real world use	Clinical practice settings in the United States	X	X	X	Nationwide, commercially insured population	Single health care system population	X	X (potential for missing data)	Limited to insured population; may not be generalizable	Limited to one health care system; may not be generalizable

Abbreviations: Admin, administration, EHR, electronic health record; IBM MS CCAE, IBM Corporation's Marketscan Commercial Claims and Encounters Database; PICOTS, population, intervention, comparator, outcome, timing, and setting; X, denotes the presence of the specified component in a given data source.

[a]See Reeder-Hayes KE, Mayer SE, Lund JL. Adherence to endocrine therapy including ovarian suppression: a large observational cohort study of US women with early breast cancer. *Cancer*. 2021;127:1220-1227.[58]

of the framework helped to solidify the study goals. Here, the intent of the authors was both descriptive—to evaluate the use of concurrent ovarian suppression and endocrine therapy—and causal—to estimate the effects of concurrent therapy versus endocrine therapy alone on endocrine therapy persistence.

Using the PICOTS framework and clearly specifying the study aims clarified the target population of interest. In step 2 of the conceptual model, the researchers identified RWD sources that could be used to address the study questions of interest. In assessing the relevance, feasibility, and accessibility of a given RWD source, researchers must consider the data contents and coverage of the RWD source. In step 2, Reeder-Hayes and colleagues considered 3 RWD types because of their accessibility: the IBM MarketScan Commercial Claims and Encounters Databases, a single health care system electronic medical records database, and a single state cancer registry, as shown in Table 3.

First, to identify the *population* of interest, the authors needed information on cancer staging, hormone receptor status, and menopausal status. None of the RWD sources could perfectly ascertain each of these components, but all would likely be able to use proxies to identify the relevant study population. Second, a critical data element for the *intervention*, *comparator*, and *outcome* is the use of endocrine therapy, which is dispensed in outpatient pharmacies and is taken by the patient at home. Cancer registries do not routinely collect the use and timing of endocrine therapy prescribing or dispensing, thus these data alone (without further linkages) would not likely be sufficient for the study question of interest. Third, follow-up for patients must be clear so that persistence to endocrine therapy can be assessed. RWD from claims and EHR can track individuals longitudinally over time. Finally, the setting of the RWD sources was assessed. The claims data capture nationwide information from individuals with commercial insurance, whereas the EHR captures all individuals (regardless of insurance status), but only within a single health care system.

In step 3 of the conceptual model, Reeder-Hayes and colleagues considered the quality of critical data elements needed to address the research question of interest. The researchers therefore ruled out the state cancer registry data because of limitations and considered the use of the claims data or EHR data. The biggest difference between these 2 sources of information was the provenance of the prescription data, in which claims reflected pharmacy dispensing data, compared with the EHR, which included only prescribing data. The EHR drug data could result in higher misclassification of drug exposure if the prescriptions were not actually filled by the patients. Although claims data capture dispensing information, it is possible for patients to pay out of pocket, and these kinds of transactions (although rare) might also result in exposure misclassification. In addition, even dispensing data do not equate to a true reflection of patient adherence, although they are regarded as a widely acceptable proxy measure.

Finally, the authors also prioritized the ability to accurately follow patients over time and observe their endocrine therapy use. In EHR, data it is possible for patients to receive care outside of the specific health care system, and such care is unlikely to be captured. The same is not true of claims data because encounters across health care systems and pharmacies are captured regardless of location. Based on an assessment of the relevance, accessibility, and quality of available RWD, Reeder-Hayes and colleagues opted to use the IBM MarketScan data. In addition to the points reviewed above, this RWD source also allowed for the ascertainment of a sufficient sample size of eligible women for the study population.

General Considerations for Using RWD for Research

Several overarching issues must be considered when using RWD for research purposes. In addition to using the proposed conceptual model discussed above when using RWD to answer research questions, it is essential to consider the question in the context of potential strengths and limitations of each distinct RWD type and specific sources within each type. The question should direct the selection of data sources rather than the converse.

Many RWD sources represent *convenience* samples and are not representative of the general population from which they are drawn. Even sources with millions of patients do not necessarily represent the entire source population or the target population of interest for the study. Exceptions include population-based registries, such as state-based cancer registries, which include all patients with cancer in a defined geographic area, and Medicare Part A, which represents inpatient coverage for nearly all individuals aged 65 years and older.

It is essential to understand the original purpose for which the RWD source was developed. For example, administrative claims data are useful for many research studies and can comprehensively capture longitudinal care across settings (ie, inpatient and outpatient). However, claims are designed for billing purposes, so treatment or service must be reimbursable, and there are associated limitations and provider coding rules for claims reimbursement that affect the specificity and accuracy of treatment or procedure capture. For example, for physician outpatient claims, the International Classification of Diseases code used must be valid, but reimbursement for these visits is based on the Common Procedure Terminology code—which may result in less accurate reporting of International Classification of

Diseases codes and greater reliability on Common Procedure Terminology codes as they undergo greater scrutiny.[61,62] Finally, data aggregators may provide a useful combination of RWD sources. However, their generalizability may be limited by the privacy-preserving linkage method (deterministic vs probabilistic) or the actual union of key sets of variables represented in the sample (eg, data subset sample and ability to follow patients longitudinally). It is also hard to trace the data provenance and assess data quality from data aggregators because the source data, by definition, are unavailable.

Another important consideration is missing or incomplete data and whether it is possible to understand the necessary patterns within the data and quantify the potential for bias in any one analysis. For most RWD sources, it is not possible to understand data missingness without a linkage to a *gold-standard* data set that includes all data for a particular set of variables. For example, when using administrative pharmacy claims data from commercial pharmacy organizations, it is difficult to know whether the patient did not receive a treatment or whether it was received from another provider who was not included in the data from the vendor. The use of such data sources to understand the uptake of specific therapies may be inappropriate unless the source includes information on the broader patient population that can be used to define the appropriate distinct denominator (for example, a specifically defined population who were eligible for the therapy and whether or not the patient is an active customer or receives other care from the pharmacy). Even population-based pharmacy data require careful evaluation. For example, Medicare Part D studies can be challenging to design because of policy designs resulting in coverage gaps (ie, the *donut hole*), in which patients must pay out of pocket for medications or drugs that are not covered and that may not be fully captured by claims.

With the exponential increase in RWD availability across various settings, another emerging approach is to assess opportunities for data linkage, including their appropriateness and feasibility.[63] Linkage might be useful to supplement a data source and enable an expanded set of research questions for analysis (eg, linkage with patient-reported outcomes). However, it requires the same unique patient identifier within each source that can be used for linking across sources. Often, each RWD source has its own unique patient identifier, but it may not permit further linkages for data security or governance reasons. The use of PII or identifiers for linking raises the issue of patient privacy and confidentiality and whether patient consent has been given for these particular uses of the data. Many linkages across disparate data sources now use *privacy-preserving patient linkages*. There are a multitude of vendors (>40) who provide this service with many different methods for matching, including deterministic, probabilistic, and combinations of both.[64]Although these types of linkages increase privacy, the accuracy and completeness of these methods have not been formally assessed. As such, rigorous evaluation and reporting of any de novo data linkage[65] should be conducted, including an assessment of the impact of false-positive and false-negative matches on the results, before using privacy-preserving patient linkages or any other linkage system.

Conclusion

The development and accessibility of RWD sources for cancer research are rapidly increasing. Therefore, it is essential to carefully consider the composition and utility of each RWD type for specific research questions. With challenges in enrollment and representativeness of patients enrolled onto cancer clinical trials, the need for RWD to address several evidence gaps is growing. Characterization and analysis of RWD for cancer and for research in other clinical areas are extremely important, especially because <5% of patients with cancer are enrolled on clinical trials, and these are not typically representative of the general population.[66] Recommendations for the use of new therapies may be based on clinical trials with limited generalizability, which may reflect outcomes under the best-case scenario; trial data do not provide information on how well these treatments may work in more diverse and complex populations, such as among older adults or those living with coexisting health conditions.

RWD are a cost-efficient, often timely source of information that have potential for answering research questions spanning the entire cancer care continuum, in addition to complementing results from clinical trials. Future efforts to integrate multiple RWD sources through transparent and robust data linkage will likely enhance the utility of RWD in cancer research. In the existing data landscape, there is no single RWD source that is likely to contain information on the entire patient trajectory. Here, using the PICOTS framework and a recent RWD case study, we demonstrate that there are several essential factors that must be evaluated in the concept, design, and analysis of RWD studies.

In summary, the appropriate use of RWD requires rigorous training of researchers, thoughtful study planning and implementation, and careful consideration of potential biases and interpretation of results to generate evidence that will reduce the cancer burden and improve the delivery of high-quality cancer care. ■

References

1. US Food and Drug Administration (FDA). Real World Evidence. Accessed August 13, 2021. fda.gov/science-research/science-and-research-special-topics/real-world-evidence
2. Giordano SH. Comparative effectiveness research in cancer with observational data. *Am Soc Clin Oncol Educ Book*. 2015;35:e330-e335.
3. Hershman DL, Wright JD. Comparative effectiveness research in oncology methodology: observational data. *J Clin Oncol*. 2012;30:4215-4222.
4. Meyer AM, Carpenter WR, Abernethy AP, Sturmer T, Kosorok MR. Data for cancer comparative effectiveness research: past, present, and future potential. *Cancer*. 2012;118:5186-5197.
5. Patient-Centered Outcomes Research Institute (PCORI) Methodology Committee. The PCORI Methodology Report. PCORI Methodology Committee; 2019.
6. Velentgas P, Dreyer NA, Nourjah P, Smith SR, Torchia MM. Developing a Protocol for Observational Comparative Effectiveness Research: A User's Guide. Publication No. 12(13)-ECH099. Agency for Healthcare Research and Quality; 2013. Accessed September 16, 2020, ncbi.nlm.nih.gov/books/NBK126190/
7. Hernan MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. *Am J Epidemiol*. 2016;183:758-764.
8. Hernan MA, Hsu J, Healy B. A second chance to get causal inference right: a classification of data science tasks. *CHANCE*. 2019;32:42-49.
9. National Academies of Sciences, Engineering, and Medicine. Examining the Impact of Real-World Evidence on Medical Product Development. Proceedings of a Workshop Series. The National Academies Press; 2019. doi:10.17226/25352
10. US Food and Drug Administration (FDA). Framework for FDA's Real-World Evidence Program. FDA; 2018. Accessed September 14, 2020. fda.gov/media/120060/download
11. Sharma M, Duan Z, Zhao H, Giordano SH, Chavez-MacGregor M. Real-world patterns of everolimus use in patients with metastatic breast cancer. *Oncologist*. 2020;25:937-942.
12. Potosky AL, Riley GF, Lubitz JD, Mentnech RM, Kessler LG. Potential for cancer related health services research using a linked Medicare-tumor registry database. *Med Care*. 1993;31:732-748.
13. Maclean JC, Halpern MT, Hill SC, Pesko MF. The effect of Medicaid expansion on prescriptions for breast cancer hormonal therapy medications. *Health Serv Res*. 2020;55:399-410.
14. Bowles EJ, Wellman R, Feigelson HS, et al. Risk of heart failure in breast cancer patients after anthracycline and trastuzumab treatment: a retrospective cohort study. *J Natl Cancer Inst*. 2012;104:1293-1305.
15. Lau EC, Mowat FS, Kelsh MA, et al. Use of electronic medical records (EMR) for oncology outcomes research: assessing the comparability of EMR information to patient registry and health claims data. *Clin Epidemiol*. 2011;3:259-372.
16. Callahan A, Shah NH, Chen JH. Research and reporting considerations for observational studies using electronic health record data. *Ann Intern Med*. 2020;172(11 suppl):S79-S84.
17. Cronin KA, Lake AJ, Scott S, et al. Annual report to the nation on the status of cancer, part I: national cancer statistics. *Cancer*. 2018;124:2785-2800.
18. Gearhart-Serna LM, Hoffman K, Devi GR. Environmental quality and invasive breast cancer. *Cancer Epidemiol Biomarkers Prev*. 2020;29:1920-1928.
19. Brown V, Partridge A, Chu L, et al. MotHER: a registry for women with breast cancer who received trastuzumab (T) with or without pertuzumab (P) during pregnancy or within 6 months prior to conception [abstract]. *J Clin Oncol*. 2013;31(15 suppl):TPS658.
20. Boffa DJ, Rosen JE, Mallin K, et al. Using the National Cancer Database for outcomes research: a review. *JAMA Oncol*. 2017;3:1722-1728.
21. McCabe RM. National Cancer Database: the past, present, and future of the cancer registry and its efforts to improve the quality of cancer care. *Semin Radiat Oncol*. 2019;29:323-325.
22. Steele JR, Wellemeyer AS, Hansen MJ, Reaman GH, Ross JA. Childhood cancer research network: a North American Pediatric Cancer Registry. *Cancer Epidemiol Biomarkers Prev*. 2006;15:1241-1242.
23. Brown JS, Maro JC, Nguyen M, Ball R. Using and improving distributed data networks to generate actionable evidence: the case of real-world outcomes in the Food and Drug Administration's Sentinel system. *J Am Med Inform Assoc*. 2020;27:793-797.
24. Khozin S, Miksad RA, Adami J, et al. Real-world progression, treatment, and survival outcomes during rapid adoption of immunotherapy for advanced non-small cell lung cancer. *Cancer*. 2019;125:4019-4032.
25. Kabadi SM, Near A, Wada K, Burudpakdee C. Treatment patterns, adverse events, healthcare resource use and costs among commercially insured patients with mantle cell lymphoma in the United States. *Cancer Med*. 2019;8:7174-7185.
26. Gajra A, Feinberg BA. A letter in support of real-world RECIST. *Adv Ther*. 2020;37:1688-1690.
27. Healthcare Delivery Research Program, Division of Cancer Control & Population Sciences, National Cancer Institute, National Institutes of Health. Surveillance, Epidemiology, and End Results (SEER)-Medicare Linked Database. Accessed September 16, 2020. healthcaredelivery.cancer.gov/seermedicare/
28. Healthcare Delivery Research Program, Division of Cancer Control & Population Sciences, National Cancer Institute, National Institutes of Health. Surveillance, Epidemiology, and End Results (SEER) Medicare Consumer Assessment of Healthcare Providers and Systems (CAHPS) Linked Data Resource. Accessed September 16, 2020. healthcaredelivery.cancer.gov/seer-cahps/.
29. Healthcare Delivery Research Program, Division of Cancer Control & Population Sciences, National Cancer Institute, National Institutes of Health. Medicare Part D Prescription Drug Data. Accessed September 16, 2020. healthcaredelivery.cancer.gov/seer-mhos/overview/
30. Davidoff AJ, Canavan ME, Feder S, et al. Patterns of pain medication use associated with reported pain interference in older adults with and without cancer. *Support Care Cancer*. 2020;28:3061-3072.
31. Clarke CA, Glaser SL, Leung R, Davidson-Allen K, Gomez SL, Keegan TH. Prevalence and characteristics of cancer patients receiving care from single vs. multiple institutions. *Cancer Epidemiol*. 2017;46:27-33.
32. Gondi S, Wright AA, Landrum MB, Zubizarreta J, Chernew ME, Keating NL. Multimodality cancer care and implications for episode-based payments in cancer. *Am J Manag Care*. 2019;25:537-538.
33. Kim E, Rubinstein SM, Nead KT, Wojcieszynski AP, Gabriel PE, Warner JL. The evolving use of electronic health records (EHR) for research. *Semin Radiat Oncol*. 2019;29:354-361.
34. Office of the National Coordinator for Health Information Technology (ONC), Department of Health and Human Services 21st Century Cures Act: Interoperability, Information Blocking, and the ONC Health IT Certification Program. Accessed August 13, 2021. federalregister.gov/documents/2020/05/01/2020-07419/21st-century-cures-act-interoperability-information-blocking-and-the-onc-health-it-certification

35. Reisman M. EHRs: the challenge of making electronic data usable and interoperable. *P T*. 2017;42:572-575.
36. Office of the National Coordinator for Health Information Technology (ONC), Department of Health and Human Services. ONC's Cures Act Final Rule. Accessed August 13, 2021. healthit.gov/curesrule/
37. Office of the National Coordinator for Health Information Technology (ONC), Department of Health and Human Services. United States Core Data for Interoperability (USCDI). Accessed August 13, 2021. healthit.gov/isa/united-states-core-data-interoperability-uscdi
38. Pines J, Selevan J, George M, McClellan M. Richard Merkin Initiative on Payment Reform and Clinical Leadership Case Study: Emergency Medicine. Brookings Institution; 2015.
39. Ross TR, Ng D, Brown JS, et al. The HMO Research Network Virtual Data Warehouse: a public data model to support collaboration. *EGEMS (Wash DC)*. 2014;2:1049.
40. Hernandez-Boussard T, Monda KL, Crespo BC, Riskin D. Real world evidence in cardiovascular medicine: ensuring data validity in electronic health record-based studies. *J Am Med Inform Assoc*. 2019;26:1189-1194.
41. Griffith SD, Tucker M, Bowser B, et al. Generating real-world tumor burden endpoints from electronic health record data: comparison of RECIST, radiology-anchored, and clinician-anchored approaches for abstracting real-world progression in non-small cell lung cancer. *Adv Ther*. 2019;36:2122-2136.
42. Wang SV, Patterson OV, Gagne JJ, et al. Transparent reporting on research using unstructured electronic health record data to generate 'real world' evidence of comparative effectiveness and safety. *Drug Saf*. 2019;42:1297-1309.
43. Kogan E, Twyman K, Heap J, Milentijevic D, Lin JH, Alberts M. Assessing stroke severity using electronic health record data: a machine learning approach. *BMC Med Inform Decis Mak*. 2020;20:8.
44. Zhang L, Hsieh MC, Petkov V, Yu Q, Chiu YW, Wu XC. Trend and survival benefit of Oncotype DX use among female hormone receptor-positive breast cancer patients in 17 SEER registries, 2004-2015. *Breast Cancer Res Treat*. 2020;180:491-501.
45. Raval MV, Bilimoria KY, Stewart AK, Bentrem DJ, Ko CY. Using the NCDB for cancer care improvement: an introduction to available quality assessment tools. *J Surg Oncol*. 2009;99:488-490.
46. Petkov VI, Miller DP, Howlader N, et al. Breast-cancer-specific mortality in patients treated based on the 21-gene assay: a SEER population-based study. *NPJ Breast Cancer*. 2016;2:16017.
47. Kurian AW, Ward KC, Howlader N, et al. Genetic testing and results in a population-based cohort of breast cancer patients and ovarian cancer patients. *J Clin Oncol*. 2019;37:1305-1315.
48. HealthVerity. Company website. Accessed August 13, 2021. healthverity.com/
49. IQVIA. Real-World Datasets. Accessed August 13, 2021. iqvia.com/solutions/real-world-evidence/real-world-data-and-insights
50. Symphony Health. Company website. Accessed August 13, 2021. https://symphonyhealth.prahs.com/
51. Flatiron Health. Company website. Accessed August 13, 2021. flatiron.com/
52. OptumLabs. Company website. Accessed August 13, 2021. optumlabs.com/
53. Carequality. Company website. Accessed August 13, 2021. carequality.org/
54. SureScripts. Company website. Accessed August 13, 2021. surescripts.com/
55. Ambra Health. Company website. Accessed August 13, 2021. ambrahealth.com/
56. Smith JY, Sow MM. Access to e-prescriptions and related technologies before and after Hurricanes Harvey, Irma, and Maria. *Health Aff (Millwood)*. 2019;38:205-211.
57. Penn CL. E-prescribing & SureScripts. The state and national landscape. *J Ark Med Soc*. 2009;106:20-21.
58. Reeder-Hayes KE, Mayer SE, Lund JL. Adherence to endocrine therapy including ovarian suppression: a large observational cohort study of US women with early breast cancer. *Cancer*. 2021;127:1220-1227.
59. Pagani O, Regan MM, Francis PA, TEXT and SOFT Investigators; International Breast Cancer Study Group. Exemestane with ovarian suppression in premenopausal breast cancer. *N Engl J Med*. 2014;371:1358-1359.
60. Saha P, Regan MM, Pagani O, et al. Treatment efficacy, adherence, and quality of life among women younger than 35 years in the International Breast Cancer Study Group TEXT and SOFT adjuvant endocrine therapy trials. *J Clin Oncol*. 2017;35:3113-3122.
61. Abraha I, Montedori A, Serraino D, et al. Accuracy of administrative databases in detecting primary breast cancer diagnoses: a systematic review. *BMJ Open*. 2018;8:e019264.
62. Shehab N, Ziemba R, Campbell KN, et al. Assessment of ICD-10-CM code assignment validity for case finding of outpatient anticoagulant-related bleeding among Medicare beneficiaries. *Pharmacoepidemiol Drug Saf*. 2019;28:951-964.
63. Rivera DR, Gokhale MN, Reynolds MW, et al. Linking electronic health data in pharmacoepidemiology: appropriateness and feasibility. *Pharmacoepidemiol Drug Saf*. 2020;29:18-29.
64. Synectics for Management Decisions, Inc. Landscape Analysis: Privacy Preserving Record Linkage Analysis (P3RL) Study (performed for Leidos Biomedical Research under Agreement 18Q110, issued as a subcontract under contract HHSN261201500003I issued by the National Cancer Institute, National Institutes of Health, Department of Health and Human Services). Synectics for Management Decisions, Inc; 2019. Accessed August 13, 2021. nih.box.com/v/P3RLSlandscape
65. Pratt NL, Mack CD, Meyer AM, et al. Data linkage in pharmacoepidemiology: a call for rigorous evaluation and reporting. *Pharmacoepidemiol Drug Saf*. 2020;29:9-17.
66. Murthy VH, Krumholz HM, Gross CP. Participation in cancer clinical trials: race-, sex-, and age-based disparities. *JAMA*. 2004;291:2720-2726.