Registries for Evaluating Patient Outcomes: A User's Guide

Addendum 2 - Tools and Technologies for Registry Interoperability

DEPARTMENT OF HEALTH & HUMAN SERVICES • USA

AHRQ
Agency for Healthcare
Research and Quality

EVIDENCE-BASED PRACTICE CENTERS®

Tools and Technologies for Registry Interoperability

Registries for Evaluating Patient Outcomes: A User's Guide 3rd Edition, Addendum 2

Prepared for:
Agency for Healthcare Research and Quality
U.S. Department of Health and Human Services
5600 Fishers Lane
Rockville, MD 20857
www.ahrq.gov

Contract No. 290-2014-00004-C

Prepared by:
L&M Policy Research, LLC, with partners OM1 and IQVIA
Washington, DC

Editors:
Richard E. Gliklich, M.D.
Chief Executive Officer and Chairman
OM1

Michelle B. Leavy, M.P.H.
Head, Healthcare Researcy & Policy
OM1

Nancy A. Dreyer, Ph.D., M.P.H., FISPE
Chief Scientific Officer
IQVIA Real-World & Analytic Solutions

AHRQ Publication No. 19(20)-EHC017-EF
October 2019

The Effective Health Care Program of the Agency for Healthcare Research and Quality (AHRQ) conducts and supports research focused on the outcomes, effectiveness, comparative clinical effectiveness, and appropriateness of pharmaceuticals, devices, and healthcare services. More information on the Effective Health Care Program can be found at www.effectivehealthcare.ahrq.gov.

This report is based on research conducted by L&M Policy Research, LLC, and partners OM1 and IQVIA, under contract to the Agency for Healthcare Research and Quality, Rockville, MD (Contract No. 290-2014-00004-C). The findings and conclusions in this document are those of the authors, who are responsible for its contents; the findings and conclusions do not necessarily represent the views of AHRQ. Therefore, no statement in this report should be construed as an official position of AHRQ or of the U.S. Department of Health and Human Services. All statements in this report, including its findings and conclusions, are solely those of the authors. The authors participated as individuals and not necessarily as representatives of their organizations. No statement in this document should be construed as an official position of any organization represented here.

None of the authors have any affiliations or financial involvement that conflicts with the material presented in this report.

Citation: Gliklich RE, Leavy MB, Dreyer NA, (eds). Tools and Technologies for Registry Interoperability, Registries for Evaluating Patient Outcomes: A User’s Guide, 3rd Edition, Addendum 2. (Prepared by L&M Policy Research, LLC under Contract No. 290-2014-00004-C.) AHRQ Publication No. 19(20)-EHC017-EF. Rockville, MD: Agency for Healthcare Research and Quality; October 2019. Posted final reports are located on the Effective Health Care Program search page. DOI: https://doi.org/10.23970/AHRQEPCREGISTRIES3ADDENDUM2.

Acknowledgments: Numerous individuals contributed to the coordination and formatting of this eBook. In addition to the authors and reviewers, we wish to acknowledge the following individuals for their significant time and dedication: Devyn Andrews, Valmeek Kudesia, Lisa Green, Laura Faas, Betsy Brigantti, Ellen Crain, Katherine Ianni, and Monica Sarmiento.

Preface

Addendum II to the Third Edition of the Registries for Evaluating Patient Outcomes: A User's Guide was performed under a contract from the Agency for Healthcare Research and Quality (AHRQ) with the purpose of presenting new, emerging themes related to designing and conducting registries. First published in 2007, the User's Guide, with translations available in Chinese and Korean, serves as a reference for planning, developing, maintaining, and evaluating registries designed to collect data about patient outcomes. The second (2010) and third (2014) editions incorporated updates to existing topics and included new chapters on methodological and technological advances in registry science. The first addendum to the Third Edition of the Registries for Evaluating Patient Outcomes, 21st Century Patient Registries, EBook addendum, was published in March 2018.

We are pleased to present five new chapters that address the changing health information technology (IT) environment in which registries operate and describe the potential of new data sources, the role of data standards, and the technologies that support interoperability of registries with other types of health IT.

Like the User's Guides, Addendum II was created with support from a large group of stakeholders representing academia, industry, government, and technology organizations. At the outset, we solicited feedback on chapter topics and outlines from AHRQ, academics, and other experts in the field. We then reached out to topic experts inviting participation in writing or reviewing the final topics selected. Once the authorship groups were established, many meetings were held to draft the chapters prior to sending for constructive feedback and editorial review to the assigned reviewer group for each paper. The collaborative efforts of contributors, reviewers, and editors resulted in a draft document that was posted for public comment on the Effective Health Care website in January 2019. This document incorporates much of the feedback received. Like previous editions, the contributors and reviewers participated as individual experts and not necessarily as representatives of their organizations. We are grateful to all those who contributed in writing, reviewing and editing this document.

We believe that these new chapters address important emerging topics in the design and development of registries in a rapidly changing health IT landscape. These topics are in an active state of development, and we offer this Addendum to aid in further development of this field.

Contents

Figures

Tables

Appendixes

Chapter 1. Health Information Technology (IT) and Patient Registries

Authors (alphabetical)

Robert S. Miller, M.D., FACP, FASCO
Medical Director, CancerLinQ
American Society of Clinical Oncology

Kristi Mitchell, M.P.H.
Senior Vice President, Center for Healthcare Transformation
Avalere Health

Rachel Myslinski, M.B.A., M.P.H.
Vice President, Practice, Advocacy & Quality
American College of Rheumatology

Josh Rising, M.D., M.P.H.
Director, Healthcare Programs
The Pew Charitable Trusts

Introduction

Health IT has evolved rapidly in the past decade. Implementation of electronic health records (EHRs) has become widespread in hospitals and ambulatory care settings.[1] Vast amounts of electronic health data are now available for use in retrospective and prospective research studies and quality improvement initiatives, and many new efforts focus on harnessing these data to inform clinical decision making, manage and improve population health, and conduct proactive safety surveillance. Electronic health data are not limited to data generated by providers; with the increasing use of applications (apps) and wearable devices, individuals are able to generate large volumes of personal health data on, for example, physical activity, heart rate, and body temperature.[2, 3]

The availability of electronic health data offers the potential for patient registries to collect deep, nuanced data on large numbers of patients at significantly reduced costs, in comparison to manual collection of data using case report forms. A patient registry is defined as "an organized system that uses observational study methods to collect uniform data (clinical and other) to evaluate specified outcomes for a population defined by a particular disease, condition, or exposure and that serves one or more pre-determined scientific, clinical, or policy purposes."[4]

While patient registries may be designed to meet a specific clinical research question, attention within the United States in recent years has focused on developing registries that can serve as central components of a learning health system and national health data infrastructure. A 2013 Institute of Medicine report defined a learning health system as a system that is 'designed to generate and apply the best evidence for the collaborative healthcare choices of each patient and provider; to drive the process of discovery as a natural outgrowth of patient care; and to ensure innovation, quality, safety, and value in healthcare.'[5] Within a learning health system, registries can support population health management, clinical decision making, quality improvement, and clinical research.[6, 7] However, registries must be connected to other registries, EHRs, and other data sources in order to meet these goals within a learning health system.

This chapter describes advances in health information technology in the past decade, explores the potential role of registries within a learning health system and national health data infrastructure, reviews the increasing role of real-world evidence in health policy, and discusses the potential for registries to leverage electronic health data to generate real-world evidence in an efficient, high-quality manner. Subsequent chapters in this eBook address the practical aspects of how patient registries can leverage electronic health data and the scientific, data quality, and legal and ethical questions that use of these data introduce.

Advances in Health Information Technology

The adoption of IT in the healthcare system has increased dramatically in the past decade. Until 2009, the U.S. healthcare system largely relied on paper medical records, which limited patients' ability to share information efficiently with other care providers and made health outcomes research difficult. To spur adoption of health information technology, Congress passed the Health Information Technology for Economic and Clinical Health (HITECH) Act as part of the American Recovery and Reinvestment Act of 2009 (ARRA). Under this legislation, the Medicare and Medicaid EHR Incentive Programs provided extensive financial support for adoption of EHRs, while other funding created and expanded Health Information Exchange (HIE) infrastructure. Prior to passage of the HITECH Act, approximately 17 percent of physicians and 9 percent of hospitals had at least a basic EHR. By 2015, 78 percent of physicians and 96 percent of hospitals had certified[a] EHR technology.[1]

Learning Health System & National Health Data Infrastructure

This rapid adoption of EHRs has created an extraordinary amount of electronic health data, and, in recent years, national attention has shifted from encouraging adoption of EHRs to building a learning health system and a national health data infrastructure that can support research and safety surveillance. Health policy consultant Lynn Etheredge first described the learning health system in 2007. In a paper in *Health Affairs*,[8] Etheredge laid out the existing knowledge gaps including a diffusion of responsibility for evidence-based medicine across Federally-funded research, especially the National Institutes of Health, life sciences companies, and the Food and

[a] *A certified EHR is EHR technology that meets the technological capability, functionality, and security requirements adopted by the Department of Health and Human Services*

Drug Administration (FDA); deficiencies in the clinical trials system; and general underinvestment in evidence-based research. He described rapid-learning opportunities that might create knowledge breakthroughs in several high-priority policy areas such as comparative benefits of drugs and biologics, the evidence base for surgical and interventional procedures, the impact of environmental factors on disease, and the health needs of minorities and patients with special needs.[8] He noted that the databases comprising the learning health system could be organized per insured population, provider type, health conditions (disease registries), age cohorts, minority populations, and others. A fully realized learning health system could modernize the Medicare/Medicaid reimbursement structure, enable a coordinated national clinical trials infrastructure, and reimagine how health technology assessments might be performed.

While a highly theoretical and aspirational vision in 2007, the concept of the learning health system has been refined through extensive national policy discussions led by the National Academy of Medicine (formerly the Institute of Medicine) and others and through numerous specialized rapid-learning pilots and initiatives.[9] For example, stakeholders participating in a workshop convened by the National Cancer Policy Forum of the Institute of Medicine in October 2009 proposed the discipline of oncology as a logical condition in which to pilot test rapid-learning principles, noting that many components of an oncology learning health system were already in place, including resources from the National Cancer Institute, state cancer registries and other databases, and a library of quality measures.[10] The American Society of Clinical Oncology (ASCO), as a professional medical society representing oncologists and other cancer clinicians, took on the task of creating an oncology learning health system by developing and implementing CancerLinQ® (Cancer Learning Intelligence Network for Quality), launching the initial version in 2016. CancerLinQ aggregates data from EHRs from U.S.-based oncology practices and delivers a suite of quality improvement tools and data visualizations for clinical care, as well as customized, fit for purpose datasets of aggregated, de-identified data for research.[11]

In parallel to the development of learning health systems, there has been an increased focus on providing patients with access to their health information electronically, to allow patients to monitor their records, contribute information, correct errors, and directly share their information with research studies. The learning health system vision has evolved to recognize that individuals must play a central role and that an individual’s health data should not be limited to what is stored in EHRs but should also include information from many sources, including technologies used by the individual.[12, 13]

Registries may play a central role in the learning health system and national health data infrastructure as sources of data that can be used to support population health and clinical decision making, quality improvement, and clinical research.[14] In particular, registries often serve as a bridge between clinical trials and clinical practice by continuing to study the effectiveness of new medical products in a real-world setting and supporting the translation of clinical practice guidelines into clinical practice. Registries are also an important source of patient-generated data, such as patient-reported outcomes.[2, 4, 15-17] Connecting patient registries to other registries, to EHRs, and to other data sources would move the nation closer to the vision of a learning health system and national health data infrastructure to support research and safety surveillance. Central to this goal are data linkage and data standardization, predicated upon strong partnerships among patients, health systems, providers, regulators, and hospitals. For example, through efforts like the Patient-Centered Clinical Research Network (PCORnet), largely comprised of 12 clinical data research networks (CDRNs) and 20 Patient-powered research networks, a national research infrastructure is being established focusing primarily on observational and comparative clinical effectiveness research (CER) studies.[18] At its core, this network has the ability to weave data from these various initiatives to address central research questions among interested groups of patients and health systems willing to share such health information.

While some progress has been made, true interoperability among EHRs and other electronic systems has yet to be achieved and remains a major technology barrier for building the learning health system. Interoperability is defined as "the ability of a system to exchange electronic health information with and use electronic health information from other systems without special effort on the part of the user."[12] Interoperability is complex, and achieving full interoperability will require the healthcare system to address many technical and organizational challenges related to security, data semantics, data format, standard services, transport techniques, and individual data matching. Recognizing the need for a national strategy to achieve full interoperability, the Office of the National Coordinator of Health Information Technology released "Connecting Health and Care for the Nation: A Shared Nationwide Interoperability Roadmap" in 2015, in which it describes the role of interoperability in building a learning health system and the key challenges that must be addressed.[12]

A full exploration of EHR interoperability is beyond the scope of this chapter. However, six key technical challenges are discussed briefly here to provide context for discussions of interoperability in subsequent chapters. More information on interfacing EHRs with registries can also be found in Chapter 4.

Interoperability Challenges

Security. Several issues related to security must be addressed to support full interoperability across the healthcare system. Interoperability requires the exchange of electronic health information, much of which is protected by patient privacy laws that govern how and with whom the data may be shared. Organizations that are transferring electronic health data must have confidence that data they send and receive are accurate and only accessed by authorized individuals. Security challenges related to interoperability include (1) ensuring that secure methods of transporting data are available; (2) preventing the unauthorized or unintended altering of data; and (3) verifying that individuals accessing data have appropriate permissions to do so.

Data Semantics. Data semantics refers to the clinical vocabularies and coding systems used to represent electronic health information. Semantic interoperability is the "ability to automatically interpret the information exchanged meaningfully and accurately in order to produce useful results as defined by the end users of both systems."[12] For example, care providers may use different terms for the same concept (e.g., Tylenol, acetaminophen). For data to be transferred and interpreted correctly, the systems exchanging the data must recognize that these terms are synonyms.

Vocabulary and terminology standards ensure that data are recorded consistently and can be interpreted correctly by other systems. Several vocabulary and terminology standards have been developed, including Systematized Nomenclature of Medicine-Clinical Terms (SNOMED CT)[19] for problems or conditions, RxNorm[20] for medications and medication allergies, and Logical Observation Identifiers Names and Codes (LOINC)[21] for laboratory tests and vital signs. Value sets are also used to identify values within standard vocabularies that are used for a specific purpose, such as identifying a population of patients for which a specific quality measure applies; the Value Set Authority Center (VSAC) is a Federal repository for value sets.[22] Common data models, such as those developed by the Sentinel Initiative[23] and the Observational Health Data Sciences and Informatics (OHDSI)[24] group are also used to standardize and facilitate the sharing of data across sites and systems. More information on data standards and common data models can be found in Chapters 3 and 5, respectively.

These tools are an important foundation for interoperability, but gaps still exist. Systems may use different coding standards, and some data that are particularly important for patient registries and other research purposes, such as radiographic images, pathology slides, and clinical notes, may not be recorded using vocabulary and terminology standards. In addition, information that is important for some research purposes may not be captured in all systems. For example, EHRs may not include information on patient characteristics, such as socioeconomic status, or patient-reported outcomes. Chapter 3 discusses the role of standardized outcome measures in supporting the consistent capture of key data elements across health information systems.

Data Format. Electronic health data formats refer to how data are structured so that data sent from one system can be interpreted, integrated and used by another system. Within the healthcare system, data are sent to and from a wide variety of health IT systems, and standard formats are needed to facilitate these exchanges. Existing standard formats include Consolidated Clinical Document Architecture (C-CDA) and HL7 v2 messaging. A newer approach is FHIR, or Fast Health Interoperability Resources. FHIR was developed by Health Level Seven (HL7) to support the exchange, integration, and retrieval of electronic health information. It is notable in that it does not require the exchange of entire documents, but rather supports the exchange of specific, clearly-defined pieces of information, thereby allowing for faster and more efficient exchanges of information. FHIR relies on application programming interfaces (APIs), which are discussed below.

Standard Secure Services. Standard, secure services support the functional capabilities that are necessary for exchanging data and require the use of service-oriented architecture (SOA). Within a SOA, application programming interfaces (APIs) define how systems interact with each other and exchange structured information. A 2014 JASON report, ‘A Robust Health Data Structure,’ recommended the creation of public APIs that are uniformly available, non-proprietary, tested by a trusted third party, and operate within a clearly defined business and legal framework.[25] Standard APIs that support exchange of data using FHIR could support greater interoperability across systems. These tools are discussed in more detail in Chapter 5.

Transport Techniques. Transport techniques describe how data are moved from one system to another system and are closely tied to security requirements. Approaches currently available are the Direct transport protocol, which uses existing email transport protocols in a secure manner; the web services approach, which typically uses the Simple Object Access Protocol (SOAP)-based web services to support transport for queries;[26] and the RESTful implementation, which is used by HL7's Fast Healthcare Interoperability Resources (FHIR) project.[27]

Individual Data Matching. Many countries with national health services assign an identification number for every individual that is recorded in his or her encounters with the healthcare system. Privacy issues notwithstanding, linkage and aggregation of an individual's information is straightforward. In the United States, in contrast, each provider typically assigns a unique identifier to each individual, and those identifiers are not used by other providers. An exception is the Department of Veterans Affairs and the Department of Defense, which assign a single personal identifier to each service member. Efforts to create a unique health identifier for each individual in the United States have been limited by privacy concerns and a 1998 prohibition on using Federal funds to investigate or create unique patient identifiers.[28] When data are exchanged across providers, individual demographic data and matching algorithms are used to match individual data. This method is not completely accurate, and some records may not be matched correctly or may not be matched at all. Errors often result from data quality issues in demographic data (such as incorrectly entered date of birth, misspelled names, changes in address, etc.). Errors in matching patient information can lead to safety issues and inappropriate treatments. Full interoperability requires that an individual's information, when transferred from one system to another, is matched and linked to the correct individual with complete accuracy. New approaches are needed to improve the quality of the data used to match patients, to improve the accuracy of deterministic and probabilistic matching algorithms, and to compare results and identify the most accurate matching algorithms for future adoption.

Many efforts, both publicly and privately funded, are underway to address these challenges and improve interoperability, so that electronic health data can be used to support a robust learning health system. In addition, policy changes, such as payment system reform and use of real-world evidence to support regulatory decision making, have created a business case for increased access to electronic health data and strengthened incentives to move toward interoperability in the healthcare system. These changes are discussed further in the following section, Availability and Use of Real-World Evidence.

Availability and Use of Real-World Evidence

As availability of electronic health data has increased, many efforts have focused on using these "real-world" data to monitor and improve patient outcomes, track the safety of medical products, and improve the value and efficiency of healthcare. The concepts of "real-world evidence" (RWE) and "real-world data" (RWD) have grown rapidly in importance in recent years in the domains of clinical practice, biomedical research, and healthcare economics and regulation. This section describes several areas where use of real-world data has evolved over the past decade. The following section discusses the role that registries play in generating real-world data to support these uses.

Despite wide usage, the exact definitions of RWE and RWD are often unclear, and they are commonly defined in the negative (i.e., RWD are data gathered from sources other than traditional clinical trials). In December 2018, the FDA released a framework for its RWE program. In this document, the FDA defines RWD and RWE as follows:

> “Real-World Data (RWD) are data relating to patient health status and/or the delivery of health care routinely collected from a variety of sources.
>
> Real-World Evidence (RWE) is the clinical evidence about the usage and potential benefits or risks of a medical product derived from analysis of RWD.
>
> Examples of RWD include data derived from electronic health records (EHRs); medical claims and billing data; data from product and disease registries; patient-generated data, including from in-home-use settings; and data gathered from other sources that can inform on health status, such as mobile devices. RWD sources (e.g., registries, collections of EHRs, administrative and medical claims databases) can be used for data collection and, in certain cases, to develop analysis infrastructure to support many types of study designs to develop RWE, including, but not limited to, randomized trials (e.g., large simple trials, pragmatic clinical trials) and observational studies (prospective or retrospective).”[29]

As discussed elsewhere in the User’s Guide, evidence obtained from prospective research, especially RCTs, has traditionally formed the foundation of the biomedical knowledge base and has informed the creation of clinical practice guidelines, treatment pathways, and other standards. RCTs have the advantage of strong internal validity, standardized interventions and pre-defined outcomes measures, and minimization of bias; therefore, they reflect “what can work” (efficacy). Nonetheless, their limitations are significant. For example, the population studied may not reflect the types of patients typically encountered in practice, who may be older and have more comorbidities. RCTs have weak external validity, are costly to perform, and are often slow to achieve results that can inform contemporary practice.

In contrast, RWD reflects “what does work” (effectiveness) with strong external validity and captures outcomes of patients commonly encountered in practice. However, RWD may be incomplete and of uncertain quality, and heterogeneity in the subject population may obscure any treatment effect.[30] Susceptibility to bias is probably the most significant shortcoming of RWD and RWE. Chapter 3 of the User’s Guide discusses the potential types of bias in patient registries and real-world data generally.

Some organizations have issued guidance statements to define best practices in the use of RWD studies for decision making,[31-33] noting both potential limitations as well as new opportunities. For example, if RWD are considered more broadly than retrospectively captured data alone, valuable prospective observational and interventional trials can be designed using “real-world” endpoints, potentially embedded in clinical practice, as pragmatic clinical trials. Pragmatic clinical trials could remedy some of the current deficiencies of evidence generation, particularly the lack of patient-centric endpoints.[34] Partnering with patients and consumers is important for many reasons, not only to better define endpoints, but also because patients themselves may be the best source of data rarely captured in EHRs, such as physical activity levels and details about socioeconomic factors and educational background.

Regulatory Decision Making

While there will continue to be a need for early stage controlled trials of therapeutics to characterize basic biology and safety, RWE can meaningfully inform regulatory endpoints as well. For the past several years, the FDA has signaled a growing openness to using RWE in regulatory decision making. Legislative requirements in the 21st Century Cures Act now obligate the FDA to consider RWE endpoints in drug label expansions, and the FDA has created a framework for evaluating the potential use of RWE in this context and in the context of supporting or satisfying postapproval study requirements.[29]

In addition, several efforts have focused on developing RWD sources to support post-market surveillance, particularly for medical devices. The Medical Device Epidemiology Network (MDEpiNET) and the National Evaluation System for health Technology (NEST) are exploring use of RWD from patient registries, EHRs, and other sources to support safety surveillance. In 2015, the Medical Device Registry Task Force called for a coordinated registry network or CRN as the “foundational architectural construct for the national system that will augment national registry development.” Through this effort, work streams in cardiovascular disease (PASSION), peripheral vascular disease (RAPID), and other areas are being launched.[35] As a central tenet to bolstering existing registry platforms, these initiatives are establishing robust partnerships, developing common data models, and implementing an infrastructure to capture EHR data.

Value-Based Care

RWD have also played a central role in the movement towards value-based care. In recent years, the healthcare industry has experienced a monumental shift in how services are valued and reimbursed. The shift from volume-based care to value-based care began in the early 2000s and continues to evolve. The definition of value-based care is essentially a financial reward system and/or payment model that values improved clinical outcomes at an efficient cost. In theory, this is a reasonable way to compensate providers that aligns incentives between patients and providers; in practice, concerns have emerged about the validity of outcome measures, the difficulty of attributing healthcare costs, and increased administrative burden.

Value-based care gained significant momentum in 2015, when the Congress replaced the Sustainable Growth Rate with the Medicare Access and CHIP Reauthorization Act of 2015 (MACRA). MACRA made significant changes to the way providers were reimbursed for Medicare patients. MACRA offers two pathways for providers: the Merit-Based Incentive Program (MIPS), which is the default pathway for the majority of providers, and an Alternative Payment Model (APM) path. MIPS requires providers to report information across four categories: quality, which requires reporting performance on a number of quality measures, including at least one outcome measure; advancing care information, which replaces the Medicare EHR Incentive Program, also known as Meaningful Use; Improvement Activities, which is a new category that looks at provider’s activities related to quality improvement in their practice; and Resource Use, which replaces the Value-based Modifier program. As of the 2018 reporting year, the Resource Use category has been weighted to zero and not implemented due to inherent problems collecting and attributing cost data in the US healthcare system. Providers who participate in an approved APM are able to avoid MIPS reporting without penalties. A subset of APMs, called Advanced APMs, allow providers to earn bonuses by taking on some risk related to their patient outcomes.[36]

In addition to the Centers for Medicare and Medicaid Services (CMS), private payers increasingly are incorporating value-based care concepts into their payment structures. For example, UnitedHealthcare's Premium Designation Program provides physician designations based on quality and cost efficiency criteria to help members make more informed and personally appropriate choices for their medical care.[37] Physicians may also use these designations when referring patients to other physicians. Blue Cross Blue Shield has the "Total Care"[38] and "Specialty Care"[39] quality initiatives, which also use quality and cost data to designate preferred providers and hospitals. These programs are an attempt to promote quality of care and cost efficiency among providers and offer more transparency to patients when selecting providers of care.

Of note, the movement toward value-based care has focused on measuring outcomes, including patient-reported outcomes, instead of relying on process measures.[40] Recognizing that patient outcomes are likely the most valuable indicator of quality of care, CMS, National Quality Forum, physician organizations, and other stakeholders have emphasized the importance of outcomes measurement in value-based care efforts. However, not all condition areas have reliable and validated outcome measures, and, even when these are available, they may not reflect what is most important to patients.[41] Patient registries can play an important role in the development of new measures by collecting data to build an evidence base for new measure concepts and providing a platform to rigorously test and validate newly specified measures. For example, the American College of Cardiology's National Cardiovascular Data Registry, which was launched in 1997,[42] encompasses a suite of registries that meets multiple purposes, including serving as a flexible platform to develop and test new quality measures across a myriad of cardiovascular related diseases and conditions.

The CMS alternative quality mechanisms for Value-based Performance Measurement (VBPM) offers a path for registries to collect data to support value-based care. Under this program, CMS allows registries to become specialty and sub-specialty reporting mechanisms under the Qualified Clinical Data Registry (QCDR) Program.[43] In this capacity, professional societies, particularly sub-specialties, have the ability to identify and embed within their patient registries, quality measures that are meaningful to their profession. For example, neurologists who primarily care for patients with multiple sclerosis had few if any measures for which to select in order to participate in CMS value-based reimbursement programs in any meaningful way. The American Academy of Neurology has recently launched the AXON Registry in 2016, with the goal of improving neurologic care and as a means to satisfy quality reporting requirements.[44] Other groups have followed a similar path; by 2019, over 125 registries were approved by CMS as QCDRs.[43]

The Evolving Role of Patient Registries

With increasing interest in RWD and RWE, a keen focus on establishing learning health systems and national research infrastructures, and the rapid digitization of healthcare, patient registries are poised to meet multiple needs for multiple stakeholders. However, to do so, patient registries must, in many cases, evolve from studies designed to meet a single purpose to reusable data infrastructures that fulfill multiple purposes.

An example of registry designed as a reusable data infrastructure is ArthritisPower. Creaky Joints, a patient advocacy group with over 80,000 members, launched ArthritisPower in partnership with University of Alabama. The registry is the first patient-led, patient -generated, patient-centered research registry for arthritis and other musculoskeletal conditions.[45] With support from PCORI, this initiative seeks to compare treatments, identify new treatments, and track long-term outcomes for patients with arthritis, while also providing a platform to develop robust patient-reported outcomes (e.g., measures of depression, fatigue, sleep disturbances). Another relevant examples is the National Institutes of Health All of Us project.[46] Embedded as part of the Federal government's Personalized Medicine Initiative, All of Us seeks to build a research cohort of over one million people. One key area of research will be the role of genetic factors in health outcomes.

As noted above, there is increasing interest in using RWE in regulatory decision making. Patient registries can serve as a platform to routinely capture more specific clinical patient information across more diverse patient population than RCTs traditionally allow. Data from patient registries may be used to build historic controls for future post hos analyses. Registry infrastructures can also serve as a platform for pragmatic clinical trials. Lastly, as discussed in the User's Guide,[4] registries remain an important tool to capture data throughout the product development lifecycle. Such registries can be product-oriented, disease-oriented, or focused on a particular patient population.

Next Steps

Despite this progress, several barriers must still be addressed before patient registries can fulfill their potential in the learning health system and national research infrastructure, as currently envisioned. Some of the key barriers include the interoperability challenges discussed above, as well as financial disincentives and concerns about patient privacy. Building a national infrastructure—and connecting to it—is expensive. Although financial incentives are changing, providers are still primarily paid on a traditional fee-for-service basis. As a result, providers are not rewarded financially for sharing information or participating in national infrastructure-building. In fact, they may even lose revenue if they reduce costs through better patient care management.

Active information blocking has also been cited as a barrier to interoperability.[47] The term 'information blocking' may be used to describe many types of activities, but generally refers to intentional attempts to interfere with the exchange or use of electronic health information. Numerous complaints about providers encountering prohibitive costs when attempting to move data from EHRs to other systems, including patient registries, have been documented. In 2015, ONC delivered a report to Congress on health information blocking that noted, "current economic and market conditions create business incentives for some persons and entities to exercise control over electronic health information in ways that unreasonably limit its availability and use. Indeed, complaints and other evidence described in this report suggest that some persons and entities are interfering with the exchange or use of electronic health information in ways that frustrate the goals of the HITECH Act and undermine broader healthcare reforms."[48] In response to concerns about information blocking, the 21st Century Cures Act provides the Office of the Inspector General (OIG) with new authority to investigate potential information blocking. One example that received widespread attention is the eClinicalWorks $155 million settlement with the OIG and Department of Justice in 2017 over allegations that eClinicalWorks' software failed to satisfy data portability requirements intended to permit healthcare providers to transfer patient data from eClinicalWorks' software to the software of other vendors."[49] Information blocking is also a prominent topic in the Notice of Proposed Rulemaking to Improve the Interoperability of Health Information, released by ONC in February 2019.[50]

Beyond technical challenges and financial disincentives, Federal and State laws may restrict how readily data may be shared between providers depending on the intended use of the data. Patients may also be reluctant to share their data. For example, close to one-third of the volunteers in the All of Us cohort did not sign the form that was needed to let researchers access their electronic health records. While full consideration of this topic is beyond the scope of this manuscript, it is critical to recognize that the basic ethical principles of respect for patient autonomy and non-maleficence, especially protection of patient privacy, are highly applicable to considerations around use of patient data.

Patients are the most important stakeholders and customers of learning health systems, although regrettably their input is not sought as often as it should be. Assessing patient attitudes around data capture and new frameworks of reuse like EHR-enabled disease registries and learning health systems can be informative. Jones et al interviewed 32 cancer patients from two distinct geographic locales regarding their perspectives on the ethical implementation of an oncology rapid-learning system. The patients expressed a range of opinions about health information privacy, with varying levels of permissiveness. While patient familiarity with EHRs and related technology influenced their comfort with specific learning health system features, their trust in the end users of the data – physicians and other members of the medical team versus pharmaceutical and insurance companies – was critical in determining their overall level of comfort. Additionally, most patients interviewed expressed a preference for a formal consent model, rather than an opt-out approach.[51] Addressing concerns will require further work as registries and the learning health system evolve.

To achieve the vision of a national health infrastructure, the cooperation of public and private entities will be necessary to overcome technical, institutional, and legal barriers. Recent public sector efforts have focused on reducing information blocking and addressing HIPAA-related concerns about data sharing. In the private sector, one notable effort is the Argonaut project, which aims to advance industry adoption of open interoperability standards, following the recommendations of several recent task force reports. In particular, this effort, which includes broad industry participation, has focused on developing the FHIR-based API and Core Data Services specification. These efforts are discussed further in Chapter 5.

As the nation focuses on using the vast amounts of electronic health data now available to build a learning health system and a national health data infrastructure, registries are likely to play a critical role. Registries will benefit from public or private sector activities that address existing barriers and make it easier access existing sources of electronic health information and share information across health systems and providers. Active participation by registry stewards and associations in these national conversations will help ensure that the needs of registries are met as the infrastructure is built.

References for Chapter 1

1. The Office of National Coordinator of Health Information Technology. U.S. Department of Health and Human Services. Health IT Dashboard. https://dashboard.healthit.gov/index.php. Accessed June 10, 2019.

2. Gliklich RE, Dreyer NA, Leavy MB, Christian JB (eds). 21st Century Patient Registries. EBook addendum to Registries for Evaluating Patient Outcomes: A User’s Guide, 3rd Edition. (Prepared by L&M Policy Research, LLC under Contract No. 290-2014- 00004-C.) AHRQ Publication No. 17(18)-EHC013-EF. Rockville, MD: Agency for Healthcare Research and Quality; February 2018. www.effectivehealthcare.ahrq.gov.

3. Piwek L, Ellis DA, Andrews S, et al. The Rise of Consumer Health Wearables: Promises and Barriers. PLoS Med. 2016;13(2):e1001953. PMID: 26836780. DOI: 10.1371/journal.pmed.1001953.

4. Gliklich R, Dreyer N, Leavy M, eds. Registries for Evaluating Patient Outcomes: A User’s Guide. Third edition. Two volumes. (Prepared by the Outcome DEcIDE Center [Outcome Sciences, Inc., a Quintiles company] under Contract No. 290 2005 00351 TO7.) AHRQ Publication No. 13(14)-EHC111. Rockville, MD: Agency for Healthcare Research and Quality. April 2014. http://www.effectivehealthcare.ahrq.gov.

5. Institute of Medicine. 2013. Best Care at Lower Cost: The Path to Continuously Learning Health Care in America. Washington, DC: The National Academies Press.

6. Institute of Medicine. 2011. Digital Infrastructure for the Learning Health System: The Foundation for Continuous Improvement in Health and Health Care: Workshop Series Summary. Washington, DC: The National Academies Press.

7. Institute of Medicine. 2015. Integrating research and practice: Health system leaders working toward high-value care: Workshop summary. Washington, DC: The National Academies Press.

8. Etheredge LM. A rapid-learning health system. Health Aff (Millwood). 2007;26(2):w107-18. PMID: 17259191. DOI: 10.1377/hlthaff.26.2.w107.

9. Etheredge LM. Rapid learning: a breakthrough agenda. Health Aff (Millwood). 2014;33(7):1155-62. PMID: 25006141. DOI: 10.1377/hlthaff.2014.0043.

10. Institute of Medicine. 2010. A Foundation for Evidence Driven Practice: A Rapid Learning System for Cancer Care: Workshop Summary. Washington, DC: The National Academies Press.

11. Miller RS, Wong JL. Using oncology real-world evidence for quality improvement and discovery: the case for ASCO's CancerLinQ. Future Oncol. 2018;14(1):5-8. PMID: 29052448. DOI: 10.2217/fon-2017-0521.

12. Connecting Health and Care for the Nation A Shared Nationwide Interoperability Roadmap. Office of the National Coordinator for Health Information Technology. 2014. https://www.healthit.gov/sites/default/files/hie-interoperability/nationwide-interoperability-roadmap-final-version-1.0.pdf. Accessed June 10, 2019.

13. The Office of the National Coordinator for Health Information Technology. U.S. Department of Health and Human Services. Improving the Health Records Request Process for Patients Insights from User Experience Research. https://www.healthit.gov/sites/default/files/onc_records-request-research-report_2017-06-01.pdf. Accessed June 10, 2019.

14. Maddox TM, Albert NM, Borden WB, et al. The Learning Healthcare System and Cardiovascular Care: A Scientific Statement From the American Heart Association. Circulation. 2017;135(14):e826-e57. PMID: 28254835. DOI: 10.1161/CIR.0000000000000480.

15. Harle CA, Lipori G, Hurley RW. Collecting, Integrating, and Disseminating Patient-Reported Outcomes for Research in a Learning Healthcare System. EGEMS (Wash DC). 2016;4(1):1240. PMID: 27563683. DOI: 10.13063/2327-9214.1240.

16. Franklin PD, Lewallen D, Bozic K, et al. Implementation of patient-reported outcome measures in U.S. Total joint replacement registries: rationale, status, and plans. J Bone Joint Surg Am. 2014;96 Suppl 1:104-9. PMID: 25520425. DOI: 10.2106/jbjs.N.00328.

17. Prodinger B, Taylor P. Improving quality of care through patient-reported outcome measures (PROMs): expert interviews using the NHS PROMs Programme and the Swedish quality registers for knee and hip arthroplasty as examples. BMC Health Serv Res. 2018;18(1):87. PMID: 29415714. DOI: 10.1186/s12913-018-2898-z.

18. PCORnet: The National Patient-Centered Clinical Research Network. Patient-Centered Outcomes Research Network. https://www.pcori.org/research-results/pcornet-national-patient-centered-clinical-research-network. Accessed June 10, 2019.

19. National Library of Medicine (NLM). SNOMED CT. 2017; https://www.snomed.org/. Accessed June 10, 2019.

20. National Library of Medicine (NLM). RxNorm. Unified Medical Language System (UMLS) https://www.nlm.nih.gov/research/umls/rxnorm/. Accessed June 10, 2019.

21. Regenstrief Institute. LOINC. https://loinc.org/. Accessed June 10, 2019.

22. Value Set Authority Center (VSAC). https://vsac.nlm.nih.gov/. Accessed June 10, 2019.

23. Sentinel Common Data Model. https://www.sentinelinitiative.org/sentinel/data/distributed-database-common-data-model/sentinel-common-data-model. Accessed June 10, 2019, 2018.

24. Observational Health Data Sciences and Informatics (OHDSI). https://www.ohdsi.org/. Accessed June 10, 2019.

25. JASON. A Robust Health Data Infrastructure. (Prepared by The MITRE Corporation under Contract No. JSR-13-700.) AHRQ Publication No. 14-0041-EF. Rockville, MD: Agency for Healthcare Research and Quality. April 2014. https://www.healthit.gov/sites/default/files/ptp13-700hhs_white.pdf.

26. The Office of the National Coordinator for Health Information Technology. U.S. Department of Health and Human Services. Understanding and Leveraging MU Stage 2 Optional Transports (SOAP). https://www.healthit.gov/sites/default/files/soapdeepdive.pdf. Accessed June 10, 2019.

27. Health Level Seven (HL7). RESTful API. https://www.hl7.org/fhir/. Accessed June 10, 2019.

28. Hillestad R, Bigelow JH, Chaudhry B, et al. IDENTITY CRISIS: An Examination of the Costs and Benefits of a Unique Patient Identifier for the U.S. Health Care System. RAND Corporation Monograph. October 2008, No. 753. http://www.rand.org/content/dam/rand/pubs/monographs/2008/RAND_MG753.pdf. Accessed June 10, 2019.

29. U.S. Food and Drug Administration. Framework for FDA's Real-World Evidence Program. https://www.fda.gov/media/120060/download. Accessed June 3, 2019.

30. Schilsky RL. Finding the Evidence in Real-World Evidence: Moving from Data to Information to Knowledge. J Am Coll Surg. 2017;224(1):1-7. PMID: 27989954. DOI: 10.1016/j.jamcollsurg.2016.10.025.

31. Berger ML, Sox H, Willke RJ, et al. Good Practices for Real-World Data Studies of Treatment and/or Comparative Effectiveness: Recommendations from the Joint ISPOR-ISPE Special Task Force on Real-World Evidence in Health Care Decision Making. Value Health. 2017;20(8):1003-8. PMID: 28964430. DOI: 10.1016/j.jval.2017.08.3019.

32. Berger M, Daniel G, Frank K., et al. A Framework for Regulatory Use of Real-World Evidence. The Margolis Center for Health Policy at Duke University website. https://healthpolicy.duke.edu/sites/default/files/atoms/files/rwe_white_paper_2017.09.06.pdf. Accessed June 10, 2019.

33. Visvanathan K, Levit LA, Raghavan D, et al. Untapped Potential of Observational Research to Inform Clinical Decision Making: American Society of Clinical Oncology Research Statement. J Clin Oncol. 2017;35(16):1845-54. PMID: 28358653. DOI: 10.1200/JCO.2017.72.6414.

34. National Academies of Sciences, Engineering, and Medicine. 2017. Real-world evidence generation and evaluation of therapeutics: Proceedings of a workshop. Washington, DC: The National Academies Press. doi: https://doi.org/10.17226/24685.

35. U.S. Food and Drug Administration (FDA). Medical Device Epidemiology Network Initiative (MDEpiNet). https://www.fda.gov/medical-devices/epidemiology-medical-devices/medical-device-epidemiology-network-initiative-mdepinet. Accessed June 10, 2019.

36. Centers for Medicare and Medicaid Services (CMS). MACRA: Delivery System Reform Medicare Payment Reform. 2017; https://www.cms.gov/medicare/quality-initiatives-patient-assessment-instruments/value-based-programs/macra-mips-and-apms/macra-mips-and-apms.html. Accessed June 11, 2019.

37. UnitedHealth Premium Program. https://www.uhc.com/health-and-wellness/take-control-of-your-care/choose-a-doctor/united-health-premium-program. Accessed June 11, 2019.

38. Blue Distinction Total Care. Blue Cross Blue Shield. https://www.bcbs.com/about-us/capabilities-initiatives/blue-distinction/blue-distinction-total-care. Accessed June 11, 2019.

39. Blue Distinction Specialty Care. Blue Cross Blue Shield. https://www.bcbs.com/about-us/capabilities-initiatives/blue-distinction/blue-distinction-specialty-care. Accessed June 11, 2019.

40. Damberg CL, Sorbero ME, Lovejoy SL, et al. Measuring Success in Health Care Value-Based Purchasing Programs: Findings from an Environmental Scan, Literature Review, and Expert Panel Discussions. Santa Monica, CA: RAND Corporation, 2014. https://www.rand.org/pubs/research_reports/RR306.html. Accessed June 11, 2019.

41. Muhlbacher AC, Juhnke C. Patient preferences versus physicians' judgement: does it make a difference in healthcare decision making? Appl Health Econ Health Policy. 2013;11(3):163-80. PMID: 23529716. DOI: 10.1007/s40258-013-0023-3.

42. Anderson HV, Shaw RE, Brindis RG, et al. A contemporary overview of percutaneous coronary interventions. The American College of Cardiology-National Cardiovascular Data Registry (ACC-NCDR). Journal of the American College of Cardiology. 2002;39(7):1096-103. PMID: 11923031.

43. Quality Payment Program Resource Library. Centers for Medicare & Medicaid Services. https://qpp.cms.gov/about/resource-library. Accessed June 10, 2019.

44. Sigsbee B, Goldenberg JN, Bever CT, Jr., et al. Introducing the Axon Registry: An opportunity to improve quality of neurologic care. Neurology. 2016;87(21):2254-8. PMID: 27694258. DOI: 10.1212/WNL.0000000000003264.

45. About ArthritisPower. CreakyJoints. https://creakyjoints.org/research/arthritispower/. Accessed June 10, 2019.

46. All of Us. National Institutes of Health. https://allofus.nih.gov/. Accessed June 11, 2019.

47. Savage LC. To Combat ‘Information Blocking,’ Look To HIPAA. Health Affairs Blog. August 24, 2017. https://www.healthaffairs.org/do/10.1377/hblog20170824.061636/full/. Accessed June 11, 2019.

48. The Office of the National Coordinator of Health Information Technology. U.S. Department of Health and Human Services. Report on Health Information Blocking. Report to Congress. April 2015. https://www.healthit.gov/sites/default/files/reports/info_blocking_040915.pdf. Accessed June 11, 2019.

49. U.S. Department of Justice. Office of Public Affairs. Electronic Health Records Vendor to Pay $155 Million to Settle False Claims Act Allegations. https://www.justice.gov/opa/pr/electronic-health-records-vendor-pay-155-million-settle-false-claims-act-allegations. Accessed June 10, 2019.

50. The Office of the National Coordinator of Health Information Technology. U.S. Department of Health and Human Services. Notice of Proposed Rulemaking to Improve the Interoperability of Health Information. https://www.healthit.gov/topic/laws-regulation-and-policy/notice-proposed-rulemaking-improve-interoperability-health. Accessed June 10, 2019.

51. Jones RD, Sabolch AN, Aakhus E, et al. Patient Perspectives on the Ethical Implementation of a Rapid Learning System for Oncology Care. J Oncol Pract. 2017;13(3):e163-e75. PMID: 28118107. DOI: 10.1200/JOP.2016.016782.

Chapter 2. Data Sources

Authors (alphabetical)

Michelle B. Leavy, M.P.H.
Head, Healthcare Research & Policy
OM1, Inc.

Anna Swenson, M.P.H.
Epidemiologist
OM1, Inc.

Introduction

Electronic health data that are relevant for registries may come from a wide variety of sources, including electronic health records (EHRs), administrative claims databases, laboratory systems, imaging systems, medical devices, and consumer devices. A 2017 survey of patient registries in the United States found that 68 percent of registries extract some data from electronic health records (EHRs), and 35 percent capture some data from other electronic data sources. While use of data from electronic data sources has grown, most registries (88 percent) still use manual data capture for at least some data.[1]

Integrating data sources with patient registries can take many forms, depending on the type(s) of data and the purpose and architecture of the registry. In some cases, registries may work directly with individual systems to integrate or link to data, while, in other cases, registries may work with sources in which the data have already been aggregated and standardized, such as clinical data warehouses and health information exchanges. The purpose of this chapter is to describe several common sources of data that may be incorporated into a patient registry and discuss the strengths and limitations of these data. Chapters 4 and 5 describe the technical approaches that may be used to incorporate these data into a patient registry, and key questions to consider when planning to incorporate data from another source are summarized in Appendix B.

When selecting data sources, registries should consider the registry purpose and the suitability of the potential data source – in terms of scope, data quality, and timeliness – for addressing that purpose. Chapter 6 of the User’s Guide provides more information on selecting data sources for use in a registry.

In addition to technical and scientific considerations, registries must pay careful attention to issues of patient privacy, informed consent, and data ownership when incorporating data from multiple sources. Registries should understand, at minimum, the purpose for which the data were collected originally (e.g., treatment, payment or healthcare operations as defined by the Health

Insurance Portability and Accountability Act Privacy Rule; for research purposes with documented individual consent; for research purposes with an Institutional Review Board (IRB) waiver of consent); the type of data contained in the data source (e.g., protected health information [PHI], sensitive information such as information about mental health conditions or infectious diseases); and who owns the data. More information on the legal and ethical framework under which data may be shared across systems in the United States can be found in Chapters 7 and 8 of the User's Guide.

Lastly, it is important to note that the following discussion focuses on sources that may contribute data *to* a registry. This discussion does not cover the issue of when or how registries should report data back to these other sources. For example, EHR data may be sent to a registry, but registry data (such as patient-reported outcome measures or data obtained from other providers for registry purposes specifically) are typically not sent back to the EHR. Many questions exist about the appropriateness and feasibility of creating these types of continuous exchanges of data, and these issues are beyond the scope of this document.

Electronic Health Records (EHRs)

An EHR is "a longitudinal electronic record of patient health information generated by one or more encounters in any care delivery setting."[2] EHRs include information on patient demographics, progress notes, problem lists, medications, vital signs, past medical history, immunizations, laboratory data, and radiology reports. While much of this information is extremely valuable for patient registries, EHRs are designed primarily support clinical care (as opposed to research). Patient registries may leverage data contained in EHRs by *integrating* with the EHR to allow for real-time or nearly real-time data exchange or by *linking* with the EHR to allow for periodic transfers of data into the registry. The decision of whether and how to incorporate data from EHR is complex and should be guided by many factors, including the purpose and scope of the registry and the availability of the necessary data elements within an EHR. These considerations are discussed in detail in Chapter 4 (Obtaining Data from EHRs).

Claims Data

Public and private medical insurers collect a wide range of data as part of evaluating coverage, tracking health services utilization, and managing billing and payment. These data, commonly referred to as 'claims data,' contain patient-specific information such as demographics, insurance coverage and copayments, healthcare provider data (e.g., specialty characteristics, locations), and treatment details such as procedures, office visits, and hospitalizations. Pharmacy claims data provide specific information on the dispensing of pharmaceutical products. Standard coding systems are used to record diagnoses, procedures, and other data; these include Current Procedure Terminology (CPT) for physician services and International Classification of Diseases (ICD) for diagnoses and hospital inpatient procedures.[3] Similarly for pharmacy claims, standard medication coding systems, such as National Drug Classification (NDC) codes, are used.

Medicare and Medicaid claims files are commonly used administrative databases in the United States. Together, the programs cover nearly 133 million people in the United States. The Medicare program covers some 59 million individuals ages 65 and older, as well as younger individuals with end-stage renal disease or who qualify for Social Security Disability.[4] Medicaid and Children's Health Insurance Program (CHIP) together cover an additional 73.8 million individuals.[5] Both programs are administered by the Centers for Medicare and Medicaid Services (CMS). Claim files for these programs can be obtained for inpatient, outpatient, physician, skilled nursing facility, durable medical equipment, hospital services, and prescription drugs. These data, which are subject to privacy rules and regulations, can be linked to other databases with appropriate permissions. The Research Data Assistance Center (ResDAC) is a CMS contractor that supports researchers interested in using Medicare and/or Medicaid data for research purposes.[6]

While Medicare and Medicaid data files are tremendously valuable for some research purposes, they are restricted to patients who are eligible for these programs. A limited number of other data sources are available at the federal and state level.[7] One such example is the Healthcare Cost and Utilization Project (HCUP) databases managed by the Agency for Healthcare Research and Quality (AHRQ). HCUP databases contain encounter-level data from all payers, dating back to 1988. The databases use a uniform format to provide longitudinal information that can be used to support research on cost and quality of health services, practice patterns, access to care, and outcomes of treatment.[8] It is important to note that many of the databases contain a sample of data, as opposed to all data. For example, the National Inpatient Sample (NIS) contains a 20% stratified systemic random sample of all discharges. In addition, linkage of these data to registry data at the individual patient level is not feasible generally; however, these data can provide useful information to inform the design of a registry and provide context for the findings of a registry. Databases available under the HCUP program are summarized in Table 2-1.[9]

Table 2-1. Databases available through the HCUP program

DATABASE NAME	DESCRIPTION
National Inpatient Sample (NIS)	Inpatient care database, with more than 7 million hospital stays each year
Kids' Inpatient Database (KID)	Pediatric inpatient care database, with approximately 3 million hospital stays each year
Nationwide Emergency Department Sample (NEDS)	National database of emergency department visits
Nationwide Readmissions Database (NRD)	Discharges for patients with and without repeat hospital visits in a year and those who have died in the hospital
State Inpatient Databases (SID)	State-specific files with all inpatient care records in participating states
State Ambulatory Surgery and Services Databases (SASD)	State-specific files with data for ambulatory surgery and other outpatient services from hospital-owned facilities (some states provide services from nonhospital-owned facilities as well)
State Emergency Department Databases (SEDD)	State-specific emergency department (ED) databases with discharge information on all emergency department visits that do not result in an admission

More recently, some efforts have focused on creating all-payer claims databases (APCDs) at the state level that can be used to produce price, resource use, and quality information for consumers. APCDs compile medical claims, pharmacy claims, dental claims, and eligibility and provider files from private and public payers, providing a more comprehensive look at healthcare services provided within a state. To date, 18 states have mandated the creation and use of APCDs; the actual implementation and use of these systems varies, as do policies for data access for research purposes.[10]

In the private sector, some companies have compiled data from private insurers and in some cases combined these data with other sources (e.g., data from EHRs). This is an area of rapid growth. While these databases may be useful in the context of patient registries, their applicability and limitations vary widely depending on the size and scope of the database and the research question(s) of interest.[7] A full review of these private sector companies is beyond the scope of this document.

Strengths & Limitations of Claims Data

In the context of patient registries, claims data offer a relatively quick way to access large volumes of health information. Use of claims data is typically less expensive and faster than longitudinal data collection directly from providers or patients.[11] Claims data may also fill gaps in data from other sources. For example, an EHR may capture detailed clinical data on a patient undergoing total joint replacement surgery, including patient characteristics and any immediate post-surgery complications. Claims data may provide information on followup care, such as physical therapy or issues that emerged later (e.g., revision surgery). Claims data are also useful for monitoring practice patterns or disease prevalence at a national or regional level, since these data often cover a wider geographic area than EHR data. It is important to distinguish between claims that are submitted by healthcare providers and the amount paid by health insurers. Health insurers have made substantial investments in claims "scrubbing" activities in which all claims are reviewed for accuracy. Aside from changes in what is paid compared to what was submitted, claims that have undergone the adjudication process employed by health insurers are considered to be more reliable than submitted claims, which have not likely undergone the same degree of curation.

Claims data have several limitations that should be considered before using these data in registry-based studies. First, claims databases are limited to individuals who were insured by a specific program (e.g., Medicare, a private payer plan). Uninsured patients are not included in these databases. Depending on the research question, the population included in a claims database may or may not be generalizable to the target population. While patients are tracked longitudinally in private payer claims data, they only remain in the dataset while they are covered by the same plan; patients typically become lost to followup when they change plans. This can limit the ability to track long-term outcomes through private payer claims data. In addition, claims databases only record billable events that are covered by the individual's plan. For example, prescriptions that were given to a patient but not filled by the patient are not included in claims databases; similarly, claims databases do not include claims for prescriptions that were dispensed but not a covered benefit. Treatments sought outside of covered settings (e.g., by a non-covered provider, alternative treatments) are also not included. Insurance plans can vary widely with respect to the services or drugs that are covered, and patients with different plans typically have different deductibles or copays. These factors can influence treatment patterns and make comparisons across plans challenging.

In addition to issues related to the scope of the data, some questions have been raised about the accuracy of claims data compared to medical records data. For example, a 2013 study examined the agreement between administrative claims data and the medical records for 13 commonly reported comorbidities and complications in patients undergoing total joint arthroplasty. The study found that the specificity of administrative claims data is generally high (greater than 92 percent for many outcomes), but the sensitivity is variable and often lower (ranging from 29 to 100 percent). The authors concluded that comorbidities and complications coded in the administrative record were highly accurate but often incomplete.[12] Data quality issues in claims data may occur due to clerical errors, different interpretations of healthcare documentation, or errors resulting from lack of education when codes change (e.g., annual updates of CPT codes, switch from ICD-9 to ICD-10).[13]

Linkages of Claims Data With Research Studies

Claims data may facilitate registry-based research in several ways. First, claims data may be useful in the registry planning phase. Claims data can provide information on treatment patterns, such as the frequency of a specific procedure, that can be helpful when planning enrollment and targeting recruitment efforts. Once registries are in the operational phase, data for all registry patients or for a subset may be linked with claims data to address a specific research question. For example, a recent study linked data from the Transcatheter Valve Therapies Registry to Medicare claims data to examine the prevalence of death, stroke, heart failure-related hospitalization, and mitral valve intervention at one year post transcatheter aortic valve replacement.[14] In another example, data from the CORRONA registry were linked to Medicare data to examine the economic savings associated with remission among rheumatoid arthritis patients.[15] Registries have also been linked with claims data to assess the generalizability of the registry population.[16] Linked datasets can be valuable tools for research; one of the largest linkages of disease registry and claims data is the SEER-Medicare dataset, which has supported a wide range of cancer-related research projects and resulted in over 1,700 publications.[17] Linkage with claims data is more difficult for registries that do not focus on the Medicare-eligible population; participants in these registries are often covered by a variety of payers, and linkage of data from multiple payers is rarely feasible.

Multiple approaches for linking registry data to claims data are available; the technical and legal aspects of these approaches are explored in the “Linking Registry Data With Other Data Sources To Support New Studies” chapter of the third edition of the User’s Guide.

Patient-Generated Health Data

In recent years, there has been increasing interest in incorporating patient-generated health data into patient registries, EHRs, and other data collection efforts. Patient-generated health data (PGHD) are defined as ‘health-related data created, recorded, or gathered by or from patients (or family members or other caregivers) to help address a health concern.’[18] PGHD may include information on the patient’s health history, treatment history, symptoms, biometric data (e.g., blood glucose reading), and lifestyle choices (activity level tracked using a wearable device). These data differ from data captured in clinical settings in two ways: patients are responsible for recording these data, and patients decide if and how to share these data with healthcare providers.

The availability of PGHD has expanded as consumers increasingly use smartphones, mobile apps, remote monitoring devices, and wearable devices that are capable of capturing health data.[19] For example, apps and wearable devices are available to track fitness,[20] sleep,[21, 22] heart rate and rhythm,[23, 24] blood pressure,[25-27] blood glucose,[28, 29] and oxygen saturation.[30] A 2016 report found more than 259,000 mobile health apps available in app stores such as Apple App Store and Google Play.[31] In addition, the growth in provider usage of EHRs and the introduction of patient portals have created new tools to connect patients and providers and to integrate PGHD into clinical care. Some devices have even received approval from the U.S. Food and Drug Administration for use in clinical workflows.[32]

Strengths of PGHD

PGHD are valuable to providers, researchers, and other stakeholders for several reasons. First, these data may supplement data collected during clinical encounters with more frequent measurements of health status, providing clinicians with a better overall picture of the patient's health status. Patients may also benefit from improved understanding of their health; for example, heart failure patients in the Connected Cardiac Care Program at Partners HealthCare reported learning more about their condition and feeling more in control of their health after regularly monitoring and sharing data on their weight, heart rate, pulse, and blood pressure.[33] Ideally, frequent monitoring could lead to timely interventions that prevent more significant complications, such as a change in prescription to reduce the likelihood of an asthma exacerbation.[34] Some evidence suggests that some PGHD, particularly from sensor data, may be more reliable than data collected in the clinic, since measures like a 6-minute walk test can be estimated through sensors that are free of the influence of healthcare professionals who may coach some patients differently than others.

Researchers are interested in using PGHD to capture important information outside of regularly scheduled visits with a provider and to follow patients over time, particularly when patients change providers or no longer need to return for followup visits (e.g., post-surgery). In addition, the ability to capture PGHD may enable researchers to recruit from a larger pool of patients efficiently, rather than relying on traditional site-based enrollment models. In fact, a recent review on the potential value of PGHD for comparative effectiveness research concluded that 'leveraging the emerging wealth of big data being generated by patient-facing technologies such as systems to collect patient-reported outcomes data and patient-worn sensors is critical to developing the evidence base that informs decisions made by patients, providers, and policy makers in pursuit of high-value medical care.'[35]

There is also a growing body of literature on the validity and utility of PGHD for pharmacovigilance. For example, in the European PROTECT (Pharmacoepidemiological Research on Outcomes of Therapeutics) Consortium, funded by the Innovative Medicines Initiative, data were collected directly from pregnant women recruited on-line from the United

Kingdom, Denmark, The Netherlands and Poland. The PROTECT study examined medication use during pregnancy using bi-weekly or monthly questionnaires administered via the Internet, with the frequency of followup determined according to the participants choosing.[36] The study compared patient-reported medication use for Danish patients with data from the Danish national prescription register and showed reasonably strong agreement; moreover, the PGHD also provided rich information about non-prescription medications (and recreational drug use) not available through other sources. It should be noted that patients consented to data linkage and provided their national identity number. The actual data linkage was accomplished through use of a trusted third party; similar approaches are being used in the United States.

Several efforts have been launched in recent years to support the use of PGHD in clinical care and research. At the Federal level, the Office of the National Coordinator (ONC) launched a project in 2015 to identify best practices, gaps, and opportunities for use of PGHD; project findings include a report on PGHD intended to inform future policy work in this area, two pilot demonstrations, and a practical guide.[31] These findings are intended to support the long-term implementation of the PGHD requirements included in the Federal Health IT Strategic Plan, the ONC Interoperability Roadmap, the 2015 Certification Rule, Stage 3 of the CMS Meaningful Use Rule, the CMS Quality Payment Program, and the Precision Medicine Initiative at the National Institutes of Health (NIH). NIH's All of Us Research Program, under the Precision Medicine Initiative, aims to collect data including PGHD from at least one million U.S. participants.[37] ONC also recently updated its Patient Engagement Playbook to include strategies for integrating PGHD into clinical care.

In the private sector, Apple released HealthKit, a common framework to support sharing of PGHD among apps, services, and providers, in 2014. The related ResearchKit was released in 2015 to provide researchers with an open source framework to build apps to support smartphone-based research. ResearchKit enables researchers to use the iPhone's sensors as well as third-party devices to monitor health variables captured in HealthKit and share those data with researchers and EHRs. In 2018, the American Medical Association and Google co-sponsored an innovation challenge aimed at improving interoperability and developing new methods of collecting and managing PGHD.[38] The Patient-Centered Outcomes Research Institute (PCORI) has also devoted funding to building a sustainable foundation to support the use of PGHD in patient-centered outcomes research.[39] More information on how to use ResearchKit and other similar tools to capture PGHD for use within a registry can be found in Chapter 5.

Limitations of PGHD

While interest in PGHD has increased, some barriers to the routine use of PGHD in research and clinical care still exist. First, as noted in the Duke-Margolis Center for Health Policy's mHealth action plan, 'interoperability as well as common data elements (and tightly bound self -defining metadata) and definitions will be critical, as disparate data streams will increasingly need to be combined to create actionable insights for maintaining an individual 's health and treating disease.'[40] Currently, PGHD sources differ in terms of what is measured, how it is measured, how data are structured, and how data may be transferred to other systems. Second, guidance on how to determine if a PGHD source is 'fit for purpose' would be useful. For example, some research has raised questions about the accuracy of some devices compared to other sources of information;[41] these concerns need to be considered in the context of the study objectives and measures of interest. In addition, multiple types of devices are available for many areas (e.g., FitBit, Jawbone for activity tracking), and it is unclear if these devices operate in the same manner and if data from these devices are interchangeable.[42] There is also little guidance on how to transform data from continuous monitoring, often over long periods, into clinically meaningful endpoints.

On the patient side, some patients may not have access to the necessary technology (e.g., a smartphone, remote monitoring devices) to generate and share PGHD. Even patients with access to the technology may be unwilling to complete the steps required to capture and share data; for example, in a pilot study involving patients with asthma, patients needed to complete a setup process that included installing and activating the MyChart app, installing and consenting to the Asthma Health app, and permitting data sharing.[34] Patients also may not recognize or understand the potential value of recording these data, or they may be reluctant to share the data with providers because of privacy concerns. Patients and other users of the data may also have different views on ownership of PGHD.[43] At the provider level, workflow changes and analytic tools may be necessary to incorporate review of PGHD and appropriate outreach to patients with concerning data. Providers may also have concerns about the accuracy and validity of PGHD from various devices and about setting realistic patient expectations for how these data are used in clinical decision making. For example, providers may be concerned about potential liability if they do not act promptly on urgent information provided through PGHD channels or if they do act on inaccurate PGHD.

Researchers also face challenges when attempting to use PGHD in the context of a clinical study. As noted above, questions about the validity of the data exist, and researchers who enroll patients remotely may have difficulty verifying participant eligibility. Once patients are enrolled, researchers must trust that the submitted PGHD were generated by the enrolled participant (and not, for example, by a family member who borrowed a device). Researchers must also address the selection bias inherent in studies that require use of a specific technology. From an ethical standpoint, researchers may encounter difficulties with Institutional Review Board (IRB) approval and the informed consent process (e.g., with regard to the data collection and privacy practices of third party developers of apps or devices), although this is a rapidly changing area.[43] The ONC report on PGHD noted that “the security and privacy protections that apply to PGHD are uneven and do not establish a consistent legal and regulatory framework.”[31] Lastly, researchers planning to collect data over a long period must address issues related to technological change and device abandonment. The PGHD landscape is changing constantly with the rapid introduction of new devices and apps and the disappearance of others, making it possible that researchers will need to modify the study protocol to accommodate these changes. Patients may also lose interest in using a device over time and stop tracking data or submitting data to the study.

Further research is needed to support the efficient and effective use of PGHD in clinical practice and research. Specifically, research is needed to identify best practices for incorporating PGHD into research studies and ideally into the patient’s EHR to inform clinical decision making. More research is also needed to understand patients’ views about sharing PGHD with providers and researchers and to address their concerns. On the technical side, standardization of common

PGHD measurements could increase the reliability and validity of these data if uniformly applied. In particular, standardized measures that could be captured through patient devices as well as in the clinical setting would increase the utility of these data for research and clinical practice. These standards could be based on existing, patient-centered standardized outcome measures, such as those developed through the AHRQ-funded Outcome Measures Framework (OMF) project (Chapter 3).

Genomic Data

Genomic data originates from an individual's DNA and may refer to both the information from genetic tests (e.g., genetic markers) as well as the actual biospecimens. Due to recent advances in genomic technology, sequencing and analysis of biospecimens has produced large amounts of genomic data that could be linked to clinical data to help diagnose diseases, identify risk factors for diseases, and monitor responses to treatment. There is significant interest in using genomic data in clinical care and in research.

In clinical care, genomic data forms the foundation for precision medicine efforts. Precision medicine refers to the use of genomic and other data to guide the selection of the appropriate drug and dosage for an individual patient. The concept has received much attention in recent years, particularly with the creation of the NIH's Precision Medicine Initiative in 2015, the passage of the 21st Century Cures Act in 2016, and the launch of NIH's All About Us research study in 2017.[37] While there is still much work to be done before precision medicine becomes broadly useful in clinical care, the practice of using genetic testing to guide treatment is already common in some areas. For example, in lung cancer, genetic testing is done to detect molecular biomarkers such as EGFR that guide treatment choices. Biomarkers also play a critical role in guiding treatment decisions for patients with invasive breast cancer.[44] Beyond oncology, genomic data are used for many purposes, such as diagnosing rare diseases and detecting chromosomal abnormalities of the fetus during pregnancy.

The interest in genomic data and precision medicine has led to substantial investments in research examining how to use genomic data across a wide range of condition areas.[45] In addition to individual research studies, several efforts have focused on creating biorepositories or biobanks to store biosamples for use in future research. One of the largest repositories of genomic data in the United States is the National Cancer Institute's Genomic Data Commons (GDC). Genomic data generated from cancer research studies are available through the GDC for re-use in new research projects, subject to controlled access terms to protect patient privacy.[46] Similarly, the RD-Connect project links genomic and phenotypic data to patient registries and other clinical databases, with the goal of streamlining multi-national rare disease research efforts.[47]

Patient registries may collect genomic data to address many research questions. For example, the American Association for Cancer Research (AACR) launched a registry in 2015 to capture genomic sequencing data from patients with late-stage cancers and link these data to clinical outcomes. The data are aggregated and analyzed to identify possible ways to improve treatment decisions and patient outcomes.[48] In another example, the Muscular Dystrophy Association (MDA) has launched the NeuroMuscular ObserVational Research (MOVR) data hub, with the goal of capturing and linking genomic data with clinical data at the national level to support research for four rare diseases: amyotrophic lateral sclerosis, spinal muscular atrophy, Duchenne muscular dystrophy, and Becker muscular dystrophy.[49] In this registry, clinical data are captured during routine care and linked with other data, such as genomic data and patient-reported outcomes.

While the use of these data has substantial potential, many barriers remain. Because genomic data contains highly sensitive information, some individuals may be unwilling to provide biosamples for research purposes.[50] Many investigators are unwilling to share genetic testing results with patients because there is no clear impact on clinical decision making; this reduces the attractiveness of study participation for patients, who may wish to acquire this information in hopes of future benefits. Concerns have also been raised about the ethical implications of genetic testing, whether information should be shared with family members who may have the same genetic risk factor, and the possibility for identification of genetic mutations unrelated to the patient's current treatment decision.[51, 52] Patient registries that intend to incorporate genomic data must consider these issues during the registry planning phase.

From an interoperability perspective, patient registries typically capture the results of genetic testing (e.g., presence of a specific mutation) within the registry dataset and, in some cases, link to a biorepository containing biosamples and more complete genomic data (see 'Biorepositories and Registries' white paper). However, as genome sequencing becomes more widespread and as the ability to store large amounts of data increases, registries may wish to store the results from both array-based sequencing and next generation sequencing, as well as new types of genomic data. Variant Call Format (VCF) files contain information only about specific genomic locations that differ whereas genomic Variant Call Format (gVCF) files contain all assayed nucleotide positions, regardless of whether they are variant.

Radiological Image Data

Imaging data include x rays, magnetic resonance imaging (MRI) scans, ultrasounds, computed tomography (CT) scans, and positron emission tomography (PET) scans that may be used for diagnosis and monitoring purposes. Increasingly, medical imaging plays an important role in guiding treatment decisions, and patient registries may wish to capture these data to support specific research objectives. When considering imaging data, it is important to distinguish between the interpretation or findings from the imaging study and the images themselves. Many registries currently store the findings from imaging studies (e.g., tumor location and size, degree of vertebral slip in lumbar spondylolisthesis). However, in some cases, registries may be interested in storing or linking to the images, as opposed to storing only the interpretation of the image. Access to the original images may be important to confirm a diagnosis, adjudicate study outcomes, or support new research questions that emerge over the course of the registry. In addition, interest in using machine learning and artificial intelligence methods to read medical images is increasing, and registries that link rich clinical data with images could be important resources as training and validation datasets.[53]

While interest in storing images is increasing, many challenges still exist. First, different imaging technology can result in incompatible imaging files, even within one healthcare setting. This issue becomes even more complicated when attempting to include images in a patient registry that captures data from multiple healthcare settings. Digital Imaging and Communications in Medicine (DICOM) is the current standard for image file format and communication profile for many types of images. This standard provides a format for metadata describing the patient, exam, and other image details, which should facilitate data exchange and interoperability. However, some researchers have noted that many fields are entered incorrectly or left blank, creating complex issues when merging datasets. Linking images from different databases can also be challenging in the absence of a master patient identifier, and direct inclusion of images in registry databases (as opposed to linkages) increases the registry data storage requirements. Further work is needed to explore how best to link or import image files into patient registries.

Clinical Data Warehouses

Clinical data warehouses (CDWs) are used for a variety of clinical, research, and administrative purposes. A CDW is a database or repository containing clinical data from a variety of sources that are standardized for use in analysis and reporting. A widely used definition of a CDW is a "subject-oriented, integrated, time-variant collection of data to support decision making."[54] Other terminology that are used to refer to CDWs are: enterprise data warehouse, medical data warehouse, biomedical data warehouse, biomedical information warehouse, healthcare data warehouse, and clinical data repository. It is important to note that the terms "clinical data warehouse" and "clinical data repository" are often used interchangeably, but they may have specific, distinct meanings within an institution.

CDWs are developed to organize and standardize data that exist in separate silos within or across organizations, enabling analysis and reporting both from a feasibility and efficiency standpoint. Within an organization, data from billing systems, registries, EHRs, pharmacy systems and laboratory systems often reside in different places. When these data are loaded into a common CDW, they can be linked together at the patient level and used in tandem to answer questions that could not be addressed within each individual data silo. For example, prescription fill data from pharmacies may be used in concert with EHR medication orders to examine patient medication use and adherence.

CDWs are designed to contain complex and heterogenous data. Ideally, CDWs have a flexible schema model that allows for the addition of new data sources and types of data at any point in the lifecycle of the CDW.[55] Most CDWs contain administrative data, such as billing data, as well as clinical data. Clinical data may come from inpatient or hospital EHR systems, disease or quality improvement registries, laboratories, pharmacies, and imaging centers. A common, unique patient identifier is required to link these disparate data sources together in the CDW. If the input data sources use different patient identifiers, a Master Patient Index must be maintained in the CDW as well.

Both structured and unstructured data may be included in the CDW. Natural language processing and other data mining techniques may be used to extract or manipulate data for inclusion in a CDW. Some examples of different types of data are provided below:

- *Pathology*: Pathology data may arrive in a report from an outside institution that is scanned and attached to a patient record in the EHR. Important information such as description of pathologic findings in tumor specimen and pathologic staging information must be extracted from these reports into a discrete element for integration into the warehouse. Various technologies are being used to accomplish this.[56]
- *Medical Imaging*: As discussed above, medical imaging data are large and complex. Special planning is required to provide the end user of a CDW with access to the image (often through a URL to a web based picture archive) while maintaining patient privacy.
- *Genomics*: As discussed above, storing genomic data, such as gVCF data, can require a huge amount of database space and may result in slower query run-time, so the end use for the data must be carefully considered when selecting the data to include in the CDW.[57] Consideration may be given to mapping variants to Human Genome Organization gene names and indexes.[56] As whole genome sequencing decreases in cost and becomes more widely available, CDW storage issues regarding the size of these data will need to be addressed.

In addition to allowing linkage of data from different sources at the patient level, CDWs are used to standardize data elements for ease of analysis. This may include ensuring that data elements are in a common format (e.g. diagnosis code from EHR in format 270.10 vs. code from claims data in format 27010) or mapping data elements to a standard terminology/ontology (e.g., ICD-10, LOINC, SNOMED). Standardization can be extremely time consuming and resource intensive, and the extent to which data are standardized within a CDW depends upon both the intended use cases for the data as well as the available resources within an organization.

In addition to use within a single organization, CDWs may be used to provide a central repository for data from multiple organizations to facilitate shared analysis and reporting.[58] CDWs that incorporate data from multiple organizations are typically organized in one of three ways. First, sites may upload their data directly into a centralized CDW that integrates and stores all of the study/registry specific data from all participating sites.[59] This model allows for efficient, centralized analyses, but resources are required to maintain a central database. This model may also trigger concerns about patient privacy, security, and data access. Alternately, individual sites may each maintain their own CDW, often with a CDM that is utilized by all participating sites. Analyses may be run at each site using shared code, since the underlying data architecture of each CDW is the same. This is known as a federated or distributed research network.[60] Lastly, individual sites may maintain their own CDW, but use a centralized server to store information for data exchange, such as the data model, controlled terminologies, and other metadata.[61]

Once implemented, CDWs support a variety of objectives. Some CDWs are enterprise wide and provide broad data services to the entire organization.[62-64] Others are narrower in scope and may exist only to meet the needs of a specific group within an organization. They may be used to generate ad hoc queries from researchers or clinicians, to run automated reports, or to identify patient populations of interest. Several examples of how CDWs are used in practice are provided below:

- *Clinical care:* Data from a CDW can be used to provide actionable feedback to clinicians. For example, the Intermountain CDW integrates data from 22 hospitals and 179 outpatient facilities that are part of the Intermountain health system. The CDW is updated daily with data from the EHR and automated reports run that identify patients who have new positive MRSA cultures and notify infection specialists to prevent transmission in the hospital or office setting.[62]
- *Precision medicine for improving treatment for individual patients:* Rutgers Cancer Institute created a CDW with a focus on integrating all data sources of importance in the treatment of an individual, including pathology, exon sequencing, radiology images and other data types which are often difficult to store and access.[56] The availability of all data points of interest in one warehouse enables clinicians to access to the full array of data needed to tailor treatment for an individual and provides a rich resource for clinical studies.
- *Research:* CDWs are used to identify patients for recruitment for clinical trials or observational research studies. The availability of diverse data elements can be used to design a targeted and efficient search strategy for appropriate patients.[65] In addition, linked data in the CDW can be used to conduct retrospective studies of populations of interest.
- *Safety reporting:* CDWs are used to identify adverse drug reactions[66] or hospital adverse events.[67]
- *Machine learning and artificial intelligence*: Machine learning algorithms use large volumes of complex data to make predictions. The data available in CDW are particularly suitable for machine learning.

Patient registries interact with CDWs in a variety of ways. The data collected by a registry may be uploaded into a CDW and then linked to additional data sources for analysis. Data supplied by other systems within the CDW may be used to validate the data reported in a registry. For example, pharmacy fill data may be used to validate reported medication use. A CDW can be used to generate or enrich a registry population and may be used to feed data in to a traditional registry electronic data capture (EDC) system from EHR, laboratory, or other data tables in the CDW.[68] Automatically feeding clinical data into a registry can reduce the time required to enter data and ensure timely availability of data elements, such as laboratory test results, within the registry. However, the registry should carefully consider the impact of automatic data feeds on registry data quality.

CDWs are powerful tools for integrating disparate data sources into a single data repository, but the design of the warehouse schema and the data standardization rules must be carefully planned, both for the initial use cases and to accommodate future use cases that may arise. While it may be desirable to have all data cleaned and mapped, doing so requires a great deal of resources on an on-going basis, and choices must be made as to what is the most efficient model for the specific CDW. As with any existing data source, the data in the CDW will only be as good as the data in the source files. Incomplete or incorrect data in the source files will be replicated in the CDW. Methods to cross-validate and error check are needed but will not be able to account for all data issues.

The patient linkage inherent in CDWs also raises concerns about patient privacy. CDWs must have the necessary access and security controls to ensure appropriate access to patient-level data. In addition, de-identified data that are transferred to the CDW may become identifiable when combined with other data in the warehouse. Data access rules and honest brokers may be necessary to protect patient privacy in these circumstances.

Health Information Exchanges

Electronic Health Information Exchange (HIE) refers to the electronic transfer of patient health information between healthcare providers. HIEs address interoperability issues and enable the bi-directional exchange of data either through a centralized data repository or through a federated network of sites. Although primarily conceptualized a means to improve the quality of care for patients and reduce healthcare costs, HIEs are tools that could be used for the creation or maintenance of a population-based registry. HIE organizations possess technical capabilities, such as data extraction from multiple organizations' health IT systems, transformation of the data into a common format, and loading of data into a common repository,[69] that are highly relevant to patient registries. Since much of the work to address interoperability issues has already been done by the HIE organization, a registry may leverage the existing infrastructure for its own purposes.

Population-based registries with state-mandated reporting requirements are increasingly interfacing with HIEs to allow providers to directly report to the registry through the HIE. For example, in Colorado, the Colorado Regional Health Information Organization (CORHIO) allows the Colorado Department of Public Health and Environment to access the network's web portal for case finding activities for the Colorado Central Cancer Registry. The cancer registry uses data in the CORHIO network portal to augment information that might be missing on cases reported through pathology lab reporting systems and to identify cancer cases that were not reported.[70] The cancer registry staff reduced time spent calling providers to obtain additional information and improved the completeness of case finding by collaborating with the HIE. Other states have developed use cases for reporting to cancer registries directly from the provider's EHR system via HIE.

Direct reporting is also common with immunization registries. For example, the Michigan Care Improvement Registry (MCIR) is a lifespan registry that contains immunization records for all Michigan residents. The Great Lakes Health Connect HIE enables participants to directly transmit immunization records to MCIR, thereby reducing provider reporting time. The MCIR also uses the HIE to improve immunization record accessibility and to exchange immunization records across state lines. Indiana healthcare providers can send immunization records for Michigan residents to the MCIR via HIE through a collaboration between the Michigan Health Information Network Shared Services and the Michiana Health Information Network. Patients living in the border areas of these two states often travel back and forth across state lines and may access healthcare in both states. Enabling providers to share health information across state lines is an important step in improving continuity of care for these patients and in improving the completeness of immunization records.[71]

To date, HIE networks and registries have largely collaborated on state-mandated registries, but there are opportunities to leverage HIE networks to create or enhance registries. In particular, HIE data may be a useful source for identifying potential patients for inclusion in a registry. For example, the Maine Health InfoNet HIE network was used to identify patients with congestive heart failure and diabetes via natural language processing[72, 73] and to predict incident essential hypertension using machine learning models.[74] These efforts could be extended to the establishment of patient registries if appropriate patient consent and data governance rules are established.

Sharing of protected health information (PHI) is a significant barrier to leveraging HIE networks for registry development. State-mandated registries, such as those discussed above, typically have legal authority for the exchange of PHI without explicit informed consent from the patient. However, most other types of registries would need to address the issue of patient consent.[75] The underlying model of an HIE may also affect its usefulness for registry activities. HIEs that use a centralized data repository are better suited for aggregating and analyzing data than models in which data are “owned” and maintained at the individual site and are not easily aggregated. However, data currency is an issue in any model where data must be “pushed” to a central repository. While HIEs represent a potential source of data for registries, further work is needed to understand barriers and to develop clear use cases beyond state-mandated registries.

Conclusion

As the ecosystem of health data expands, registries increasingly are interested in linking to or integrating data from other sources to minimize the burden of data entry and to address specific research objectives. Beyond EHRs, many sources of relevant data exist. However, incorporation of these data sources is challenging in many cases. Registries should carefully consider the purpose of the registry and the suitability of the data source for achieving that purpose, as well as the legal and ethical implications of incorporating other data sources, as a first step before addressing the technical interoperability challenges discussed in the next two chapters of this document. Further research to develop tools to help registries understand the quality of other data sources and the potential impact of incorporating these data into the registry database also would be useful to help inform these decisions.

References for Chapter 2

1. Blumenthal S. The Use of Clinical Registries in the United States: A Landscape Survey. EGEMS (Wash DC). 2017;5(1):26. PMID: 29930965. DOI: 10.5334/egems.248.

2. HIMSS. Electronic Health Records Definition. https://www.himss.org. Accessed June 10, 2019.

3. Virnig B, Parsons H. Strengths and Limitations of CMS Administrative Data in Research. Research Data Assistance Center (ResDAC). https://www.resdac.org/articles/strengths-and-limitations-cms-administrative-data-research. Accessed June 10, 2019.

4. Centers for Medicare & Medicaid Services. Medicare Enrollment Dashboard. https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Dashboard/Medicare-Enrollment/Enrollment%20Dashboard.html. Accessed June 10, 2019.

5. Centers for Medicare and Medicaid Services. March 2019 Medicaid & CHIP Enrollment Data Highlights. https://www.medicaid.gov/medicaid/program-information/medicaid-and-chip-enrollment-data/report-highlights/index.html. Accessed June 10, 2019.

6. Research Data Assistance Center (ResDAC). https://www.resdac.org/cms-data. Accessed June 10, 2019.

7. Doshi JA, Hendrick FB, Graff JS, et al. Data, Data Everywhere, but Access Remains a Big Issue for Researchers: A Review of Access Policies for Publicly-Funded Patient-Level Health Care Data in the United States. EGEMS (Wash DC). 2016;4(2):1204. PMID: 27141517. DOI: 10.13063/2327-9214.1204.

8. Healthcare Cost and Utilization Project (HCUP). Agency for Healthcare Research and Quality. https://www.hcup-us.ahrq.gov. Accessed June 10, 2019.

9. Healthcare Cost and Utilization Project (HCUP). Databases. Agency for Healthcare Research and Quality. https://www.hcup-us.ahrq.gov/databases.jsp. Accessed June 10, 2019.

10. All-Payer Claims Databases. Agency for Healthcare Research and Quality. https://www.ahrq.gov/professionals/quality-patient-safety/quality-resources/apcd/index.html. Accessed June 24, 2019.

11. Hammill BG, Hernandez AF, Peterson ED, et al. Linking inpatient clinical registry data to Medicare claims data using indirect identifiers. Am Heart J. 2009;157(6):995-1000. PMID: 19464409. DOI: 10.1016/j.ahj.2009.04.002.

12. Bozic KJ, Bashyal RK, Anthony SG, et al. Is administratively coded comorbidity and complication data in total joint arthroplasty valid? Clin Orthop Relat Res. 2013;471(1):201-5. PMID: 22528384. DOI: 10.1007/s11999-012-2352-1.

13. Alluri RK, Leland H, Heckmann N. Surgical research using national databases. Ann Transl Med. 2016;4(20):393. PMID: 27867945. DOI: 10.21037/atm.2016.10.49.

14. Joseph L, Bashir M, Xiang Q, et al. Prevalence and Outcomes of Mitral Stenosis in Patients Undergoing Transcatheter Aortic Valve Replacement: Findings From the Society of Thoracic Surgeons/American College of Cardiology Transcatheter Valve Therapies Registry. JACC Cardiovasc Interv. 2018;11(7):693-702. PMID: 29622149. DOI: 10.1016/j.jcin.2018.01.245.

15. Curtis JR, Chen L, Greenberg JD, et al. The clinical status and economic savings associated with remission among patients with rheumatoid arthritis: leveraging linked registry and claims data for synergistic insights. Pharmacoepidemiol Drug Saf. 2017;26(3):310-9. PMID: 28028867. DOI: 10.1002/pds.4126.

16. Reeves MJ, Fonarow GC, Smith EE, et al. Representativeness of the Get With The Guidelines-Stroke Registry: comparison of patient and hospital characteristics among Medicare beneficiaries hospitalized with ischemic stroke. Stroke. 2012;43(1):44-9. PMID: 21980197. DOI: 10.1161/STROKEAHA.111.626978.

17. National Cancer Institute. SEER-Medicare Publications by Journal & Year. https://healthcaredelivery.cancer.gov/seermedicare/overview/pubs_jour_year.php. Accessed June 10, 2019.

18. The Office of the National Coordinator of Health Information Technology. U.S. Department of Health and Human Services. Definition of Patient-Generated Health Data. https://www.healthit.gov/topic/scientific-initiatives/patient-generated-health-data. Accessed June 10, 2019.

19. Reeder B, David A. Health at hand: A systematic review of smart watch uses for health and wellness. J Biomed Inform. 2016;63:269-76. PMID: 27612974. DOI: 10.1016/j.jbi.2016.09.001.

20. Henriksen A, Haugen Mikalsen M, Woldaregay AZ, et al. Using Fitness Trackers and Smartwatches to Measure Physical Activity in Research: Analysis of Consumer Wrist-Worn Wearables. J Med Internet Res. 2018;20(3):e110. PMID: 29567635. DOI: 10.2196/jmir.9157.

21. de Zambotti M, Baker FC, Colrain IM. Validation of Sleep-Tracking Technology Compared with Polysomnography in Adolescents. Sleep. 2015;38(9):1461-8. PMID: 26158896. DOI: 10.5665/sleep.4990.

22. de Zambotti M, Goldstone A, Claudatos S, et al. A validation study of Fitbit Charge 2 compared with polysomnography in adults. Chronobiol Int. 2018;35(4):465-76. PMID: 29235907. DOI: 10.1080/07420528.2017.1413578.

23. Garabelli P, Stavrakis S, Po S. Smartphone-based arrhythmia monitoring. Curr Opin Cardiol. 2017;32(1):53-7. PMID: 27875477. DOI: 10.1097/HCO.0000000000000350.

24. Macinnes M, Martin N, Fulton H, et al. Comparison of a smartphone-based ECG recording system with a standard cardiac event monitor in the investigation of palpitations in children. Arch Dis Child. 2019;104(1):43-7. PMID: 29860228. DOI: 10.1136/archdischild-2018-314901.

25. Melville S, Teskey R, Philip S, et al. A Comparison and Calibration of a Wrist-Worn Blood Pressure Monitor for Patient Management: Assessing the Reliability of Innovative Blood Pressure Devices. J Med Internet Res. 2018;20(4):e111. PMID: 29695375. DOI: 10.2196/jmir.8009.

26. Chandrasekhar A, Kim CS, Naji M, et al. Smartphone-based blood pressure monitoring via the oscillometric finger-pressing method. Sci Transl Med. 2018;10(431). PMID: 29515001. DOI: 10.1126/scitranslmed.aap8674.

27. Milani RV, Lavie CJ, Bober RM, et al. Improving Hypertension Control and Patient Engagement Using Digital Tools. Am J Med. 2017;130(1):14-20. PMID: 27591179. DOI: 10.1016/j.amjmed.2016.07.029.

28. Heintzman ND. A Digital Ecosystem of Diabetes Data and Technology: Services, Systems, and Tools Enabled by Wearables, Sensors, and Apps. J Diabetes Sci Technol. 2015;10(1):35-41. PMID: 26685994. DOI: 10.1177/1932296815622453.

29. Basatneh R, Najafi B, Armstrong DG. Health Sensors, Smart Home Devices, and the Internet of Medical Things: An Opportunity for Dramatic Improvement in Care for the Lower Extremity Complications of Diabetes. J Diabetes Sci Technol. 2018;12(3):577-86. PMID: 29635931. DOI: 10.1177/1932296818768618.

30. Garde A, Dehkordi P, Wensley D, et al. Pulse oximetry recorded from the Phone Oximeter for detection of obstructive sleep apnea events with and without oxygen desaturation in children. Conf Proc IEEE Eng Med Biol Soc. 2015;2015:7692-5. PMID: 26738074. DOI: 10.1109/EMBC.2015.7320174.

31. The Office of the National Coordinator for Health Information Technology. U.S. Department of Health and Human Services. Conceptualizing a Data Infrastructure for the Capture, Use, and Sharing of Patient-Generated Health Data in Care Delivery and Research through 2024. White Paper. [Prepared by Accenture Federal Services for the under Contract No. HHSP233201500093I, Order No. HHSP23337001T]. January 2018. https://www.healthit.gov/sites/default/files/onc_pghd_final_white_paper.pdf. Accessed June 10, 2019.

32. Powell AC, Landman AB, Bates DW. In search of a few good apps. JAMA. 2014;311(18):1851-2. PMID: 24664278. DOI: 10.1001/jama.2014.2564.

33. Partners HealthCare: Connecting Heart Failure Patients to Providers through Remote Monitoring. Commonwealth Fund, January 30, 2013. https://www.commonwealthfund.org/publications/case-study/2013/jan/partners-healthcare-connecting-heart-failure-patients-providers. Accessed June 10, 2019.

34. Genes N, Violante S, Cetrangol C, et al. From smartphone to EHR: a case report on integrating patient-generated health data. npj Digital Medicine. 2018;1(1):23. DOI: 10.1038/s41746-018-0030-8.

35. Howie L, Hirsch B, Locklear T, et al. Assessing the value of patient-generated data to comparative effectiveness research. Health Aff (Millwood). 2014;33(7):1220-8. PMID: 25006149. DOI: 10.1377/hlthaff.2014.0225.

36. Dreyer NA, Blackburn SC, Mt-Isa S, et al. Direct-to-Patient Research: Piloting a New Approach to Understanding Drug Safety During Pregnancy. JMIR Public Health Surveill. 2015;1(2):e22. PMID: 27227140. DOI: 10.2196/publichealth.4939.

37. All of Us. National Institutes of Health. https://allofus.nih.gov/. Accessed June 11, 2019.

38. Bresnick J. AMA, Google Launch Health Data Interoperability, PGHD Challenge. Health IT Analytics. April 9, 2018.

39. Patient-Centered Outcomes Research Institute. Using Patient Generated Health Data to Transform Healthcare. https://www.pcori.org/research-results/2017/using-patient-generated-health-data-transform-healthcare. Accessed June 10, 2019.

40. Duke-Margolis Center for Health Policy. Mobilizing mHealth Innovation for Real-World Evidence Generation. September 2017. https://healthpolicy.duke.edu/sites/default/files/atoms/files/mobilizing_mhealth_innovation_for_real-world_evidence_generation.pdf. Accessed June 10, 2019.

41. Murakami H, Kawakami R, Nakae S, et al. Accuracy of Wearable Devices for Estimating Total Energy Expenditure: Comparison With Metabolic Chamber and Doubly Labeled Water Method. JAMA Intern Med. 2016;176(5):702-3. PMID: 26999758. DOI: 10.1001/jamainternmed.2016.0152.

42. Wood WA, Bennett AV, Basch E. Emerging uses of patient generated health data in clinical research. Mol Oncol. 2015;9(5):1018-24. PMID: 25248998. DOI: 10.1016/j.molonc.2014.08.006.

43. Bietz MJ, Bloss CS, Calvert S, et al. Opportunities and challenges in the use of personal health data for health research. J Am Med Inform Assoc. 2016;23(e1):e42-8. PMID: 26335984. DOI: 10.1093/jamia/ocv118.

44. Duffy MJ, Harbeck N, Nap M, et al. Clinical use of biomarkers in breast cancer: Updated guidelines from the European Group on Tumor Markers (EGTM). European journal of cancer (Oxford, England : 1990). 2017;75:284-98. PMID: 28259011. DOI: 10.1016/j.ejca.2017.01.017.

45. Ginsburg GS, Phillips KA. Precision Medicine: From Science To Value. Health Aff (Millwood). 2018;37(5):694-701. PMID: 29733705. DOI: 10.1377/hlthaff.2017.1624.

46. National Cancer Institute. About Genomic Data Commons. https://gdc.cancer.gov/about-data/data-sources. Accessed June 10, 2019.

47. Gainotti S, Torreri P, Wang CM, et al. The RD-Connect Registry & Biobank Finder: a tool for sharing aggregated data and metadata among rare disease researchers. Eur J Hum Genet. 2018;26(5):631-43. PMID: 29396563. DOI: 10.1038/s41431-017-0085-z.

48. AACR Project GENIE: Powering Precision Medicine through an International Consortium. Cancer Discovery. 2017;7(8):818.

49. Howell RR, Zuchner S. MOVR-NeuroMuscular ObserVational Research, a unified data hub for neuromuscular diseases. Genet Med. 2019;21(3):536-8. PMID: 29934516. DOI: 10.1038/s41436-018-0086-5.

50. Adams JU. Genetics: Big hopes for big data. Nature. 2015;527(7578):S108-9. PMID: 26580158. DOI: 10.1038/527S108a.

51. Khan A, Capps BJ, Sum MY, et al. Informed consent for human genetic and genomic studies: a systematic review. Clin Genet. 2014;86(3):199-206. PMID: 24646408. DOI: 10.1111/cge.12384.

52. Liebeskind DS. Innovative Interventional and Imaging Registries: Precision Medicine in Cerebrovascular Disorders. Interv Neurol. 2015;4(1-2):5-17. PMID: 26600792. DOI: 10.1159/000438773.

53. Kohli MD, Summers RM, Geis JR. Medical Image Data and Datasets in the Era of Machine Learning-Whitepaper from the 2016 C-MIMI Meeting Dataset Session. J Digit Imaging. 2017;30(4):392-9. PMID: 28516233. DOI: 10.1007/s10278-017-9976-3.

54. Inmon WH. Building the data warehouse: John Wiley & Sons; 2005.

55. Huser V, Cimino JJ. Desiderata for healthcare integrated data repositories based on architectural comparison of three public repositories. AMIA Annu Symp Proc. 2013;2013:648-56. PMID: 24551366.

56. Foran DJ, Chen W, Chu H, et al. Roadmap to a Comprehensive Clinical Data Warehouse for Precision Medicine Applications in Oncology. Cancer Inform. 2017;16:1176935117694349. PMID: 28469389. DOI: 10.1177/1176935117694349.

57. Horton I, Lin Y, Reed G, et al. Empowering Mayo Clinic Individualized Medicine with Genomic Data Warehousing. J Pers Med. 2017;7(3). PMID: 28829408. DOI: 10.3390/jpm7030007.

58. Ajayi OJ, Smith EJ, Viangteeravat T, et al. Multisite Semiautomated Clinical Data Repository for Duplication 15q Syndrome: Study Protocol and Early Uses. JMIR Res Protoc. 2017;6(10):e194. PMID: 29046268. DOI: 10.2196/resprot.7989.

59. Kunjan K, Toscos T, Turkcan A, et al. A Multidimensional Data Warehouse for Community Health Centers. AMIA Annu Symp Proc. 2015;2015:1976-84. PMID: 26958297.

60. Davies M, Erickson K, Wyner Z, et al. Software-Enabled Distributed Network Governance: The PopMedNet Experience. EGEMS (Wash DC). 2016;4(2):1213. PMID: 27141522. DOI: 10.13063/2327-9214.1213.

61. Skripcak T, Belka C, Bosch W, et al. Creating a data exchange strategy for radiotherapy research: towards federated databases and anonymised public datasets. Radiother Oncol. 2014;113(3):303-9. PMID: 25458128. DOI: 10.1016/j.radonc.2014.10.001.

62. Evans RS, Lloyd JF, Pierce LA. Clinical Use of an Enterprise Data Warehouse. AMIA Annual Symposium Proceedings. 2012;2012:189-98. PMID: PMC3540441.

63. Danciu I, Cowan JD, Basford M, et al. Secondary use of clinical data: the Vanderbilt approach. J Biomed Inform. 2014;52:28-35. PMID: 24534443. DOI: 10.1016/j.jbi.2014.02.003.

64. Jannot AS, Zapletal E, Avillach P, et al. The Georges Pompidou University Hospital Clinical Data Warehouse: A 8-years follow-up experience. Int J Med Inform. 2017;102:21-8. PMID: 28495345. DOI: 10.1016/j.ijmedinf.2017.02.006.

65. Weng C, Bigger JT, Busacca L, et al. Comparing the effectiveness of a clinical registry and a clinical data warehouse for supporting clinical trial recruitment: a case study. AMIA Annu Symp Proc. 2010;2010:867-71. PMID: 21347102.

66. Zhang Q, Matsumura Y, Teratani T, et al. The application of an institutional clinical data warehouse to the assessment of adverse drug reactions (ADRs). Evaluation of aminoglycoside and cephalosporin associated nephrotoxicity. Methods Inf Med. 2007;46(5):516-22. PMID: 17938772.

67. O'Leary KJ, Devisetty VK, Patel AR, et al. Comparison of traditional trigger tool to data warehouse based screening for identifying hospital adverse events. BMJ Qual Saf. 2013;22(2):130-8. PMID: 23038408. DOI: 10.1136/bmjqs-2012-001102.

68. Connolly D, Adagarla B, Nair M, et al. SEINE: Methods for Electronic Data Capture and Integrated Data Repository Synthesis with Patient Registry Use Cases. 2014.

69. Harris AH, Chen C, Rubinsky AD, et al. Are Improvements in Measured Performance Driven by Better Treatment or "Denominator Management"? J Gen Intern Med. 2016;31 Suppl 1:21-7. PMID: 26951270. DOI: 10.1007/s11606-015-3558-1.

70. State Public Health Department Enhancing Disease Reporting With HIE Data. http://www.corhio.org/news/2015/9/30/547-state-public-health-department-enhancing-disease-reporting-with-hie-data. Accessed June 10, 2019.

71. Michigan, Indiana in Full Production for Interstate Health Information Exchange. https://detroit.cbslocal.com/2013/03/04/michigan-indiana-in-full-production-for-interstate-health-information-exchange/. Accessed June 10, 2019.

72. Wang Y, Luo J, Hao S, et al. NLP based congestive heart failure case finding: A prospective analysis on statewide electronic medical records. Int J Med Inform. 2015;84(12):1039-47. PMID: 26254876. DOI: 10.1016/j.ijmedinf.2015.06.007.

73. Zheng L, Wang Y, Hao S, et al. Web-based Real-Time Case Finding for the Population Health Management of Patients With Diabetes Mellitus: A Prospective Validation of the Natural Language Processing-Based Algorithm With Statewide Electronic Medical Records. JMIR Med Inform. 2016;4(4):e37. PMID: 27836816. DOI: 10.2196/medinform.6328.

74. Ye C, Fu T, Hao S, et al. Prediction of Incident Hypertension Within the Next Year: Prospective Study Using Statewide Electronic Health Records and Machine Learning. J Med Internet Res. 2018;20(1):e22. PMID: 29382633. DOI: 10.2196/jmir.9268.

75. Mello MM, Adler-Milstein J, Ding KL, et al. Legal Barriers to the Growth of Health Information Exchange-Boulders or Pebbles? Milbank Q. 2018;96(1):110-43. PMID: 29504197. DOI: 10.1111/1468-0009.12313.

Chapter 3. Data Standards

Authors (alphabetical)

Richard E. Gliklich, M.D.
Chief Executive Officer and Chairman
OM1, Inc.

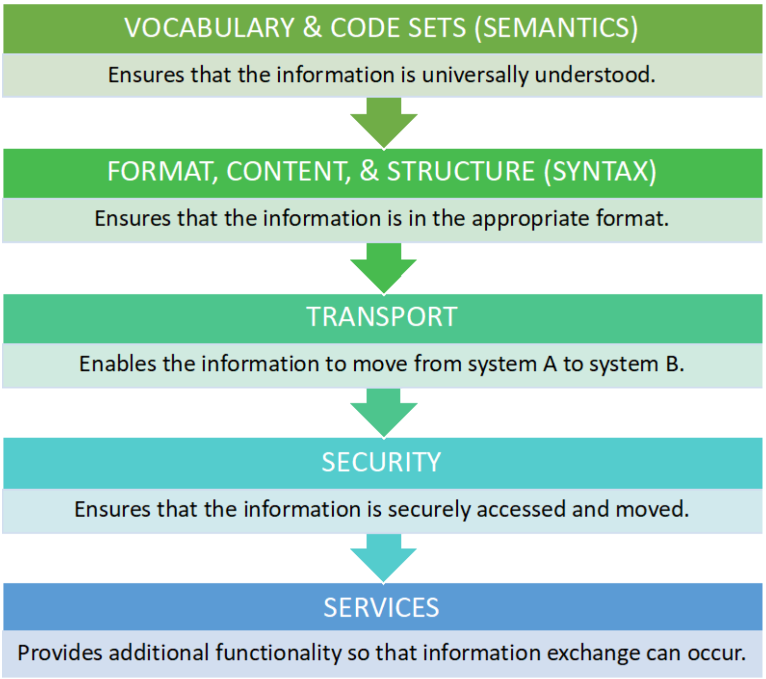
Michelle B. Leavy, M.P.H.
Head, Healthcare Research & Policy
OM1, Inc.

Introduction

Interoperability is defined as the ability of a system to exchange electronic health information with and use electronic health information from other systems without special effort on the part of the user.[1] As discussed in Chapter 1, interoperability is complex and requires consideration of multiple factors, including the data to be shared, the format of the data, the necessary permissions to protect patient privacy, and method of transferring the data.

Technical standards are an important foundation for exchanging data. The standards relevant for interoperability can be grouped into five categories, as shown below in Figure 3-1.

Figure 3-1. Categories of standards relevant for interoperability*



VOCABULARY & CODE SETS (SEMANTICS)

Ensures that the information is universally understood.

FORMAT, CONTENT, & STRUCTURE (SYNTAX)

Ensures that the information is in the appropriate format.

TRANSPORT

Enables the information to move from system A to system B.

SECURITY

Ensures that the information is securely accessed and moved.

SERVICES

Provides additional functionality so that information exchange can occur.

**Adapted from Connecting Health and Care for the Nation A Shared Nationwide Interoperability Roadmap. Office of the National Coordinator for Health Information Technology.[1]*

Data standards, such as vocabularies and code sets, are a critical building block for the interoperability of electronic health information. Data standards support semantic interoperability, meaning the ability for systems exchanging the data to interpret the data correctly. For example, different health systems may use different terms for the same concept (e.g., Tylenol, acetaminophen). When these data are exchanged, the systems must recognize these terms as synonyms and not different medications.

The purpose of this chapter is to discuss the role of data standards in supporting semantic interoperability, describe how registries currently use data standards, and discuss the applicability for registries of recent efforts to promote data harmonization and standardization.

Vocabulary and Terminology Standards

Vocabulary and terminology standards provide a consistent approach for documenting electronic health data across providers and sites to support clinical care and healthcare operations. Several vocabulary and terminology standards have been developed for different purposes, as shown in Table 3-1.

Table 3-1. Examples of vocabulary and terminology standards*

STANDARD	ACRONYM	DESCRIPTION	DEVELOPER
Current Procedural Terminology	CPT®	Medical service and procedure codes commonly used in public and private health insurance plans and claims processing.	American Medical Association
International Classification of Diseases	ICD-10	International standard for classifying diseases and other health problems recorded on health and vital records. The ICD is also used to code and classify mortality data from death certificates in the United States.	World Health Organization
Systemized Nomenclature of Medicine	SNOMED CT	Clinical healthcare terminology that maps clinical concepts with standard descriptive terms.	International Health Terminology Standards Development Organization
National Drug Code	NDC	Unique 3-segment number used as the universal identifier for human drugs.	U.S. Food and Drug Administration
RxNorm	RxNorm	Standardized nomenclature for clinical drugs. The name of a drug combines its ingredients, strengths, and/or form. Links to many of the drug vocabularies commonly used in pharmacy management and drug interaction software.	National Library of Medicine
World Health Organization Drug Dictionary	WHODRUG	International drug dictionary	World Health Organization
Logical Observation Identifiers Names and Codes	LOINC®	Concept-based terminology for lab orders and results.	Regenstrief Institute for Health Care

**Adapted from "Data Elements for Registries," in Registries for Evaluating Patient Outcomes: A User's Guide.*[2]

Within a standard vocabulary, a subset of terms and codes that is used for a specific purpose may be grouped into a value set. For example, a value set may be created to identify all patients with heart failure for the purposes of calculating a heart failure quality measure. The Value Set Authority Center (VSAC) is a Federal repository for value sets.[3]

Despite progress in the use of vocabulary and terminology standards, challenges still exist. Multiple standards are still used for some areas (e.g., medications), and some systems that capture electronic health data use local terminologies instead of existing standards. In addition, some types of electronic health data, such as radiographic images, pathology slides, and clinical notes, may not be recorded using vocabulary and terminology standards.

Where feasible, patient registries should consider defining data elements based on existing vocabulary and terminology standards, particularly if the registry intends to incorporate data from other health IT systems. As an example, diagnoses that are required for entry into the registry may be defined in terms of ICD-10 codes.

Common Data Elements

Common data elements (CDEs) are standardized data elements that can be used in multiple clinical studies. In the context of registries, use of CDEs would support linkages and aggregations with data from other studies and potentially improve efficiency in data capture, as data could be captured once and used in multiple studies.[4] General or core CDEs are intended to be relevant across therapeutic and disease areas, while disease-specific CDEs are intended to be used within a specific therapeutic and disease area. General and disease-specific CDEs are often grouped into a minimum or core set of data elements to be collected in all studies of a specific type (e.g., diabetes patient registries). To facilitate consistent capture of data across sites and over time, CDEs typically include standard definitions, code lists, and instructions.

Many sources of CDEs exist. The National Institutes of Health (NIH) manages a Common Data Element (CDE) Repository that contains CDEs developed through NIH-supported efforts as well as other efforts. The CDE Repository includes both general and disease-specific CDEs. Where possible, CDEs are linked to standardized value sets in VSAC. The Clinical Data Interchange Standards Consortium (CDISC) also has created a set of general CDEs suitable for use in clinical research. Specific to registries, the Common Healthcare Data Interoperability Project is a collaboration between the Duke Clinical Research Institute and The Pew Charitable Trusts to advance interoperability across EHRs and registries by standardizing frequently used clinical concepts as CDEs. These and other CDE efforts are summarized in Appendix A.

Standardized Outcome Measures

While standardized vocabularies and CDEs have facilitated the consistent capture of health data in many areas, they have not addressed one of the key challenges associated with use of electronic health data for clinical research, quality improvement, population health management, and value-based care. That challenge is the lack of common, standardized patient outcome measures across medical conditions and consistent definitions for the data elements that must be collected to determine these outcomes. By patient outcomes, we refer to that condition-specific information that is relevant to both patients and providers, and which clearly describes whether an individual’s disease or condition has improved or worsened. For registries that seek to evaluate patient outcomes, adopting standardized outcomes allows different registry holders to compare and/or aggregate results. Further, use of standardized outcomes paves the way for more standardized collection and exchange of the underlying key variables within and between health information technology systems.

Standardized outcome measures provide a higher-level grouping of standardized data elements, such as CDEs or value sets, into an outcome measure definition that can be captured consistently across providers and care settings. Currently, standardized outcome measures have not been developed or are not widely used in most condition areas. Instead, registries, clinical trials, quality improvement initiatives, and other data collection efforts frequently measure different outcomes or use different definitions of the same outcome measure. For example, in 2016, the Agency for Healthcare Research and Quality (AHRQ) conducted a technology assessment to determine the safety and efficacy of retinal prosthesis systems for halting disease progression in patients with retinitis pigmentosa. The 11 studies included in the systematic review reported 74 different outcome measures. Only three of the 74 outcome measures were reported by three or more studies, and only four of the outcome measures had evidence of validity and reliability. In addition, even when studies reported the same acuity test, the data were reported in different ways, making it difficult to aggregate and compare the results. The authors of the review noted that little consensus exists among researchers studying retinal prosthesis systems as to which outcomes to measure. Due to the inconsistencies in evidence, the report made no conclusions about the likelihood of patient benefit from these devices.[5] This type of variation in the selection of outcome measures is common across condition areas and has been well-documented in the literature.[6-9]

Variation in the definition of a specific outcome measure concept is equally problematic. Consider, for example, the definitions of bleeding that are used in cardiovascular research. A systematic review and meta-analysis published in 2014 found that 10 different definitions of major bleeding are currently used in clinical trials and patient registries for patients undergoing percutaneous coronary intervention (PCI). The definitions include different clinical events (e.g., blood transfusion, hemorrhage), different laboratory parameters, and different outcomes (e.g., mortality), and the incidence of major bleeding, naturally, varies depending on the definition used by the study. In one example cited by the authors, non-coronary artery bypass graft related major bleeding occurred in 0.87% of patients according to one definition but in 3.1% of the same patients according to another definition. While PCI studies are measuring the same outcome, comparison across studies is challenging because of the variations in definition.[10] An earlier review, published in 2007, identified the same issue with bleeding definitions in PCI studies, leading the authors to conclude that “different bleeding definitions can lead to markedly different conclusions about the safety of an antithrombotic regimen.”[11]

Technical interoperability will not address issues such as these that result from measurement of different concepts or use of different definitions. To fully realize the promise of health data interoperability, it is essential to standardize the outcome measures that are being collected across health IT systems, so meaningful comparisons and aggregations can be made. This requires both consensus on which outcome measures should be captured in each condition area, as well as standardization of the definitions and data elements that make up these key outcome measures.

Many consensus-based efforts with different intended uses and scopes have been launched to address these issues. Some efforts have focused on standardizing the definition of a single outcome, such as myocardial infarction,[12] while others have focused on harmonizing the outcome measure concepts captured across studies in a specific disease area. OMERACT (Outcome Measures in Rheumatology), a long-standing, independent, and international initiative, is an example of the latter type of effort. Over the past 20 years, OMERACT has developed core sets of outcome measures for use in rheumatoid arthritis, osteoarthritis, psoriatic arthritis, fibromyalgia, and other rheumatic disease research through a well-documented, repeatable process that has served as a model for other efforts.[13] Some efforts have focused on improving the methodology used to develop and report on consensus-based standards[14, 15] or increasing access to standards that have already been developed. For example, the COMET Initiative provides a searchable database of harmonization efforts published in the peer-reviewed literature.[16] Although a full review of existing efforts is beyond the scope of this chapter, more information on many relevant efforts can be found in Appendix A.

Despite myriad efforts, many new research studies, including patient registries, do not use existing standardized measures or data elements. In some cases, standards have not yet been developed for a specific condition area. In other cases, researchers may not be aware of existing standards, may disagree with the standards or wish to measure different outcomes, or may be uncertain about the quality or value of using the existing standards.[6] A 2016 report from The Pew Charitable Trusts examined barriers to use of existing data standards in patient registries and found that registry stewards frequently have not participated in the development of data standards, resulting in standards that may not meet the needs of registries and their stakeholders. In addition, use of data standards is not required, and, although some registries use them, much of the value of standardization can only be realized if the majority of registries use the standards. That tipping point has not yet been reached.[4] Finally, even when researchers would like to adopt a data standard for new projects, concerns about preserving the continuity of evidence from earlier studies and the challenges of mapping data from previous datasets to new data elements may lead them to use legacy data elements or measures.

Standardization is a critical building block in a national research infrastructure and learning health system, and efforts are needed to address these barriers and spur increased adoption and use of standardized data elements and outcome measures. In particular, inclusion of patient registries in the development of data standards is essential, given the central role that registries play in the learning health system and national research infrastructure (see Chapter 1). The following sections describe the Outcome Measures Framework project, an effort funded by the Agency for Healthcare Research and Quality (AHRQ) to address several of these barriers by developing standardized outcome measures in five condition areas for use in patient registries and other health information technology systems across the learning health system.

Development of the OMF

Over the past eight years, AHRQ has supported a series of projects to understand the variation in outcomes measurement in patient registries and to develop tools to support harmonization of outcome measures. This work launched in 2011 with a series of stakeholder meetings designed to gather information on how outcome measures were collected in existing patient registries and how stakeholders would like to see information on outcome measures presented. In parallel, background research was conducted to identify existing models or systems designed to categorize and/or present information on data elements, outcome measures, or quality measures. Based on the background research and stakeholder feedback, the Outcome Measures Framework (OMF) was created in early 2012. The OMF is a conceptual model for classifying outcomes that are relevant to patients and providers across most conditions. The OMF was revised following a series of web-based meetings and document review cycles with stakeholders and finalized in December 2012.[17]

The second phase of the OMF project began in 2013 with a systematic literature review of systems used to standardize language and definitions for outcome measures and other data elements, including systems for registries, clinical trials, electronic health records (EHRs), and quality reporting systems. The literature review identified 61 publications on three major topics: harmonizing data elements, key components of outcome measures, and governance plans for existing models. Many of the publications described efforts to harmonize data elements or create core sets of outcome measures; these efforts were identified as useful models for developing standardized outcome measures through a consensus-driven process. At the time this review was completed (2014), no existing efforts with the same or substantially similar goals as the OMF project were identified.[18]

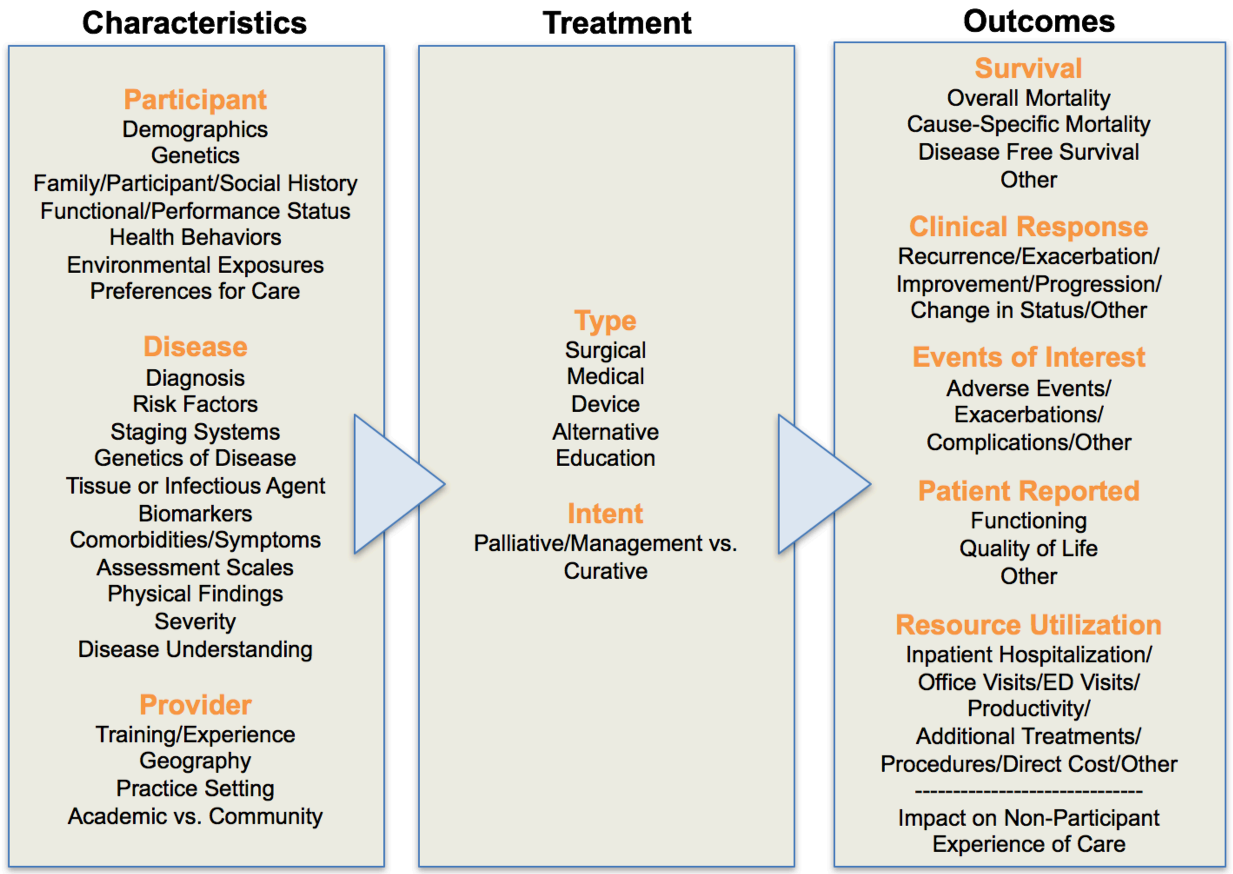
In preparation for using the OMF to support the development of standardized outcome measures, a qualitative analysis was conducted in 2015 to test the robustness of the OMF and identify any areas for improvement. Outcome measures from four diverse condition areas – depression, asthma, rheumatoid arthritis, and cardiac surgery – were abstracted from patient registries listed on ClinicalTrials.gov in June 2015 and mapped to the OMF. The condition areas were selected to represent different types of conditions, treatment options, providers, care settings, and patient populations. Two of the condition areas (rheumatoid arthritis and cardiac surgery) were selected for further analysis, and additional outcome measures were abstracted from patient registry-run websites and the published literature and mapped to the OMF. Across the four condition areas, 416 outcome measures were identified and reviewed. Most measures mapped directly to the OMF; analysis of the measures that did not map directly to the OMF resulted in minor modifications to the framework. The analysis demonstrated the robustness of the OMF for classifying a diverse group of outcome measures and highlighted its potential for supporting the development of standardized outcome measures in a range of condition areas.[19]

Throughout each phase of the development of the OMF, stakeholder feedback has been actively sought and incorporated into the framework. Over 400 stakeholders representing registry stewards, healthcare provider organizations, professional societies, academia, research and consulting organizations, government agencies, patient/consumer organizations, journal editors, payers, and pharmaceutical and medical device companies have participated in the various meetings and review activities. More information on the development of the OMF can be found in the 2014 publication on development of the OMF[17] as well as AHRQ reports on the literature review and analysis of ClinicalTrials.gov data.

Structure of the OMF

The OMF (Figure 3-2) is a hierarchy with three levels: domains, subcategories of data elements, and data elements. The domains – characteristics, treatments, and outcomes – represent the process by which characteristics of the participant, disease, and provider influence treatment, and by which characteristics and treatment together influence outcomes. The process may be iterative, in that outcomes of one treatment may determine additional courses of treatment. At the second level, subcategories of data elements are presented to help guide the definition of an outcome measure. For example, information on the intent of a treatment (palliative vs. curative vs. management) is important when determining the appropriate outcomes to measure. Lastly, at the third level are the categories of data elements that would be used to define an outcome measure, such as data elements to capture the patient demographics and diagnosis. These categories are intentionally broad so that the framework can be used across condition areas; not all categories will be relevant in a specific condition area.

Figure 3-2. Outcome Measures Framework[21]



In the Outcomes domain, outcome measures are grouped into five main categories: survival, clinical response or status, events of interest, patient-reported, and resource utilization. These categories represent both final outcomes, such as mortality, as well as intermediate outcomes, such as clinical response. While final outcomes may be most important in some condition areas, inclusion of intermediate outcomes such as clinical response makes the framework applicable to chronic conditions such as asthma or diabetes, where tracking patient-reported outcomes and disease progression over time is critical. It is also important to note that outcome measures may fit in more than one category. As an example, patient-reported outcomes may be used to assess clinical response (or status) for some conditions (e.g., depression).

Finally, two categories – Experience of Care and Impact on Non-Participant – are included below the Outcomes domains section. These measures fall outside of the structure of the OMF, in that they do not reflect an outcome of treatment for an individual patient; however, these are important concepts to capture in some condition areas. For example, a registry may wish to capture a birth outcome for a woman receiving treatment during pregnancy. Registries also may wish to understand patients' experiences of care, particularly as they relate to specific issues encountered during treatment, such as care coordination and provider communication in oncology.

The framework is a common model intended to be applied to specific conditions in potentially differing ways. For that reason, recommendations for measurement frequency are not specified in the model, but should be specified when applying the OMF to specific condition areas. Different timeframes and measurement frequencies may be appropriate depending on the condition area and outcome measure of interest. Further, some timepoint data collection decisions are made by registries today with a goal to minimize administrative and respondent burden. As these data elements are incorporated into interoperable health IT systems, those limitations may become fewer, allowing for new timepoints for some measures to be added (e.g., longer followup).

Use of the OMF To Support Measure Harmonization

The OMF is now being used for its intended purpose of serving as a content model for developing standardized outcome measures for use in patient registries and other primary data collection efforts. AHRQ led this effort in collaboration with the Food & Drug Administration and the National Library of Medicine. A key goal of the initiative, which was supported by the Assistant Secretary of Planning and Evaluation (ASPE) Patient-Centered Outcomes Research Trust Fund,[b] was to standardize the definitions of the components that make up the outcome measures, so that the measures could be captured consistently within EHRs and so that those using the measures can understand the level of comparability across different systems and studies.

For this exercise, standardized outcome measures were developed for five condition areas using a reproducible process involving registry sponsors and other stakeholders, such as clinicians and representatives from patient advocacy organizations, payers, funding agencies, regulatory bodies, and research organizations. The five condition areas – atrial fibrillation, asthma, depression, non-small cell lung cancer, and lumbar spondylolisthesis – were selected to represent different types of conditions (chronic, acute, mental health), treatment modalities, care providers and care

[b] *This work was supported by the Office of the Secretary Patient-Centered Outcomes Research Trust Fund under Interagency Agreement #16-566R-16.*

settings, and patient populations. Within each condition area, workgroups made up of registry sponsors and other stakeholders produced a minimum set of standardized measures that could be captured in future registries as well as in clinical practice in the condition area of interest; workgroups also identified characteristics of the patient, disease, and provider that are necessary to support appropriate risk adjustment for the measures included in the minimum set. The minimum measure sets and fully populated, condition-specific frameworks are published elsewhere,[20, 21] but the key lessons learned and insights are summarized below.[22]

The condition areas selected for this project were chosen intentionally to present different challenges, with the goal of testing the robustness of the OMF as a tool for supporting harmonization across condition areas. Several differences were observed across the workgroups, and some of these differences had an impact on the harmonization effort. First, the number of registries participating in each of the five workgroups was similar, but the number of outcome measures collected in each condition area ranged from 27 in depression to 112 in atrial fibrillation, as shown in Table 3-2, which naturally influences the likelihood that an OMF categorization would be covered.

Table 3-2. Categorization of outcome measures in five condition areas

	ATRIAL FIBRILLATION	ASTHMA	DEPRESSION	LUNG CANCER	LUMBAR SPONDYLOLISTHESIS
# of Participating Registries	13	13	11	15	10
# of Outcome Measures	112	46	27	66	57
OMF Category: Survival	11 (10%)	2 (4%)	2 (7%)	6 (9%)	2 (4%)
OMF Category: Clinical Response	4 (4%)	9 (20%)	11 (41%)	10 (15%)	21 (37%)
OMF Category: Events of Interest	81 (72%)	3 (7%)	2 (7%)	8 (12%)	9 (16%)
OMF Category: Patient-Reported	6 (5%)	14 (30%)	10 (37%)	30 (45%)	21 (37%)
OMF Category: Resource Utilization	10 (9%)	17 (37%)	2 (7%)	9 (14%)	3 (5%)
OMF Category: Experience of Care	N/A	1 (2%)	N/A	3 (5%)	1 (2%)

The categorization of the measures also differed across condition areas. In each condition area, all outcome measures captured in the registries were collected and categorized according to the OMF. The measures were then reviewed and prioritized by the workgroups, with the goal of identifying a minimum set of measures that are broadly relevant. The minimum measure set is intended for use as a core set of outcomes that will be collected in all future registries and would also be suitable for use in clinical practice in the specific condition area; some studies may collect additional outcomes using other definitions to meet specific purposes. While the categorization of all measures differed by condition area, the categorization of the measures in the minimum measure sets is relatively consistent across condition areas (see Table 3-3). It should be noted that, as discussed above, some measures may fit into more than one category (e.g., patient-reported outcomes that are used to measure clinical response).

Table 3-3. Minimum measure sets – categorization of outcome measures

	ATRIAL FIBRILLATION	ASTHMA	DEPRESSION	LUNG CANCER	LUMBAR SPONDYLOLISTHESIS
# of Measures in Minimum Set	18	19	11	8	14
Survival	3 (17%)	1 (5%)	2 (18%)	3 (38%)	1 (7%)
Clinical Response	3 (17%)	5 (26%)	3 (27%)	2 (25%)	5 (36%)
Events of Interest	7 (39%)	5 (26%)	2 (18%)	2 (25%)	4 (29%)
Patient-Reported	2 (11%)	4 (21%)	1 (9%)	*	3 (21%)
Resource Utilization	3 (17%)	3 (16%)	2 (18%)	1 (13%)	1 (7%)

**The group agreed on the importance of capturing patient-reported outcomes, but did not reach consensus on a specific domain to capture.*

In addition, the condition areas differed with respect to existence of consensus statements and data standards. Multiple consensus statements and data standards were identified to support the work of the atrial fibrillation workgroup. Data standards and consensus definitions also exist for lung cancer, and, to some extent, asthma. However, very few relevant consensus statements or data standards were identified depression and lumbar spondylolisthesis. In areas where standard definitions are already in use, the workgroups were generally able to reach consensus on definitions more quickly. Existing quality measures also played an important role in some workgroups. In asthma and depression, quality measures that require use of specific validated instruments (e.g., asthma control tests, depression screening tool) are already widely implemented, and the workgroup chose to recommend use of the same instruments to further encourage harmonization across providers and data collection initiatives.

Survival Measures

Across the condition areas, all-cause mortality was generally identified as an important outcome measure. Discussion focused on cause-specific mortality, with emphasis on the difficulty of ascertaining cause of death in a more consistent and accurate fashion than what is listed on a death certificate. An example of a cause-specific mortality measure is procedure-related death, which the atrial fibrillation workgroup defined as, “all-cause mortality within 30 days of the procedure or during the index procedure hospitalization (if the postoperative length of stay is > than 30 days). Procedure-related deaths include those related to a complication of the procedure or treatment for a complication of the procedure.”[23]

Clinical Response or Status

Clinical Response measures capture the clinician’s assessment of whether the patient is responding to treatment – meaning improving, worsening, or remaining stable – or, for patients not receiving treatment, whether the patient’s clinical status is changing. These measures were challenging for many of the workgroups for the reason that, in many clinical conditions, a uniform approach to assessing clinical response has not been clearly articulated by the providers who treat those conditions. Moreover, clinicians freely admit that it can be difficult to date the onset and resolution of exacerbations of chronic diseases. It should also be noted that in some condition areas, different outcomes may be used depending on the intent of treatment (e.g., anticoagulation vs. ablation in atrial fibrillation). Timeframes for measuring clinical response, such as timeframes for measuring remission and response in depression, were also challenging areas for agreement as there is limited research to guide optimal time points, now how to achieve such optimal measurement intervals in real-world data collection. An example of a clinical response measure for asthma is exacerbation, defined as: “exacerbations of asthma are episodes characterized by an increase in symptoms of shortness of breath, cough, wheezing or chest tightness and decrease in lung function, i.e. they represent a change from the patient’s usual status that is sufficient to require a change in treatment. Exacerbation includes any of the following: (a) prescribed systemic steroids (defined as 2 or more days of oral steroids or a steroid injection) or increasing the oral steroid dose from dose at baseline; (b) an asthma-related hospitalization, ED visit, urgent care center visit, or unscheduled office visit requiring prescription of systemic corticosteroids; (c) documentation by provider of acute asthma exacerbation.”

Events of Interest

Outcome measures included in this category often captured complications or adverse events related to treatment or events associated with disease progression. In some cases, events in this category overlapped with events included in the resource utilization category (e.g., hospitalization or emergency department visit). An example of an event of interest for depression is suicide ideation and behavior, defined as, “selection of ‘several days’, ‘more than half the days’ or ‘nearly every day’ option on PHQ-9 item 9 (“Thoughts that you would be better off dead or of hurting yourself in some way”). Supplemental assessments of suicide ideation and behavior should be completed for patients who screen positive for suicide ideation on the PHQ-9. Supplemental assessments should be completed using an appropriate, brief, validated instrument, such as the Concise Health Risk Tracking (CHRT) scale.”

Patient-Reported Outcomes

The category of patient-reported outcomes was one of the most challenging areas across the workgroups. Most workgroups reached consensus quickly on the importance of capturing patient-reported outcomes, but consensus on the domains to measure and instrument(s) to use was much harder to obtain. Patient-reported domains of interest varied depending on the intent of treatment, stage of disease, population under study, and purpose of the registry. Even when workgroup members agreed on a domain (e.g., atrial fibrillation-related quality of life), agreement on a specific instrument was difficult. In some cases, this was due to use of various instruments in clinical practice, while in other cases, no single instrument was routinely used in clinical practice or there may be many aspects of quality of life that were of interests, e.g., ability to take care of oneself and others, to work, etc. In addition, this project restricted recommendations of specific instruments to instruments that are validated, publicly available, and have strong psychometric properties. Instruments that met these criteria were not available to measure all domains of interest identified by the workgroups.

The experience of the workgroups in this area reflects the need for further development and validation of patient-reported outcome measures that are appropriate for routine use in clinical practice as well as research. In addition, development of crosswalks from legacy instruments to newer instruments is essential to allow for comparisons and data mapping across studies.

Resource Utilization

Most workgroups focused on all healthcare utilization related to the condition, with some groups calling out specific events of interest (e.g., hospitalization). Recognizing the inherent difficulty of determining whether an event was truly related to a condition or just coincidental is difficult under the best of circumstances. For some conditions, impact on productivity and missed days of school were also identified as outcomes of interest. An example of a resource utilization measure from depression is work productivity, defined as, "work productivity loss (overall work impairment/absenteeism plus presenteeism), as measured by the Work Productivity and Activity Impairment Questionnaire (WPAI), reported in 12-month intervals." (As mentioned previously, measures may fit into more than one category; here, the WPAI is a patient-reported outcome, but it is being used to measure resource utilization in this context.)

Translation to Standardized Terminologies

As noted above, a key goal of this project was to improve the comparability and consistency of outcome measures collected in different studies and different systems. Narrative definitions, such as those produced by the workgroups, still allow for inconsistency in data collection, particularly when the data are abstracted from different EHRs or other data collection systems. As a final step in this project, the narrative definitions produced by the workgroups were translated into standardized terminologies (e.g., ICD-10, SNOMED, LOINC) to reduce burden of implementation and to facilitate consistent capture within an EHR.

For each measure, clinical informaticists defined the recommended reporting period, initial population for measurement, outcome-focused population, and data criteria and value sets. Because EHR data often do not contain all required components of a measure definition, the standardized definitions focus on gathering the clinician's assertion of an outcome condition and as much supporting evidence as possible, so that some structure evidence is available. Relationships between events are often not directly asserted within an EHR (e.g., a specific complication resulting from a procedure). Where possible, relationships are inferred based on time stamps and intervals; where this is not possible (e.g., cause of death), the expression logic requires an asserted relationship.

For each outcome, the following were defined:

- An object representing the outcome condition itself: In many cases, the only structured data will be an assertion of an outcome, with all the supporting evidence being present in the narrative.
- Fast Healthcare Interoperability Resources (FHIR) resources for evidence for the outcome: These include labs, diagnostic imaging, etc.
- FHIR resources for additional relevant events: These might include procedures, encounters, etc.
- Temporal aspects for all events: These allow for inferred relationships.

In addition, this project aimed to leverage existing resources and build connections across initiatives. To support that goal, existing common data elements and value sets were identified through review of four sources (eCQI Resource Center,[24] Value Set Authority Center [VSAC],[3] Consolidated CDA [C-CDA],[25] and NIH Common Data Elements Repository[26]) and used wherever possible. Results of the comparisons were documented in the Introductory document for each library of common data definitions, and existing common data elements and value sets were used where appropriate.

The final standardized data libraries will be made available publicly on the AHRQ website.

Conclusion

Data standardization creates the foundation for sharing data across health IT systems. Consistent and widespread use of data standards will improve semantic interoperability and facilitate more meaningful exchanges of data. Many data standardization efforts to date have focused on *how* to collect commonly captured data (e.g., demographics, medications, procedures) using standardized vocabularies or CDEs that are relevant across condition areas. More recently, standardization efforts have shifted to focus on *what* to collect within a specific condition area, for example by defining core sets of disease-specific CDEs or standardized outcome measures. Standardization of both how and what data are collected will support broader efforts to build a national research infrastructure and learning health systems.

Despite the promise of standardization, many challenges remain. Registries may be unaware of standards or may find that existing standards are not suitable for their purposes. In many condition areas, disease-specific CDEs and standardized outcome measures have not been developed. Further work is also needed to explore the willingness of registry developers to adopt the standards, such as standardized outcome measures, in future research projects and to identify incentives to encourage widespread implementation of these standards within other data collections systems, such as EHRs.

References for Chapter 3

1. Connecting Health and Care for the Nation: A Shared Nationwide Interoperability Roadmap. Office of the National Coordinator for Health Information Technology. 2014. https://www.healthit.gov/sites/default/files/hie-interoperability/nationwide-interoperability-roadmap-final-version-1.0.pdf. Accessed June 10, 2019.

2. Gliklich R, Dreyer N, Leavy M, eds. Registries for Evaluating Patient Outcomes: A User's Guide. Third edition. Two volumes. (Prepared by the Outcome DEcIDE Center [Outcome Sciences, Inc., a Quintiles company] under Contract No. 290 2005 00351 TO7.) AHRQ Publication No. 13(14)-EHC111. Rockville, MD: Agency for Healthcare Research and Quality. April 2014. http://www.effectivehealthcare.ahrq.gov.

3. Value Set Authority Center (VSAC). https://vsac.nlm.nih.gov/. Accessed June 10, 2019.

4. Next Steps to Encourage Adoption of Data Standards for Clinical Registries. The Pew Charitable Trusts. November 2016. https://www.pewtrusts.org/en/research-and-analysis/fact-sheets/2016/11/next-steps-to-encourage-adoption-of-data-standards-for-clinical-registries. Accessed June 10, 2019.

5. Fontanarosa J TJ, Samson DJ, VanderBeek BL, Schoelles, K. Retinal Prostheses in the Medicare Population. AHRQ Project ID: RPST0515. Rockville, MD: Agency for Healthcare Research and Quality; 2016.

6. Tunis SR, Clarke M, Gorst SL, et al. Improving the relevance and consistency of outcomes in comparative effectiveness research. J Comp Eff Res. 2016;5(2):193-205. PMID: 26930385. DOI: 10.2217/cer-2015-0007.

7. Curtis JR, Jain A, Askling J, et al. A comparison of patient characteristics and outcomes in selected European and U.S. rheumatoid arthritis registries. Semin Arthritis Rheum. 2010;40(1):2-14 e1. PMID: 20674669. DOI: 10.1016/j.semarthrit.2010.03.003.

8. Kirkham JJ, Clarke M, Williamson PR. A methodological approach for assessing the uptake of core outcome sets using ClinicalTrials.gov: findings from a review of randomised controlled trials of rheumatoid arthritis. BMJ. 2017;357:j2262. PMID: 28515234. DOI: 10.1136/bmj.j2262.

9. Maddox TM, Albert NM, Borden WB, et al. The Learning Healthcare System and Cardiovascular Care: A Scientific Statement From the American Heart Association. Circulation. 2017;135(14):e826-e57. PMID: 28254835. DOI: 10.1161/CIR.0000000000000480.

10. Kwok CS, Rao SV, Myint PK, et al. Major bleeding after percutaneous coronary intervention and risk of subsequent mortality: a systematic review and meta-analysis. Open Heart. 2014;1(1):e000021. PMID: 25332786. DOI: 10.1136/openhrt-2013-000021.

11. Steinhubl SR, Kastrati A, Berger PB. Variation in the definitions of bleeding in clinical trials of patients with acute coronary syndromes and undergoing percutaneous coronary interventions and its impact on the apparent safety of antithrombotic drugs. Am Heart J. 2007;154(1):3-11. PMID: 17584547. DOI: 10.1016/j.ahj.2007.04.009.

12. Thygesen K, Alpert JS, Jaffe AS, et al. Third universal definition of myocardial infarction. Journal of the American College of Cardiology. 2012;60(16):1581-98. PMID: 22958960. DOI: 10.1016/j.jacc.2012.08.001.

13. OMERACT. Outcome Measures in Rheumatology. https://omeract.org/. Accessed June 10, 2019.

14. Kirkham JJ, Davis K, Altman DG, et al. Core Outcome Set-STAndards for Development: The COS-STAD recommendations. PLoS Med. 2017;14(11):e1002447. PMID: 29145404. DOI: 10.1371/journal.pmed.1002447.

15. Kirkham JJ, Gorst S, Altman DG, et al. Core Outcome Set-STAndards for Reporting: The COS-STAR Statement. PLoS Med. 2016;13(10):e1002148. PMID: 27755541. DOI: 10.1371/journal.pmed.1002148.

16. The COMET (Core Outcome Measures in Effectiveness Trials) Initiative. http://www.comet-initiative.org/. Accessed June 4, 2019.

17. Gliklich RE, Leavy MB, Karl J, et al. A framework for creating standardized outcome measures for patient registries. J Comp Eff Res. 2014;3(5):473-80. PMID: 25350799. DOI: 10.2217/cer.14.38.

18. L&M Policy Research, LLC, Quintiles Outcome. Registry of Patient Registries Outcome Measures Framework: Literature Review Findings and Implications. OMF Literature Review Report. (Prepared under Contract No. 290-2014-00004-C.) AHRQ Publication No. 16-EHC036-EF. Rockville, MD: Agency for Healthcare Research and Quality; September 2016. www.effectivehealthcare.ahrq.gov/reports/final/cfm.

19. Gliklich RE BK, Eisenberg F, Hanna J, Leavy MB, Campion D, Christian JB. Registry of Patient Registries Outcome Measures Framework: Information Model Report. Methods Research Report. (Prepared by L&M Policy Research, LLC, under Contract No. 290-2014-00004-C.) AHRQ Publication No. 17(18)-EHC012-EF. Rockville, MD: Agency for Healthcare Research and Quality; February 2018. www.effectivehealthcare.ahrq.gov/reports/final/cfm. DOI: https://doi.org/10.23970/AHRQROPRMETHODS.

20. Calkins H, Gliklich RE, Leavy MB, et al. Harmonized outcome measures for use in atrial fibrillation patient registries and clinical practice: Endorsed by the Heart Rhythm Society Board of Trustees. Heart Rhythm. 2019;16(1):e3-e16. PMID: 30449519. DOI: 10.1016/j.hrthm.2018.09.021.

21. Gliklich RE, Castro M, Leavy MB, et al. Harmonized outcome measures for use in asthma patient registries and clinical practice. J Allergy Clin Immunol. 2019. PMID: 30857981. DOI: 10.1016/j.jaci.2019.02.025.

22. Leavy MB, Schur C, Kassamali FQ, Johnson ME, Sabharwal R, Wallace P, Gliklich RE. Development of Harmonized Outcome Measures for Use in Patient Registries and Clinical Practice: Methods and Lessons Learned. Final Report. (Prepared by L&M Policy Research, LLC under Contract No. 290-2014-00004-C) AHRQ Publication No. 19-EHC008-EF. Rockville, MD: Agency for Healthcare Research and Quality; February 2019. DOI: https://doi.org/10.23970/AHRQEPCLIBRARYFINALREPORT.

23. Kappetein AP, Head SJ, Genereux P, et al. Updated standardized endpoint definitions for transcatheter aortic valve implantation: the Valve Academic Research Consortium-2 consensus document. J Thorac Cardiovasc Surg. 2013;145(1):6-23. PMID: 23084102. DOI: 10.1016/j.jtcvs.2012.09.002.

24. eCQI Resource Center. https://ecqi.healthit.gov/. Accessed June 10, 2019.

25. Consolidated C-CDA. http://www.hl7.org/implement/standards/product_brief.cfm?product_id=408. Accessed June 10, 2019.

26. NIH Common Data Elements (CDE) Repository. National Library of Medicine. National Institutes of Health. https://cde.nlm.nih.gov/. Accessed June 4, 2019.

Chapter 4. Obtaining Data From Electronic Health Records

Authors (alphabetical)

Vera Ehrenstein, M.P.H., D.Sc.
Professor
Department of Clinical Epidemiology
Aarhus University Hospital

Hadi Kharrazi, M.H.I. M.D. Ph.D. (lead author)
Assistant Professor
Johns Hopkins School of Public Health
Department of Health Policy and Management
Johns Hopkins School of Medicine
Division of Health Sciences Informatics
Assistant Director
Center for Population Health IT

Harold Lehmann, M.D. Ph.D.
Associate Professor
Johns Hopkins School of Medicine

Casey Overby Taylor, Ph.D.
Assitant Professor of Medicine
Johns Hopkins University

Introduction

There is growing interest in using data captured in electronic health records (EHRs) for patient registries. Both EHRs and patient registries capture and use patient-level clinical information, but conceptually, they are designed for different purposes. A patient registry is defined as "an organized system that uses observational study methods to collect uniform data (clinical and other) to evaluate specified outcomes for a population defined by a particular disease, condition, or exposure and that serves one or more predetermined scientific, clinical, or policy purposes."[1]

An EHR is an electronic system used and maintained by healthcare systems to collect and store patients' medical information.[c] EHRs are used across clinical care and healthcare administration to capture a variety of medical information from individual patients over time, as well as to manage clinical workflows. EHRs contain different types of patient-level variables, such as demographics, diagnoses, problem lists, medications, vital signs, and laboratory data. According

[c] *EHRs are sometimes referred to as Electronic Medical Records (EMRs). This chapter uses both terms interchangeably.*

to the National Academies of Medicine, an EHR has multiple core functionalities, including the capture of health information, orders and results management, clinical decision support, health information exchange, electronic communication, patient support, administrative processes, and population health reporting.[2]

In summary, registries are patient-centered, purpose-driven, and designed to derive information on defined exposures and health outcome. In contrast, EHRs are visit-centered and transactional. Despite these differences, EHRs capture a wealth of data that is relevant to patient registries. EHRs also may assist in certain functions that a patient registry requires (e.g., data collection, data cleaning, data storage), and a registry may augment the value of the information collected in an EHR (e.g., comparative safety, effectiveness and value, population management, quality reporting).[3]

EHRs provide a unique opportunity for health systems to develop internal registries or contribute to external registries. Within a health system, registries are often developed by integrating registry functionalities with existing EHR platforms (i.e., EHR-integrated registries); however, these registries are limited to the health system's patient population and may be unable to capture longitudinal data from different provider settings. Registries that capture EHR data from multiple health systems typically interface with EHRs to receive data on an interval basis (i.e., EHR-linked or EHR-reported registries), although automating such efforts and creating a bidirectional exchange of information are still challenging.

The Meaningful Use program (see Chapter 1) has propelled the development of both EHR-linked and EHR-integrated registries. For example, EHR-integrated registries have expanded to meet EHR certification requirements and to help health systems meet requirements for workflow efficiency and quality improvement to achieve value-based criteria (e.g., improving population health). EHR-linked registries have grown as the Meaningful Use program specifically requires the reporting of EHR data to external registries (e.g., public health registries, quality reporting registries).[4] Meaningful Use Stage-1 provided an optional objective (which became a mandatory objective in Meaningful Use Stage-2) for eligible hospitals and professionals to submit EHR-extracted electronic data to immunization registries.[5] Meaningful Use Stage-2 further expanded EHR reporting to cancer registries and other specialized registries (e.g., birth defects, chronic diseases, and traumatic injury registries).[6]

Driven in large part by Meaningful Use, EHR vendors and clinical providers are incentivized to develop processes that would facilitate the design and launch of EHR-based registries in the United States. Yet, despite these incentives, the practice of using EHR-based registries is still relatively immature and, like all evolving research programs, faces many challenges.[7]

The purpose of this chapter is to describe the opportunities and challenges related to fully integrating or linking EHRs and patient registries. The chapter reviews common and emerging EHR data types that can be incorporated in registries, provides sample use cases of integrating EHRs and registries, and proposes a series of hypothetical technical architectures to link or integrate a registry with an EHR. The chapter closes with a discussion of possible future directions for EHR-registry integration. Key questions to consider when planning to incorporate data from EHRs as well as other sources are provided in Appendix B.

Common and Emerging EHR Data Types

EHRs provide various types of data that can be linked, integrated, or merged directly into a registry. The Meaningful Use program has led to the collection of a Common Clinical Data Set (CCDS) across most providers. These data are now generally available in EHRs; the data that are commonly available will likely continue to expand as Office of the National Coordinator, under the 21st Century Cures Act, moves toward building Core Data for Interoperability (USCDI) requirement.[8] EHRs can also provide data types of emerging interest to registries. Both types are described in Tables 4-1 and 4-2.

Table 4-1. Common data types of EHRs that can be integrated/interfaced with internal/external registries

DATA TYPE	EXAMPLE
Demographics	age, sex/gender, race
Diagnoses	diagnosis, severity, medical history
Problem List	active diagnosis, resolved diagnosis
Family History	familial disorders, risk factors
Allergies	food and medication allergy, anaphylaxis
Immunization	DTaP, HepB, IPV
Medications	Prescriptions written
Procedures	inpatient, outpatient
Lab Orders/Values	CBC results, HbA1C levels
Vital Signs	BMI (weight and height), blood pressure
Reports	radiology, pathology and other reports
Utilization	cost, hospitalization
Emerging Data Types	
Biosample Data	meta data about a biological sample
Genetic Information	genome sequence data
Social Data	income level, education, employment status
Patient-Generated	mHealth, patient communications
Community	community specifications
Geo-spatial	neighborhood built environment
Surveys	HRA, PHQ9, PROs
Free Text	various markers (e.g., specific test results)
Other Data Types	clinical workflow data

BMI = body mass index; CBC = Complete Blood Count; DTaP = Diphtheria, Tetanus, & acellular Pertussis; HbA1c = Hemoglobin A1c; HepB = Hepatitis B; HRA = health risk assessment; IPV = Inactivated poliovirus; PHQ = patient health questionnaire; PRO = Patient Reported Outcome

Table 4-2. Emerging data types of EHRs that can be integrated/interfaced with internal/external registries

DATA TYPE	EXAMPLE
Biosample Data	age, sex/gender, race
Genetic Information	diagnosis, severity, medical history
Social Data	active diagnosis, resolved diagnosis
Patient-Generated	familial disorders, risk factors
Community	food and medication allergy, anaphylaxis
Geo-spatial	DTaP, HepB, IPV
Surveys	Prescriptions written
Free Text	inpatient, outpatient
Other Data Types	CBC results, HbA1C levels

CBC = Complete Blood Count; DTaP = Diphtheria, Tetanus, & acellular Pertussis; HbA1c = Hemoglobin A1c; HepB = Hepatitis B; IPV = Inactivated

In addition to these data, EHRs capture a considerable amount of unstructured data (e.g., clinical notes) that can be further processed to extract specific data of importance to a registry (e.g., specific information extracted from radiology reports to determine eligibility).

Data types commonly extracted from EHRs and imported into registries are patient identifiers, demographics, diagnoses, medications, procedures, laboratory results, vital signs, and utilization events. These are discussed further below.

Patient Identifiers

EHRs are designed to facilitate the identification of individual patients in clinical workflows. Patient identifiers include patient's full name, date of birth, contact information such as address and phone numbers, name and contact information of the next of kin, emergency contact information, and other personal information deemed necessary for healthcare delivery operations (e.g., employer information, insurance information). For internal operations, EHRs generate a unique patient ID (i.e., medical record number) that is used within the care setting to identity a specific patient. Organizations that provide care at multiple facilities (e.g., a health system with multiple hospitals and outpatient facilities) often have a second patient identifier that can be used to find a patient across the entire health network (i.e., master patient record). If a health system is connected to a statewide or regional health information exchange (HIE), the EHR may include a third patient identifier that has been issued by the HIE (i.e., statewide master patient index).[9]

Conditional to receiving proper consents and adhering to Health Insurance Portability and Accountability Act (HIPAA) policies,[10] patient identifiers stored in EHRs can be used to merge patient EHR records with a patient registry. For example, a registry may collaborate with a statewide HIE to locate the master patient indexes of all registry patients and then ask multiple providers to locate the EHR records of those individuals using the HIE-issued patient master indexes. However, many registries do not have the option of acquiring master patient indexes from an HIE. These registries typically use alternative methods for matching patient identifiers and importing EHR data. Potential mistakes in matching registry patients with EHR patients may lead to quality issues such as incomplete or inaccurate data.

Demographics

EHRs generally contain patient demographic information such as age, gender, and ethnicity/race. These data are needed for clinical operations and are mandated by the Meaningful Use objectives. The quality of data on age and gender is often acceptable because of the various mandates to collect them accurately.[11-13] However, the quality of demographics data may be affected by other factors including mode of measurement, user mistakes, and data conversion issues.[14] EHRs often have a moderate to high missing data rate for non-essential demographic information such as income, marital status, education, employment status, and nationality.[15, 16]

Coding standards for demographic data have been published but are not always used. Demographic data such as education and nationality are often not coded in a standardized approach. Age data are governed by HIPAA and have sharing limitations if they contain a certain level of granularity (e.g., age represented by the exact date of birth or if ages above a certain limit).[17] Demographic data are often used by registries to match patient records across data sources. Thus, legal limitations to sharing demographic data may hinder the development of multi-source/multi-site EHR-based registries that require demographic data for these purposes.

Diagnoses

Diagnosis often is a key variable to evaluate a patient for inclusion in a registry. The quality of diagnosis data is often acceptable, in part due to various mandates to collect these data accurately.[10-12] EHRs also include problem lists as a way to capture active versus non-active diagnoses, but the quality of data found in problem lists may need further validation.

Some established vocabulary standards are available to encode diagnosis data. These include the International Classification of Diseases (ICD),[18] International Classification of Primary Care (ICPC),[19] Systematized Nomenclature of Medicine (SNOMED),[20] Diagnostic and Statistical Manual of Mental Disorders (DSM),[21] and Read Codes.[22] In the U.S., ICD is the most commonly used system to capture diagnostic data in both EHRs and registries. Mapping diagnostic data from one coding system to another is challenging; even mapping diagnoses from one version of a coding system to another version is difficult (e.g., mapping ICD-9 to ICD-10). In addition, certain diagnostic codes – such as HIV status and mental illness diagnoses – are protected by various federal and state-level laws[23] that may limit the ability to extract these codes for use in external registries.

Medications

In addition to diagnosis, registries often use medication data as eligibility criteria. Many registries also capture medication data to study treatment effect and/or safety. EHRs contain information on prescriptions that are written, while pharmacy claims data contain information on prescriptions that were filled. When EHR medication data are coupled with pharmacy claims data, a number of important constructs, such as medication adherence and reconciliation rates (e.g., medication regimen complexity index)[24] can be derived and reported to a registry.

The quality of EHR medication data is often acceptable due to various mandates to collect medication data in EHRs. Common vocabulary standards for medications include National Drug Codes (NDCs),[25] RxNorm,[26] Systematized Nomenclature of Medicine's (SNOMED)[20] Chemical axis, Anatomical Therapeutic Chemical Classification System (ATC),[27] and a number of commercial drug codes such as MediSpan®, Multum®, Generic Product Identifier® (GPI), and First Databank® (FDB). Each coding standard addresses different aspects of a medication (e.g. drug class, ingredients, dosage).

Potential semantic interoperability issues may arise when medication data are combined from multiple sources and mapped from one coding system to another. For example, an RxNorm code (drug class) may map to multiple NDC codes (packaged drug). Furthermore, some EHR-derived medication information may not be specific enough for research purposes (e.g., data on generics, like biosimilars, generally do not reflect which generic product was supplied to the patient).

Procedures

Procedure data include clinical procedures such as surgery, radiology, pathology, and laboratory. Procedure data can be extracted directly from EHRs and reported to registries; however, procedures reported from one EHR generally only include those procedures taking place within the premises of a provider using the same EHR and may not include procedures that occurred elsewhere.

Vocabulary standards for procedures include International Classification of Diseases' Clinical Modification (ICD-CM),[18] Current Procedural Terminology (CPT)[28] and Healthcare Common Procedure Coding System (HCPCS).[29] Each coding system is designed to capture procedures within a specific clinical context (e.g., primary care, hospital facility). EHR-based procedure data may not have the level of detail necessary for a registry (e.g., techniques used in a clinical procedure such as a surgical process). These procedure nuances are often entered as unstructured data that usually do not accompany structured EHR-extracts for registries.

Laboratory Data

Currently, the best sources of laboratory data are the information systems used by standalone laboratories, which are frequently but not always incorporated into the EHR. Laboratory data include both lab orders and lab results. Coding standards for lab orders and lab results include the Logical Observation Identifiers Names and Codes (LOINC),[30] the Systematized Nomenclature of Medicine (SNOMED),[20] and the Current Procedural Terminology (CPT).[28] Currently, there are no mandated laboratory coding system for certified EHRs, and the majority of healthcare providers rely on local coding systems for lab orders/results. This limits the interoperability of multi-site EHR-derived lab data for registries.

In addition, different healthcare facilities may use different laboratory tests to measure the same analyte, each of which has a different laboratory code. Discussion is needed across the provider network on how to link lab items, preferably using automated tools and not a manual process, so that a single query across the network will return all the desired data from multiple EHRs for a single registry. In addition, certain lab results are protected by federal and state laws (e.g., lab tests revealing HIV status) and thus might be missing from EHR-extracts reporting to external registries. Further, some laboratory data are accessible to clinicians without incorporation into the EHR; in fact, some lab data require active steps by the clinician to import into the EHR. Inaccurate interpretations may be made without understanding why some lab data are missing from an EHR.

Vital Signs

EHRs are a primary source of vital sign data. Vital sign data include physiological variables such as height, weight, body mass index, pulse rate, blood pressure, respiratory rate, and temperature. LOINC is the common coding standard for vital signs. Most provider organization, however, do not actively use LOINC codes to capture vital signs in their EHRs as it is not mandated by the Meaningful Use program.

The completeness of EHR-derived vital signs such as height and weight is often acceptable for use in registries. Issues with human errors and units of measurement may affect data quality; thus, data cleaning is essential before use for registries.[31] For example, weight and height data may include incorrect units (e.g., pounds reported as kilograms). EHR also may lack proper meta-data that are important for the clinical interpretation of the data (e.g., sitting versus standing blood pressure measurements).

Utilization/Cost

Utilization data can be extracted from EHRs especially when insurance claims data are not available. Note that EHR-level utilization data are limited to events that have occurred within a particular provider’s facilities and often do not contain utilization data from other providers. Utilization can be defined as cost, hospitalization, readmission, emergency room admission or other significant healthcare events. The quality and completeness of utilization data are often acceptable due to reimbursement guidelines.[12]

There are no specific standard utilization coding terminologies for EHRs; however, most EHRs adhere to the utilization guidelines of claims submission policies. A number of reimbursement policies recommend specific reference-coding systems to encode utilization events. Certain utilization events are protected by various federal and state-level laws (e.g., mental health visit), and a registry may not receive utilization data related to those conditions from an EHR.

Surveys

Survey data are usually collected from self-reported questionnaires; however, clinical data captured by surveys are increasingly stored within EHRs for various purposes. Some EHRs provide standardized surveys that can be accessed via patient portals to capture patient reported outcomes or symptoms (i.e., Patient-Reported Outcomes Measurement Information System or PROMIS).[32] Risk factors and self-reported behaviors often are important to registries, and such data can be derived from EHR-integrated surveys (e.g., smoking status, socioeconomic status, housing condition). Also, registries may add and integrate their own customized questionnaires in EHRs so that patients can directly enter the necessary information needed for a registry (e.g., determine eligibility; collect additional data for a study).

EHR-integrated surveys are prone to sampling, selection, response, and social-desirability biases. The quality of EHR-integrated survey data varies considerably depending on the questionnaire, and the validity and reliability of custom-built EHR-integrated surveys are often difficult to measure in the context of a clinical practice.

Surveys cover variable domains and often do not adhere to coding standards. Indeed, surveys measuring the same concept may code their variables differently. One approach to reduce bias and error in survey-collected EHR data is to use standardized questionnaires across EHRs and healthcare providers. Some of the many standardized questionnaires include the Patient Reported Outcomes Measures (PROMs), Patient Health Questionnaires (PHQ), Health Risk Assessments (HRA), Life Event Checklist (LEC), and Generalized Anxiety Disorder (GAD) screening tools.

Social Data

Social data include variables ranging from individual-level factors to community-level elements (e.g., smoking status, socio-economic status, housing condition). Social variables are often considered important factors in registries as these variables enable researchers to understand the underlying social context and potential disparities associated with the outcome of interest. As an example, social data captured within a registry can be used to assess treatment affordability or understand heterogeneity of treatment effects. Although increasingly recognized as important variables, social and behavioral data are not routinely captured in EHRs.[14] EHR-derived social data are often incomplete and limited to a few data types.[33] Moreover, social determinants of health that could be imported from data sources such as social services organizations are usually missing in EHRs and registries due to the lack of interoperability.[34]

Although a number of coding standards have been proposed to standardize social data, most EHRs use proprietary coding vocabularies. Social data are often of low quality, mainly due to incomplete survey responses and the subjective nature of many social questions. Although most social data are not subject to HIPAA, they can still be subject to other privacy rules such as the Family Education Rights and Privacy Act (FERPA).[35] Establishing linkages among patient-level EHR records, social service records, and registries has faced both technical and regulatory challenges in the past.

Patient-Generated Data

Patient-generated data can include a wide array of variables (e.g., physical activity, sleep patterns, self-reported sign and symptoms, uploaded blood sugar levels) and may be captured within an EHR through various means (e.g., integrated personal health records, mobile-health exchange platforms, wearable device interfaces).[36] EHR-based patient-generated data are highly customized and inconsistent across EHRs. Standards are becoming more available for mobile health and wearables devices,[37] but have not yet been widely adopted for patient-generated data captured within EHRs. Although the quality of the data collected by mobile health and wearable devices is improving, accuracy and comparability are still challenging when such data are collected using different devices. Self-entered data collected via surveys (e.g., entering physical activity types) are subject to a variety of selection factors and errors (e.g., overestimating recall of time spent exercising). Data interoperability may become more challenging as more non-standardized devices enter the market. Additionally, consenting processes via internet and mobile health solutions may be complex, and the creation of large EHR-integrated registries using patient-generated data requires careful attention to legal and regulatory issues.[38, 39]

Sample Use Cases and Architecture of EHR-Based Registries

Registries that incorporate EHR data may use a variety of IT system architectures. Registry architects must consider the number of participating sites (single-site or multi-site), variety of underlying EHRs (one enterprise-level EHR, multiple EHR installations of the same vendor, multiple EHRs from different vendors), existence and connectivity to Health Information Exchanges (HIEs) (centralized, federated or distributed), and other factors that affect interoperability.

Following are examples of three “hypothetical” EHR-based registry types, each with a different combination of stakeholders and IT infrastructures (Table 4-2). Registries designed to support clinical care are often based on single enterprise-level EHRs, while registries designed for research are often hosted external to EHRs but may receive EHR extracts from multiple sources. Public health registries, similar to registries designed for research, are often hosted by health departments outside of a single EHR environment but receive EHR reports on a regular basis. Note, these are generalized examples; actual IT infrastructure and features may vary.

Table 4-3. IT infrastructure and other features of sample registry types using EHR data

TYPE/SPECS	REGISTRY TO SUPPORT CLINICAL CARE	REGISTRY DESIGNED FOR RESEARCH	PUBLIC HEALTH REGISTRY
Scope	Depending on the provider network's size: Local City State Regional National (e.g., VHA)	Depending on the research aim/goal: State Regional National International	Depending on the public health authority: City County State National/Federal
Stakeholders	providers (usually within the network) biopharmaceutical and medical device companies [optional] employers [optional] payers [optional]	research institutes government (local, state or federal) non-profit organization disease associations biopharmaceutical and medical device companies employers (e.g., professional sports)	government (local, city, county, state or federal)
Sources of Data	mainly EHRs sometimes EHR-based patient portals sometimes merged insurance claims infrequently EHR-embedded surveys	surveys custom EHR extracts other data sources (e.g., biobanks, administrative health insurance claims, eCRFs, other registries)	surveys EHRs other data sources (e.g., HIEs, environmental, health department, surveillance systems)
Number of EHRs	Usually one enterprise-level EHR	Multiple/various EHRs	Multiple/various EHRs
EHR Interoperability Requirement	low	high	medium
EHR Integration Level	usually automated	sometimes manual frequently semi-automated rarely fully automated	automated for some (e.g., immunization records) manual for others (e.g., non-MU registries)
EHR Integration Tools	EHR built-in tools sometimes custom-built APIs / extracts	custom-built APIs / extracts	EHR build-in tools for select reporting (MU program) Custom-built tools/APIs for non-MU registries
System Dependency	EHR-based	May not be EHR-based	not EHR-based
Common Architecture	Often Centralized	Centralized Distributed Federated	Centralized Distributed Federated

TYPE/SPECS	REGISTRY TO SUPPORT CLINICAL CARE	REGISTRY DESIGNED FOR RESEARCH	PUBLIC HEALTH REGISTRY
Dominant Hosting DB	EHR-embedded (e.g., EHR registry data warehouse)	Could be distributed and likely to include non-EHR DB (i.e., research-specific data collection)	non-EHR DB (i.e., public health database)
Alternative Names	clinical quality registry improvement/measure registry chronic disease management registry high risk [population] registry	product or disease registry clinical research registry research network registry	outbreak registries vaccination registries disease surveillance (e.g., cancer)
Typical Functions	clinical workflow management disease/cohort management (e.g., care coordination) population health management (e.g., case management)	evidence for effectiveness, comparative effectiveness, safety and/or value for clinicians, patients and payers natural history of disease studies	public health services outbreak surveillance syndromic surveillance epidemiological research biopharmaceutical research, e.g., vaccine effectiveness and safety
Timeliness	usually real-time sometimes periodic (daily extracts)	mostly periodic (daily, weekly or monthly extracts) sometimes real-time	sometimes real-time sometimes periodic (daily, weekly or monthly extracts)
Scalability	limited to individual EHR vendors	depends on registry architectures adopting EHR interoperability standards	depends on adopting interoperability standard and future stages of MU for public health reporting

API = Application Programming Interface (see Chapter 5); DB = Database; eCRF = electronic Case Report Forms; HER = Electronic Health Record; HIE = Health Information Exchange (see Chapter 2); MU = Meaningful Use (see Chapter 1); VHA = Veteran Health Affairs

In a fully interoperable ecosystem, registry-specific functionality could be presented in a software-as-a-service or middleware model, interacting with the EHR as the presentation layer on one end and the registry database on the other.[3] In this ideal model, the EHR is a gateway to multiple registries and clinical research activities through an open architecture that leverages best-in-class functionality and connectivity. Full interoperability would enable registries to interact across multiple EHRs, and EHRs to interact with multiple registries. Comprehensive interoperability, however, has not yet been realized, and customized IT architectures are required to facilitate the integration and interfacing of EHRs with registries.[3] The following are examples of IT architectures that could support EHR-integrated/linked registries for clinical operations, research projects, and public health missions.

EHR-Integrated Registries To Support Clinical Care

Healthcare providers often develop and manage EHR-based registries that are used to support clinical care and meet operational goals (referred to here as ‘clinical registries’). To develop clinical registries, providers typically use EHR-based tools that are developed by EHR vendors. These EHR-based registries can facilitate clinical workflow, monitor quality metrics, enable disease/cohort management, and offer population health management features. In particular, the

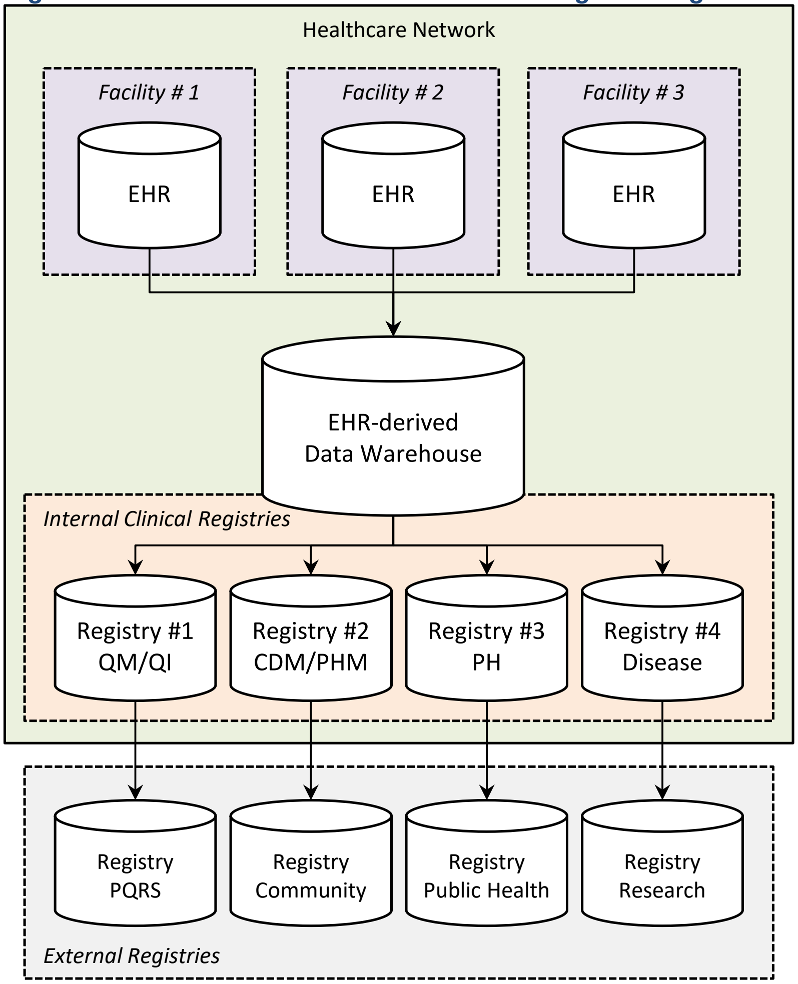
Triple Aim of care, health and cost has provided a framework to achieve value-based care while reducing cost.[40] This framework promotes ‘population health’ while enhancing the individual’s experience of care and lowering cost.[41] Effective population health management is essential to ensuring that resources are directed towards improving health outcomes of patients at the highest risk for developing undesired outcomes. The notion of population health management necessitated that health providers develop EHR-based registries to focus on high-risk subpopulations (e.g., patients at high risk for mortality and morbidity, cost, hospital and emergency room admission or who have a chronic condition that requires direct management, such as diabetes).[42, 43]

A major challenge with EHR-integrated clinical registries is the lack of out-of-network data in a health network’s EHR.[44] In other words, data generated during patient encounters with out-of-network providers, who may not be using the same EHR, will be missed in the registry resulting in incomplete and sometimes outdated data. Individual health networks often complement their EHR data with insurance claims to generate a more complete picture of a patient’s health status; however, use of insurance claims is not always practical given that a large patient population of a health delivery network may use dozens, if not hundreds, of different insurers. Many challenges of EHR-based population health registries are derived from the overarching challenges within the broader domain of population health informatics.[45]

Clinical registries usually use a centralized architecture and often have an EHR data warehouse as their backbone along with multiple data marts containing various registry data. The centralized architecture accumulates and manages data in a single and centralized repository. The advantages of a centralized model are: simplicity and efficiency; greater data consistency; and easier patient linkage if the same patient identifiers are used across the healthcare network. Potential disadvantages of a centralized model include: data capture that is limited to users of a single EHR vendor across the healthcare network (e.g., trouble with integrating a different EHR vendor if a new facility joins the network); and difficult data exchange with registries developed by other networks due to a lack of interoperability.

Healthcare networks often develop clinical registries based on their underlying enterprise-wide EHR architecture (Figure 4-1). Data collected at different facilities of a healthcare delivery network (e.g., hospitals and outpatient clinics) are aggregated in a common data repository such as an EHR’s data warehouse. Facilities not using the same EHR platform face extra work to harmonize and standardize their data before feeding it into the data warehouse. Data warehouses can be used to develop multiple data marts feeding into various registries for different purposes such as quality measures, disease management, population health management, and public health reporting. Internal clinical registries are sometimes linked to external registries for reporting purposes (e.g., PQRS reporting),[46] although interoperability challenges may limit such exchanges.

Figure 4-1. Common architecture of EHR-integrated registries to support clinical care*



CDM = Chronic Disease Management; HER = Electronic Health Record; PH = Public Health; PHM = Population Health Management; PQRS = Physician Quality Reporting System; QI = Quality Improvement; QM = Quality Measure

**Harmonization, standardization, and data quality control will be executed at the data warehouse level (not shown in the diagram as a separate component)*

EHR-Linked Registries Designed for Research

Registries designed for research purposes (referred to here as ‘research registries’) may use EHR data on a variety of levels. At the low end, research registries may use EHR data to identify and enroll eligible patients into studies that use supplementary registry-specific data collection. In this scenario, EHR data are used to identify eligible patients (based on the registry’s inclusion and exclusion criteria), and minimal EHR data (e.g., family history of breast cancer) are imported into the registry. The remaining registry-specific data are captured through another means, usually a dedicated data repository that allows for entry of eCRFs and web-based survey forms. On the other end of the spectrum, some research registries have been built entirely using EHR data (e.g., California Cancer registry).[47] Many other research registries use a combination of self-reported and EHR data (e.g., Autism treatment network).[47] Registries in which EHR-based extracts are merged with registry data on a periodic basis are referred to here as EHR-linked registries.

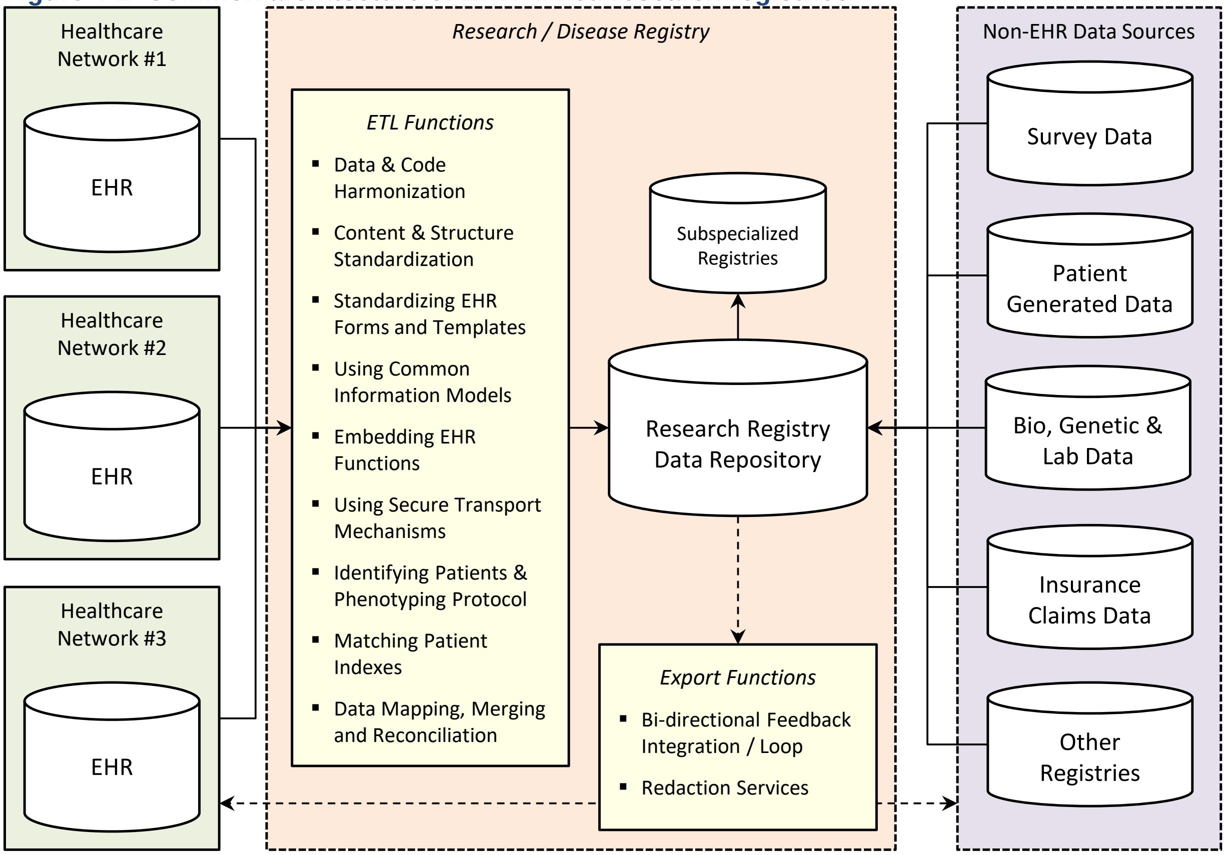
The increasing semantic and syntactic interoperability among healthcare providers is a major driver for EHR-linked registries. EHR-linked research registries often use application programing interfaces hosted by healthcare providers to extract and share standardized EHR data and then use semi-automated approaches to merge the EHR data with existing registry records. Moreover, bi-directionally interoperable EHR-linked registries may also serve an important role by delivering relevant information from a registry back to a clinician (e.g., natural history of disease, safety, effectiveness, and quality).

EHR-linked research registries collect EHR data using a variety of mechanisms, ranging from automated EHR-embedded push protocols to manual ad-hoc EHR-database pulls. Triggers for EHR data extraction include standardized protocols that follow the inclusion and exclusion criteria of the research registry (i.e., phenotyping queries; retrieve protocols). After receiving the EHR data, research registries use a multi-phase process to import incoming EHR data (Figure 4-2). Extract, transform, and load functions may include data curation activities such as data preparation, data standardization, secure data transfer, data mapping, data redaction, data integration/merging, and data reconciliation. Various organizations such as the Clinical Data Interchange Standards Consortium (CDISC) and the Standards and Interoperability (S&I) Framework have introduced detailed mechanisms to automate and standardize the incorporation of EHR data for other purposes including registries (e.g., CDISC Link Initiative[48]). Additionally, the growing number of common data models have enabled registry developers to adhere to specific predefined standards that facilitate integration of EHR-based data as well as data sharing among registries (e.g., Clinical Information Modeling Initiative's (CIMI) Reference Model,[49] FDA Sentinel Initiative,[50] and Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM)).[51] Chapter 5 describes common data models in more detail.

Importing and merging data from EHRs into research registries is challenging. Automating the data imports requires high degrees of interoperability, data curation, and post-hoc harmonization as well as attention to data quality. For example, if inclusion criteria are encoded differently in different EHRs, the comparability of data may be impacted, creating artificial distortion between outcomes measured by different EHRs.[52] Merging EHR data-imports with existing patient data in a registry also requires reliable master patient indexing to avoid inaccurate patient-matching which would compromise any inferences drawn from the data.[9, 53] Data curation is critical, as integration of EHR data can expose data quality issues that may affect research findings.[54]

Data governance must be considered as well. Registries designed for research may be funded and managed by a broad range of organizations (e.g., federal, state, non-profit, private). Although patient privacy is safeguarded and protected under federal and state laws,[55] data governance policies vary, resulting in different barriers for different registries when importing and integrating EHR data.[56] Additionally, the incentives and liabilities associated with extracting and pushing data from an EHR to an internal or external registry are not always clear for healthcare providers.[57]

Figure 4-2. Common architecture of EHR-linked research registries*


Healthcare Network #1
EHR
Healthcare Network #2
EHR
Healthcare Network #3
EHR
Research / Disease Registry
ETL Functions
Data & Code Harmonization
Content & Structure Standardization
Standardizing EHR Forms and Templates
Using Common Information Models
Embedding EHR Functions
Using Secure Transport Mechanisms
Identifying Patients & Phenotyping Protocol
Matching Patient Indexes
Data Mapping, Merging and Reconciliation
Subspecialized Registries
Research Registry Data Repository
Export Functions
Bi-directional Feedback Integration / Loop
Redaction Services
Non-EHR Data Sources
Survey Data
Patient Generated Data
Bio, Genetic & Lab Data
Insurance Claims Data
Other Registries

HER = Electronic Health Record; ETL = Export, Transform, and Load

**Harmonization, standardization, and data quality control will be executed during the ETL process (not shown in the diagram as a separate component)*

EHR-Linked Public Health Registries

Public health agencies have long used registries for surveillance and tracking purposes. For example, local and state public health departments usually maintain immunization registries that receive information from clinicians and other entities such as schools and pharmacies. Other common public health registries include syndromic surveillance and specialized registries such as birth defects, chronic diseases, and traumatic injury registries. In recent years, coincident with the rising EHR adoption among providers, public health entities began to link various registries with EHRs. A significant driver of increased EHR integration has been the Meaningful Use program, which incentivized clinicians to share EHR immunization and syndromic surveillance data with public health agencies.[7] Other drivers have included the maturation of data standards (both semantic and syntactic) for automating and improving the transmission of EHR data to public health registries (e.g., distributed population queries),[58] and the increased interest of value-based care provider organizations in assessing the needs and improving the health of the communities they serve (e.g., community health needs assessment).[59] Most EHR-linked public health registries have relied on semi-automated processes; only recently have more automated mechanisms been introduced and adopted (e.g., vaccination registries). EHR-linked public health registries follow a similar architecture to that of EHR-linked research registries (Figure 4-2); however, the methods used to collect data from EHRs may vary as not all public health registries require patient-level data (e.g., counts are sufficient for some purposes). Methods used include but are not limited to: (1) semi-automated forms/templates to collect public health specific information about patients that fit a certain criteria (e.g., S&I Framework SDC);[60] (2) data exchange protocols for receiving case reports from certified EHRs (e.g., MU public health reporting objectives);[7] (3) tools to mine EHR and HIE data for signs and symptoms relevant to public health emergencies and outbreaks (e.g., ESSENCE Syndromic Surveillance System);[61] and, (4) distributed data network queries to collect aggregated data from multiple providers when the identity of patients is not relevant (e.g., PopMedNet).[62]

Some public health agencies have been able to directly integrate their registries with the EHRs of clinicians who provide care in their jurisdiction. The prime example of such a fully-integrated EHR-linked public health registry is the New York City (NYC) Population Health Registry.[63] This registry collects information from NYC's eligible healthcare professionals across several domains (e.g., Influenza-like-Illnesses). The NYC's Population Health Registry has been successful as most eligible professionals in NYC use the same EHR system, one which is capable of reporting data in real-time to local public health agencies. The Population Health Registry is part of NYC Macroscope Hub,[64] a surveillance system for tracking conditions managed by primary care practices (e.g., obesity, diabetes, hypertension, and smoking).

Technical Issues and Operational Challenges of EHR-Based Registries

EHR-based registries fulfill different purposes and use different IT system architectures, but many technical issues and operational challenges are common across the range of registries. This section describes several common challenges, such as identification of eligible patients; data quality; unstructured data; interoperability; data sharing and patient privacy; data access and patient privacy; and human resources.

Identifying Eligible Patients

Retrieval protocols and phenotyping methods are commonly applied against EHR data to define the denominator of interest and identify eligible patients for screening, clinical trials, and inclusion in registries.[52] Computational phenotyping involves operationalizing process, outcome and case definitions as a set of measures that can be captured during regular episodes of clinical care and that are stored in the EHR. General categories of data that are drawn for computational phenotyping from EHRs include medications, laboratory tests, and diagnoses.[52] Operationalized definitions can be used for a number of applications including cohort screening and identification to enable clinical research; assessments of current healthcare delivery processes and outcomes; and, changes due to new healthcare practices and interventions. Common for any of these applications is a need to evaluate the operational definitions that are used. Given that EHR data are collected for the purpose of documentation and are collected at various points in time for each patient, there are a number of opportunities for potential biases to arise and for data to be missing. As such, a sound evaluation of the measurement approach is required prior to the use of those measures for secondary analyses of cohort screening and identification. To date, there have been a significant number of studies requiring cohort identification that report common measures such as positive and negative predictive values, sensitivity, and specificity prior to conducting downstream analyses. The evaluation of measure results depends in part on the intended use of an operational definition and EHR data source(s). Some frameworks have been developed to assist investigators in characterizing potential limitations to the use of operational definitions with EHR registry data so that when analyses are performed the confidence level of those findings can be quantified.[52, 65]

Various challenges with denominator and variables selections exist when extracting data from EHRs for registries. Ambiguous phenotyping algorithms and lack of standardized retrieval protocols often result in selecting a denominator of patients from an EHR that is irrelevant, skewed, or biased for a registry. Multiple factors can be used to modify and refine the definition of a population denominator (e.g., age, gender, diagnoses, medications, lab results, radiologic findings, special conditions such as disability, and administrative information such as insurance coverage). Selecting the timeframes of the EHR data extract is also complex and may result in incomplete temporal data represented in registries. Despite the higher interoperability of EHR data and standardization of phenotyping protocols, fine details of EHR data may affect the selection results. Some of the challenges include:

- *Process of Care*: different providers or clinical workflows generate different data values for the same event or fact; hence, the same fact or event might be represented differently in the same EHR.
- *Nature of Intervention*: different interventions with different levels of risk may be encoded similarly, meaning EHR does not contain the true risk factors for those interventions.[43, 66]

Data Quality

As a basic good practice, registries should use some form of data curation to review and assess data quality. In the context of EHR-based registries, data quality issues stem from the fact that data extracted from EHRs often requires extensive cleaning and preparation before being imported into registries. EHRs are designed to manage the transaction of healthcare and support clinical workflow and documentation for billing. The purpose of an EHR is not to conduct research, and EHRs are not designed to systemically collect research-grade longitudinal data. As a result, data captured by EHRs are of variable quality.[14, 45] For example, EHRs often house reliable laboratory and medication data for clinical purposes, but EHRs typically lack consistent and sufficiently detailed data on risk factors, levels of education, or socioeconomic status.[16] The quality of source data can affect both the underlying data as represented in a registry and the results generated using such data. Thus, EHR data may not be appropriate for some research purposes.

Data quality can be defined in various perspectives. The most impactful aspects of data quality for registries are:[14]

- *Accuracy:* the extent to which data captured in EHR accurately reflects the state of interest, which is often complex to measure because the true value of a given variable remains unknown.
- *Completeness*: the level of missing data for a particular data element in the EHR for the population of interest; this is commonly measured as a data quality indicator for EHR-integrated registries. It is important to note that for research purposes, a distinction is made between “must-have’ and “nice-to have” data, recognizing that completeness of “must-have” data is most important.
- *Timeliness:* the length of time between the initial capture of a value and the time the value becomes available in the EHR.

It is important to note that data quality varies across EHRs used by different healthcare organizations. Moreover, changes may be made to EHR systems “behind the scenes” that affect data quality. For example, upgrades intended to improve performance or add features may inadvertently result in poor record linkage or may require updating record extraction protocols. Evaluating data quality, completeness and accuracy should be conducted as an on-going process and not a one-time exercise.

Unstructured Data

EHRs contain a considerable amount of unstructured data, such as progress notes. The loosely structured nature of typed text (also known as ‘free text’) is effective in day-to-day clinical workflows but presents a major challenge for automating EHR-based registries. The unstructured data may contain key patient information missing in structured data, extra information complementing structured data, or even data that may contradict information represented by structured data. The complexities of unstructured data, along with the fact that existing text mining tools and natural language processing applications have limited accuracy in extracting information from free text,[67] have prompted some registries to ask for a manual chart review of individual patients before final inclusion in the registry. Unstructured data limits the application of automated computational phenotyping methods and increases the likelihood of low data quality (e.g., missing data) when data are extracted from structured EHR data only.

Many EHRs also allow a choice of places where important data may be entered. For example, some EHR have been set up to facilitate quick entry of “easy treatments” that then results in fragmented storage of treatment information. Treatment information may also be buried in clinical notes, which may not be accessible for research purposes since notes often include a patient’s name and other personally identifiable information that can be difficult to spot and redact systematically.

Interoperability

Interoperability is defined as the ability of a system to exchange electronic health information with, and use electronic health information from other systems without special effort on the part of the user.[68] Interoperability requires multiple stages, 'sending', 'receiving', 'finding' and eventually 'using' the data.[68] As discussed in Chapter 1, interoperability spans multiple dimensions of standards: regulatory, contractual, privacy, exchange formats, content, and technology.[68, 69] In the context of EHRs and registries, syntactic interoperability is the ability of heterogeneous health information systems to exchange data with a registry, and semantic interoperability implies that the registry understands the data exchanged at the level of defined domain concepts.

From an EHR/registry perspective, functional interoperability could be described as a standards-based solution that achieves the following set of requirements: "The ability of any EHR to exchange valid and useful information with any registry, on behalf of any willing provider, at any time, in a manner that improves the efficiency of registry participation for the provider and the patient, and does not require significant customization to the EHR or the registry system."[3]

Although interoperability of EHRs with other EHRs and health IT systems has increased over the last decade,[70] most health systems do not share in-depth EHR-level data with other health systems. Lack of interoperability is a major limiting factor for the extraction, integration, and linkage of EHR data for registries. Most EHRs are not fully interoperable in the core functions that would enable them to participate in various registries without a significant effort.[3] This deficiency is directly related to a combination of technical and economic barriers to EHRs' adoption and deployment of standards-based interoperability solutions.[3] EHR vendors also provide heavily customized versions of their own systems for each client thus creating additional barriers to interoperability.[3] Since registries seek data across large and generalizable populations, making EHRs interoperable across providers is a key step in facilitating EHR-based registry efforts.

Data sharing and interoperability challenges are not limited to incoming EHR data for a registry. In a learning health system, a bidirectional registry shares its findings with providers that have shared their EHR data. In such a reciprocal model, the findings are turned into knowledge and can effectively be used to change the delivery of care and improve outcomes across all participating providers. Currently, there are no common standards on how to distribute registry findings while protecting the identity of individual healthcare providers. Sharing the findings about data quality issues with data providers is challenging as well as it may result in legal ramifications (e.g., individual providers might become liable when data is captured inaccurately).

Linking and integrating various EHR data sources for registries also requires matching patients across databases. HIEs are sometimes required to generate master patient indexes (MPIs) to match patients across diverse EHR data sources. Developing and utilizing an MPI is a complex process and may introduce error and bias in registries despite many tools being available to accomplish this process.[9] It is worth noting that most of the data elements needed to create MPI are considered protected health information according to HIPAA regulations and may not be available for registries to complete the matching process.

EHR Infrastructure and Deployment

EHRs may provide IT infrastructure and tools to support the development of an EHR-based registry, but they typically do not provide turnkey solutions for functional registries. Over the last decade, a variety of EHR tools have been developed that could form the building blocks of EHR-based registries. For example, EHR-based clinical data warehouses collect and store EHR data across an entire health network. These system-wide data warehouses often serve as the backbone of data products that eventually support an EHR-integrated registry (see Chapter 2). However, challenges with updating, maintaining, scaling, and sharing such tools across healthcare providers still hinders development of registries.

In addition, the architecture of an EHR deployment within a healthcare delivery system may influence the usefulness of EHR for different registry applications. For example, a health system that lacks an enterprise-level EHR architecture may find it challenging to develop a system-wide EHR-integrated registry, as each of its entities operates a standalone EHR with no interoperable solution to share data among them.

Data Access, Privacy, and Use

Data access and privacy challenges are complex in multi-site EHR-based registries. Chapters 7 and 8 of the User’s Guide provide more information on ethics, informed consent, and protecting patient privacy. Data sharing is an additional concern in the context of EHR-based registries. Decisions must be made about whether a single institutional review board (IRB) will suffice or whether all sites will require local IRB approval. Governance is also challenging as the rules around sharing of data (identifiable or de-identified) vary depending on the organizations involved and the purpose of the research.

Human Resources

Most healthcare providers, especially small office-based practices, do not have adequate staff time or even the necessary expertise to solve all potential challenges with EHR-registry integration/linkage. Indeed, several types of expertise are needed, such as:

- Regulatory/ethics – what data can we share?
- Scientific – what question is important?
- Research design – how do we answer the question?
- Clinical – do the data mean what we think they mean?
- Informatics – do the data maintain their epistemological integrity from clinical collection to analysis?
- Information technology (IT) – how do we curate and manage the data?
- Statistics and epidemiology – how do we answer the question with the data obtained?

In addition, although EHRs may offer cost-effective solutions for registry use, the need to capture comprehensive data for registries may counter this cost-effectiveness balance (e.g., requiring costly changes to the clinical workflow). Assuming that all data objectives for a registry can be met within an EHR, data collection for EHR-based registries hypothetically could be achieved at the time of a clinical encounter, thus reducing the cost of data collection; however, this has yet to be achieved on a widespread basis.

Other Factors

Other factors may also affect the usefulness of EHRs as a foundation for internal registries and/or for contributing to external registries. These include challenges with collecting patient consent within clinical workflows, incorporating patient-reported data, and safeguarding the security of the data.[71]

International Perspective on EHR-Based Registries

Some international registries are derived from national data collected in the context of national health insurance programs. In the Nordic countries, the unique constellation of universal coverage, a network of population-wide registries and databases, and individual-level linkage[72] make registries optimally suited for observational medical research in multiple clinical domains[73] and, increasingly for pragmatic trials.[74-76] In some countries, EHRs can be readily linked with the registry data using nationwide individual identifiers. For example, Nordic countries maintain a wide network of continuously updated databases, which collectively cover most health events, which can be linked on individual level in combinations dictated by the needs of a given study. In the United Kingdom, the Clinical Practice Research Datalink (CPRD)[77] and The Health Improvement Network (THIN) are important sources of routinely collected data, originating in EHRs. Both CPRD and THIN capture information routinely gathered in the course of daily operations of participating general practices. The data undergo a set of built-in data checks before being available for research. In some instances, additional data are linked (e.g., hospital records, or basic socioeconomic data). All patients registered with the participating practices, regardless of their disease, are included in the resulting dataset as long as they are enrolled in a participating practice.

Similarly, routine records are also being collected in some form in many countries in Europe though generally with less national coverage than in England, with non-exhaustive list including Netherlands,[78, 79] Italy,[79, 80] Scotland,[81] Germany,[82] France,[83] and Spain.[84] In North America, routine health records from a single-payer system are maintained by provinces in Canada;[85] and, increasingly, in Asia, including South Korea,[86] and Taiwan.[87] Although not originally established for research, routine data have been playing an increasingly important role in studies of health and disease, including post-marketing risk-management commitments.

The Future of EHR-Based Registries

The true promise of EHRs for registries is in facilitating the achievement of a practical, scalable, and efficient means of collecting registry data for multiple purposes. Scalability constraints on patient registries can be dramatically reduced by using digitized information.[3] Paper records are inherently limited because of the associated difficulty of systematically identifying eligible patients for research activities and the effort required to re-enter information into a database.[3] Digitized information has the potential to make it easier to meet both of these requirements, enabling larger, more diverse patient populations and avoiding duplication of effort by participating clinicians and patients.[3] However, duplication of effort can be reduced only to the extent that EHRs capture data elements and outcomes with specific, consistent, and interoperable definitions — or that data can be found and transformed by other processes and technologies (e.g., natural language processing) into standardized formats that match registry specifications.[3]

Despite the challenges and barriers of using EHRs for registries, EHRs will likely play a key role in expanding and developing existing and future registries. Multiple factors are poised to increase the role of EHRs in registries in the near future such as:

- increasing adoption of light-weight and efficient interoperability standards (e.g., HL7 FHIR);[88]
- new methods to measure EHR interoperability;[69]
- innovative technical frameworks to harmonize the extraction of data from EHRs (e.g., S&I Framework SDC);[60]
- introduction of new EHR-embedded tools to develop EHR-integrated registries (e.g., define and apply retrieval protocols; additional EHR-integrated forms for registries);
- incentivizing healthcare providers to share EHR data with registries (e.g., Meaningful Use);[89]
- aligning value-based efforts and population health management goals with reporting of EHR data to registries across providers (e.g., MACRA);[90] and
- providing additional clarifications about the application of HIPAA and other privacy protection rules in the context of EHR-based registries for both operational purposes and research.[91]

EHRs can be linked or integrated with registries in many formats or various purposes. Future research should focus on developing and disseminating additional guidelines and technical documentations about registry integration with EHRs for public use. Finally, achieving a fully interoperable EHR-based registry, so that EHRs and patient registries function seamlessly with one another, is unlikely to be accomplished in the near future.[3] However, it is critical that a level of interoperability be achieved to prevent the creation of information silos within proprietary informatics systems that make it difficult or impossible to develop large EHR-based registries and conduct research across diverse practices and populations.

References for Chapter 4

1. Gliklich R, Dreyer N, Leavy M, eds. Registries for Evaluating Patient Outcomes: A User's Guide. Third edition. Two volumes. (Prepared by the Outcome DEcIDE Center [Outcome Sciences, Inc., a Quintiles company] under Contract No. 290 2005 00351 TO7.) AHRQ Publication No. 13(14)-EHC111. Rockville, MD: Agency for Healthcare Research and Quality. April 2014. http://www.effectivehealthcare.ahrq.gov.

2. Aspden P, Corrigan JM, Wolcott J, et al. Key Capabilities of an Electronic Health Record System: Letter Report. Institute of Medicine of the National Academies; 2004.

3. Gliklich RE, Dreyer NA, Leavy MB. Interfacing Registries With Electronic Health Records. Registries for Evaluating Patient Outcomes: A User's Guide. 2. Third ed. Rockville, MD: Agency for Healthcare Research and Quality (AHRQ); 2014. p. 3-22.

4. Dixon BE, Gibson PJ, Grannis SJ. Estimating Increased Electronic Laboratory Reporting Volumes for Meaningful Use: Implications for the Public Health Workforce. Online Journal of Public Health Informatics. 2014;5(3):225. PMID: 24678378. DOI: 10.5210/ojphi.v5i3.4939.

5. Electronic Health Record Incentive Program, 42 CFR 412, 413, 422, 495 (2010).

6. Centers for Disease Control and Prevention (CDC). Summary of Public Health Objectives in Stage 2 Meaningful Use ONC and CMS Final Rules Version 1.1. 2014; https://www.cdc.gov/ehrmeaningfuluse/docs/summary-of-ph-objectives-in-stage-2-mu-onc-and-cms-final-rules_04_01_2014.pdf. Accessed August 15, 2019.

7. Centers for Disease Control and Prevention (CDC). Meaningful Use. https://www.cdc.gov/ehrmeaningfuluse/. Accessed August 15, 2019.

8. U.S. Core Data for Interoperability. Office of the National Coordinator for Health Information Technology. Version 1. https://www.healthit.gov/isa/us-core-data-interoperability-uscdi. Accessed June 18, 2019.

9. The Office of the National Coordinator for Health IT (ONC). Patient Identification and Matching: Final Report. https://www.healthit.gov/sites/default/files/patient_identification_matching_final_report.pdf. Accessed August 15, 2019.

10. U.S. Department of Health and Human Services (DHHS). Part 164-Security and Privacy. https://www.gpo.gov/fdsys/pkg/CFR-2011-title45-vol1/pdf/CFR-2011-title45-vol1-part164.pdf. Accessed August 16, 2019.

11. Jha AK, Burke MF, DesRoches C, et al. Progress Toward Meaningful Use: Hospitals' Adoption of Electronic Health Records. Am J Manag Care. 2011;17(12 Spec No.):SP117-24. PMID: 22216770.

12. Centers for Medicare and Medicaid Services (CMS). Health Information Technology: Standards, Implementation Specifications, and Certification Criteria for Electronic Health Record Technology, 2014 edition. Fed Regist. 2012 Sep 4; 77(171):54163-292. PMID: 22946139.

13. Centers for Medicare and Medicaid Services (CMS). Acute Care Hospital Inpatient Prospective Payment System. Medicare Learning Network. December 21, 2012.

14. Chan KS, Fowles JB, Weiner JP. Review: electronic health records and the reliability and validity of quality measures: a review of the literature. Med Care Res Rev. 2010;67(5):503-27. PMID: 20150441. DOI: 10.1177/1077558709359007.

15. Madden JM, Lakoma MD, Rusinak D, et al. Missing clinical and behavioral health data in a large electronic health record (EHR) system. J Am Med Inform Assoc. 2016;23(6):1143-9. PMID: 27079506. DOI: 10.1093/jamia/ocw021.

16. Mendelsohn AB, Dreyer NA, Mattox PW, et al. Characterization of Missing Data in Clinical Registry Studies. Therapeutic Innovation & Regulatory Science. 2015;49(1):146-54. PMID: 30222467. DOI: 10.1177/2168479014532259.

17. Health Insurance Portability and Accountability Act of 1996 (HIPAA), Pub. L. No. 104-191, 110 Stat. 139 (1996) (codified as amended in scattered sections of 42 U.S.C.). HIPAA Privacy Rule Regulations codified at 45 CFR pts. 160 & 164 (2010).

18. (CDC) CfDCaP. International Classification of Diseases Tenth Revision Clinical Modification (ICD-10-CM). https://www.cdc.gov/nchs/icd/icd10cm.htm. Accessed August 16, 2019.

19. World Health Organization (WHO). International Classification of Primary Care, Second Edition (ICPC-2). http://www.who.int/classifications/icd/adaptations/icpc2/en/. Accessed August 16, 2019.

20. National Library of Medicine (NLM). SNOMED CT. 2017; https://www.snomed.org/. Accessed June 10, 2019.

21. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders (DSM-5®): APA Publishing; 2013.

22. National Library of Medicine. Unified Medical Language System (UMLS). https://www.nlm.nih.gov/research/umls/index.html. Accessed August 15, 2019.

23. 42 CFR Part 2 - Confidentiality of Alcohol and Drug Abuse Patient Records. https://www.govinfo.gov/app/details/CFR-2010-title42-vol1/CFR-2010-title42-vol1-part2. Accessed August 16, 2019.

24. George J, Phun YT, Bailey MJ, et al. Development and validation of the medication regimen complexity index. Ann Pharmacother. 2004;38(9):1369-76. PMID: 15266038. DOI: 10.1345/aph.1D479.

25. Food and Drug Administration (FDA). National Drug Code Directory. 2017; https://www.fda.gov/drugs/drug-approvals-and-databases/national-drug-code-directory. Accessed August 16, 2019.

26. National Library of Medicine (NLM). RxNorm. Unified Medical Language System (UMLS) https://www.nlm.nih.gov/research/umls/rxnorm/. Accessed June 10, 2019.

27. World Health Organization (WHO). The Anatomical Therapeutic Chemical Classification System with Defined Daily Doses (ATC/DDD). http://www.who.int/classifications/atcddd/en/. Accessed August 16, 2019.

28. (AMA) AMA. Current Procedural Terminology (CPT). https://www.ama-assn.org/amaone/cpt-current-procedural-terminology. Accessed August 16, 2019.

29. Centers for Medicare and Medicaid Services (CMS). HCPCS - General Information. https://www.cms.gov/Medicare/Coding/MedHCPCSGenInfo/index.html. Accessed August 16, 2019.

30. Regenstrief Institute. LOINC. https://loinc.org/. Accessed June 10, 2019.

31. Townsend N, Rutter H, Foster C. Improvements in the data quality of a national BMI measuring programme. Int J Obes (Lond). 2015;39(9):1429-31. PMID: 25869597. DOI: 10.1038/ijo.2015.53.

32. National Institute of Health (NIH). Patient-reported Outcomes Measurement Information System (PROMIS). http://www.healthmeasures.net. Accessed August 16, 2019.

33. Committee on the Recommended Social and Behavioral Domains and Measures for Electronic Health Records; Board on Population Health and Public Health Practice; Institute of Medicine. Capturing Social and Behavioral Domains and Measures in Electronic Health Records: Phase 2. Washington (DC): National Academies Press (US); 2015 Jan 8. Abstract. https://www.ncbi.nlm.nih.gov/books/NBK269341/. Accessed August 16, 2019.

34. Adler NE, Stead WW. Patients in context--EHR capture of social and behavioral determinants of health. N Engl J Med. 2015;372(8):698-701. PMID: 25693009. DOI: 10.1056/NEJMp1413945.

35. Education USDo. Family Educational Rights and Privacy Act (FERPA). https://www2.ed.gov/policy/gen/guid/fpco/ferpa/index.html. Accessed August 16, 2019.

36. Workman TA. Engaging Patients in Information Sharing and Data Collection: The Role of Patient-Powered Registries and Research Networks. AHRQ Community Forum White Paper. AHRQ Publication No. 13-EHC124-EF. Rockville, MD: Agency for Healthcare Research and Quality; September 2013.

37. Health Level Seven (HL7). Mobile Health. http://www.hl7.org/Special/committees/mobile/. Accessed August 16, 2019.

38. Arora S, Yttri J, Nilse W. Privacy and Security in Mobile Health (mHealth) Research. Alcohol Res. 2014;36(1):143-51. PMID: 26259009.

39. Dreyer NA, Blackburn S, Hliva V, et al. Balancing the Interests of Patient Data Protection and Medication Safety Monitoring in a Public-Private Partnership. JMIR Med Inform. 2015;3(2). PMID: 25881627. DOI: 10.2196/medinform.3937.

40. Berwick DM, Nolan TW, Whittington J. The triple aim: care, health, and cost. Health Aff (Millwood). 2008;27(3):759-69. PMID: 18474969. DOI: 10.1377/hlthaff.27.3.759.

41. Improvement IfH. The IHI Triple Aim. http://www.ihi.org/Engage/Initiatives/TripleAim/Pages/default.aspx. Accessed August 16, 2019.

42. Eggleston E, Klompas M. Rational Use of Electronic Health Records for Diabetes Population Management. Current Diabetes Reports. 2014;14(4):479. PMID: 24615333. DOI: 10.1007/s11892-014-0479-z.

43. Duncan I. Healthcare Risk Adjustment and Predictive Modeling. Winsted, CT: ACTEX Publications; 2011.

44. Kharrazi H, Weiner JP. IT-enabled Community Health Interventions: Challenges, Opportunities, and Future Directions. EGEMS (Wash DC). 2014;2(3):1117. PMID: 25848627. DOI: 10.13063/2327-9214.1117.

45. Kharrazi H, Lasser EC, Yasnoff WA, et al. A proposed national research and development agenda for population health informatics: summary recommendations from a national expert workshop. J Am Med Inform Assoc. 2017;24(1):2-12. PMID: 27018264. DOI: 10.1093/jamia/ocv210.

46. Centers for Medicare and Medicaid Services. Physician Quality Reporting System (PQRS). https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/PQRS/Downloads/PQRS_OverviewFactSheet_2013_08_06.pdf. Accessed June 20, 2019.

47. Patient-Centered Outcomes Research Institute. Comprehensive Inventory of Research Networks. https://www.pcori.org/funding-opportunities/research-support-funding-opportunities/research-support-funding/comprehensive. Accessed August 16, 2019.

48. Clinical Data Interchange Standards Consortium (CDISC). The CDISC Healthcare Link Initiative. https://www.cdisc.org/system/files/all/standard_category/application/pdf/healthcare_link_chapter.pdf. Accessed August 16, 2019.

49. Clinical Information Modeling Initiative (CIMI). CIMI Reference Model. https://wiki.hl7.org/index.php?title=Proposed_CIMI_Reference_Model. Accessed August 16, 2019.

50. Sentinel Common Data Model. https://www.sentinelinitiative.org/sentinel/data/distributed-database-common-data-model/sentinel-common-data-model. Accessed June 10, 2019, 2018.

51. Observational Health Data Sciences and Informatics (OHDSI). https://www.ohdsi.org/. Accessed June 10, 2019.

52. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. J Am Med Inform Assoc. 2013;20(1):117-21. PMID: 22955496. DOI: 10.1136/amiajnl-2012-001145.

53. Dreyer NA, Rodriguez AM. The fast route to evidence development for value in healthcare. Curr Med Res Opin. 2016;32(10):1697-700. PMID: 27314301. DOI: 10.1080/03007995.2016.1203768.

54. Arts DG, De Keizer NF, Scheffer GJ. Defining and improving data quality in medical registries: a literature review, case study, and generic framework. J Am Med Inform Assoc. 2002;9(6):600-11. PMID: 12386111. DOI: 10.1197/jamia.m1087.

55. 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, 45 CFR Parts 160 and 164 (2013).

56. The Office of the National Coordinator for Health IT (ONC). Health Information Exchange Governance. https://www.healthit.gov/sites/default/files/governancehitweekpresentation.pdf. Accessed August 16, 2019.

57. Fiks A, Grundmeier R, Steffes J. Comparative Effectiveness Research Through a Collaborative Electronic Reporting Consortium. Pediatrics. 2015;136(1):e215-24. PMID: 26101357. DOI: 10.1542/peds.2015-0673.

58. The Office of the National Coordinator for Health Information Technology (ONC). Distributed Population Queries. https://www.healthit.gov/sites/default/files/052412_hitsc_queryhealthpresentation.pdf. Accessed August 19, 2019.

59. Centers for Disease Control and Prevention (CDC). Community Health Assessments and Health Improvement Plans. https://www.cdc.gov/publichealthgateway/cha/plan.html. Accessed August 16, 2019.

60. S&I Framework. Structured Data Capture Initiative. https://www.healthit.gov/topic/scientific-initiatives/pcor/research-evaluation/structured-data-capture-sdc. Accessed August 19, 2019.

61. Lombardo JS, Burkom H, Pavlin J. ESSENCE II and the framework for evaluating syndromic surveillance systems. MMWR Suppl. 2004;53:159-65. PMID: 15714646.

62. Massachusetts eHealth Institute. PopMedNet: Distributed Data Network. http://mehi.masstech.org/programs/past-programs/mdphnet-project/popmednet-distributed-data-network. Accessed August 16, 2019.

63. New York State Department of Health. Population Health Registry. https://www.health.ny.gov/health_care/medicaid/redesign/ehr/registry/phr.htm. Accessed August 19, 2019.

64. New York City Health Department. The NYC Macroscope. http://www1.nyc.gov/site/doh/data/health-tools/nycmacroscope.page. Accessed August 16, 2019.

65. Pathak J, Kho AN, Denny JC. Electronic health records-driven phenotyping: challenges, recent advances, and perspectives. J Am Med Inform Assoc. 2013;20(e2):e206-11. PMID: 24302669. DOI: 10.1136/amiajnl-2013-002428.

66. Richesson RL, Hammond WE, Nahm M, et al. Electronic health records based phenotyping in next-generation clinical trials: a perspective from the NIH Health Care Systems Collaboratory. J Am Med Inform Assoc. 2013;20(e2):e226-31. PMID: 23956018. DOI: 10.1136/amiajnl-2013-001926.

67. Ford E, Carroll JA, Smith HE, et al. Extracting information from the text of electronic medical records to improve case detection: a systematic review. J Am Med Inform Assoc. 2016;23(5):1007-15. PMID: 26911811. DOI: 10.1093/jamia/ocv180.

68. Benson T. Principles of Health Interoperability HL7 and SNOMED. Health Informatics. 2010.

69. Samarath A, Sorace J, Patel V. Measurement of Interoperable Electronic Health Care Records Utilization. US Department of Health and Human Services https://aspe.hhs.gov/pdf-report/measurement-interoperable-electronic-health-care-records-utilization. Accessed August 16, 2019.

70. Swain M, Charles D, Patel V, et al. Health Information Exchange Among U.S. Non-federal Acute Care Hospitals: 2008-2014. ONC Data Brief. 2014(17).

71. Kharrazi H, Chi W, Chang H-Y, et al. Comparing Population-based Risk-stratification Model Performance Using Demographic, Diagnosis and Medication Data Extracted From Outpatient Electronic Health Records Versus Administrative Claims. Medical Care. 2017;55(8):789-96. PMID: 28598890. DOI: 10.1097/MLR.0000000000000754.

72. Frank L. Epidemiology. When an Entire Country Is a Cohort. Science. 2000;287(5462):2398-9. PMID: 10766613. DOI: 10.1126/science.287.5462.2398.

73. Krueger WS, Anthony MS, Saltus CW, et al. Evaluating the Safety of Medication Exposures During Pregnancy: A Case Study of Study Designs and Data Sources in Multiple Sclerosis. Drugs Real World Outcomes. 2017;4(3):139-49. PMID: 28756575. DOI: 10.1007/s40801-017-0114-9.

74. Raungaard B, Jensen LO, Tilsted HH, et al. Zotarolimus-eluting durable-polymer-coated stent versus a biolimus-eluting biodegradable-polymer-coated stent in unselected patients undergoing percutaneous coronary intervention (SORT OUT VI): a randomised non-inferiority trial. Lancet. 2015;385(9977):1527-35. PMID: 25601789. DOI: 10.1016/S0140-6736(14)61794-3.

75. Frobert O, Lagerqvist B, Olivecrona GK, et al. Thrombus aspiration during ST-segment elevation myocardial infarction. N Engl J Med. 2013;369(17):1587-97. PMID: 23991656. DOI: 10.1056/NEJMoa1308789.

76. Shurlock B. Randomization Within Quality Registries: A Cost-Effective Complement to Classical Randomized Trials. European heart journal. 2014;35(1):1-2. PMID: 24382633. DOI: 10.1093/eurheartj/eht493.

77. Herrett E, Gallagher AM, Bhaskaran K, et al. Data Resource Profile: Clinical Practice Research Datalink (CPRD). Int J Epidemiol. 2015;44(3):827-36. PMID: 26050254. DOI: 10.1093/ije/dyv098.

78. Pharmo Record Linkage System. https://www.pharmo.nl/. Accessed August 16, 2019.

79. Coloma PM, Schuemie MJ, Trifiro G, et al. Combining electronic healthcare databases in Europe to allow for large-scale drug safety monitoring: the EU-ADR Project. Pharmacoepidemiol Drug Saf. 2011;20(1):1-11. PMID: 21182150. DOI: 10.1002/pds.2053.

80. Avillach P, Coloma PM, Gini R, et al. Harmonization process for the identification of medical events in eight European healthcare databases: the experience from the EU-ADR project. J Am Med Inform Assoc. 2013;20(1):184-92. PMID: 22955495. DOI: 10.1136/amiajnl-2012-000933.

81. Alvarez-Madrazo S, McTaggart S, Nangle C, et al. Data Resource Profile: The Scottish National Prescribing Information System (PIS). Int J Epidemiol. 2016;45(3):714-5f. PMID: 27165758. DOI: 10.1093/ije/dyw060.

82. The German Pharmacoepidemiological Research Database. https://www.bips-institut.de/en/research/research-infrastructures/gepard.html. Accessed August 16, 2019.

83. Boudemaghe T, Belhadj I. Data Resource Profile: The French National Uniform Hospital Discharge Data Set Database (PMSI). Int J Epidemiol. 2017;46(2):392-d. PMID: 28168290. DOI: 10.1093/ije/dyw359.

84. Bolíbar B, Fina Avilés F, Morros R, et al. [SIDIAP Database: Electronic Clinical Records in Primary Care as a Source of Information for Epidemiologic Research]. Medicina Clínica. 2012;138(14):617-21.

85. Garies S, Birtwhistle R, Drummond N, et al. Data Resource Profile: National electronic medical record data from the Canadian Primary Care Sentinel Surveillance Network (CPCSSN). Int J Epidemiol. 2017;46(4):1091-2f. PMID: 28338877. DOI: 10.1093/ije/dyw248.

86. Cheol Seong S, Kim Y-Y, Khang Y-H, et al. Data Resource Profile: The National Health Information Database of the National Health Insurance Service in South Korea. International Journal of Epidemiology. 2016;46(3):799-800.

87. Chen Y-C, Yeh H-Y, Wu J-C, et al. Taiwan's National Health Insurance Research Database: Administrative Health Care Database as Study Object in Bibliometrics. Scientometrics. 2011;86(2):365-80.

88. Health Level 7 (HL7). Welcome to FHIR. 2017; https://www.hl7.org/fhir/. Accessed June 10, 2019.

89. Centers for Medicare & Medicaid Services (CMS). Stage 3 Program Requirements for Providers Attesting to their State's Medicaid EHR Incentive Program. Regulations and Guidance 2016; https://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/Stage3Medicaid_Require.html. Accessed August 16, 2019.

90. Centers for Medicare and Medicaid Services (CMS). MACRA: Delivery System Reform Medicare Payment Reform. 2017; https://www.cms.gov/medicare/quality-initiatives-patient-assessment-instruments/value-based-programs/macra-mips-and-apms/macra-mips-and-apms.html. Accessed June 11, 2019.

91. U.S. Department of Health and Human Services (DHHS). Clinical Data Repositories - OHRP Correspondence. 2015; https://www.hhs.gov/ohrp/regulations-and-policy/guidance/june-25-2015-letter-to-robert-portman/index.html. Accessed August 16, 2019.

Chapter 5. Obtaining Data From Other Sources

Authors (alphabetical)

Allison Bryant, M.P.H.
Senior Epidemiologist
OM1, Inc.

Michelle B. Leavy, M.P.H.
Head, Healthcare Research & Policy
OM1, Inc.

Introduction

While integration of electronic health record (EHR) data with registries is perhaps most common, many registries wish to incorporate data from other sources, such as medical or consumer devices, imaging databases, or biorepositories. Previous chapters in this document describe the types of data that may be obtained from those sources as well as the operational, ethical and legal, and scientific challenges associated with those sources. This chapter describes approaches to obtaining and integrating of electronic health data with other data sources, such as patient registries. The first section describes approaches to requesting and exchanging data, while the second section discusses the role of common data models in supporting the integration of data from multiple sources. Key questions to consider when planning to incorporate data from other sources are summarized in Appendix B.

Tools and Technologies for Obtaining Data

Data may be obtained from other systems in many ways. At the most basic level, data may be extracted from one system, transformed, and loaded into another system; this process is known as ‘extract, transform, and load’ or ETL. Data are typically extracted at regular intervals for all patients, rather than on a real-time basis for individual patients. Chapter 11 of the User’s Guide discusses considerations related to ETL in more detail.

Application Programming Interfaces (APIs)

More recently, substantial effort has been devoted to developing tools to transfer data using application programming interfaces (APIs). An API is a set of tools and resources that allows two applications to communicate with each other; in other words, the API delivers a request from one application to another application and then returns the response to the first application. APIs are widely used in many areas. For example, travel websites use APIs to request flight availability and price information from the airline websites and return that information to the website user.

In healthcare, use of APIs is relatively new. The 21st Century Cures Act requires the Federal government to improve the interoperability of health information, and the Office of the National Coordinator (ONC) has identified APIs as an important tool to support the exchange of health data.[1] APIs are envisioned as a tool to enable patients to easily request and retrieve data from many different EHRs to manage their own healthcare or the healthcare of a family member. APIs may also be used to request and retrieve data for research purposes. While APIs are already used in many areas, APIs developed for health data must meet additional requirements, such as those related to authentication, authorization, encryption, and patient selection.

The Fast Healthcare Interoperability Resources (FHIR) initiative builds on the idea of using APIs by providing a set of specifications and an API to support healthcare interoperability. The specifications describe a standard for restructuring health data from disparate sources into a single format to facilitate interoperability. FHIR was introduced by Health Level Seven (HL7) as a proposed interoperability standard.[2]

Using FHIR, software developers can build applications (‘apps’) to interface with any EHR or other health IT system to request and/or send back information, thus eliminating the need for custom-build solutions for each EHR. Unlike some earlier standards such as the Consolidated Clinical Document Architecture (C-CDA), FHIR allows for transfer of defined sets of data, rather than entire medical records; this contributes to efficiency. For example, a physician may need to access a single immunization record for a new patient; rather than transfer the entire medical record, FHIR can be used to request just that piece of information.

FHIR also has the potential to enable patients to access their own longitudinal health records by providing tools to integrate data from multiple systems into a usable format. For example, Apple announced a pilot project in early 2018 that combines patient-generated data from the iPhone Health app with data from the individual’s electronic medical record into a personal health record. The project uses the FHIR specification.[3]

Apps for Interoperability

APIs and FHIR have the potential to reshape how registries interact with other data sources to obtain data. For example, the registry may be able to request and retrieve the most relevant data at the individual patient level on a regular basis from an EHR or other source. Several organizations are developing these tools, and researchers are exploring the potential of using new apps designed to facilitate research. Some examples of these efforts are described below.

SMART Health IT

The SMART Health IT project builds on the FHIR specifications to create an open, standards-based app platform that allows developers to build apps that run across the healthcare system. FHIR defines ‘core’ data models, and SMART applies a set of ‘profiles’ on top of those data models to specify the vocabularies that are used to express common types of clinical data, such as medications, laboratory results, and diagnoses. SMART also includes an authorization model based on the OAuth standard. In EHR systems or clinical data warehouses that have implemented the SMART platform, clinicians and other authorized users can exchange data using apps. The SMART project includes a gallery of apps that can be used to improve clinical care, support research, and facilitate population health management.[4] SMART apps are currently used at several healthcare institutions, including Boston Children's Hospital and Duke Medicine.

Apple HealthKit and ResearchKit

In 2014, Apple released HealthKit, a common framework to support sharing of patient-generated health data (PGHD) among apps, services, and providers. The related ResearchKit was released in 2015 to provide researchers with an open source framework to build apps to support smartphone-based research. ResearchKit enables researchers to use the iPhone's sensors as well as third-party devices to monitor health variables captured in HealthKit and share those data with researchers and EHRs. Asthma researchers used ResearchKit to conduct a prospective observational asthma study entirely remotely using a smartphone platform. The study aimed to characterize the cohort of patients enrolled via a mobile platform and to assess the feasibility of this approach. Over 7,000 U.S. participants enrolled in the Asthma Mobile Health Study using an iPhone app built on the Apple ResearchKit framework, and data were collected on demographics, socioeconomic status, medication use, asthma control, and resource utilization.[5, 6] In a related pilot study, four participants in the Asthma Mobile Health Study completed the setup process to share their study data with their care provider via the Epic MyChart app for use in their clinical care.[7]

In another example, researchers at the University of California, San Francisco enrolled a remote cohort of 9,750 participants and used smartwatches to obtain heart rate and step count data. The Health eHeart Study aimed to develop and validate a deep neural network to detect atrial fibrillation using smartwatch data. The study captured more than 139 million heart rate measurements and found that smartwatch photoplethysmography coupled with a deep neural network is able to detect atrial fibrillation passively.[8] Recruitment is now underway for the Apple Heart Study, which aims to evaluate whether the Apple Heart Study App can use data from the Apple Watch to identify irregular heart rhythms, including those caused by atrial fibrillation. The study aims to recruit 500,000 participants.[9]

Sync for Science

Sync for Science (S4S) also builds on the FHIR standard and the OAuth security profiles to help participants in research studies share their data with researchers. The S4S project, run by Harvard Medical School with funding from the National Institutes of Health (NIH) and ONC, works with major EHR systems, such as Epic, Allscripts, and Cerner, to allow individuals to request their electronic health record data and share it securely with a research study. S4S leverages the Common Clinical Data Set, as defined under Meaningful Use (see Chapter 4), and works with EHR vendors to implement open standards that support the exchange of this dataset with patients and patient-selected apps.[10]

The S4S technology is being used in the All of Us Research Program, supported the NIH's Precision Medicine Initiative (see Chapter 1). With S4S, All of Us can obtain medical record data from individuals who enroll in the study outside of participating center. The technology is also being used in a National Evaluation System for health Technology (NEST) demonstration project (see Chapter 1 for more information on NEST). The 'Using a Novel mHealth Platform to Obtain Real-World Data for Post-Market Surveillance' project aims to test the feasibility of 'using a novel mobile health platform to provide real-world data that can be used for post-market surveillance of patients after either bariatric surgery (sleeve gastrectomy or gastric bypass) or catheter-based atrial fibrillation ablation.'[11] The project is enrolling 60 participants and using a mobile app, HugoPHR, to aggregate data from EHRs, pharmacy portals, wearable and sync-able devices, and questionnaires/patient-reported outcome measures.

Other Efforts

In November 2018, the U.S. Food and Drug Administration (FDA) released a mobile app that is designed to collect patient-reported data and store them in a central location for use in clinical research studies. The platform includes a web-based configuration portal and a single mobile app, MyStudies. The app does not currently push or pull data to/from an HER, but this capability could be added by organizations that wished to devote resources to modifying the MyStudies system for a specific use. The goal of the MyStudies system is to provide a tool that research sponsors and developers can configure for different studies and different therapeutic areas, while remaining compliant with FDA requirements for data authenticity, integrity, and confidentiality. The data storage environment supports the auditing that is required under 21 CFR Part 11 and the Federal Information Security Management Act. Two versions of the app are available – one built on Apple's ResearchKit framework, and one built on the open source ResearchStack framework (e.g., for use on Google's Android operating system).[12]

The Centers for Medicare and Medicaid Services (CMS) is also working to provide Medicare beneficiaries with the ability to connect their claims data to other programs, such as research studies. CMS has released Blue Button API, which enables developers to build beneficiary-facing applications that allow a beneficiary to grant access to their claims data for another purpose (e.g., a patient registry, clinical trial, or other project). Blue Button API also builds on the FHIR standard and OAuth 2.0 security profile.[13]

Next Steps

The use of APIs and the FHIR standard are introducing many new approaches to data transfer, with rapid innovation, app development, and iteration based on pilot study findings. This area is likely to continue to develop quickly in the near future. While these tools are not widely used in registries to date, they may become more common and useful as they mature. Further research is needed to explore use cases for patient registries and to inform the development of best practices in this area.

Common Data Models

Data models are important tools for integrating data from multiple sources. A data model specifies the definitions, structure, and relationships of data elements and can be used in single site studies as a framework to organize and define data elements, or more commonly, when there is a need to integrate and share data across disparate sources. A common data model (CDM) is used to “standardize and facilitate the exchange, pooling, sharing, or storing of data from multiple sources.”[14] As each database has a distinct physical format and may use different terminologies or coding standards, a CDM can help to minimize variability and facilitate a common interpretation across the underlying data sources. This is achieved by implementing a common data format, applying standardized data transformation rules and assumptions to the data, and developing common definitions and terminology during the data preparation process.[15]

The use of CDMs in observational research gained traction following the passage of the Food and Drug Administration (FDA) Amendments Act (FDAAA) of 2007, which mandated the development of “validated methods for the establishment of a post-market risk identification and analysis system to link and analyze safety data from multiple sources.”[16] In response, the FDA launched the Sentinel Initiative in May 2008 to establish a national electronic system for monitoring post-market drug safety. The system enables automated monitoring of product safety information from healthcare data systems such as registries, electronic health record systems (EHRs), and administrative claims databases. Sentinel uses a distributed data approach in which the participating data partners transform their data using a standardized data structure referred to as the Sentinel Common Data Model. Once local data are transformed according to the CDM, data queries can be executed in a distributed fashion and the results then pooled into a common database, known as the Sentinel Distributed Database (SDD).[17] The distributed model employed by Sentinel allows data to remain with the data holder rather than be pooled centrally.

In addition to the Sentinel CDM, other CDMs have been developed for use in observational research. Four representative examples are described in Table 5-1.

Rationale for Use of CDMs

The primary advantage of using a CDM to facilitate registry development is the ability to integrate data from multiple sources into a standard format. Note, for some types of research work, such as with Sentinel, a CDM can be used as part of a distributed research network, but this is more the exception than the rule. Healthcare databases used in research often serve distinct purposes; EHRs reflect clinical data captured at the point of care, whereas administrative claims data are used to support billing and reimbursement. As a result, the information captured across these databases differs in terms of content and structure. Furthermore, one system may use a different information model than another system,[18] and the way in which healthcare providers interact with and record data within each system may differ.

Table 5-1. Selected CDMs used in observational research

CDM	CDM DEVELOPER	PURPOSE	CONSIDERATIONS
Sentinel CDM[19]	FDA, in collaboration with Sentinel Data Partners	Created in 2008 in response to the FDA's efforts to create a national electronic system for monitoring the safety of FDA-regulated products CDM is a "standard data structure that allows Data Partners to quickly execute distributed programs against local data"	Structured largely based on administrative and claims data from health insurers. Efforts to incorporate other data types such as electronic health record (EHR) data are underway Designed primarily to support safety surveillance Extensive library of documentation, analytic code, and software toolkits publicly available
Observational Medical Outcomes Partnership (OMOP) CDM[20]	Observational Health Data Sciences and Informatics (OHDSI)	Established in 2008 to inform the appropriate use of observational healthcare databases for studying the effects of medical products Developed as an open-source, community standard for observational healthcare data Designed to include all observational health data elements that are relevant for analysis use cases to support the generation of reliable scientific evidence about the natural history of disease, healthcare delivery, effects of medical interventions, the identification of demographic information, healthcare interventions, and outcomes	Designed to accommodate a wide variety of observational health data, including from electronic health records and administrative claims data Relies on standardized vocabularies containing relevant corresponding healthcare concepts and reuses existing vocabularies when possible Contains standardized derived elements which contain information about the clinical events of a patient that are not obtained directly from the source data Open source tools and resources, including an active online community of users
The National Patient-Centered Clinical Research Network (PCORnet) CDM[21]	Research partnership comprising clinical data research networks (CDRNs) and patient-powered research networks (PPRNs)	Based on the Sentinel CDM, the PCORnet CDM was developed in 2014 to define a standard organization and representation of data for the PCORnet Distributed Research Network, which facilitates multi-site patient-centered research across participating members PCORnet uses EHR data from clinical data research networks (CDRNs), along with patient-generated data from patient-powered research networks (PPRNs)	Leverages standard terminologies and coding systems for healthcare to enable interoperability with evolving data standards Primarily suited to accommodate EHR data and patient-reported data
Informatics for Integrating Biology and the Bedside	Partners HealthCare System	Funded by the NIH as a National Center for Biomedical Computing, the i2b2 Center	The back-end infrastructure (the "Hive") defines the structure of the underlying data repository

CDM	CDM DEVELOPER	PURPOSE	CONSIDERATIONS
(i2b2) data model[22]		developed an informatics framework in 2007 Designed to bridge clinical research data and basic science research data to better understand the genetic bases of complex disease	The “workbench” is an application suite of query and mining tools that allows users to ask questions about the data

Similarly, coding practices and terminologies may vary across sources, as discussed above. While several standardized terminologies exist across clinical coding domains (e.g., medications, conditions, procedures, laboratory tests), many healthcare systems use local or proprietary terminology, particularly for medication coding. To address differences in medication coding, the National Library of Medicine (NLM) developed RxNorm, a naming system for drugs that supports semantic interoperability between existing drug terminologies. Similarly, SNOMED CT is a comprehensive clinical healthcare terminology system used within EHR systems that supports mapping to other standard disease coding systems, such as ICD-9 and ICD-10, that may be used within a local healthcare system. Working with different data sources requires knowledge of the local terminologies and methodologies to map codes to a common standard. A CDM can provide a standard database format and mapping standards to harmonize disparate coding systems and support the ability to compare results across data sources.[23] For example, a CDM may provide tools to support mapping from one vocabulary standard to another (e.g., NDC to RxNorm) or from a local terminology to a standard vocabulary via name/synonym matching.

Once data are mapped to a common format, CDMs can introduce efficiencies in the analysis process. Historically, analyses conducted within or across registries or other healthcare databases have required the development of *ad hoc* analysis plans to define data transformation rules and assumptions applied to the data. The creation and validation of these analytic programs is labor intensive, and programs are rarely reusable across different databases or research questions. A CDM can help address this problem by defining transformation rules for standardized derived elements that are built and stored within the CDM and are broadly applicable to a variety of research questions. One example of a standardized derived element is the concept of a period of drug use, or a span of time that a patient is assumed to be exposed to a particular medication; OMOP refers to this as a “drug era”.[24] While source data may contain individual medication exposures (i.e., prescription and/or fill records), a standardized derivation can establish rules to collapse these individual exposures into a drug era. This derivation can be applied to all medication records in the model and stored in a table within the CDM. Defining and deriving this concept that can be systematically used across different medication types and research questions *a priori* precludes the need for study-level determination of how to handle individual medication exposures.

In addition, the use of a CDM enables standardized queries and analytic programs to be directly shared across organizations using that same CDM. The Sentinel Initiative CDM provides an example of the ability to perform rapid analytics on large volumes of healthcare data by reusing validated, standardized queries which can be shared and run by different data providers.[25] The

Routine Querying System provides SAS programs that are designed to run against the Sentinel CDM.[26] The Observational Medical Outcomes Partnership (OMOP) CDM maintains a publicly available repository of analytic code for users to implement as well as an active online community of members who post and discuss questions relating to implementation of the model. Following a CDM where a standard data structure, terminology, and analytic programs are shared ultimately contributes to increased transparency and reproducibility of research.

Limitations of CDMs

Transforming and integrating data from multiple data sources into a CDM, often in large volumes, is not without challenges. The initial implementation of a CDM requires a substantial time commitment, in-depth understanding of the model structure and standard vocabularies, and thorough knowledge of the local structure and assumptions that apply to each data source to be included in the model. A common challenge is ensuring the CDM is flexible enough to support the relevant research needs (e.g., safety, comparative effectiveness) and data types (e.g., EHR, claims, patient-reported) while still maintaining the granularity required to answer complex questions. Furthermore, the standardization process may actually introduce some degree of systematic error based on how the definition was constructed. Also, the data transformation rules dictated by the data model may not be appropriate for all circumstances or the vocabulary mapping (e.g., for a particular condition) may not be consistent with an end user's interpretation.[27] In addition, certain assumptions are not dictated by the data model; for example, data cleaning to eliminate duplicate records or invalid birth years must be performed at the discretion of the user based on knowledge of the source data. Thus, the quality of data in the model is driven by the quality of data that is input to the model.

A common concern with transforming data into a CDM is information loss. Only the data for which there are equivalent data fields and tables in the model, and for which source codes can be mapped to standardized coding systems, are available for analysis.[28] For example, a model that does not have a place to store patient reported outcomes (PRO) data will not contain this data. From a terminology perspective, information loss may occur when a more granular terminology (e.g., SNOMED) is mapped to a less specific terminology (e.g., ICD-9).[29] A clear understanding of the extraction, transformation, and loading (ETL) process that determines how the source data are mapped to the CDM can help the user understand the level of effort required to implement a CDM and help to prevent information loss.[30]

Another challenge is the lack of harmonization across existing CDMs. A new effort, launched in 2017 and run by the FDA, is working to harmonize existing CDMs, including Sentinel, PCORnet, OHDSI, and i2b2, to improve the utility and interoperability of data from each of these networks.[31] Work on this effort is ongoing.[32]

Selecting an Appropriate CDM

Various factors, such as the suitability of available models and availability of resources to implement the model, can inform the choice to use a CDM.[33] While most CDMs are developed with the expectation that a user will maintain the database schema and data elements of the model, such that data can be shared or standardized analytic programs can be run, there is often the flexibility for modifications such as incorporating additional content to suit the user's needs. Several criteria should be considered when evaluating the suitability of existing CDMs for a specific registry purpose. In particular, the criteria developed by the Scalable Architecture for Federated Translational Inquiries Network (SAFTI Net) project are highly relevant here.[33] The SAFTI Net project was an AHRQ-funded project to develop a scalable, distributed network to support comparative effectiveness research. The criteria provided in Table 5-2 are adapted from the SAFTI Net criteria to apply to apply to patient registries generally.

Table 5-2. Criteria to evaluate the suitability of a CDM for a specific purpose, adapted from the SAFTI Net Project[33]

CRITERIA	DEFINITION	CONSIDERATIONS
Data Coverage	The ability to accommodate the necessary data elements and domains to address the registry's research questions.	Are the necessary data domains and data elements available in the supported version of the data model?
Extensibility	The methods used to expand the data model for more data elements, data types and new data domains.	Can new elements be incorporated by adding new values to existing data elements (e.g., adding "outpatient" as a new value to VISIT_TYPE_) versus needing to add new tables or columns? Larger scale domain extension may require changes to the database technical platform for tuning, indices, and efficiencies
Standardization	Use of standardized vocabularies	Does the model support use of standardized terminologies? Are tools available to support mapping from one standard to another, or from local terminologies to a standard terminology?
Scalability	The ability of the model to support datasets of different sizes.	Can the model be sized to smaller or larger datasets? What size of datasets have been supported in actual field use?
Adaptability	Willingness of existing user community to accept and incorporate data model additions and changes.	How broad a variety of data domains may be modeled? Are active user groups available to support sustainability?
Understandability	The effort required for technical staff to understand the data model and for data analysts to understand how to construct a query.	Are adequate training materials and supporting documentation available? Are active user groups available to help address questions?
Efficiency	The growth of the database with absent data (how null values are handled).	How are nulls handled in the database?
Current Usage	The number and diversity of uses of the data model and the size of the community using and supporting the model.	Have other researchers successfully used the model in a manner similar to the intended use? Is there an active user community to engage with researchers on the intended project?
Stability	The number of changes to the data model over the past 12–24 months.	Is the data model under review and revised periodically? A nonstable model may require underlying infrastructure changes to maintain.

CRITERIA	DEFINITION	CONSIDERATIONS
Cost	Licensing, staffing, costs of infrastructure.	If not public domain, what are the licensing costs? What are the staffing and resource costs? What are the infrastructure costs required to run the model?

Conclusions

Registries may use a variety of approaches to obtain data and integrate data from other sources. In many cases, the approach will be customized for the needs of a specific registry. However, many efforts are underway to develop standards-based tools, such as mobile applications, that can be adapted relatively easily to meet the specific needs of a study. Registries may also benefit from use of common data models to facilitate integration of data from disparate sources. Further research is needed to explore how these tools may be used in patient registries and to inform the development of best practices in this area.

References for Chapter 5

1. Rucker D. APIs: A Path to Putting Patients at the Center. Health IT Buzz. The Office of the National Coordinator of Health IT. April 24, 2018. https://www.healthit.gov/buzz-blog/interoperability/apis-path-putting-patients-center. Accessed June 10, 2019.

2. Health Level 7 (HL7). Welcome to FHIR. 2017; https://www.hl7.org/fhir/. Accessed June 10, 2019.

3. Comstock J. Apple to launch Health Records app with HL7's FHIR specifications at 12 hospitals. Healthcare IT News. January 24, 2018. https://www.healthcareitnews.com/news/apple-launch-health-records-app-hl7s-fhir-specifications-12-hospitals, June 10, 2019.

4. What is SMART? SMART Health IT. https://smarthealthit.org/an-app-platform-for-healthcare/about/. Accessed June 10, 2019.

5. Chan YY, Bot BM, Zweig M, et al. The asthma mobile health study, smartphone data collected using ResearchKit. Sci Data. 2018;5:180096. PMID: 29786695. DOI: 10.1038/sdata.2018.96.

6. Chan YY, Wang P, Rogers L, et al. The Asthma Mobile Health Study, a large-scale clinical observational study using ResearchKit. Nat Biotechnol. 2017;35(4):354-62. PMID: 28288104. DOI: 10.1038/nbt.3826.

7. Genes N, Violante S, Cetrangol C, et al. From smartphone to EHR: a case report on integrating patient-generated health data. npj Digital Medicine. 2018;1(1):23. DOI: 10.1038/s41746-018-0030-8.

8. Tison GH, Sanchez JM, Ballinger B, et al. Passive Detection of Atrial Fibrillation Using a Commercially Available Smartwatch. JAMA Cardiol. 2018;3(5):409-16. PMID: 29562087. DOI: 10.1001/jamacardio.2018.0136.

9. ClinicalTrials.gov. Apple Heart Study: Assessment of Wristwatch-Based Photoplethysmography to Identify Cardiac Arrhythmias. https://clinicaltrials.gov/ct2/show/NCT03335800. Accessed June 10, 2019.

10. Sync for Sciences. http://syncfor.science/. Accessed June 10, 2019.

11. ClinicalTrials.gov. Post-Market Surveillance With a Novel mHealth Platform. NCT03436082. https://clinicaltrials.gov/ct2/show/NCT03436082. Accessed June 10, 2019.

12. U.S. Food and Drug Administration. FDA MyStudies App. https://www.fda.gov/Drugs/ScienceResearch/ucm624785.htm. Accessed June 10, 2019.

13. Centers for Medicare and Medicaid Services. Blue Button 2.0. https://bluebutton.cms.gov/. Accessed June 10, 2019.

14. Ajayi OJ, Smith EJ, Viangteeravat T, et al. Multisite Semiautomated Clinical Data Repository for Duplication 15q Syndrome: Study Protocol and Early Uses. JMIR Res Protoc. 2017;6(10):e194. PMID: 29046268. DOI: 10.2196/resprot.7989.

15. Kunjan K, Toscos T, Turkcan A, et al. A Multidimensional Data Warehouse for Community Health Centers. AMIA Annu Symp Proc. 2015;2015:1976-84. PMID: 26958297.

16. Food and Drug Administration Amendments Act of 2007, Pub. L. No. 110-85.

17. Sentinel Initiative. Background. https://www.sentinelinitiative.org/background. Accessed June 12, 2019.

18. OHDSI Data Standardization. https://www.ohdsi.org/data-standardization/. Accessed June 10, 2019.

19. Sentinel Common Data Model. https://www.sentinelinitiative.org/sentinel/data/distributed-database-common-data-model/sentinel-common-data-model. Accessed June 10, 2019, 2018.

20. Observational Health Data Sciences and Informatics (OHDSI). https://www.ohdsi.org/. Accessed June 10, 2019.

21. PCORnet Common Data Model (CDM). http://pcornet.org/pcornet-common-data-model/. Accessed June 10, 2019.

22. i2b2 Research Data Warehouse. FAQs. https://i2b2.cchmc.org/faq. Accessed June 10, 2019.

23. Saitwal H, Qing D, Jones S, et al. Cross-terminology mapping challenges: a demonstration using medication terminological systems. J Biomed Inform. 2012;45(4):613-25. PMID: 22750536. DOI: 10.1016/j.jbi.2012.06.005.

24. OHDSI Design Principles. 20 June 2017:https://github.com/OHDSI/CommonDataModel/wiki/Design-Principles. Accessed 06 March 2018.

25. Chrischilles EA, Gagne JJ, Fireman B, et al. Prospective surveillance pilot of rivaroxaban safety within the US Food and Drug Administration Sentinel System. Pharmacoepidemiol Drug Saf. 2018;27(3):263-71. PMID: 29318683. DOI: 10.1002/pds.4375.

26. Sentinel Initiative. Routine Querying Tools (Modular Programs). https://www.sentinelinitiative.org/sentinel/surveillance-tools/routine-querying-tools. Accessed June 10, 2019.

27. Reisinger SJ, Ryan PB, O'Hara DJ, et al. Development and evaluation of a common data model enabling active drug safety surveillance using disparate healthcare databases. J Am Med Inform Assoc. 2010;17(6):652-62. PMID: 20962127. DOI: 10.1136/jamia.2009.002477.

28. Voss EA, Makadia R, Matcho A, et al. Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. J Am Med Inform Assoc. 2015;22(3):553-64. PMID: 25670757. DOI: 10.1093/jamia/ocu023.

29. Ogunyemi OI, Meeker D, Kim HE, et al. Identifying appropriate reference data models for comparative effectiveness research (CER) studies based on data from clinical information systems. Med Care. 2013;51(8 Suppl 3):S45-52. PMID: 23774519. DOI: 10.1097/MLR.0b013e31829b1e0b.

30. Rijnbeek PR. Converting to a common data model: what is lost in translation? : Commentary on "fidelity assessment of a clinical practice research datalink conversion to the OMOP common data model". Drug Saf. 2014;37(11):893-6. PMID: 25187018. DOI: 10.1007/s40264-014-0221-4.

31. OHDSI. Real World Data and the PCORTF Common Data Model Harmonization Project. https://www.ohdsi.org/wp-content/uploads/2015/04/Overview_RWD-and-PCORTF-projectMay232017.pdf. Accessed June 10, 2019.

32. The Office of the National Coordinator for Health Information Technology. U.S. Department of Health and Human Services. Building Data Infrastructure to Support Patient Centered Outcomes Research (PCOR). https://www.healthit.gov/topic/scientific-initiatives/building-data-infrastructure-support-patient-centered-outcomes-research. Accessed June 10, 2019.

33. Kahn MG, Batson D, Schilling LM. Data model considerations for clinical effectiveness researchers. Med Care. 2012;50 Suppl:S60-7. PMID: 22692260. DOI: 10.1097/MLR.0b013e318259bff4.

List of Reviewers

Reviewers (alphabetical)

Regan W. Bergmark, M.D.
Associate Surgeon, Otolaryngology-Head and Neck Surgery
Patient Reported Outcomes, Value and Experience (PROVE) Center
Center for Surgery and Public Health
Brigham and Women's Hospital
Instructor, Department of Otolaryngology
Harvard Medical School

Kathleen Blake, M.D., M.P.H.
Vice President, Healthcare Quality
American Medical Association

Stephan Fihn, M.D., M.P.H.
Director, VHA Office of Analytics & Business Intelligence
U.S. Department of Veterans Affairs

Dan Levy, M.S.
VP, Data Solutions
OM1, Inc.

Fang Li, M.D.
Senior Director, Clinical Informatics
OM1, Inc.

Vandana Menon, M.D.
Vice President, Research
OM1, Inc.

James E. Tcheng, M.D.
Professor of Medicine
Professor of Community and Family Medicine (Informatics)
Duke University Health System

Appendix A. Data Harmonization and Standardization Efforts

Table A-1. General/multi-condition initiatives

NAME OF INITIATIVE	TYPE OF INITIATIVE	OBJECTIVES/WORK PRODUCT
Clinical Data Interchange Standards Consortium (CDISC) Clinical Data Acquisition Standards Harmonization (CDASH)	Data harmonization	Provides basic recommended data elements for 18 domains (e.g., demographics, adverse events) that are common to most therapeutic areas and most phases of clinical research.[1]
National Institute of Neurological Disorders and Stroke (NINDS) Common Data Elements (CDE) Project	Data harmonization; Repository	Develops data standards for use in clinical research within the neurological community and maintains a catalog of these data standards.[2,3]
Core Outcome Measures in Effectiveness Trials (COMET)	Data harmonization; Repository	Collects resources relevant to core outcome measure sets to facilitate the exchange of information and foster new research.[4]
Agency for Healthcare Research and Quality (AHRQ) Common Formats	Data harmonization; Repository	Provides common definitions and reporting formats for to help providers uniformly report patient safety events. Also includes metadata registry with data element attributes and technical specifications.[5]
European Clinical Research Infrastructures Network (ECRIN) Database	Repository	Provides database of outcomes related to specific medical devices, taken primarily from health technology assessments (HTAs) and other relevant publications, such as systematic reviews and horizon scans.[6]
Rare Diseases Registry Program (RaDaR).	Data harmonization	Formerly known as Global Rare Diseases Patient Registry and Data Repository (GRDR). Aims to build and develop data standards and best practices for implementation across the international rare disease community.[7]
International Consortium for Health Outcomes Measurement (ICHOM)	Data harmonization	Develops standard sets of outcome measures for specific condition areas, resulting in published standard sets for multiple conditions.[8]
National Quality Forum (NQF)	Endorsement body; Repository	Endorses consensus standards for performance measurement and provides searchable catalog of quality measures.[9]
National Quality Registry Network (NQRN)	Registry network	Now part of PCPI, NQRN is a network of private and public registries and stakeholders working towards improving patient outcomes through registries.[10]
The National Patient-Centered Clinical Research Network (PCORnet)	Data harmonization (planned)	Developing a national infrastructure for patient-centered clinical research, using multiple data sources from multiple networks, which will require inter-network data harmonization.[11]
Consensus Measures for Phenotypes and eXposures (PhenX)	Measure development; Repository	Develops standardized measures of phenotypes and exposures for use in Genome-wide Association Studies (GWAS) and other research; provides a searchable catalog of measures.[12]
Patient Registry Item Specifications and Metadata for Rare Diseases (PRISM)	Repository	Developed library of questions used in rare disease registries to support re-use and eventually facilitate standardization efforts.[13, 14]

NAME OF INITIATIVE	TYPE OF INITIATIVE	OBJECTIVES/WORK PRODUCT
Patient Reported Outcomes Measurement Information System (PROMIS)	Measure development; Repository	Develops standardized measures of patient–reported health status for physical, mental, and social well-being.[15]
TREAT-NMD Registry of Outcome Measures (ROM)	Repository	Provides database of outcome measures suitable for inclusion in neuromuscular disease studies.[16]
NIH Toolbox for Assessment of Neurological and Behavioral Function	Measure development	Developed standard measures that can be used to assess cognitive, sensory, motor and emotional function across diverse study designs and settings.[17]
United States Health Information Knowledgebase (USHIK)	Infrastructure	Provides database of healthcare-related metadata, specifications, and standards.[18]
National Library of Medicine (NLM) Value Set Authority Center (VSAC)	Infrastructure	Serves the central repository for the official versions of value sets that support Meaningful Use 2014 Clinical Quality Measures (CQMs).[19]
Common Healthcare Data Interoperability Project between Pew Charitable Trusts (Pew) and Duke Clinical Research Institute (DCRI)	Data harmonization	Identifies and develops common data elements based on clinical concepts across registries, building on US Core Data for Interoperability and Health Level 7 standards.[20]
Agency for Healthcare Research and Quality (AHRQ) Outcomes Measures Framework (OMF)	Data harmonization	Assesses the feasibility of classifying consistent outcomes measures for: atrial fibrillation, asthma, depression, lung cancer, and lumbar spondylolisthesis.[21]

Table A-2. Condition-specific initiatives

CONDITION-SPECIFIC INITIATIVES	TYPE OF INITIATIVE	OBJECTIVES/WORK PRODUCT
Bleeding Academic Research Consortium	Data harmonization	Developed standardizing bleeding definitions for cardiovascular disease clinical trials.[22]
American College of Cardiology/American Heart Association Task Force on Clinical Data Standards	Data harmonization	Develops data standards for multiple areas (e.g., heart failure, cardiac imaging, atrial fibrillation, electrophysiological procedures).[23]
National Cancer Institute (NCI) Cancer Data Standards Repository (caDSR)	Repository	Provides a repository of common data elements (CDEs), metadata, and data standards used in cancer research.[24]
Diabetes Data Strategy (Diabe-DS)	Data harmonization	Created common data elements for Type 1 diabetes using a disease-specific domain analysis model.[25]
Division of Tuberculosis Elimination, Centers for Disease Control and Prevention	Data harmonization	Developed standardized treatment outcomes for multi-drug resistant tuberculosis.[26]
European Hematology Association (EHA) Scientific Working Group on Thrombocytopenias	Data harmonization	Developed standardized data definitions for treatment response for Primary Immune Thrombocytopenic Purpura (ITP).[27]
Federal Interagency Traumatic Brain Injury Research (FITBIR)	Data harmonization; Repository	Provides data dictionary based NINDS CDE project, with ability for investigators to submit alternate terms and translation rules for the same element.[28]
Grid-Enabled Measures (GEM)	Infrastructure	Facilitates virtual community of investigators to promote the use of standardized measures that are tied to theoretically-based constructs and facilitate the ability to share resulting harmonized data.[29]
Harmonizing Outcome Measures for Eczema (HOME)	Data harmonization	Continues to update “roadmap” and release publications on the development and implementation of core sets of outcome measurements, including quality of life measurements.[30]
North American Association of Central Cancer Registries	Data harmonization	Develops and promotes the use of uniform data standards for cancer registries.[31]
National Cardiovascular Research Infrastructure (NCRI)	Data harmonization	Developed harmonized cardiovascular data definitions for clinical research, patient registries, and patient care by using existing data elements and creating new data elements, when necessary.[32]
National Database of Autism Research (NDAR)	Repository	Provides a data dictionary with pre-defined data structures, as well as tools to support the development of community data standards.[33]
Outcome Measures in Rheumatology (OMERACT)	Data harmonization	Develops core sets of outcome measures for use in rheumatic diseases using a documented, reproducible process.[34]

References for Appendix A

1. Clinical Data Interchange Standards Consortium (CDISC). Clinical Data Acquisition Standards Harmonization (CDASH). https://www.cdisc.org/standards/foundational/cdash. Accessed August 19, 2019.

2. Stone K. NINDS common data element project: a long-awaited breakthrough in streamlining trials. Ann Neurol. 2010;68(1):A11-3. PMID: 20583225. DOI: 10.1002/ana.22114.

3. NINDS Common Data Elements. https://www.commondataelements.ninds.nih.gov/. Accessed August 19, 2019.

4. The COMET (Core Outcome Measures in Effectiveness Trials) Initiative. http://www.comet-initiative.org/. Accessed June 4, 2019.

5. Agency for Healthcare Research and Quality. Common Formats. http://www.pso.ahrq.gov/common. Accessed August 19, 2019.

6. European Clinical Research Infrastructure Network (ECRIN). Medical Device Outcome Measure Database. http://outcome-measure.ecrin.org/. Accessed August 19, 2019.

7. National Institutes of Health. National Center for Advancing Translational Sciences. Rare Diseases Registry Program (RaDaR). https://ncats.nih.gov/radar. Accessed August 15, 2019.

8. International Consortium for Health Outcomes Measurement (ICHOM). https://www.ichom.org/. Accessed June 10, 2019.

9. National Quality Forum. http://www.qualityforum.org. Accessed August 19, 2019.

10. National Quality Registry Network. . https://www.thepcpi.org/page/NQRN. Accessed January 17, 2019.

11. PCORnet: The National Patient-Centered Clinical Research Network. Patient-Centered Outcomes Research Network. https://www.pcori.org/research-results/pcornet-national-patient-centered-clinical-research-network. Accessed June 10, 2019.

12. RTI International. PhenX Consensus Measures for Phenotypes and Exposures. https://www.phenx.org/. Accessed August 19, 2019.

13. Richesson R, Shereff D, Andrews J. PRISM Library: Patient Registry Item Specifications and Metadata for Rare Diseases. J Libr Metadata. 2010;10(2-3):119-35. PMID: 21057650. DOI: 10.1080/19386389.2010.506385.

14. Richesson RL, Shereff D, Andrews JE. Standardization of Questions in Rare Disease Registries: The PRISM Library Project. Interact J Med Res. 2012;1(2):e10. PMID: 23611924. DOI: 10.2196/ijmr.2107.

15. National Institute of Health (NIH). Patient-reported Outcomes Measurement Information System (PROMIS). http://www.healthmeasures.net. Accessed August 16, 2019.

16. TREAT-NMD. Registry of outcome measures. http://www.researchrom.com/. Accessed August 19, 2019.

17. National Institutes of Health and Northwestern University. NIH Toolbox: For the assessment of neurological and behavioral function. http://www.healthmeasures.net/explore-measurement-systems/nih-toolbox. Accessed August 19, 2019.

18. Agency for Healthcare Research and Quality. United States Health Information Knowledgebase (USHIK). https://ushik.ahrq.gov/mdr/portals. Accessed August 19, 2019.

19. Value Set Authority Center (VSAC). https://vsac.nlm.nih.gov/. Accessed June 10, 2019.

20. Registry Data Standards. Duke Clinical Research Institute. https://dcri.org/registry-data-standards/. Accessed June 10, 2019.

21. Agency for Healthcare Research and Quality. Outcome Measures Framework. https://effectivehealthcare.ahrq.gov/topics/registry-of-patient-registries/outcome-measures-framework. Accessed June 10, 2019.

22. Mehran R, Rao SV, Bhatt DL, et al. Standardized bleeding definitions for cardiovascular clinical trials: a consensus report from the Bleeding Academic Research Consortium. Circulation. 2011;123(23):2736-47. PMID: 21670242. DOI: 10.1161/CIRCULATIONAHA.110.009449.

23. Boris JR, Beland MJ, Bergensen LJ, et al. 2017 AHA/ACC Key Data Elements and Definitions for Ambulatory Electronic Health Records in Pediatric and Congenital Cardiology: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Data Standards. Journal of the American College of Cardiology. 2017;70(8):1029-95. PMID: 28716477. DOI: 10.1016/j.jacc.2017.06.027.

24. National Cancer Institute. Cancer Data Standards Registry and Repository (caDSR). https://wiki.nci.nih.gov/display/cadsr/cadsr+wiki. Accessed August 19, 2019.

25. Barton C, Kallem C, Van Dyke P, et al. Demonstrating "collect once, use many"--assimilating public health secondary data use requirements into an existing Domain Analysis Model. AMIA Annu Symp Proc. 2011;2011:98-107. PMID: 22195060.

26. Laserson KF, Thorpe LE, Leimane V, et al. Speaking the same language: treatment outcome definitions for multidrug-resistant tuberculosis. Int J Tuberc Lung Dis. 2005;9(6):640-5. PMID: 15971391.

27. Rodeghiero F, Stasi R, Gernsheimer T, et al. Standardization of terminology, definitions and outcome criteria in immune thrombocytopenic purpura of adults and children: report from an international working group. Blood. 2009;113(11):2386-93. PMID: 19005182. DOI: 10.1182/blood-2008-07-162503.

28. National Institutes of Health. Federal Interagency Traumatic Brain Injury Research (FITBIR). Defining Data. https://fitbir.nih.gov/content/data-definition. Accessed August 19, 2019.

29. National Cancer Institute. Grid-Enabled Measures (GEM). https://www.gem-beta.org/Public/Home.aspx. Accessed August 19, 2019.

30. Harmonising Outcome Measures for Eczema. http://www.homeforeczema.org/. Accessed August 19, 2019.

31. North American Association of Central Cancer Registries. Data Exchange Standards and Record Description, Volume I. https://www.naaccr.org/data-exchange-standards-record-description/. Accessed August 19, 2019.

32. Anderson HV, Weintraub WS, Radford MJ, et al. Standardized cardiovascular data for clinical research, registries, and patient care: a report from the Data Standards Workgroup of the National Cardiovascular Research Infrastructure project. Journal of the American College of Cardiology. 2013;61(18):1835-46. PMID: 23500238. DOI: 10.1016/j.jacc.2012.12.047.

33. National Database for Autism Research. https://nda.nih.gov/about.html. Accessed August 19, 2019.

34. OMERACT. Outcome Measures in Rheumatology. https://omeract.org/. Accessed June 10, 2019.

Appendix B. Key Questions When Planning To Obtain Data From Other Sources

Selecting Data Sources

- Is this data source appropriate? In other words, are the data relevant for the registry purpose? Are the data reliable, meaning captured with sufficient consistency and accuracy to meet the registry objectives? Are the data sufficiently complete for the registry purpose?
- Is it feasible to obtain the data at the appropriate frequency for the registry purpose (e.g., once, annually, quarterly, in real-time, etc.)?
- What ethical and legal requirements must be met to obtain these data?
- Will de-identified or identifiable data be obtained?
- What is the cost of obtaining these data?

Technical Considerations

- Are any data standards used?
- Is a data dictionary available?
- Will the data be transformed to align with the registry data model?
- Will any variables be derived?
- Will any data be extracted using unstructured data (e.g., extraction of data elements from unstructured text using natural language processing)?
- How will data be linked to the appropriate patient within the registry?

Operational Considerations

- What approach will be used by registry stewards to train sites/individuals to submit data?
- How will the registry monitor data quality over time?
- How often will the data be transferred?
- How often do the data change? How will changes in data be handled within the registry?
- How will conflicting data be handled?
- How will missing data be handled?