The Office of the National Coordinator for Health Information Technology

National Health IT Priorities for Research

BACKGROUND REPORT

Prepared on behalf of the U.S. Department of Health and Human Services (HHS), Office of the National Coordinator for Health Information Technology (ONC) under Contract: HHSP233201600021I, RTI International.

October 2020

Table of Contents

Table of Figures

Table of Tables

EXECUTIVE SUMMARY

Introduction. The Office of the National Coordinator for Health Information Technology (ONC) Chief Scientist Division (CSD) is charged with developing and evaluating ONC's overall scientific efforts and activities.

To guide its priorities, CSD will develop a scientific framework and action plan focused on the work needed to advance the nation's health information technology (IT) infrastructure over the next 3 to 5 years in support of advancements in biomedical and health services research.

This background report is intended to support that effort by reviewing the following:

- Current federal and nonfederal work to advance the national health IT infrastructure to support biomedical and health services research
- Research initiatives under way and anticipated that drive national health IT infrastructure requirements
- Gaps in the health IT infrastructure that may impede those initiatives

The overall aim is to identify needs and gaps regarding the use of health IT to support the biomedical and health services research enterprise.

Organization of this Report. This background report presents an introduction and methods in Chapter 1, federal work led by ONC and others to advance the health IT infrastructure to support biomedical and health services research in Chapter 2, selected research initiatives and challenges they address in Chapter 3, selected peer-reviewed and grey literature in Chapter 4, and high-priority gap areas in Chapter 5.

Methods. The RTI team reviewed selected published and grey literature that would inform an understanding of advances in the health IT infrastructure to support biomedical and health services research. We used the ONC's definition of health IT (https://www.healthit.gov/policy-researchers-implementers/glossary) and defined infrastructure as the underlying framework or features of a system or organization (http://www.dictionary.com/browse/infrastructure?s=t). Research is defined broadly to include biomedical and health services research since health IT plays a significant role in advancing each.

The overall health IT focus is broad and includes technologies used as part of care delivery to capture, store, index, search, or extract health-related data, including data that can be used in research. However, we do not include as "health IT" systems that are designed and used primarily to conduct research in a research setting.

Literature was selected from materials provided by ONC experts, the project team, and a focused scan of the published and grey literature. Potential keywords, citations, and citing literature of relevance to the main topic were identified, along with federal reports, the peer-reviewed literature, and the grey literature. Literature was prioritized, reviewed, and summarized in Chapters 2-4, and gaps areas were synthesized in Chapter 5.

Review Findings. ONC and other federal agency priorities are summarized in Chapter 2, including ONC authorities under the Health Information Technology for Economic and Clinical Health and 21st Century Cures Act, ONC work to support Patient-Centered Outcomes Research, the Precision Medicine Initiative (PMI), and a review of Health IT Standards Committee recommended standards to support precision medicine. Additional federal initiatives at the Food and Drug Administration, National Institutes of Health, National Cancer Institute, National Library of Medicine, and Veterans Health Administration are identified.

Notable research initiatives are described in Chapter 3, including the *All of Us* research program, Million Veteran Program, PrecisionFDA for next generation sequencing, and several research platforms for extracting and using electronic health record (EHR) data for research. Activities sponsored by organizations including professional societies, foundations, non-profit organizations, and commercial organization are also briefly described.

Chapter 4 reviews selected peer-reviewed and grey literature, including several National Academy of Medicine reports that address research and the role of health IT, key documents focused on principles for data exchange and patient-centered outcomes research, consent, and common data models.

Gap Areas. Chapter 5 identifies six gap areas important for the advancement of research that intersect with ONC's authority and are not currently addressed sufficiently through ongoing health IT initiatives. Gap areas include specific gaps that may related to governance, policy, services, data, and/or standards. The following gap areas were identified:

1. **Limited health IT prioritization of research**
   As health IT is incrementally adapted to address new scientific discoveries, operational and workflow needs, evolving business and clinical priorities, organizational shifts, and updated technologies, adaptations should also reflect research priorities.
2. **Limited health IT production of research data**
   Health IT routinely produces tremendous amounts of data that are increasingly being leveraged for research. However, systems do not consistently perform data collection in ways that would better support research, consistently identifying the data captured, its timing, its level of granularity, how it was captured, and the context during collection.
3. **Limited health IT support for research platforms**
   Health IT serving as the data source for a research platform should provide data and metadata that supports the platform's major functions, including: receiving and processing multiple data streams, matching and linking the data, honoring data use agreements, identifying redundant data, managing updates to data and metadata, and working with varying data formats.
4. **Limited health IT support for research functions**
   Specialized functions for researchers, such as locating specific data, searching multiple data sources, indexing data of interest, querying for matching records, and identifying consenting and randomization status are relevant not only to researchers, but to other stakeholders including providers and patients. In addition, there is a growing need for health IT to support functions that easily incorporate research evidence into practice through decision support triggers and rules, application programming interfaces with third-party functions, and IT-driven changes in workflow.
5. **Limited health IT support for patient and family engagement in research**
   There is a growing opportunity to use health IT to directly engage the patient and family in research, whether by inviting individual-contributed data, supporting patient and family review of the accuracy of data, requesting individual permission to use data, or by enabling patient and family research leadership.
6. **Lack of a robust health IT architecture to support research**
   A consistent software architecture across different health IT components and technologies at many different organizations would advance research through better access to data, improved understanding of data context, more powerful tools for researchers, and more transparent coordination across disparate systems.

In summary, this background report describes federal and other initiatives, peer-reviewed and grey literature, and gap areas relevant to the development of a scientific framework to help guide ONC priorities for advancing the health IT infrastructure to support biomedical and health services research.

INTRODUCTION AND METHODS

Introduction

The Office of the National Coordinator for Health Information Technology (ONC) Chief Scientist Division (CSD) is charged with developing and evaluating ONC's overall scientific efforts and activities.

To guide its priorities, CSD is developing a scientific framework and action plan focused on the work needed to advance the nation's health information technology (IT) infrastructure over the next 3 to 5 years in support of advancements in biomedical and health services research.

This background report is intended to support that effort by reviewing the following:

- Current federal and nonfederal work to advance the national IT infrastructure to support biomedical and health services research
- Research initiatives underway and anticipated that drive national health IT infrastructure requirements
- Gaps in the health IT infrastructure that may impede those initiatives

The overall aim is to identify needs and gaps regarding the use of health IT to support the biomedical and health services research enterprise.

Report Organization

Chapter 1 introduces the purpose and methods for this background report. Chapter 2 describes work led by ONC and others to advance the health IT infrastructure to support biomedical and health services research. Chapter 3 describes selected research initiatives that leverage health IT, and the challenges they face. Chapter 4 reviews selected peer-reviewed and grey literature that identify challenges at the intersection of health IT and research. Chapter 5 presents high-priority gap areas for future work based on the materials reviewed.

Methods

To support scientific framework development, we identified selected published and grey literature that would inform an understanding of advances in the health IT infrastructure to support biomedical and health services research.

Definitions

We define health IT using ONC's definition:

> *The application of information processing involving both computer hardware and software that deals with the storage, retrieval, sharing, and use of health care information, data, and knowledge for communication and decision making.*
>
> https://www.healthit.gov/policy-researchers-implementers/glossary

Health IT infrastructure refers to

> *the underlying framework or features of a system or organization.*
> http://www.dictionary.com/browse/infrastructure?s=t

Research is defined broadly for this background report. It includes biomedical and health services research because health IT plays a significant role in advancing each.

The scope of this report includes technologies used as part of care delivery to capture, store, index, search, or extract real-world data used in research. We consider IT that is designed and used primarily to conduct research, in a research setting, as out of scope for this definition. We also realize that there is some flexibility in the definition, and some examples may be context-dependent.

A functional definition is favored to accommodate shifts in research needs, changes in the data and technologies that are relevant, and potential overlap between non-research and research activities. The broad concept of infrastructure includes not only software and hardware, but also policies, people, and culture.

The focus of this report is health IT infrastructure that supports a broad "work system" that produces advances in health care, biomedical research, and health services research.

Literature Selection

The aim of this background report is to focus on health IT infrastructure gaps and needs informed by the peer-reviewed and grey literature.

We began with materials provided by several industry experts from ONC and the project team that were gathered during initial project planning. These materials were used to identify potential keywords, citations, and citing literature of relevance to the main topic. The team then searched federal reports, the peer-reviewed literature, and the grey literature to gather information about the agencies and initiatives important for advancing the health IT infrastructure (see Figure 1-1).

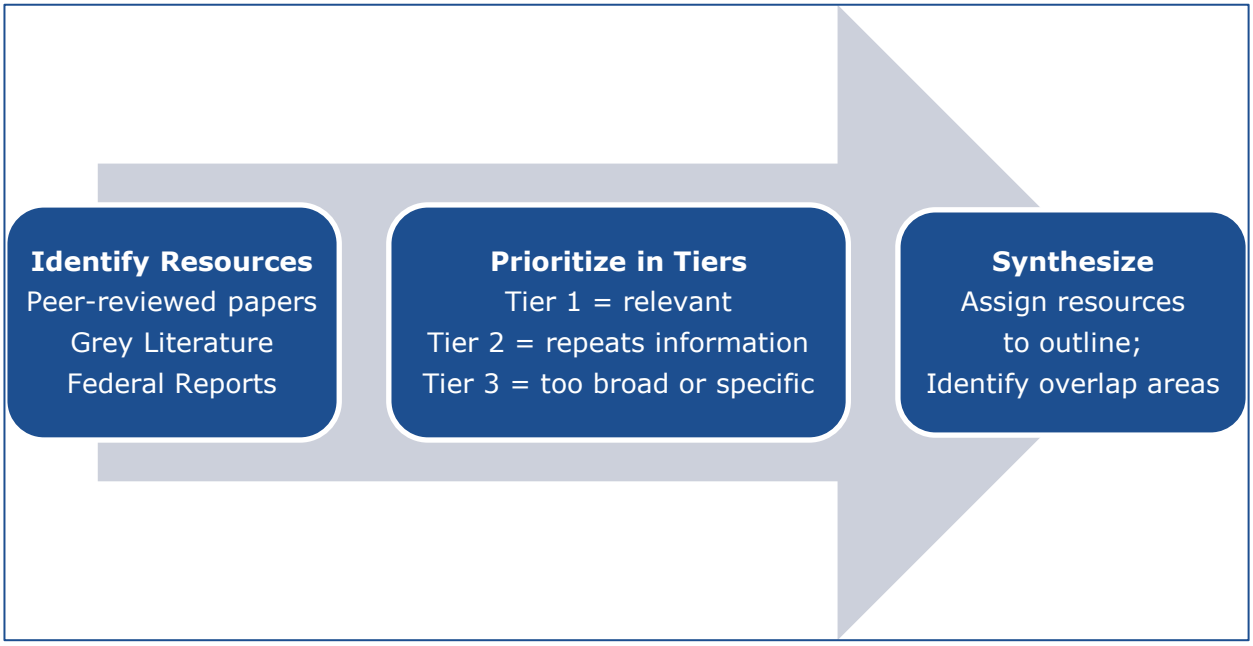


Figure 0-1. Background Report Development Steps

Sources known to publish in this area were examined closely, including *Health Affairs*, *Generating Evidence & Methods* to improve patient outcomes (eGEMS), and the *Journal of the American Medical Informatics Association* (JAMIA) in particular. We performed targeted searches using Google Scholar and Google. We used keywords such as interoperability, health information exchange, security, privacy, patient matching, and data infrastructure, and located relevant related resources based on article references and citing articles to explore specific topics and initiatives. Our team reviewed government websites from key agencies including ONC, the National Institutes of Health (NIH), the Agency for Healthcare Research and Quality (AHRQ), the Food and Drug Administration (FDA), and the National Library of Medicine (NLM).

Literature Review

Subject matter experts categorized the resources into three groups relevant to this work:

1. Current health IT infrastructure needs and goals,

2. Emerging health IT infrastructure needs and goals related to new areas of research such as the Precision Medicine Initiative and the Million Veteran Program, and
3. Specific health IT infrastructure technology and policy-related challenges and opportunities anticipated in the next 5 years.

To prioritize articles and reduce redundancy, the team reviewed groups of articles in each of the three categories listed above and assigned a category of *relevant*, *repeats information*, and *too broad or specific*. The final set of resources were reviewed in detail and abstracted for the report.

FEDERAL HEALTH IT PRIORITIES

This chapter describes current Office of the National Coordinator for Health Information Technology (ONC) work and other federal work to advance the health information technology (IT) infrastructure to support the needs of researchers.

ONC Priorities

Since its creation in 2004, ONC's work to develop and advance the national health IT infrastructure has been focused on coordinating the national effort to implement and use health IT and facilitating the electronic exchange of health information. ONC's role as the principal federal entity at the forefront of health IT infrastructure efforts (1) was strengthened by the Health Information Technology for Economic and Clinical Health (HITECH) Act (2), which tasked ONC with working to develop and implement health IT policies and standards to support a nationwide health IT infrastructure and the electronic exchange of health information to improve health care quality and safety (3).

Electronic health record (EHR) adoption sharply increased in response to the American Recovery and Reinvestment Act of 2009 (ARRA) investment of funds and the HITECH Act to incentivize EHR adoption. As of 2016, 78% of office-based physicians and 96% of hospitals are using certified EHR technology (4). Although a major accomplishment, EHR adoption marks a commencement rather than an endpoint to transform the practice of health care and the conduct of research through the robust use of health IT.

Passage of the 21st Century Cures Act (Cures Act) (5) in 2016 strengthened ONC's mandate to improve the interoperability of health information and reduce provider burden associated with using EHRs. Each of the requirements under Title IV of the Cures Act contributes to more robust exchange of health information and greater information transparency among providers, patients, caregivers, and other consumers of health information including researchers. ONC is working to implement Cures Act provisions focused on improving the flow and exchange of electronic health information, including advancing interoperability; prohibiting information blocking; and enhancing the usability, accessibility, and privacy and security of health IT.

ONC Projects Supporting PCOR and PMI

In addition to leading work to advance interoperability and reduce EHR burden, ONC leads and contributes to several scientific initiatives (Table 2-1) that support federal health IT goals. Five ONC-led projects are helping to develop the data infrastructure that will support patient-centered outcomes research (PCOR): the Data Access Framework (DAF); structured data capture (SDC); patient-generated health data (PGHD); patient matching, aggregating, and linking (PMAL); and privacy and security framework for PCOR (PSP). In addition, ONC is collaborating on three cross-agency projects: Coordinated Registry Network (CRN) for Women's Health Technologies, Common Data Model Harmonization (CDMH), and Patient-Reported Outcomes through Health IT (PRO).

Table 2-1. ONC Projects Supporting PCOR and PMI

ONC Scientific Initiatives*

• ONC projects supporting the PCOR data infrastructure
  ○ Data Access Framework (DAF)
  ○ Structured Data Capture (SDC)
  ○ Patient-Generated Health Data (PGHD)
  ○ Patient Matching, Aggregating, and Linking (PMAL)
  ○ Privacy and Security Framework for PCOR (PSP)
  ○ Coordinated Registry Network (CRN) for Women's Health Technologies
  ○ Common Data Model Harmonization (CDMH)
  ○ Patient-Reported Outcomes through Health IT (PRO)
• ONC projects supporting the PMI
  ○ Sync for Science (S4S)
  ○ Sync for Genes (S4G)
  ○ S4S Privacy and Security

Key: ONC = Office of the National Coordinator for Health IT; PCOR = Patient-Centered Outcomes Research; PMI = Precision Medicine Initiative

*Accessed on June 11, 2018 at: https://www.healthit.gov/topic/scientific-initiatives

ONC Projects Supporting the PCOR Data Infrastructure

The DAF (6) project was focused on the identification, testing, and validation of the standards necessary to access and extract data from within an organization's health IT systems, from an external organization's health IT systems, or from health IT systems across multiple organizations (7). ONC's SDC (6, 8) project promoted the collection of structured data by helping to develop the standards necessary to retrieve, display, and fill a structured form or template, and then store the completed form in a repository or submit it to an external system.

An ONC-led PGHD (9) project conducted two pilot demonstrations, disseminated a white paper and issued a practical guide that identified best practices, gaps, and opportunities to advance the collection and use of PGHD that may improve health outcomes and lower costs. The PMAL (10) project is actually a collection of projects that are aimed at improving patient matching across research, claims, and clinical data sets (11).

The PSP project is developing tools and resources that guide readers through the responsible use and protection of EHR data for PCOR and address some of the technical barriers to the exchange of patient consent for research and treatment, payment, and health care operations (12). One of the components of the PSP project, is aimed at enabling interoperable exchange of patient consent, also known as electronic consent management (13). An ONC-sponsored landscape assessment and challenges analysis (14) documented the complexity of federal, state, and organizational requirements and the persistence of paper forms as a result. The PCTP recommended the use of "basic" yes/no choices for treatment, payment, and health care operations, and for research in general; and the use of more "granular" choices for legally sensitive EHR data such as mental health information or HIV/AIDS status. The project's scope includes work to identify, test, and validate technical standards that support an individual's consent preferences.

The CRN project will establish a standards-based clinical research network connecting three clinical registries focused on women's health and develop tools to facilitate the collection of data within these registries so that the data can be reused.

The CDMH project will harmonize the common data models from Sentinel, PCORnet, OHDSI, and i2b2 to further advance the utility and interoperability of the data within these networks for use in PCOR.

The PRO project will standardize the integration of structured PRO data into EHRs and other health IT solutions to support interoperable exchange of this information.

ONC Projects Supporting the Precision Medicine Initiative

ONC plays an important role in the Precision Medicine Initiative (PMI), which launched in 2015. ONC's role in the PMI focus is to (a) accelerate opportunities for innovative collaboration around pilots and testing of standards that support health IT interoperability for research, (b) adopt policies and standards to support privacy and security of cohort participant data, and (c) advance standards that support a participant-driven approach to patient data contribution. The Health IT Standards Committee (HITSC) identified existing standards that were ready to support the PMI, emerging standards that would be important, and gaps in available standards (Table 2-2). In addition, the HITSC identified "accelerators"—opportunities to advance and improve the standards they identified (15). These recommendations set priorities for ONC work in support of the PMI.

ONC, in partnership with the National Institutes of Health (NIH), launched Sync for Science (S4S) (16), S4S Privacy and Security, and Sync for Genes (S4G) (17, 18). S4S is a collaboration among software vendors, health care organizations, ONC, NIH, and researchers to develop and pilot read-only application programming interfaces (APIs) that permit a patient to direct the sharing of their EHR data with the *All of Us* Research Program's clinical data repository for research purposes. The S4G effort aims to develop genomic data standards and improve the use of application programming interfaces (APIs) to access and share standardized genomic data. The S4S Privacy and Security project identified safeguards and protections for privacy and security when health care providers and health IT developers would like to adopt and use APIs in their workflow.

Table 2-2. HITSC Recommended Standards to Support PMI (15)

Health IT Standards Committee Recommendations
Readily Applicable Standards for Precision Medicine • Precision medicine efforts should align to standards currently referenced in the 2015 Interoperability Standards Advisory where they are included in current regulation, including EHR Incentive Program and Health IT Certification Rules (C). • Use standards to capture and represent family health history such as SNOMED CT and the HL7Version 3 Implementation Guide: Family History/Pedigree for familial relationships to express as a pre-coordinated or post-coordinated code (B). • Leverage HL7 DIGITizE Actions Collaborative draft LOINC specification for pharmacogenomics by supporting ongoing IOM Genomic Roundtable efforts (B).

Health IT Standards Committee Recommendations
Promising Standards for Precision Medicine • Support HL7 Clinical Genomics WG standards development, including CDA R2 Clinical Genetics Reporting, Clinical Genomics Pedigree Model, HL7 Genetic Testing Results Message (V2), and Clinical Sequencing Domain Analysis Model (DAM) (B). • Open ID Connect, OAuth and UMA should be considered for authorization and authentication; further piloting and testing should be considered (B). • Include more complete authorization standards (e.g., Integrating the Healthcare Enterprise cross-enterprise user assertion); ensure authorization standards are compatible across disparate networks (C). • Support Global Alliance for Genomics and Health (GA4GH) work to address computable consent in research context (C).
Standards Gaps for Precision Medicine • ONC should convene a stakeholder group to address granular, dynamic computable consent. There are existing standards in this space, but without clear implementation guidance, and alignment between HIPAA and Common Rule should be addressed (A). • Race and ethnicity: OMB Standard may be suitable for some purposes but inadequate for precision medicine and directing therapy or clinical decisions (A). • ONC should work with stakeholders to define what is the minimum data set and/or means required to make precision medicine data useful in an EHR and in a clinical setting (A). • Microbiome, exposome data standards (C). • Capture of sexual orientation and gender identity remain challenging; ONC should consider how to advance recent efforts of the Fenway Institute in this area (B).
Accelerators • 2016 Precision Medicine Initiative pilots: ONC should invest in pilots of FHIR to support individual data donation use case, which is centrally important to the Precision Medicine Initiative and NIH Cohort (B). • Support incorporation of Human Phenotype Ontology in the Unified Medical Language System (UMLS) Metathesaurus and connections between the Human Phenotype Ontology and SNOMED CT (C). • Support ongoing work Online Mendelian Inheritance in Man (OMIM): Codes for phenotypes, genotypes and links between the two (C). • Support dbSNP and ClinVar opportunity to develop a service that would get consumer data from these sources and synthesize so it is digestible for a clinical information system (C).

Health IT Standards Committee Recommendations

Key:

(A) ONC should form an additional Task Force to advance this recommendation for PMI

(B) ONC should apply various tools (e.g., S&I Initiative, pilot project, policy guidance) to advance these existing or emerging standards

(C) ONC should follow and monitor existing standards development processes

Abbreviations: HIPAA = Health Information Portability and Accountability Act; SNOMED CT = Systematized Nomenclature of Medicine—Clinical Terms; dbSNP = Database of Short Genetic Variations; ClinVar = Clinical Variants database; LOINC = Logical Observation Identifiers Names and Codes

Earlier ONC Work to Support Interoperability

To address the challenge of accessing EHR data using third-party applications, a Strategic Health IT Advanced Research Projects (SHARP, 2010-2014) project called Substitutable Medical Applications and Reusable Technologies (SMART) worked to develop new capabilities (19, 20). Modeled conceptually around the iOS platform, the project piloted the development of software "wrappers" for legacy EHRs that defined standard interactions using APIs and a well-specified data model that fostered consistent access to EHR data without the need to adapt each API to different EHR products or installed instances. SMART contributed to the development of the Health Level 7® (HL7) Fast Healthcare Interoperability Resources® (FHIR) standard, developed an open-source FHIR API server, and catalyzed the creation of starter applications (e.g., Patient Selector, Pediatric Growth Charts, Blood Pressure Centiles, and Cardiac Risk) to demonstrate the feasibility of the architecture. A SMART App Gallery (21) has been created to allow third-party developers to list their apps and EHR users to view and compare them.

Other Federal Organizations and Initiatives

A number of other federal initiatives that leverage key data or components provided by the health IT infrastructure are shown in Table 2-3. Several initiatives are further described in Chapter 3.

NIH is the largest funder of federally supported biomedical research, one of many agencies and organizations that fund research and/or components of medical research activities. The health IT infrastructure supports a variety of technologies and wide-ranging activities, from EHR phenotyping (22); to data science innovations such as machine learning (23, 24); to taxonomies used in clinical, billing, and finance applications.

Table 2-3. Selected Federal Initiatives

Federal Agency	Initiative
Food and Drug Administration (FDA)	• PrecisionFDA https://precision.fda.gov/ • PreCert https://www.fda.gov/MedicalDevices/DigitalHealth/DigitalHealthPreCertProgram/default.htm • Sentinel Initiative (25) https://www.fda.gov/safety/fdassentinelinitiative/ucm2007250.htm

Federal Agency	Initiative
National Institutes of Health (NIH)	• *All of Us* Research Program https://allofus.nih.gov/ • BRAIN Initiative https://www.braininitiative.nih.gov/ • DataScience@NIH https://datascience.nih.gov/
National Cancer Institute (NCI)	• Cancer Moonshot https://www.cancer.gov/research/key-initiatives/moonshot-cancer-initiative
National Library of Medicine (NLM)	• Partners with ONC, Centers for Medicare & Medicaid Services (CMS), FDA, VA, and other government agencies to accomplish nationwide goals and activities in the area of health IT. • Funds Logical Observation Identifiers Names and Codes (LOINC) development to promote greater interoperability of lab test results and patient assessment instruments. • Maintains the Unified Medical Language System (UMLS) ontology and works closely with Health-Level 7 (HL7) and other collaborates in the standards development process. https://www.nlm.nih.gov/
Department of Veterans Health Affairs Office of Research & Development (VA)	• Million Veteran Program (MVP) https://www.research.va.gov/mvp/

By leveraging EHR clinical data generated through routine care delivery, many research initiatives aim to produce evidence that, in turn, promotes better care. To reduce barriers to research and improve the generalizability of research findings, NIH is building a large cohort as part of the *All of Us* Research Program, which is part of the PMI (26). *All of Us* seeks to enroll a cohort of more than one million people with representation from key subpopulations in an unprecedented effort to link their biospecimen data, including genetic testing information, with their patient record information.

A related initiative is under way at the Veterans Health Administration (VHA), the Million Veteran Program (MVP). This work also uses a precision medicine approach with approximately one million veterans (and some active duty military) willing to share their biospecimen data and patient record information to investigate genetic, behavioral, and environmental effects on health among that population. Partners to this include the Department of Defense (DoD) and the Department of Energy (DOE) (27).

SELECTED RESEARCH INITIATIVES LEVERAGING HEALTH IT

A growing number of innovative research initiatives aim to leverage health information technology (IT) data and the health IT infrastructure at an unprecedented scale. Initiatives in this chapter highlight current health IT infrastructure areas that need to be strengthened, and suggest future work that will be needed as research initiatives accelerate.

The *All of Us* Research Program

With $1.5 billion in 21st Century Cures Act (Cures Act) funding for the *All of Us* Research Program (*All of Us*) over a 10-year period (28), the National Institutes of Health (NIH) is funding development of a voluntary national research cohort of one million or more volunteers. Its vision is to propel understanding of health and disease and create a foundation for a new way of doing research through engaged participants and open, responsible data sharing as part of the Precision Medicine Initiative (PMI).

Enrollment into the *All of Us* Research Program began on May 6, 2018. *All of Us* seeks to build a one million-strong study cohort of volunteers throughout the United States who electronically share their health-related data, activities, and genetic profiles (26). To achieve its goal, *All of Us* requires specialized infrastructure and considerable use of existing health IT infrastructure to recruit and enroll volunteers directly and through health provider organizations (HPOs); gain their consent; collect and track their biospecimens; and link biospecimen data to the numerous forms of electronic health data from electronic health records (EHRs), insurance systems, and personal health technologies including sensors and wearable devices (29). Beyond the physical infrastructure are policies and procedures that address how volunteers' data will be secured and protected, how volunteers enroll and withdraw from *All of Us*, and the conditions for using the data. A conceptual model (Figure 3-1) illustrates the multiple components of *All of Us* and the ways in which components interact within the infrastructure. HPO enrollees will share their EHR data directly from their provider organization, whereas direct volunteers will individually share copies of their health data, in some cases after obtaining it from their provider. HPO data contributions and direct volunteer contributions to *All of Us* require the preparation of an extract of data that is standardized for transmittal.

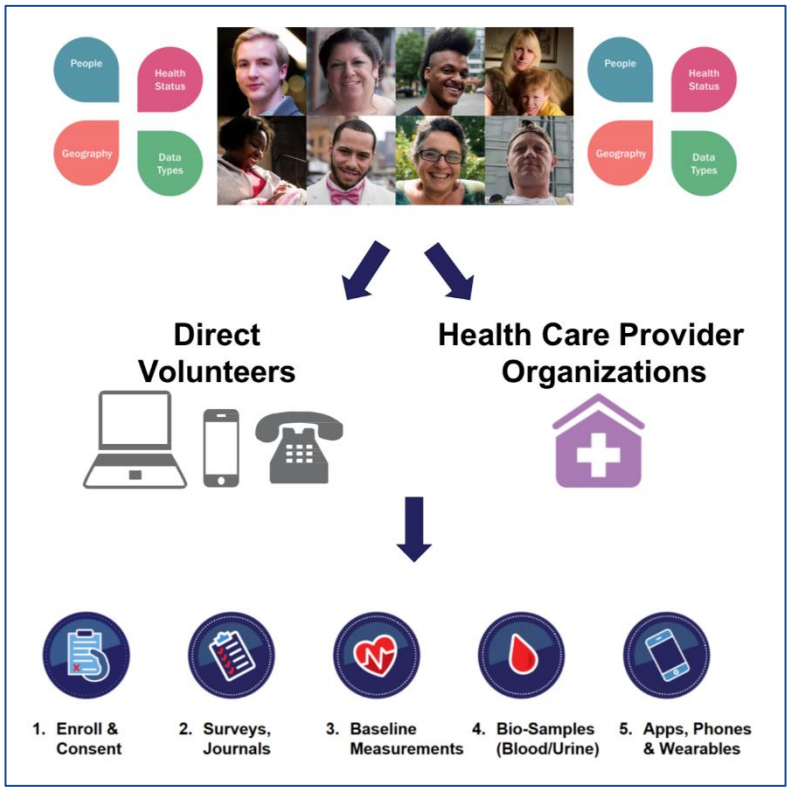


Figure 0-1. All of Us Research Program Conceptual Model (30)

The types of data collected for *All of Us* are broad, including clinical and insurance claims data; survey and demographic data; genomic and other biospecimen-derived data; and mobile, implantable, or other device

data—all of which may be stored electronically or on paper. *All of Us* will need robust data feeds from EHRs and other health IT for participants recruited by providers including those working for the Veterans Health Administration (VA), so that longitudinal data can be transferred from source systems to research databases. Direct volunteers will need their own data feeds from EHRs and other health IT, and a mechanism for ongoing sharing of their data with the *All of Us* infrastructure. Health IT systems will need ways of storing information about which data to share, how often, and with what restrictions (if any). Flexibility will be needed to accommodate changes in data content and permissions, such as novel data types added, existing data permissions modified, or data taxonomies changed (with preservation of the prior classification).

A data security policy principles and framework (31) was developed for the PMI through a broad collaboration of government stakeholders and security experts, modeled on the National Institute of Standards and Technology (NIST) risk mitigation steps of identify, protect, detect, respond, and recover. The goal of building and retaining trust among researchers and participants is paramount. Governance is a major focus because providing reliable and clear oversight to privacy and security, as well as having policies for communicating potential benefits and risks for participants (and would-be participants) to consider, are critical to ensuring participation in the long term, and to facilitating interoperability (32). It is unclear how the privacy and security framework used for the PMI aligns with broader health IT privacy and security policies and practices. Work may be needed to reconcile them.

A strong principle of *All of Us* is sharing data back with participants (33). Research findings based on a variety of selected assays and analyses may be shared back with HPOs, with participant permission. Rules are being developed to guide policies for sharing individual results with participants directly. Since many results (especially genetic results) are uncertain, and may require a skilled counselor to interpret, a thoughtful and flexible policy is desired. A related question is whether health IT will be able to receive research findings that are "shared back," which likely depends on both permissions (i.e., organizational policy for receiving and documenting third-party information) and data standards (i.e., how the information can be stored).

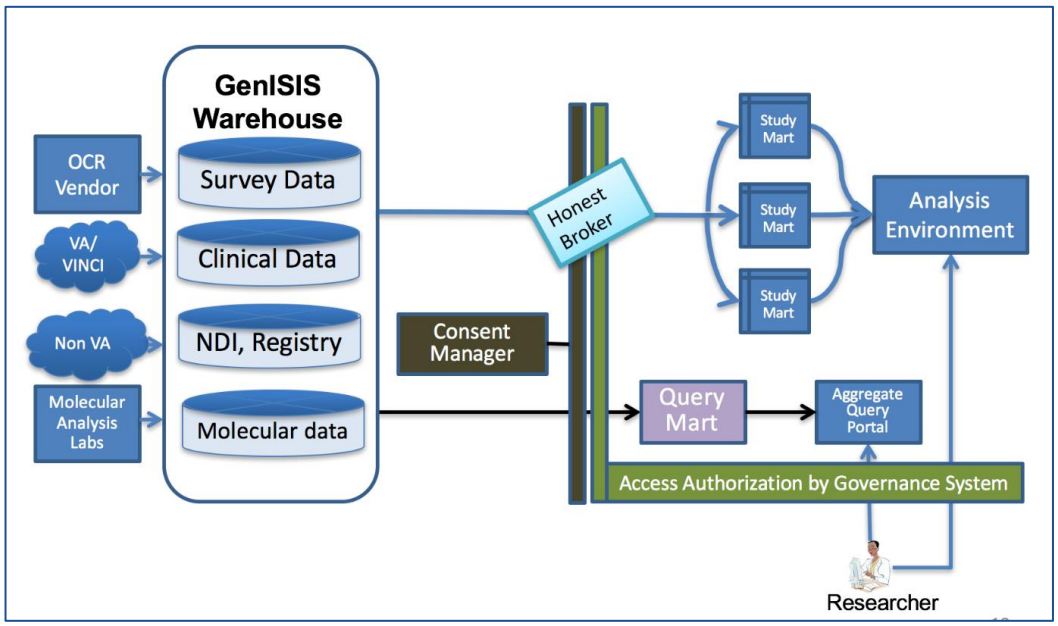
The *All of Us* Research Program will provide participants with tools allowing them to review, annotate, and contextualize the clinical data provided by them for the purposes of *All of Us* and research use, not clinical care. This brings up several questions for consideration. Should health IT systems in general be readied to accept contextualized information from participants? Record linkage technologies will be needed by HPOs to identify records they send to PMI, even after assignment of a PMI research unique ID. Is a standard approach to record linking feasible or desirable as part of a broad national health IT infrastructure? How well does this support PMI feedback to an HPO, especially if a participant changes providers?

Million Veteran Program

A prototype for developing a large biobank cohort of patients, along with associated electronic data, is the Million Veteran Program (MVP) that began in 2010, led by the U.S. VA Office of Research & Development. MVP is a VA project to recruit 1,000,000 veterans or active duty Department of Defense (DoD) volunteers willing to donate biospecimens and clinical data for research to determine how health is affected by genetics, behavior, and environmental factors (34, 35). Although this predates and is distinct from the NIH's *All of Us* research program, they both intend to follow participants longitudinally, collect genetic specimens, and capture many other kinds of data. A large group of MVP participants are anticipated to join *All of Us*.

The informatics infrastructure for storage, integration, retrieval, and analysis of MVP data, the genomic information system for integrative science (GenISIS), has four main functions: assist in recruitment activities, track MVP-collected specimens, provide secure data storage, and maintain a secure computing environment for future research analyses. GenISIS also includes additional tools for data extraction from the VA EHR to be incorporated into analytical efforts. Figure 3-2 illustrates how source data are pooled in the GenISIS data

warehouse, are made available in a de-identified form for researcher queries via the Query Mart extract, and are analyzed under an approved research study.



Key: OCR = Optical character recognition; VA = Veterans Health Affairs; VINCI = VA informatics and computing infrastructure; GenISIS = Genomic information system for integrative science; NDI = National death index

Figure 0-2. MVP Genomic Analytic Environment (GenISIS) (36)

The GenISIS system serves as a prototype for *All of Us*, offering several years of practical experience for recruiting participants, harvesting their health IT data, and developing policies for privacy, security, consent, and other areas. This review did not identify new implications for the health IT infrastructure when considering the PMI cohort program. Continuity of the MVP program will require researchers to work through the conversion from the Veterans Health Information System and Technology Architecture (VistA) EHR to the newly purchased Cerner EHR system at some point in the future. Both *All of Us* and MVP will, in general, have to plan for changes in data source systems over time, either due to system upgrades or system replacement.

High-Performance Computing

The Department of Energy (DOE) plays an important role in precision medicine research with the National Cancer Institute (NCI) following a 2015 Executive Order directing creation of a cohesive, multi-agency strategic vision and federal investment strategy in high-performance computing (HPC). The DOE leads a National Strategic Computing Initiative, which brings expertise in HPC that can reach exascale (one billion calculations per second) processing speeds and uses DOE capabilities to perform complex predictive modeling and other processer-intensive analytic tasks. A specific cancer precision medicine pilot aims to harness population and citizen science to improve understanding of cancer and patient response; gather key population-wide data on treatment, response, and outcomes; and develop novel avenues for patient consent, data sharing, and participation. The collaboration is aimed at specific innovations such as machine analysis of unstructured text; predictive modeling of cancer risk, response to treatment, and recurrence; and automated monitoring and modeling of disease (37). The long-range goal is surveillance data captured on cancer patients that includes their demographics, pathology, molecular characterization, initial and subsequent treatment, progression/recurrence data, and survival outcome or cause of death. Modeling based on this real-world data is anticipated to accelerate improved treatment outcomes and prospectively support development of new diagnostics and treatments.

Cancer Moonshot

The Cures Act authorizes a budget of $1.8 billion over 7 years for the Cancer Moonshot, which is accelerating cancer research to make more therapies available to more patients, while also improving the ability to prevent cancer and detect it at an early stage. Molecular biomarkers and the genetic basis for tumors, their response to treatment, their risk of recurrence, and their prevention are being uncovered at an accelerating pace as new tools are used for DNA sequencing and associations between genotypes and phenotypes, drug response, cancer recurrence, and other factors. A growing scientific literature on genetic variants and how they impact human physiology is being matched with patient tumor information to suggest tailored approaches to treatment.

The informatics infrastructure to support these efforts requires many components used in other areas, along with specialized tools for capturing and analyzing genomic, proteomic, and other "-omic" data. The cBioPortal for Cancer Genomics, originally developed at Memorial Sloan Kettering Cancer Center, offers a public site that provides visualization, analysis, and download of large-scale cancer genomics data sets. By providing open-access, open-source interactive exploration of multidimensional cancer genomics data sets, the cBioPortal significantly lowers the barriers between complex genomic data and cancer researchers who want rapid, intuitive, and high-quality access to molecular profiles and clinical attributes from large-scale cancer genomics projects and empowers researchers to translate these rich data sets into biologic insights and clinical applications (38).

PrecisionFDA, Real-World Data, and Regulation of Low-Risk Devices

Three FDA initiatives with relevance to the use of health IT infrastructure are described in his section.

PrecisionFDA

PrecisionFDA began in 2015 as a community platform for next-generation sequencing (NGS) assay evaluation and regulatory science exploration. The FDA is responsible for assuring accurate and reliable genetic test results, which can vary for a particular sample depending on the laboratory and methods used. The steps typically involved in processing a genetic test to produce a report and support clinical decisions have great complexity and uncertainty (39) at many points in the process (Figure 3-3). PrecisionFDA invites community members to share innovations, reference materials, and software they develop, and it provides documentation tools as well as private and public access areas so researchers, clinicians, patients, and others can access and use data and tools that advance accurate and precise genetic data and its use.

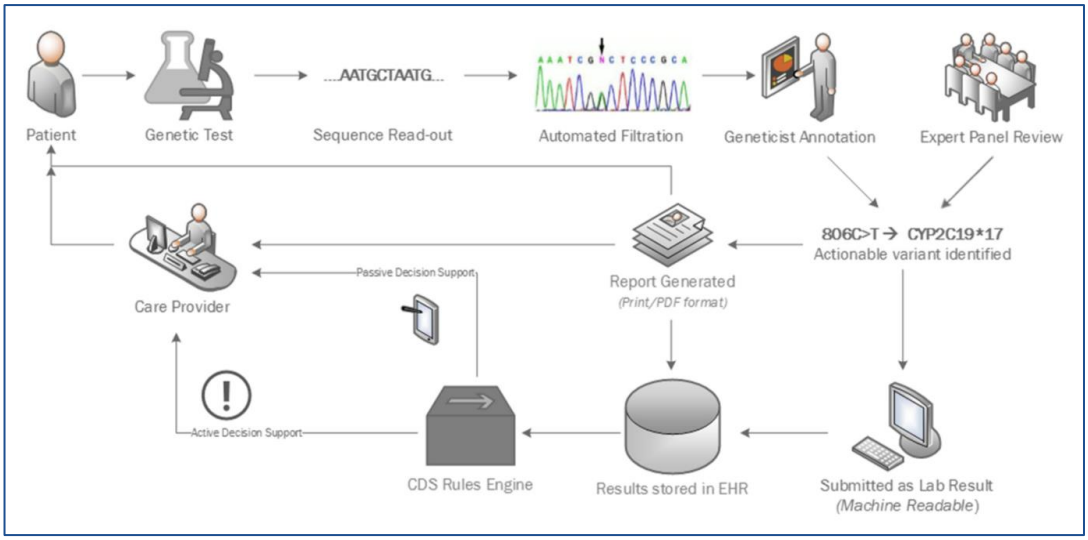


Figure 0-3. Prototypic Workflow of Genetic Testing to EHR Data (39)

Real-World Data

Since the FDA is responsible not only for pre-market approval of medical devices and drugs, but also for post-market safety surveillance, there is great interest in capturing non-clinical-trial data, otherwise known as real-world data (RWD), to inform regulatory decisions.

Collecting and using RWD is anticipated to speed premarket approvals, because the FDA believes that RWD can reduce premarket approval requirements and strengthen post-market surveillance. This creates incentives for device makers to improve and commit to RWD collection and analysis during post-market device use to speed pre-market clearance from the FDA under their risk-based regulatory determinations.

Since regulatory decisions are based on scientific evidence that is often very costly to produce, the FDA and medical device industry have a strong interest in leveraging new methods for data capture, storage, aggregation, and analysis use of EHRs and many other traditional and novel data sources to produce real-world evidence (RWE) (40). More diverse designs and strategies for conducting biomedical and health services research such as large simple trials, pragmatic randomized controlled trials (RCTs), observational studies, and the n-of-1 clinical trial are placing new demands on health IT. These approaches rely on data collected during routine care as well as potentially through research and non-care activities. Stakeholders and their roles have expanded, and now include researchers, clinicians, patients, citizen scientists, and others providing data, receiving data, analyzing data, and engaging in governance oversight to protect patient privacy as well as ensure access to data for its authorized use (40).

A good example of using RWD is the development of the FDA Sentinel System, an effort to improve the FDA's system for post-market surveillance of the safety of previously approved products (41). With new authorities in 2007 to oversee the safety of drugs after approval, the Sentinel System consists of a network of databases from 18 data partners—primarily national and regional health insurers, but also integrated delivery systems and one large hospital network—to which FDA can submit queries to evaluate product risks (42). In 2015, Sentinel received data from 4 of the largest health insurers, a number of registries, approximately 88 hospitals, and had at least some information on about 178 million people.

A concern about Sentinel has been its heavy reliance on claims data that do not reflect the details of clinical care found in the EHR, or patient outcomes directly. It also has had limited use for rapid assessment of drug safety problems because of its data sources, update frequency, and sometimes conflicting data from different databases. When a query is triggered, the central coordinating center, Harvard Pilgrim Health Care Institute, requests each data partner to run the query locally using Sentinel tools. Data partners share their results with the coordinating center, which aggregates the results across data partners. Data at each location is updated quarterly at most, with some updates lagging by a year or more. Improvements in Sentinel capabilities are anticipated with the addition of new data partners and sources, including the Innovation in Medical Evidence Development and Surveillance program, the National Patient-Centered Clinical Research Network housed at the Patient-Centered Outcomes Research Institute, and the NIH Collaboratory Distributed Research Network (42).

FDA Proposed Regulation for Low-Risk Medical Devices

To accelerate regulatory approval of innovative low-risk medical devices that capture RWD, the FDA issued a Digital Health Innovation Action Plan (43). The Plan points out that "under the Cures Act, certain medical software, including certain software that supports administrative functions, encourages a healthy lifestyle, serves as electronic patient records, assists in displaying or storing data, or provides limited clinical decision support, is no longer considered to be and regulated as a medical device." The FDA has received public comments on its draft clinical decision support guidance issued in December 2017. It also announced the intention to "pre-certify" organizations that met quality systems requirements and agreed to routinely collect real-world post-market data and apply advanced analytics as part of surveillance.

i2b2

Informatics for Integrating Biology and the Bedside (i2b2) (44) is an NIH-funded National Center for Biomedical Computing (NCBC) initiative based at Partners HealthCare System in Boston, MA, and established in 2004. Freely available, i2b2 is a scalable computational framework to address the need to translate genomic findings and hypotheses in model systems relevant to human health. By distributing the computational tools, methodologies, biomedical data sets, and educational materials widely within the biomedical and computational research communities, researchers can identify which institutions have suitable participants for a study based on inclusion/exclusion criteria such as age, gender, laboratory tests, diagnosis, consent level, and anonymized narrative reports. New methods continue to be developed within the i2b2 community, such as integrated genomic and clinical queries (45).

REDCap

Research Electronic Data Capture (REDCap) is a secure web application for building and managing online surveys and databases that can be used to collect virtually any type of data and can support 21 CFR Part 11, the Federal Information Security Management Act, and Health Information Portability and Accountability Act (HIPAA)-compliant environments. REDCap was originally developed for researchers at Vanderbilt University Medical Center, before being offered worldwide under a free license to universities and nonprofits. REDCap includes a survey library with over 150 curated survey instruments, can support offline or online data capture, can import EHR data for use in survey branching logic, and supports computer adaptive testing to limit user responses based on predictive response-pattern models (46). Since REDCap is used in over 120 countries with 500,000 projects (47), it has become a common component in the research and health IT environment for a very large number of organizations. REDCap uses a study-specific data dictionary that accommodates standardized common data models and/or a researcher-configured data dictionary, and survey data can be collected online or offline using desktop software or mobile apps.

OHDSI / OMOP

Observational Health Data Sciences and Informatics (OHDSI) (48) is a consortium to leverage the Observational Medical Outcomes Partnership (OMOP) common data model (CDM; currently v5) to study large data sets to detect drug effects observed using EHR data. Using a distributed data model and standardized queries executed in the local environment of a data holder, OHDSI participants can leverage tools and approaches designed by a community of individuals and organizations to perform health care data analytics using a variety of data types and sources. OMOP uses standardized clinical data, health systems data (e.g., providers, care sites), health economics data (e.g., claims data, cost data), and a number of OMOP-derived elements (e.g., cohort, drug era) to provide a consistent, standardized dataset. OHDSI's federated query capability and use of a common data model to generate RWD for use by not only researchers, but many other stakeholders, is an example of an important approach that researchers and patients anticipate the health IT infrastructure will support.

PCORnet

PCORnet is a large, highly representative, national "network of networks" that collects data routinely gathered in a variety of health care settings, including hospitals, doctor's offices, and community clinics. PCORnet empowers individuals and organizations to use data to answer practical questions that help patients, clinicians, and other stakeholders make informed health care decisions (49). With a coordinating center for 13 clinical data research networks, 20 people-powered research networks, and two health plan research networks, PCORnet uses a CDM based on the mini-sentinel CDM to leverage\ health IT used routinely for care to support research. Their CDM continues to evolve and includes an extensive set of standard data elements. PCORnet is funded by the Patient-Centered Outcomes Research Institute.

PCORnet uses a Distributed Research Network (DRN) infrastructure in which each partner network securely collects and stores data in a standardized way within their own institutions. When a researcher or user (called a requestor) submits a research question through an online access point called the *front door*, that question is reviewed by the coordinating center. The coordinating center then taps the data of the individual partner networks through a specialized query format. A response to the original question is generated and sent back to the researcher (Figure 3-4). The goal of PCORnet is to significantly reduce the time and effort required to start studies and build the necessary research infrastructure to conduct them. PCORnet supports a range of study designs, including large, simple clinical trials and studies that combine an experimental component, such as a randomized trial, with a complementary observational component.

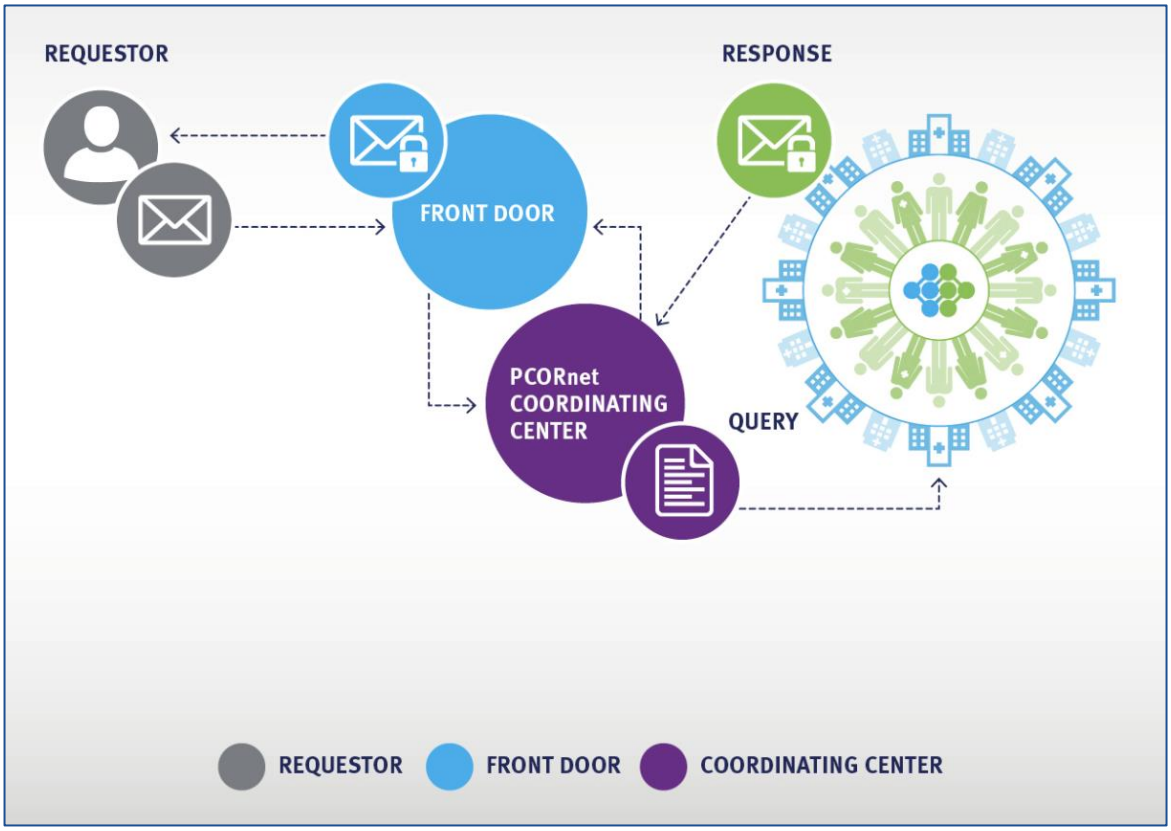


Figure 0-4. The PCORnet Common Data Model (50)

Advancing the health IT infrastructure to align with PCORnet needs will advance the kinds of research that can be conducted by patients and many stakeholders.

BioSense

The BioSense program was launched in 2003 with the aim of establishing a nationwide integrated public health surveillance system for early detection and assessment of potential bioterrorism-related illness. The program has matured over the years from an initial Centers for Disease Control and Prevention (CDC)–centric program to one focused on building syndromic surveillance capacity at the state and local levels. The uses of syndromic surveillance have also evolved, from an early focus on alerts for bioterrorism-related illness to situational awareness and response, to various hazardous events and disease outbreaks. Future development of BioSense (now the National Syndromic Surveillance Program) in the short-term includes a focus on data quality with an emphasis on stability, consistency, and reliability and, in the long term, increased capacity and innovation, new data sources and system functionality, and exploration of emerging technologies and analytics.

Data standards include Logical Observation Identifiers Names and Codes (LOINC), International Classification of Disease (ICD9/10), HL7 v2.5.1 Sex, and CDC codesets for race and ethnicity. The capability to submit electronic syndromic surveillance data to public health agencies and actual submission according to applicable law and practice was included as an objective in the Medicare and Medicaid EHR Incentive Programs, and illustrates the secondary use of EHR data for public health surveillance (51).

Professional Society, Foundation, and Nonprofit Activities

A number of nongovernment organizations conduct activities (Table 3-1) that either draw upon or contribute to the health IT infrastructure to support research.

Table 3-1. Selected Foundation and Public-Private Partnership Initiatives

Organization	Activities

American Medical Association (AMA)	• Integrated Health Model Initiative (IHMI) https://www.ama-assn.org/integrated-health-model-initiative-ihmi
	• Healthcare Interoperability and Innovation Challenge Sponsored by Google https://www.ama-assn.org/ama-google-launch-health-care-interoperability-innovation-challenge
	  ○ Import (or transfer) Patient-Generated Health Data (PGHD) from a mobile device or a mobile application into one or more phases of clinical care
	  ○ Extract (or transfer) data from one or more phases of clinical care and send it back to a mobile application or mobile device

The Pew Charitable Trusts	• Health Information Technology http://www.pewtrusts.org/en/projects/health-information-technology ○ Research focus on evidence-based policies that strengthen patient matching, promote data integration to enhance health information systems' interoperability, and improve the usability of EHRs
Medical Device Innovation Consortium	• Nonprofit 501(c)3 public-private partnership created with the sole objective of advancing medical device regulatory science http://mdic.org/about-us/overview/ • National Evaluation System for health Technology Coordinating Center (NESTcc) serves as a catalyst for the generation of real-world evidence (RWE) of sufficient quality for regulatory, coverage, patient, and clinical decision-making http://mdic.org/cc/

For example, the American Medical Association (AMA) has launched an Integrated Health Model Initiative, which includes development of a data model for organizing and exchanging information with a focus on semantics for improving clinical understanding. More recently, it has launched a Challenge Competition focused

on bringing patient-generated health data (PGHD) from a device into use during a clinical care visit, and data from the visit back to the device. This provides real-world experience in leveraging the health IT components and demonstrating data transfer in the care setting that can also support research.

The Pew Charitable Trusts continues to focus on health IT, and issued several reports in 2017 focused on patient safety, health IT usability, and the intersection of the two areas (52). Pew has provided feedback to the Office of the National Coordinator for Health Information Technology (ONC) in 2018 urging better patient matching, use of data standards to support interoperability, development of the Trusted Exchange Framework and Common Agreement (TEFCA), a robust set of data elements in the U.S. Core Data for Interoperability (USCDI), and the formation of a Health IT Safety Collaborative (53).

The Medical Device Innovation Consortium (MDIC) (54) is a nonprofit organization founded in 2012 to advance medical device regulatory science through a partnership of device makers, FDA, payers, providers, patients, and other stakeholders. A key MDIC initiative funded through an FDA grant and medical device user fees, the National Evaluation System for health Technology (NEST), was formed as a public-private partnership with device makers, insurers, researchers, consumers, the FDA, and other stakeholders to work toward a more balanced regulatory approach that would help accelerate the devices available to consumers and covered (when appropriate) by their insurers.

The NEST Coordinating Center (NESTcc) (55) was designed to serve as a catalyst in establishing functional and efficient pathways for key stakeholders to generate lower-cost, nearer real-time RWE of sufficient quality for regulatory, coverage, patient, and clinical decision-making. NESTcc specifically aims to develop, verify and operationalize methods of evidence generation and data use, demonstrate scalability across health care systems and device types and manufacturers, and build out critical functions and processes to help sustain the NESTcc initiative.

Commercial Activities

Several large clinical research initiatives work with organizations in the commercial sector that lead the use and management of electronic health data in research applications. Three key areas in which commercial organizations have been prominent are in application programming interface (API) development and use, secure cloud-based infrastructure for storing and computing clinical research data, and artificial intelligence (AI) development. Several of these are highlighted in Table 3-2 below.

Table 3-2. Selected Commercial Initiatives

Company	Initiative	Engagement with Health IT
Amazon AWS	▪ From website: "store and compute your data, collaborate with peers, and integrate your findings into clinical practice" (56) https://aws.amazon.com/	▪ Ability to process genomic data at scale for research initiatives and provide clinical integration of genomic sequencing through flexible APIs.
Apple	▪ Apple Heart Study and iPhone and iPad compatible medical devices ▪ Apple Health consolidates health data from multiple EHRs and trackers https://www.apple.com/ios/health	▪ Using consumer wearables devices and mobile apps for data collection in research studies ▪ Leveraging interoperability standards advanced by the Argonaut Project

Company	Initiative	Engagement with Health IT
Google/ Verily	▪ Verily supports the *All of Us* Research Program, collaborating with Vanderbilt and the Broad Institute on the Data and Research Center ▪ Leads projects to tailor treatments for Parkinson's, for example, including the Study Watch (physiological and environmental data), retinal imaging with AI, and Project Baseline for 4 years of data collection for 10,000 individuals (with Duke and Stanford) https://verily.com/projects	▪ Facilitating electronic data capture and management in large cohort research studies ▪ Using wearable devices and sensors for data collection in large cohort research studies
IBM	▪ Watson Health https://www.ibm.com/watson/health/	▪ Implementing AI strategies to match potential research participants to clinical trials.
Microsoft	▪ Partnering with UPMC Enterprises, a for-profit division of nonprofit health care delivery service University of Pittsburgh Medical Center (57) https://www.geekwire.com/2018/microsoft-healthcare/	▪ Leveraging AI and cloud-based platforms (Azure) to enhance the security of health data as well as facilitating efficiencies in clinical care and research
Nuance	▪ Nuance develops voice recognition technologies for use in the clinical space that convert provider dictation into codified documentation driven by EHR documentation tools and logic https://www.nuance.com/content/nuance/en_US/healthcare/clintegrity/documentation-improvement/computer-assisted-physician-documentation.html	▪ Using AI and natural language processing (NLP) to support speech-to-codified-text that enhances the use of EHR documentation in research ▪ Partnerships with over 150 EHR platforms (including Cerner, MEDITECH, and Epic)
Optum Labs	▪ Commercial subsidiary of UnitedHealth Group that is merging EHR, consumer, and payor data https://www.optumlabs.com/	▪ Leveraging big data and predictive analytics for advancements in clinical care and research

Company	Initiative	Engagement with Health IT
SalesForce	▪ Platform used by the University of California – San Francisco in Wisdom research study to develop a 100,000-person cohort to develop personalized schedules to screen for breast cancer https://wisdom.secure.force.com/portal/ https://www.salesforce.com/customer-success-stories/uc-health/	▪ Leveraging platform for participant management in large cohort research study

Among these are Amazon Web Services (AWS), which provides the back-end for high throughput batch-processing genomics at scale (56) for the PrecisionFDA initiative; IBM, which is engaged in big data projects using Watson (58); and several other technology companies (NVIDIA, Nuance, and GE) (59) that participate in a variety of artificial intelligence initiatives. In addition, Microsoft is partnering with the University of Pittsburgh Medical Center with a focus on digital hospital development spanning multiple projects (57).

New approaches in AI (59-61) and cognitive computing from companies such as IBM and Microsoft offer promising tools to help researchers, clinicians, and patients themselves wade through data to make patient-specific or study-specific decisions. These new approaches extend beyond structured and unstructured text within EHRs to digital imaging that enables providers and organizations to reduce the number of repeat studies or to use image libraries along with AI to better identify and track disease. Coupled with speech-recognition data entry tools for more efficient reporting, commercial solutions seek to integrate data collection, data analysis, and human decision-making.

In addition, the growing number of clinical- and research-grade consumer wearables are accelerating the collection, storage, and sharing of personal health data for precision medicine. For example, Apple is funding its first-ever medical research study, in partnership with Stanford Medicine, on heart health. Commercial entities are working to bring together clinical data stored in EHRs, consumer data stored in mobile devices and wearables, and claims data stored in insurance databases. Optum Labs, for example (62), offers services and products through its data assets and partnerships (e.g., AARP and Mayo Clinic) to improve health care using novel metrics based on analytics of disparate data types.

International Activities

To follow best practices where they exist and to be positioned to collaborate beyond the United States, ONC is also participating in and monitoring international initiatives. Table 3-3 highlights some of these. Exemplars in the international arena on the use of health IT for medical research include the 100,000 Genomes and Genomics England projects in the United Kingdom (UK). Similar initiatives in the European Union include EUnetHTA to promote health information exchange and related Horizon 2020 open science solutions (63).

Table 3-3. Selected European Initiatives

Organization	Initiative
Elixir	• ELIXIR is an intergovernmental organization that brings together life science resources from across Europe including databases, software tools, training materials, cloud storage, and supercomputers to form a single infrastructure for scientists to find and share data, exchange expertise, and agree on best practices. https://www.elixir-europe.org/about-us (64)
Joint Action EUnetHTA	• Promotes health information exchange across EU member states http://www.eunethta.eu/about-us (65)
The Global Alliance for Genomics and Health (GA4GH)	• An international nonprofit: "community is working together to create frameworks and standards to enable the responsible, voluntary, and secure sharing of genomic and health-related data." https://www.ga4gh.org/ (66)
EU Framework Programme for Research and Development	• Horizon 2020 https://ec.europa.eu/programmes/horizon2020/ (67)

Many countries are investing in health IT and precision medicine. For example, the UK is working to utilize genetic information for individualized health care delivery, including genetic training for providers and pharmacists. Its 100,000 Genomes Project has sequenced over 50,000 genomes as of February 2018. Through the project, the UK has developed public-private partnerships for advancing precision medicine including establishing an organization, Genomics England; establishing 13 National Health Service regional "genomic medicine centres;" contracting with Illumina, a genetic sequencing company; and contracting with companies such as Congenica and Fabric Genomics, which provide clinical interpretation services (68). These investments are advancing precision medicine in the UK by sequencing genes across the population and making the data available for use in clinical care and by pharmacies.

Across Europe, multiple agencies are driving regulatory changes to promote personalized and precision medicine (69). Personal data protection policies are being updated through the European Council's General Data Protection Regulation. And the EUnetHTA is promoting health information exchange as part of precision medicine infrastructure development (63). These and other activities are being operationalized through Horizon 2020 which, for its final 2 years of funding, supports multiple research programs and pilots that promote open science solutions to personalized medicine (70).

These research activities comprise a cross-cutting sample that are using the current health IT infrastructure and also are advancing the limits of that infrastructure.

SELECTED PEER-REVIEWED AND GREY LITERATURE

This chapter reviews selected reports from the peer-reviewed and grey literature that identify needs and challenges for the health information technology (IT) infrastructure to support research. Selected articles describe a future vision, summarize requirements, and/or discuss emerging challenges.

Early Vision for Health IT

Health IT Infrastructure Vision, Requirements, Challenges

- The health IT infrastructure should support six health system aims for improvement: safety, effectiveness, patient-centeredness, timeliness, efficiency, and equity.
- The health care system should match care to science.
- The health care system should reduce delays, wasted time, and wasted effort.
- Health IT serves the needs of patients, microsystems (practices), and organizations in a complex environment.

The landmark National Academy of Medicine (NAM; formerly known as the Institute of Medicine) report in 1999, To Err Is Human (71), alarmed the public and medical communities with news that errors in medicine that lead to harm in patients occurred frequently, and most often when systems failed rather than individuals. It called upon physicians and all others who perform within those systems to address the root causes for errors and build resilience into work systems.

The report identified health IT as sometimes helpful in reducing risk, but, in other cases, contributing to increased risk. A subsequent 2001 NAM report, Crossing the Quality Chasm (72, 73), addressed a broader range of quality problems that included patient safety. This report has been widely referenced in identifying six overarching aims for improvement: safety, effectiveness, patient-centeredness, timeliness, efficiency, and equity. Many of these aims directly relate to health IT, and to research.

The effectiveness aim, for example, highlights the importance of scientific research and feedback: "The health care system should match care to science, avoiding both overuse of ineffective care and underuse of effective care." The efficiency aim identifies the erosive burden of delays, both for patients and those who care for them, and calls for a continuous reduction of wasted effort and time. The timeliness aim describes how delays in recognition of a diagnosis or an effective treatment can be life-changing. Health IT interoperability plays a major role in the efficiency and timeliness aims, as well as the effectiveness aim.

Don Berwick, an author of both NAM reports, explained (73) the underlying four-level framework for the report. He stated that improvements in the experience of patients (level A) are largely influenced and determined by the functioning of small units of care delivery (microsystems) (level B), which reflect the organizations that support them (level C), which operate within an environment of policy, payment, regulation, accreditation, and other factors (level D). Health IT designed for level A use must satisfy requirements at all levels.

Computational Technology for Effective Health Care

Health IT Infrastructure Vision, Requirements, Challenges

- There is an overarching need for technologies that provide cognitive support to providers, to patients, and to caregivers.
- Immediate steps should focus on faithfully recording available data.
- Future-oriented steps should focus on capturing more data with greater context.
- There is a need to architect information and workflow systems to accommodate disruptive change.

The shift to an IT-based information management paradigm was examined by a 2009 NAM committee that reported on "Computational Technology for Effective Health Care— Immediate Steps and Strategic Directions." (74) The report focused on two areas: near-term improvement and the design of future systems. An interdisciplinary committee of experts found, based on site visits to eight medical centers across the United States, that there was a strong future need for "cognitive support"—tools and systems that offer clinicians and patients assistance for thinking about and solving problems related to health care.

The report identified non-IT- and IT-related factors that make automation challenging, including

- Complex medical workflows, decisions made despite a great deal of uncertainty, and a time- and resource-pressured environment;
- A confusing landscape of payers, coverage plans, and incentives that are sometimes conflicting; and
- Monolithic, inflexible systems that often automate workflows based on paper processes rather than computer-based workflows.

The committee highlighted information-intensive aspects of future health care, including the following:

- The need for comprehensive data on patients' conditions, treatments, and outcomes
- The need for cognitive support for providers and patients to integrate patient-specific data, evidence-based practice guidelines, and research results into daily practice
- Instruments and tools to manage portfolios of patients and highlight problems as they arose
- Rapid integration of new instrumentation, biological knowledge, and treatment modalities that analyze all patient experience as experimental data for continuous "learning"
- Expanded care settings that include work, home, and remote locations for monitoring and treatment
- Empowerment of patients and families in managing health decisions, with access to their own medical information and support for communication with providers

The report concluded with principles for two kinds of recommended change: evolutionary change for near-term gains, and radical change to align systems with visionary care. The latter principles include archiving data for subsequent re-interpretation (anticipating future advances in biomedical knowledge) and technologies that clarify the context of data.

Toward Precision Medicine

Health IT Infrastructure Vision, Requirements, Challenges

- Advances toward precision medicine will produce a new taxonomy of disease, new types of data, and new relationships between data.
- The concept of an information commons with constant updates, including copies of the "old" taxonomy as a reference, is critical for these advances to produce a new knowledge network.
- Health systems must not only produce data for research, but incorporate research-generated knowledge.
- In some cases, shared IT resources among researchers and clinicians are advantageous.

NAM's 2011 Toward Precision Medicine (75) report described a vision of a new taxonomy of disease based on molecular and mechanistic knowledge. The report describes new approaches to research that leverage routine collection of data outside of a research setting that uses health IT. It describes the health IT infrastructure needed to support development of a new taxonomy and lays the groundwork for precision medicine.

A new taxonomy of disease based on molecular medicine would change the current taxonomy of disease, the International Classification of Diseases (ICD), by introducing a relational structure in addition to (or instead of) the current hierarchical structure, and by requiring more frequent updates: the ICD was updated in 1992 for ICD-10, and in 1977 for ICD-9. As new diseases identified by their molecular markers change our understanding of certain diseases, it is important to retain the current observations used in ICD-10, such as physical findings, imaging, laboratory results, tissue pathology reports, reported symptoms, and various other kinds of information researched over many decades.

A new taxonomy informed by molecular medicine could reveal connections between outcomes and many other attributes, including the patient's genes (genome), other factors that influence gene expression (epigenome), the microscopic organisms that coexist inside and around the human body (microbiome), the particular physical findings they demonstrate and symptoms they experience, other reported experience (patient-generated health data [PGHD] or patient-reported outcomes [PRO]), their environmental exposure (exposome), or any other type of data that relates in some way to the patient (e.g., social determinants of health [SDOH]) that can be measured and recorded.

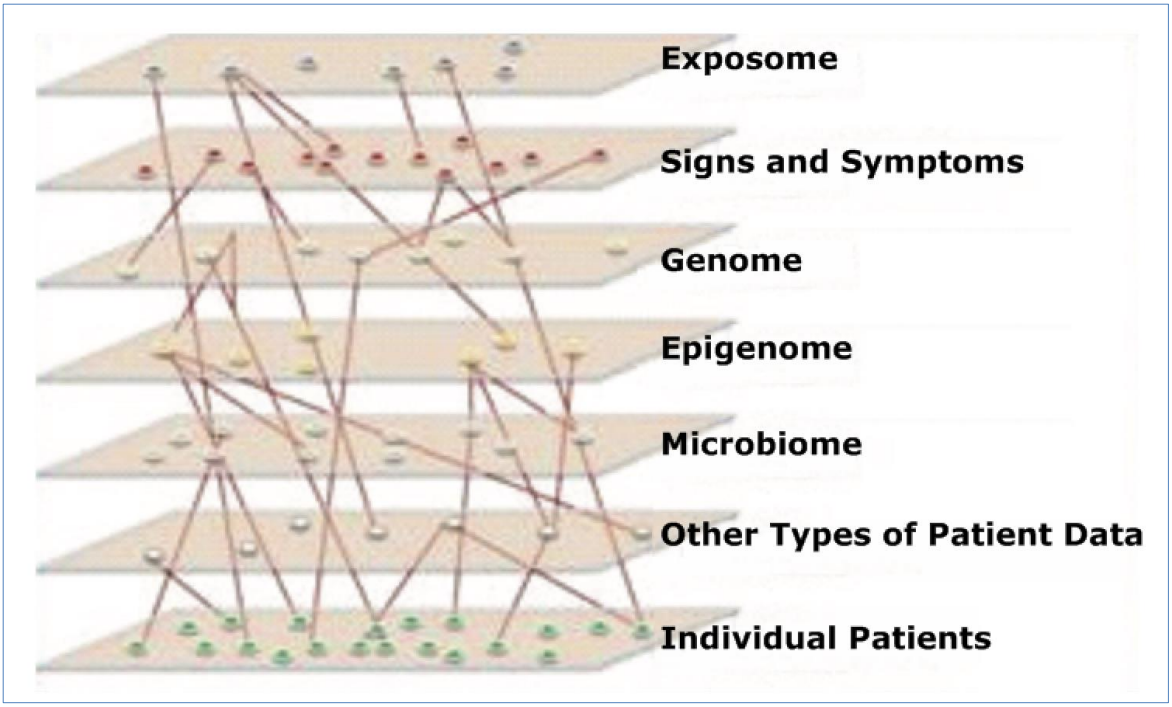


Figure 0-1. Information Commons (75)

Figure 4-1 illustrates different types of data (each layer) that relate to a set of individual patients (bottom layer), with connecting lines indicating data relationships such as co-occurrence. Figure 4-2 shows the broader knowledge network (red shaded box), surrounding the information commons. The knowledge network organizes many different data sources (circles) including EHRs, and leads to new discoveries (bottom) in biomedical research (left) and clinical medicine (right), while also enhancing understanding of disease classification (center).

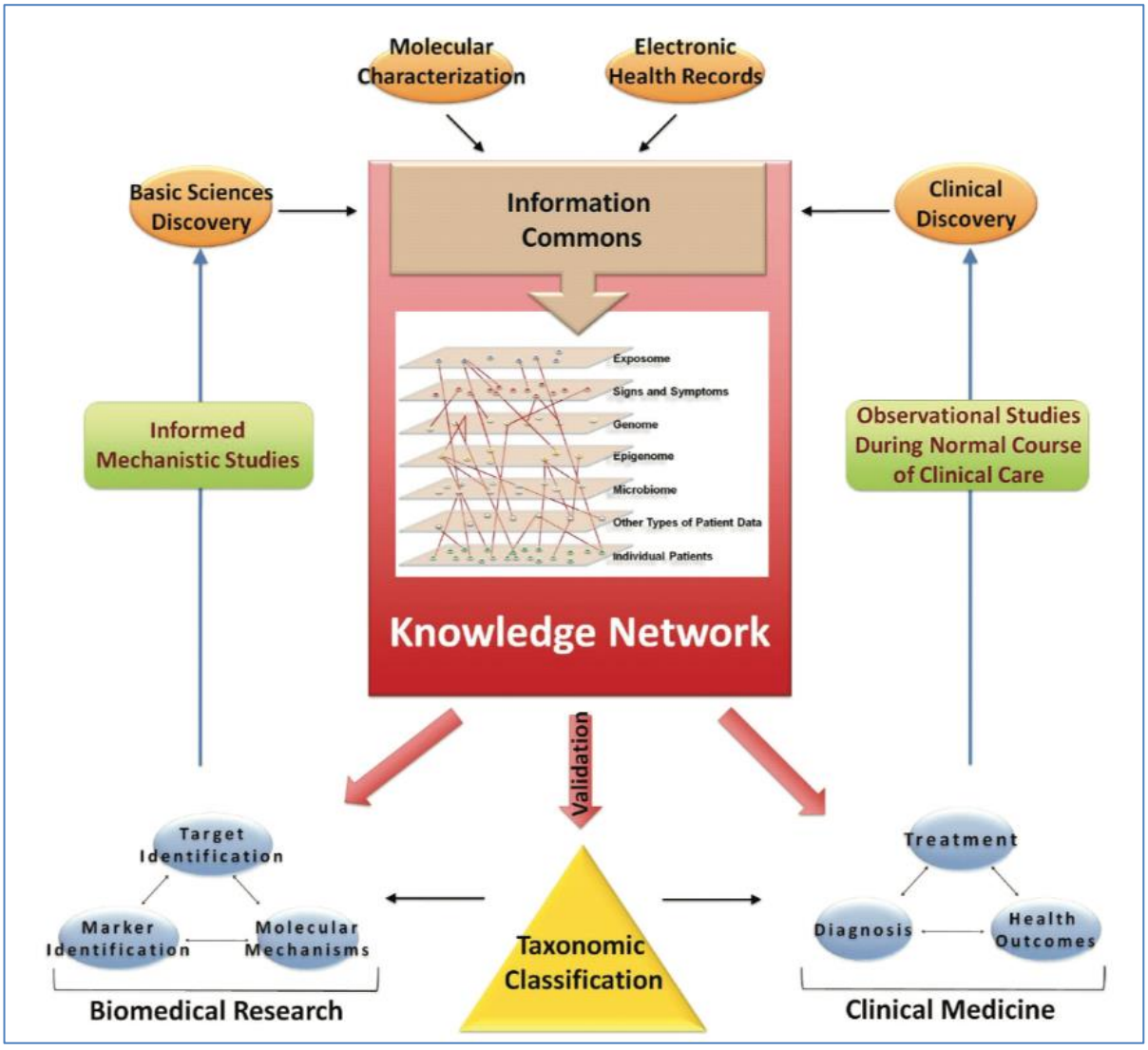


Figure 0-2. Knowledge Network (75)

The conceptual model in the *Toward Precision Medicine* report raises the questions about what parts of the information commons are part of "research IT" and what parts are part of health IT, and if they overlap. Another question relates to the growing amounts of "-omic" data that might potentially populate the EHR, along with consumer-generated data, financial data, medical and consumer device data, and other types of data. What would a scalable health IT architecture need to look like to house these diverse kinds and large amounts of data?

Furthermore, how should health IT users (e.g., clinicians and patients) document new molecular data, new disease definitions, or related scientific information if their health IT tools and systems are not yet "ready" for the new data? What would promote a flexible enough approach to accommodate new structured and coded information? What tools might clinicians and patients need to explore the information commons and knowledge network when exploring an uncertain diagnosis, treatment plan, or monitoring plan, and how might the health IT infrastructure support that?

As patient data are re-evaluated or re-interpreted as the scientific literature evolves, what mechanisms are needed to trigger re-interpretations, and to communicate new relevant findings to the patient and his/her care

team? Finally, as patient data flow between different EHRs and other health IT systems, how will different classifications used in different places be maintained and reconciled?

The report also pointed out that separately funded and staffed research projects made it more difficult to align clinical care priorities and research activities, recruit large enough sample sizes, and avoid the sometimes redundant infrastructure of research and clinical settings. It envisioned opportunities for open research systems in which researchers and clinicians worked more closely with one another, and greater opportunities for long-term follow-up of patients and feedback on clinically relevant results into a patient's clinical care. Careful design and use of a national health IT infrastructure may help to address some of these challenges.

The six final recommendations to help realize an information commons, knowledge network, and new taxonomy were as follows:

1. Conduct pilot studies that begin to populate the information commons with data.
2. Integrate data to construct a disease knowledge network.
3. Initiate a process within an appropriate federal agency to assess the privacy issues associated with the research required to create the information commons.
4. Ensure data sharing.
5. Develop an efficient validation process to incorporate information from the knowledge network of disease into a new taxonomy.
6. Incentivize partnerships.

Social Determinants of the Health

Health IT Infrastructure Vision, Requirements, Challenges

- SDOH can be used to predict future events such as the need for referrals for social services using a combination of EHR and other data.
- By leveraging health IT for standardized data collection and linkage with other key data elements, powerful predictive models can be developed.
- Predictive models are sensitive to the prevalence of the predicted event in a population.
- Predictive model use must be tailored for different settings.

A NAM report in 2014 focused on the opportunity to use EHR data to improve treatment, population management, and research to better understand SDOH, including the social and behavioral factors in the onset, progression, and treatment of disease (76). The committee made the following five recommendations to address measures and domains to capture, EHR capabilities to require, and areas of further study:

- Four measures should be captured under the Medicare and Medicaid EHR Incentive Program to demonstrate "meaningful use": race/ethnicity, tobacco use, alcohol use, and residential address.
- Eight social and behavioral domains should be captured to demonstrate meaningful use: (1) educational attainment, (2) financial resource strain, (3) stress, (4) depression, (5) physical activity, (6) social isolation, (7) intimate partner violence (for women of reproductive age), and (8) neighborhood median-household income.
- EHR certification should include appraisal of a product's ability to acquire, store, transmit, and download self-reported data germane to the SDOH and behavioral determinants of health.
- The National Institutes of Health (NIH) should develop a plan for advancing research using social and behavioral determinants of health collected in EHRs.

- The U.S. Department of Health and Human Services (HHS) should convene a task force within the next 3 years to review advances in the measurement of SDOH and behavioral determinants of health and make recommendations for new standards and data elements for inclusion in EHRs. Task force members should include ONC, Center for Medicare and Medicaid Innovation (CMMI), Agency for Healthcare Research and Quality (AHRQ), Patient-Centered Outcomes Research Institute, NIH, and research experts in social and behavioral science.

A recent study (77) captured the promise and complexity of using EHR data in combination with public and community data to predict SDOH. The authors predicted the need for social service referrals over a 5-year period with 60% to 75% sensitivity, specificity, and accuracy measures.

Real-World Evidence

Health IT Infrastructure Vision, Requirements, Challenges

- Effective data linkages are critical.
- Data collection during routine care delivery depends on cost/time commitments of the organization that performs the work.
- Health IT should support a direct patient role in capture, sharing, and reuse of real-world data (RWD).
- Unique device identifiers must be stored routinely by health IT.
- Reducing data fragmentation is a priority for patients, systems, and all stakeholders.
- Heath IT data will need to support many different types of studies.

The NAM workshop, (78) "Examining the Impact of Real-World Evidence on Medical Product Development," identified several challenges when health IT used in care delivery, payment, and operations is leveraged for research. Real-world evidence (RWE) derived from RWD is especially attractive among drug and device makers because it may help streamline the regulatory approval process (Figure 4-3).

Effective data linkage is essential for use of RWD to generate RWE. Different projects have used Sentinel data alone, linked with adjudicated medical records, linked with state registry data, linked with EHRs (ADAPTABLE trial), (79) linked with patient-generated data, and as a platform for randomized controlled trials (RCTs) through IMPACT-Afib.

The workshop found that with a shift toward evidence generation that occurs during routine care delivery, the cost and time requirements must make sense to the organization that owns the process. Collaborative relationships among stakeholders (data generators and data consumers) are key. Other workshop findings are summarized below:

- Payor coverage decisions impact patient access. Financial and economic data have a direct impact on patient treatment decisions.
- Decision makers (e.g., FDA, integrated health systems, pharmaceutical companies) fit the research design to the question being answered. Data are used to support many different research designs.

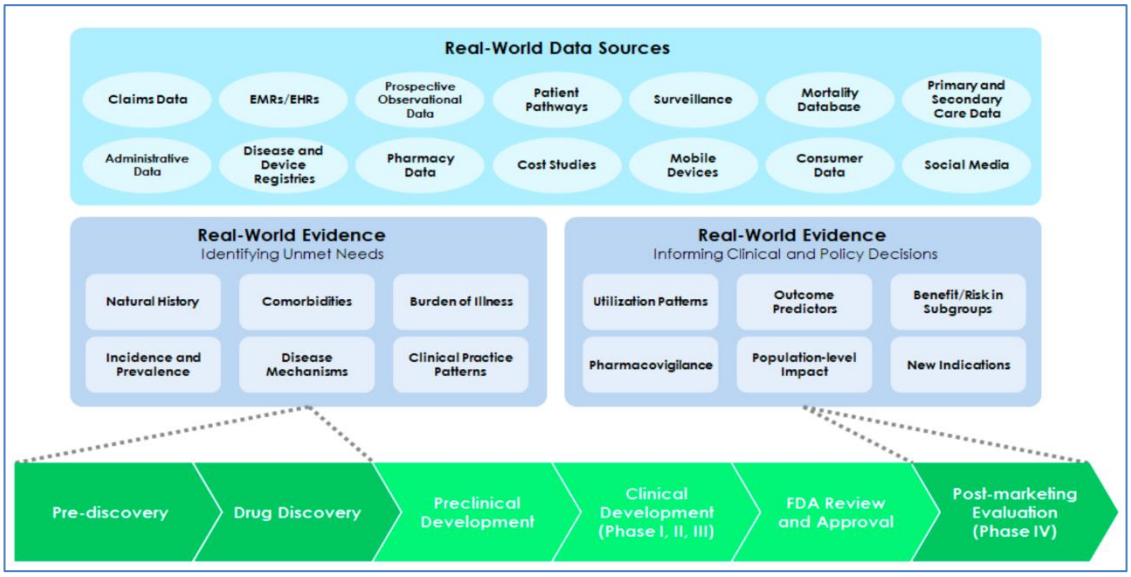


Figure 0-3.Real-World Data and Real-World Evidence Supports Drug Development and Post-Marketing Surveillance (78)

- Health information exchange (HIE) may serve as a de facto clinical data warehouse for community hospitals. Community hospital health information system interoperability with HIEs is important.
- Patients will play a direct role in the capture, sharing, and reuse of RWD.
- Successful, large sample trials (e.g., Salford Lung Studies) (80, 81) can be expensive and require investment in data processing, error management, data linkages, and other infrastructure. To adapt the study infrastructure for future studies, stakeholders converted the platform to individual, cloud-based, configurable, and modular applications.
- Unique device identifiers (for implantable and non-implantable medical devices) are helpful if stored in EHRs or claims data. Health IT should be capable of storing and retrieving unique IDs.
- Registries are useful because often they contain high-quality, fit-for-purpose, curated data that can be easily linked with other data sources through coordinated registry networks. However, they also have significant drawbacks in that they are expensive to develop and maintain, and are impractical in some situations. They vary in data quality and methods, they can pose significant administrative challenges, and they must safeguard patient privacy and security.
- Overcoming data fragmentation is a major challenge; the patient is an important beneficiary of defragmenting the information spread across an EHR or multiple instances of health IT; institutions that collect data should share it and adhere to core principles and use data only for purposes for which it is fit, to protect patient privacy and security, and follow other core principles.
- Hierarchical rating of evidence should be replaced by evidence “grades” by analytical method instead of by data source. Health IT should be able to show strength of evidence in a variety of ways.
- The Observational Health Data Sciences and Informatics (OHDSI) program relies on data aggregation from four different sources and is routinely collected to minimize bias and allow for easy comparisons. Linkages are critical.
- RCTs are necessary for certain types of studies to detect adverse effects or moderately beneficial effects of new treatments, and to establish causality. However, hybrid RCT-RWD approaches are promising, along with other non-RCT approaches. Health IT infrastructure must support the right method of research, whatever that is determined to be.

PCORTF Research Data Infrastructure Framework

Health IT Infrastructure Vision, Requirements, Challenges

- Essential components that make data more readily usable for research are (1) standards, (2) services, (3) policies, and (4) governance structures.
- A security and privacy policy framework is needed to improve researchers' ability to access and query clinical data.

In 2010 the Patient Protection and Affordable Care Act (ACA) mandated that relevant federal agencies coordinate to build data capacity for patient-centered outcomes research (PCOR), including the development and use of clinical registries and health outcomes research data networks. It specified that the aim was to "develop and maintain a comprehensive, interoperable data network to collect, link, and analyze data on outcomes and effectiveness from multiple sources, including electronic health records" (82). A portfolio of 31 projects funded by the Office of the Secretary, Patient-Centered Outcomes Research Trust Fund (OS-PCORTF) from 2012 to 2016 was examined in a 2017 report (83) that conceptualized a data infrastructure framework (Figure 4-4) based on analysis of the portfolio with level 1 (bottom) to level 4 (top). The framework identified four essential infrastructure components that make data more readily usable for research (Figure 4-4, level 2) and include:

1. Standards, or accepted specifications that ensure that the data used for research, are consistent and usable across different sources and for different uses;
2. Services, such as programming protocols and interfaces, that allow for the capture, storage, linkage, analysis, and exchange of clinical data or evidence;
3. Policies that address how data are used and ensure that data are protected and secure; and
4. Governance structures to support data sharing among organizations.

The framework identified a large variety of data resources (level 1) important for the PCOR projects studied, and five core functionalities (level 3) that health IT must perform to support the conduct of research.

The analysis found that significant progress has been made in standards adoption and use, particularly for common data elements, and in services that encompass the resources to capture, store, and exchange data. They also identified further work that was most needed: to implement policies that oversee data use, security, and privacy; and to create governance structures that support the efficient use of data. The report found significant progress has also been made toward the core functionalities (level 3) of use of clinical data for research and standardized collection of standardized clinical data, with only modest progress made in the other areas. The following potential areas for future research were identified:

- Developing technical services and standards for services that allow patient data to be securely linked to other data sources.
- Developing standards, services, and policies to assure data quality for research.
- Creating a policy framework that preserves security and privacy while improving the ability to access and query clinical data by researchers.
- Developing a better understanding and methods to address the socio-legal challenges that arise with using patient data for research.
- Engaging in dissemination efforts to promote greater awareness of OS-PCORTF initiatives and products among members of the research community.

Overall, they reported greater data capacity advances through successful structuring, linking, and sharing of electronic health information across patient groups and repositories throughout the health care ecosystem (83). They also suggested that disseminating the knowledge gained is a high priority, since researchers within the federal government and outside of it constantly need to structure, collect, link, and analyze new data types. They recommended (1) a publicly funded meta-data system or catalogue of past and present federally funded data capacity-building projects, (2) methods to continually improve data quality, (3) effective governance mechanisms among agencies, research entities, and health systems, and (4) assistance designing e-health data systems.

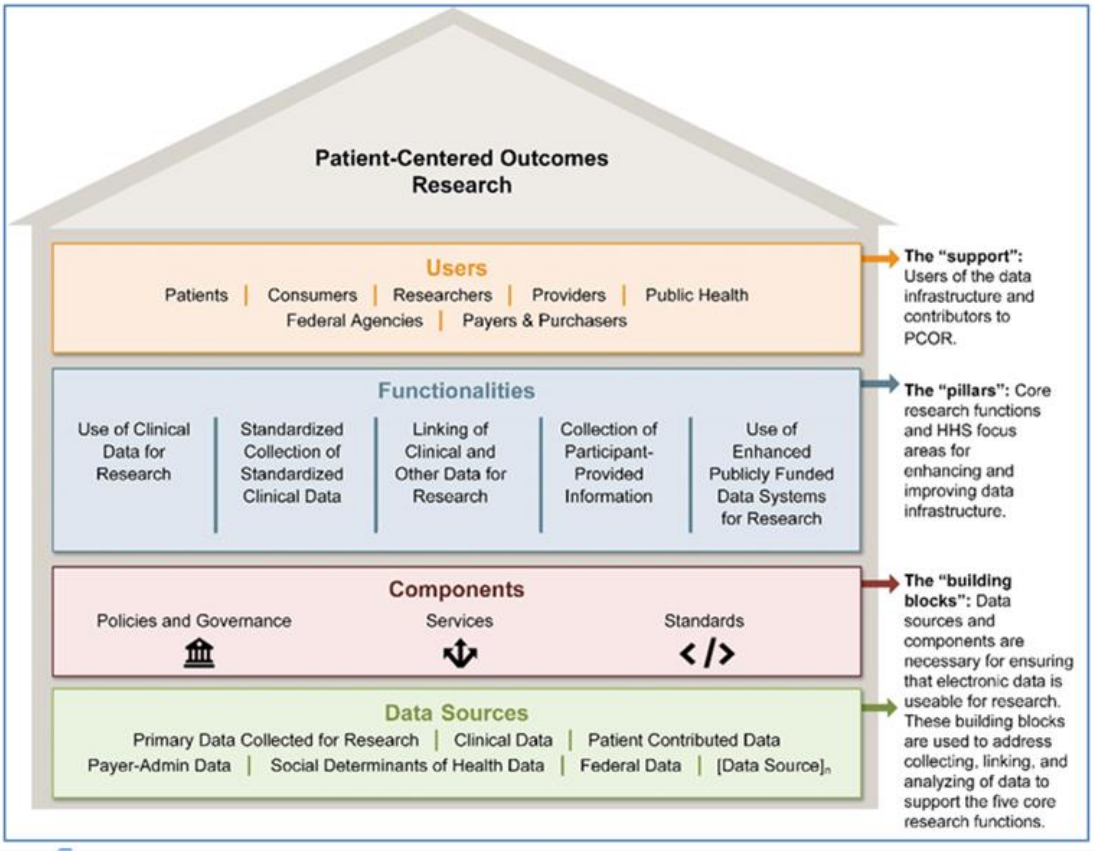


Figure 0-4. A Strategic Framework for PCOR Data Infrastructure (83)

JASON Health Care Reports

Health IT Infrastructure Vision, Requirements, Challenges

- Unifying software architecture for the exchange of health information.
- Expanded use of certified APIs for reading and writing medical record data from EHRs.
- Artificial intelligence (AI) is promising and requires high-quality training data as well as large-scale data collection to be most effective.

ONC and other agencies and foundations commissioned three JASON studies that focused on advancing the nation's health IT infrastructure and capabilities. The April 2014 JASON report entitled "A Robust Health Data

Infrastructure" (84) found lack of interoperability to be a major impediment to care and research, and reported the need for technical solutions that would repurpose clinical data for research use (85). These experts proposed moving to a "unifying software architecture for the exchange of health information" that frees data from organization and system silos and increases incorporation of diverse data, not only from different kinds of data sources, but also from individuals in the population who are under-represented in research datasets. They propose expanded use of certified APIs for reading and writing medical record data from EHRs, and definition of an overarching health IT architecture by ONC to guide infrastructure work.

Another report released in November 2014 entitled "Data for Individual Health" (86) called for an HHS framework to measure progress in advancing the health IT infrastructure toward a Learning Health System. The report also recommended "race to the top" challenges and adoption of policies that support open APIs and less rigid FDA regulatory requirements for innovation using medical devices. The report reiterated the call for HHS to lead, calling for adherence to the principles of the ONC 10-Year Vision (87) for an interoperable health IT infrastructure. Open APIs are proposed as a scalable approach to accessing data from any EHR "without special effort" as required under the Cures Act (88).

A more recent JASON report published in January 2018 (61, 89, 90) addresses the use of AI in health and health care. It focuses on the technical capabilities, limitations, and applications of AI that could be realized in the next 10 years. It noted the importance and promise of AI, the critical need for high-quality training data (91), the importance of large-scale data collection, and the need for peer review of AI methods as part of the process for accepting them as a basis for health decisions.

FAIR Principles

Health IT Infrastructure Vision, Requirements, Challenges

- Findability, accessibility, interoperability, and reusability (FAIR) are essential principles needed to ensure that data are useful to researchers.
- Humans and machines should be able to use data.
- FAIR levels of adherence provide potential benchmarks.

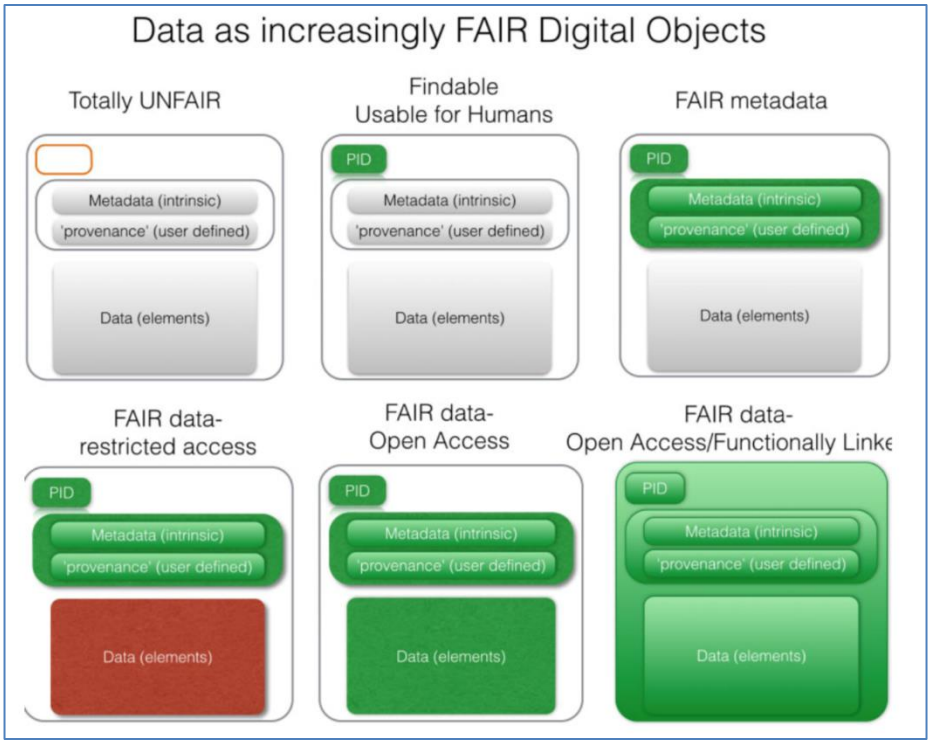
Recognizing common challenges across industries and settings, an international community of stakeholders representing academia, industry, funding agencies, and scholarly publishers began working in 2011 on the challenge of ensuring that researchers would be able to find and reuse electronic research data reliably and in a scientifically valid way in the future. The group held a workshop in 2014, and published a concise and measurable set of principles in 2016, the FAIR Data Principles (92). They identified findability, accessibility, interoperability, and reusability as essential principles needed to ensure that data are useful to researchers, and specifically address the need for both humans and their machines to be able to use data.

The FAIR principles are listed in Table 4-1. A description of what constitutes data, and how it is constructed, is found in the Guiding Principles document on the FORCE11 website (93). Many, if not all, of the FAIR principles apply to data used in biomedical and health services research. The characteristics, norms, and practices that data resources, tools, and infrastructures should exhibit to be considered 'FAIR' can be achieved with a wide range of technologies and implementations, which were deliberately not specified.

Table 4-1. FAIR Guiding Principles (92)

Principles
To Be Findable: F1. (meta)data are assigned a globally unique and persistent identifier F2. data are described with rich metadata (defined by R1 below) F3. metadata clearly and explicitly include the identifier of the data it describes F4. (meta)data are registered or indexed in a searchable resource
To Be Accessible: A1. (meta)data are retrievable by their identifier using a standardized communications protocol A1.1 the protocol is open, free, and universally implementable A1.2 the protocol allows for an authentication and authorization procedure, where necessary A2. metadata are accessible, even when the data are no longer available
To Be Interoperable: I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation I2. (meta)data use vocabularies that follow FAIR principles I3. (meta)data include qualified references to other (meta)data
To Be Reusable: R1. meta(data) are richly described with a plurality of accurate and relevant attributes R1.1. (meta)data are released with a clear and accessible data usage license R1.2. (meta)data are associated with detailed provenance R1.3. (meta)data meet domain-relevant community standards

The FAIR Guiding Principles describe and accommodate different levels of adherence (Figure 4-5), starting with a persistent identifier (PID), followed by shared metadata (intrinsic to the data object), modifiable provenance information, machine-understandable data elements (that may be locked or unlocked), and fully public metadata and data under public license. This approach provides relevant benchmarks to consider in an advanced health IT infrastructure that supports diverse researchers using data from a broad set of data sources.



Source: https://www.force11.org/sites/default/files/images/FAIR-digital-objects.jpg

Figure 0-5. FAIR Levels of Adherence (93)

Electronic Clinical Data and Patient-Centered Outcomes Research

Health IT Infrastructure Vision, Requirements, Challenges

- It is challenging to conduct research in the absence of standardized ontologies and data collection.
- There are unique data governance concerns related to the transfer, storage, deidentification, and access to electronic clinical data (ECD).
- There are gaps in the literature on topics such as the use of clinical informatics for cohort identification, cloud computing, and single-point access to research data.

A systematic literature review (94) was conducted in 2012 to characterize a new body of literature on PCOR and clinical informatics and to identify cross-cutting themes and gaps in the literature. Using medical subject heading (MeSH) search terms in PubMed, a manual review, citations from a portfolio of AHRQ study proposals, and websites from 12 initiatives, one hundred thirty-two articles were selected to be included in the review, and fell into three categories. They (1) provided historical context or frameworks for using clinical informatics for research, (2) described platforms and projects, or (3) discussed issues, challenges, and applications of natural language processing.

Two cross-cutting themes were also identified: the challenges of conducting research in the absence of standardized ontologies and data collection; and unique data governance concerns related to the transfer,

storage, de-identification, and access to ECD. A list of concepts derived from the selected articles is shown in Table 4-2.

Table 4-2. Concepts Derived from Systematic Review on Comparative Patient-Centered Outcomes Research and Clinical Informatics (94)

Systematic Review Concepts		
• Context • Clinical informatics platforms • Clinical informatics projects • Natural language processing • Data use and quality • Research networks	• Standardized data collection • Identifiers and de-identification • Security • Metadata • Patient involvement • Institutional Review Boards (IRBs) • Governance	• Library of phenotypes • The Learning Health System and Comparative Effectiveness Research (CER) • Gaps identified (no articles focus primarily on these) • Single-point access • Cloud computing • Cohort identification

Pharmacogenics and Clinical Decision Support

Health IT Infrastructure Vision, Requirements, Challenges

- A national CDS infrastructure should include five core components.
- CDS knowledge will require updating over time.
- There are challenges due to busy clinicians, commercial EHRs, and processor-intensive cloud-based logic.
- PGx CDS requires ongoing maintenance of drug and genomic information, rules to prioritize display of information, and modular applications that can function within or beside commercial EHR systems.

Because bringing clinical recommendations to the point of care is essential, the health IT infrastructure should support this capability system-wide. Inherent cognitive limits of clinicians and patients, compounded by rapidly growing knowledge and tailoring that cause recommendations to change over time, are important challenges to address and overcome.

Kawamoto et al. propose a framework for CDS (95) and described five core components of a national CDS infrastructure: (1) centrally managed repositories of computer-processable medical knowledge; (2) standardization of the associated CDS information for genomic and personalized medicine; (3) standardized representation of genomic and non-genomic patient data; (4) standard approaches for leveraging genomic knowledge repositories and patient data to guide clinical care; and (5) a standard approach for retrieving relevant patient data across health information systems.

A national CDS infrastructure will be needed to guide the appropriate use and interpretation of new genomic assays, including five core components. Although many efforts are under way, there are challenges due to busy clinicians, commercial EHRs, and processor-intensive cloud-based logic. Non-clinician-driven PGx-CDS, also with many constraints, may offer an alternative avenue for applying this knowledge.

A recent paper highlights the importance of design when considering how to introduce scalable CDS with a focus on usability. Khelifi and colleagues (96) found it challenging to apply PGx data to medication ordering using most

current health IT systems, and suggest non-embedded CDS to enhance the usability and design of commercial EHR systems.

They reviewed several projects: eMERGE (Electronic Medical Records and Genomics) (97) to bring genome and phenome data together for research, PharmGKB (98) to provide an online knowledge base of PGx publications, and CSER (Clinical Sequencing Exploratory Research) (99) to invite stakeholders to collaborate and share best practices and innovations in the field. They developed a model PGx CDS system prototype and design suggestions when searching, selecting, and personalizing a medication prescription. They proposed contextual display of PGx test results, other treatment options, and relevant patient information to reduce cognitive burden and improve clinician task performance.

Khelifi et al. introduced a contextual display to integrate PGx test results into clinician workflow. Current commercial EHR systems may not be able to implement semigraphical CDS as designed. Consideration must be given to whether modular designs would ever be able to accommodate this approach. They recognized that the medication taxonomy and PGx information are changing, making it difficult to identify relevant information for every drug-task combination. They identified the need for ongoing maintenance of drug and genomic information, along with rules to help prioritize the display of information to avoid overwhelming the clinician with too many details.

Common Data Models

Health IT Infrastructure Vision, Requirements, Challenges

- EHR data may require additional contextual information for field transformations to occur and data conflicts to be resolved.

A 2013 paper in *Medical Care* (100) compared four common data models: OMOP (EHR data), Mini-Sentinel Common Data Model (MSCDM) (EHR and claims data), Clinical Data Interchange Standards Consortium's Analysis Data Model (CDISC-ADaM) (clinical trials data), and Biomedical Research Integrated Domain Group (BRIDG) model (clinical trials and preclinical protocol-driven research data). They extracted data based on a PCOR study scenario from an EHR (Epic) into a local Epic-associated clinical data warehouse for research (CDRW). They then mapped the scenario data in the CDRW to the four common data models they studied to compare and contrast the fit.

Although many fields mapped easily, some field transformations required experts (data source designers and administrators) knowledgeable about the source data codes and context, which creates difficulty when trying to scale the effort.

Sometimes the same data from different fields need to be reconciled, which can create conflicts, such as a provider associated with a procedure (according to a billing code) versus the provider associated with a patient visit. Mechanisms to understand the context and resolve the conflict are needed.

Genome-Wide and Phenome-Wide Association Studies

Health IT Infrastructure Vision, Requirements, Challenges

- Automated phenotyping is challenging but important because GWAS and PheWAS studies are so powerful.
- Automated text abstraction using deep learning AI techniques is being explored, and may potentially be useful not only in the research context, but for care delivery purposes.
- Health IT plays an important role in coordinating reanalysis of genomic data as interpretations evolve over time.

The Electronic Medical Records and Genomics (eMERGE) Network (101) (www.gwas.org), an NIH-funded consortium of five institutions with DNA data linked to EHRs, assessed the utility of EHRs as a consistent and reliable source of phenomic data. The project has produced some important lessons:

1. eMERGE has shown that patient data obtained during the normal course of clinical care is a valid source for replicating genome-phenome associations that were previously detected only in carefully qualified research cohorts.
2. The genome-wide associations examined required more participants than any single institution had for sufficient power to detect an association.
3. High-quality EHR-derived phenotypes required free-text in addition to codes (including ICD codes, though codes have to be repeated multiple times to gain validity), laboratory-medicine results, and medication histories. Natural language processing of physician comments was essential to get high predictive values.
4. High specificity and precision was maintained across differing EHRs at the five sites.

Policy-related challenges were, overall, greater than technical challenges, particularly in achieving meaningful data sharing and respect for patient privacy concerns. A simplified data use agreement helped address these needs (97, 102, 103).

In a different study using historical data from large numbers of patient records and genetic data from biobanks that store patient DNA samples, associations between phenotypes and genotypes are used to help researchers understand drug mechanisms, efficacy, toxicity, repurposing, and many other associations (104). The two research methods, GWAS and PheWAS (Figure 4-6), are complementary. Researchers can rapidly identify genetic risk factors that were previously unknown for a given condition (phenotype). Conversely, they can also identify a variety of different conditions and other phenotypic findings associated with a particular genetic abnormality (105, 106). A particular phenotype may be defined based on ICD-9 billing codes, a medication on their problem list, free-text data, or a problem-list entry such as obesity, heart disease, or smoking history.

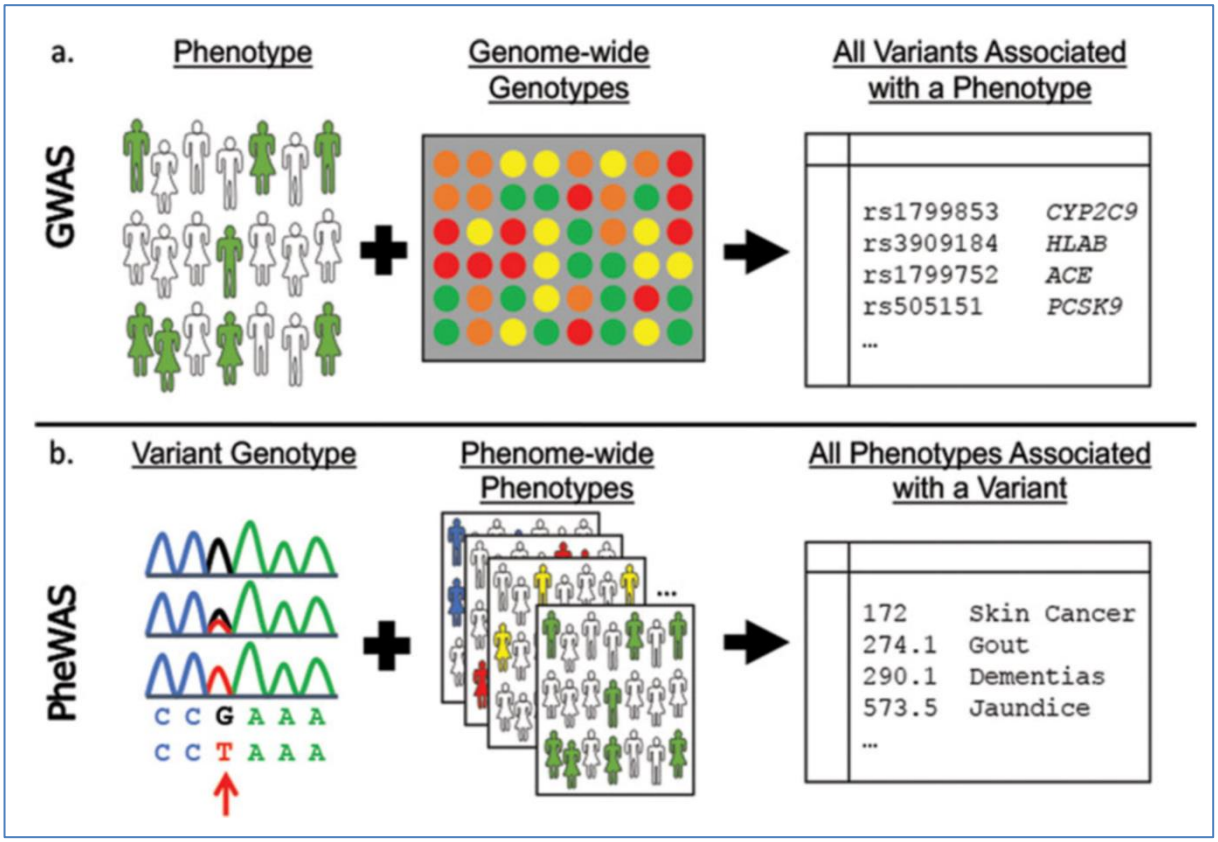


Figure 0-6. Genome-Wide Association Studies and Phenome-Wide Association Studies (104)

Using automated phenotyping instead of using manual-trained coders has been only partially successful so far. Scanned handwritten or typed reports imported as images into an EHR, dictated progress notes containing unstructured narrative, and incomplete contextual information can lead to errors such as the appearance of an incorrect diagnosis or procedure, or missing information that would change the phenotype.

To illustrate the power of automated phenotyping, Barnado et al. (107) developed an approach using diagnosis codes, medication data, and lab data to identify patients with systemic lupus erythematosus (SLE) based on their EHR data. To train their algorithms they began using a database with 2.5 million subjects, identified nearly 6,000 with at least one SLE ICD9 code, and picked sets of 100 random patient records to train their algorithms. Their best algorithm performed well, with a positive predictive value (PPV) at 91% in the validation set based on one or more counts of the SLE ICD-9 code (diagnosis code), antinuclear antibody (ANA) positive (‡1:40) (lab data), and ever use of both disease-modifying antirheumatic drugs and steroids (medication data).

A recent publication by Shickel et al. (108) described a growing variety and number of AI deep learning techniques being applied in different contexts to EHR records. They were used for information extraction, representation learning, outcome prediction, phenotyping, and de-identification. Unlike the approach used in supervised machine learning, in which hand-crafted representations created by experts using available data elements are tested by trial-and-error, unsupervised learning allows the data to direct the representation.

The article identified several limitations such as model interpretability, data heterogeneity, and lack of universal benchmarks. It describes the need to harmonize across multiple standards and a large number of schemata given (a) each local organization's variations in the use of terminologies such as ICD, CPT, LOINC, and RxNorm (Figure 4-7); (b) the partial mappings maintained by UMLS and SNOMED CT; and (c) the multiple EHR data types

that include numerical quantities, date-time objects, categorical values, natural language free-text, and derived time series.

Shickel et al. describe several common narrative text extraction goals including single concept extraction, temporal event extraction, relation extraction, and abbreviation extraction. They also describe methods for assessing the performance of machine learning techniques for different applications of machine learning to EHR data. The benchmarks could help show that automated de-identification of EHR data satisfies HIPAA requirements.

The authors also talk about the importance of trust and transparency in the health care context when algorithms that improve patient safety, care quality, information privacy, and offer other valued services are used. Establishing standards of interpretability, especially for deep-learning algorithms, is needed to make these methods applicable in routine care.

Shirts et al. (101) studied how the source, route of entry, and display of genetic information in the EHR impacts clinician use, sometimes making it difficult to find this information and take appropriate action. They found several different ways to enter and display genetic information, ranging from lab feeds, to fields in the EHR that were labeled or unlabeled, to problem lists, to clinician notes. More consistent ways of storing and finding information in the EHR, as well as standardizing the content, reduce the burden on clinicians of not being able to find information or take action based upon it, and benefit the patient. Interoperable systems help not only with data sharing but also may reduce user exposure to systems with which they are wholly unfamiliar.

Aronson et al. (109) highlighted the importance of providing CDS for genetic testing within the physician workflow, which ideally would be provided through the EHR. They believe this requires EHRs to be extended with specialized ancillary systems that help clinicians receive and interpret genetic test results and stay up-to-date on changing genetic knowledge. EHRs can potentially use SMART on Fast Healthcare Interoperability Resources (FHIR) for results display, if source data systems are also SMART-enabled.

Schema	Number of Codes	Examples
ICD-10 *(Diagnosis)*	68,000	- J9600: Acute respiratory failure - I509: Heart failure - I5020: Systolic heart failure
CPT *(Procedures)*	9,641	- 72146: MRI Thoracic Spine - 67810: Eyelid skin biopsy - 19301: Partial mastectomy
LOINC *(Laboratory)*	80,868	- 4024-6: Salicylate, Serum - 56478-1: Ethanol, Blood - 3414-0: Buprenorphine Screen
RxNorm *(Medications)*	116,075	- 161: Acetaminophen - 7052: Morphine - 1819: Buprenorphine

Figure 0-7. Example Classification Schema for Diagnoses, Procedures, Laboratory Tests, and Medications (108)

Research IT Driven by Secondary Use of Clinical Data

Health IT Infrastructure Vision, Requirements, Challenges

- Organizations that have health IT and mature research IT can leverage and optimize their health IT infrastructure to better support research.

Danciu et al. at Vanderbilt University School of Medicine describe in detail (110) a rich research data warehouse that stores copies of enterprise EHR, administrative system, and ancillary data in a relational database with a de-identified synthetic derivative (SD) and an identified research derivative (RD) (Figure 4-8) that supports a wide range of inquiries, data extraction capabilities, and methods development opportunities.

The architectural and operational requirements were developed with knowledge of other robust approaches at leading academic institutions such as Intermountain Healthcare, Columbia University Medical Center, Massachusetts General Hospital, Brigham and Women's Hospital, Stanford Medical Center, and others. A growing need among researchers for clinical data reuse, participant recruitment, data and knowledge management, adherence to IRB and financial requirements, and direct engagement of patients/consumers allows larger organizations to fund, develop, and maintain a dedicated research infrastructure.

Danciu et al. report a number of benefits from this approach. Researchers benefit from the data, tools, and expertise available to them, regardless of their training or seniority. Patients benefit from being able to participate in research through cohort identification, and potentially to expand their role in research. Clinicians may benefit from optimization of workflow and research findings that directly impact their daily work. Leadership benefits from increased grants and contracts and recruiting new faculty. Intersecting quality improvement and research needs for data tools and knowledge enhance operational work, clinical care, and the research being carried out.

They make an interesting observation in their environment. "While researchers benefit from operational data collected routinely in EHRs and adjacent systems, they feel it is a research function, not the responsibility of clinicians and staff, to process and maintain data in a meaningful way for research." They highlight the value of the combination of deep clinical knowledge and research domain knowledge when conducting research.

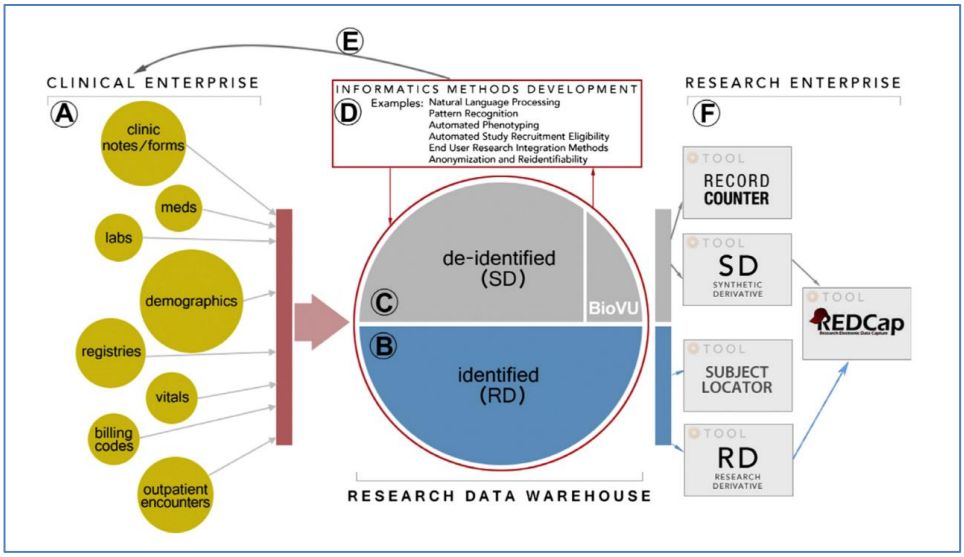


Figure 0-8. Vanderbilt Clinical and Research Informatics Environment (110)

Ethical Framework for Consent and Data Sharing

Health IT Infrastructure Vision, Requirements, Challenges

- Dynamic consent and portable legal consent are two emerging alternatives to one-time consent.
- Establishing and maintaining the public trust is a high priority when asking volunteers to donate their data, especially over time.

Changes in the use of health IT, and consumer technologies more generally, have generated new avenues for patient engagement including the direct sharing of health data by patients in a variety of circumstances. A recent paper (111) describes the need for a clearer ethical framework that addresses several new modes of data sharing by patients: crowdsourcing, social networking platforms, and dynamic consent.

In traditional research, patients/subjects play a specific role, donate data for use by researchers under a one-time consent, and are typically protected by IRB oversight. Newer modes of engagement differ in several ways, from citizen scientists who may have less research leadership experience, to ongoing relationships between patients and those receiving their data, to the direct engagement of the research participant as patients use web-based platforms to search for other patients, find clinical trials, network with scientists, and sometimes exchange personal data directly.

These shifts have raised questions about what expectations an individual should have about sharing data, receiving benefits for doing so, and protecting it.

For example, patients who wish to share continuously generated and analyzed personal device data with a health care system, in exchange for real-time monitoring and alerting, may not clearly understand the commitments and freedoms of the parties in this arrangement and who exactly bears what risk if a harm occurs. A concept of dynamic consent is emerging in which patients can set detailed preferences about how they are contacted, what data classes they are willing to share and under what conditions, and what rights they retain when revoking or changing these permissions. Another concept, portable legal consent, is being developed as a common approach to documenting, sharing, and activating access rules to support secondary uses and protections of data that was initially shared with a one-time consent (112). Researchers and data donors recognize that establishing and maintaining the public trust is a high priority so that volunteers who are asked to supply data and help drive the research agenda feel comfortable doing so.

Potential Observation Bias Reflected in EHR Data

Health IT Infrastructure Vision, Requirements, Challenges

- Observational data can be biased, and much of the data in EHRs is observational.
- Research expertise is needed to avoid overinterpreting observational data that reflects the health care process in addition to the health status of the patient.

Agniel et al. (113) discuss the opportunities and challenges associated with using EHR data in large observational studies—exactly the kinds of studies that are becoming more common with GWAS and PheWAS research to find associations between genomic and phenomic data. The authors point out that the majority of EHR data is essentially “observational,” meaning that it reflects not only the health of the patient, but the fact that someone

decided to put that information into the EHR, reflecting processes that are distinct from the health state of the patient.

They point out that this can have consequences, such as finding associations that might be misinterpreted unless the researcher understands this source of bias. For example, the ability to afford treatment, available ordering panels, time of day, and other factors impact the likelihood that a health finding will be documented in the EHR. To illustrate their point, the authors calculated the risk of death associated with a laboratory test being ordered—any laboratory test that has any result—in a cohort of patients who visited two hospitals during a 12-month period during 2005 to 2006. They found that the odds ratio of being alive 3 years later was highly correlated with having a lab test ordered, and even more strongly correlated when the time of day for the order was factored into the model. They found for example that if any laboratory test was ordered, it explained 68% of the variation in a multiple logistic regression model.

They conclude that researchers should be aware of observational bias, and recommend that researchers should work to understand "context" if possible when conducting research that involves observational process data from the EHR.

Adoption of NIH Computing Tools

Health IT Infrastructure Vision, Requirements, Challenges

- Sustained value of research IT to users reinforces their use of IT. This, along with flexibility when suggesting the use of standards, and defining short-term project accomplishments, were several critical factors in IT adoption that may generalize to many forms of health IT.

With a focus on ways to improve the health IT infrastructure to support research, it is important to consider environmental factors that often play an important role. Masys et al. (114) compared three large NIH-funded technology advancement projects, caBIG, REDCap, and i2b2, and believe that the challenges and nonsustainability of caBIG were likely due to several factors. One was the "enterprise" approach of caBIG, which provided pre-established data models, vocabularies, software systems, standards, and funding to participate for NCI Cancer Centers, but saw very limited adoption. In contrast, the REDCap and i2b2 initiatives provided software tools for free and produced highly configurable systems and tools that their customers could adapt.

The authors concluded that standards were important and especially useful when viewed as a value-added activity, but were not more important than serving high-priority user needs. They found that keeping scope smaller and focused on first release (in months, not years) was important. They also recognized that regardless of an organization's initial decision to use costly or free software, sustained funding was needed for ongoing support costs. The implication for the health IT infrastructure is to focus adoption strategies on providing sustained value to users and purchasers.

Precision Medicine in Cancer Care

Health IT Infrastructure Vision, Requirements, Challenges

- Precision medicine in cancer care requires new standards for the collection, analysis, and sharing of samples and data from cancer patients.
- Infrastructure to produce, store, link, and share genomics data has a broad impact on prevention, tumor classification, and treatment.
- Genomic information collection will dramatically expand from only a minority of patients to almost all patients.
- International cooperation around the production, analysis, interpretation, and sharing of cancer patient samples and mutation databases is needed.
- EHRs serve as a broad platform that merge a variety of patient information and expert advice to facilitate coordinated cancer care.

Multidisciplinary work to identify challenges and opportunities within the field of precision medicine in cancer care found multiple dependencies on health IT. There is a strong need for investments in data infrastructure (as well as public education around precision medicine) to more effectively and efficiently "produce, store, link, and share [precision medicine] data" (115). Sequencing technology equipment, secure high-throughput computing infrastructures, and reliable and standardized EHR systems integrating genomic and phenotypic patient information are also identified as critical factors. The opportunity to merge cancer registries and EHR data to find the best treatment options for a cancer patient have broad implications for health IT as a platform for advances in care (116).

Changes in practice are also needed to fully integrate precision medicine in cancer care (115), including the use of liquid biopsies to speed cancer detection, the reduction of turnaround time from tumor sample collection to actionable results, and better understanding mechanisms of treatment resistance. Standards are needed to improve identification and matching of published information with a specific genetic variant. With high variation in the way results are published, tissue samples are collected, samples are prepared and analyzed, molecular analysis is performed, and population genomic data is shared, automated approaches to linking patient findings to research findings often lack precision. Variable genetic testing practices and mismatched intervention endpoints can also make diagnosis and treatment decisions more difficult and uncertain.

An article by Cohen explores from multiple perspectives—those of regulators, payers, and drug developers—the regulatory challenges of bringing to market any personalized therapeutics and concomitant diagnostics (117). He notes that approvals for a personalized therapy may not co-occur with approvals for the biomarker used to test for its appropriate use and changes in the coverage policies by insurers, causing delays in its use even if it is effective.

There are also are key privacy considerations in how data from health IT can be used or shared within a large-scale precision medicine infrastructure, although privacy wishes do vary. In one example, the Personal Genome Project (PGP) at Harvard University, over 5,000 participants who contributed biological samples and made their personal genetic information publicly available, completed consents and affirmed that they knew that they can be identified from their data. Technology is likely to play a big role in evolving standards and expectations of personal privacy, through enhanced methods of informed consent, and technology that may provide more effective privacy (118, 119).

Another focus of precision medicine cancer research pertains to helping patients navigate and interpret the complex and confusing information that health IT and precision medicine generate. Giuse et al. report patients' improved comprehension of melanoma symptoms and treatments when given access to a multimodal consumer-friendly decision support tool, as compared to patients with standard information tools and patient controls (120). Precision medicine in cancer care illustrates that health IT plays an important role in supporting not only basic capabilities such as storing and retrieving patient clinical data, but also in helping patients and providers navigate the nuanced interpretation of complex data.

HEALTH IT GAPS IDENTIFIED

Earlier chapters describe initiatives, peer-reviewed literature, and grey literature that identify how the health IT infrastructure has been leveraged to support research, and challenges that need to be addressed.

This chapter describes six gap areas (Table 5-1) that are important for the advancement of research, are not addressed sufficiently through ongoing health IT initiatives, and intersect with ONC's authority. Each gap area includes specific gaps related to one or more areas including governance, policy, services, data, and standards.

Limited Health IT Prioritization of Research

The high-level goals for the health system as a whole, improving quality, cost-effectiveness, efficiency, and the provider experience, are powerful drivers for change to health IT. Research is typically a secondary goal or is not explicitly listed at all. Efforts to increase the stature of research as a priority during changes to the health IT infrastructure would help to close this gap, and would focus greater efforts on understanding the potential impact of changes in health IT for the conduct of research.

Figure 5-1, label A, shows the change process from the current state (Health IT^0) to a future state (Health IT^1). For example, improvements in interoperable data sharing between two or more systems could also lead to improved data sharing for research purposes if designed with that purpose, whereas a reduction in provider documentation to streamline care could impact the capture of important information for research.

In a more global context, there's a tradeoff between incremental health IT adaptations that address near-term needs, and substantial health IT re-design to align with a more integrated future vision of health IT that prioritizes advances in health care *and* research. As scientific knowledge and practice evolve, they drive the need for incremental health IT changes such as updated disease taxonomies, changing drug catalogues, new laboratory tests, new uses of existing tests and therapies, decision support changes, updated directories for providers/facilities and other key resources, new sources and types of data relevant to care (e.g., genomics and mobile devices), increasing data granularity and metadata, and changes in workflow.

Disruptive changes in scientific knowledge and practice, such as the shift to molecular medicine, value-based care, powerful machine learning, massive amounts of patient data, and growing cognitive support needs among professionals and patients, are driving the need for more substantial architectural and design changes in health IT. For example, a lung cancer diagnosis based on molecular medicine may be different from the code used for billing. Health IT developers are increasingly open to third-party software tools that help them adapt quickly to new research evidence. Genomic interpretation is a case in point, since the "meaning" of a result can change over time as new research is conducted.

Limited Health IT Production of Research Data

The second gap area (label B) is an expectation that health IT will routinely produce data for research purposes as a byproduct of normal technology use during care delivery, payment, operations, and other functions. Health IT would ideally reduce the researcher's challenge to find specific data, access it, index the available data,

validate the accuracy of data and its context, find essential and useful metadata, and consistently know the taxonomy and/or ontology applied during data capture and storage.

Data production ranges from basic, such as when data for a single individual are extracted and shared during a care transition (e.g., a hospital discharge, or a change in primary care providers), to complex, such as when a large database is queried to monitor for adverse effects of a treatment or a combination of risk factors in a population. Organizational experience in producing data varies widely, and many entities do not have much expertise producing research data. The specific challenges and gaps associated with producing data for research are varied, including policy-related (e.g., permissions management), technical (e.g., deployment of trusted services), data-related (e.g., having insufficient metadata), and others. This gap area is anticipated to expand dramatically in the coming years as a growing number of researchers seek genomic data, medical and consumer device data, and behavioral and environmental data and use powerful machine learning methods

Researchers need to capture, save, and retrieve metadata to interpret health data, whether collected during a patient visit, provided remotely by the patient, obtained from tissue or imaging studies, or generated during care delivery. Metadata can be diverse. Examples include the taxonomies and ontologies in use during data collection, the purpose for capturing the data, the capture methods, provenance, and permissions. It may also be necessary to request additional information *after* data has been analyzed. Identifiers for data sources, the individual they describe, and any other attribute used for matching purposes (e.g., a provider, a setting) are especially important, since unclear identifiers create difficulties when aggregating and analyzing data.

Many health IT systems holding patient data are not preconfigured to handle a research request for data, especially an electronic request, although the TEFCA aims to establish policies that will streamline this process. Even so, there is a need to develop and agree on standardized services, standardized data, and clearly referenced taxonomies and ontologies to enable robust use of research data.

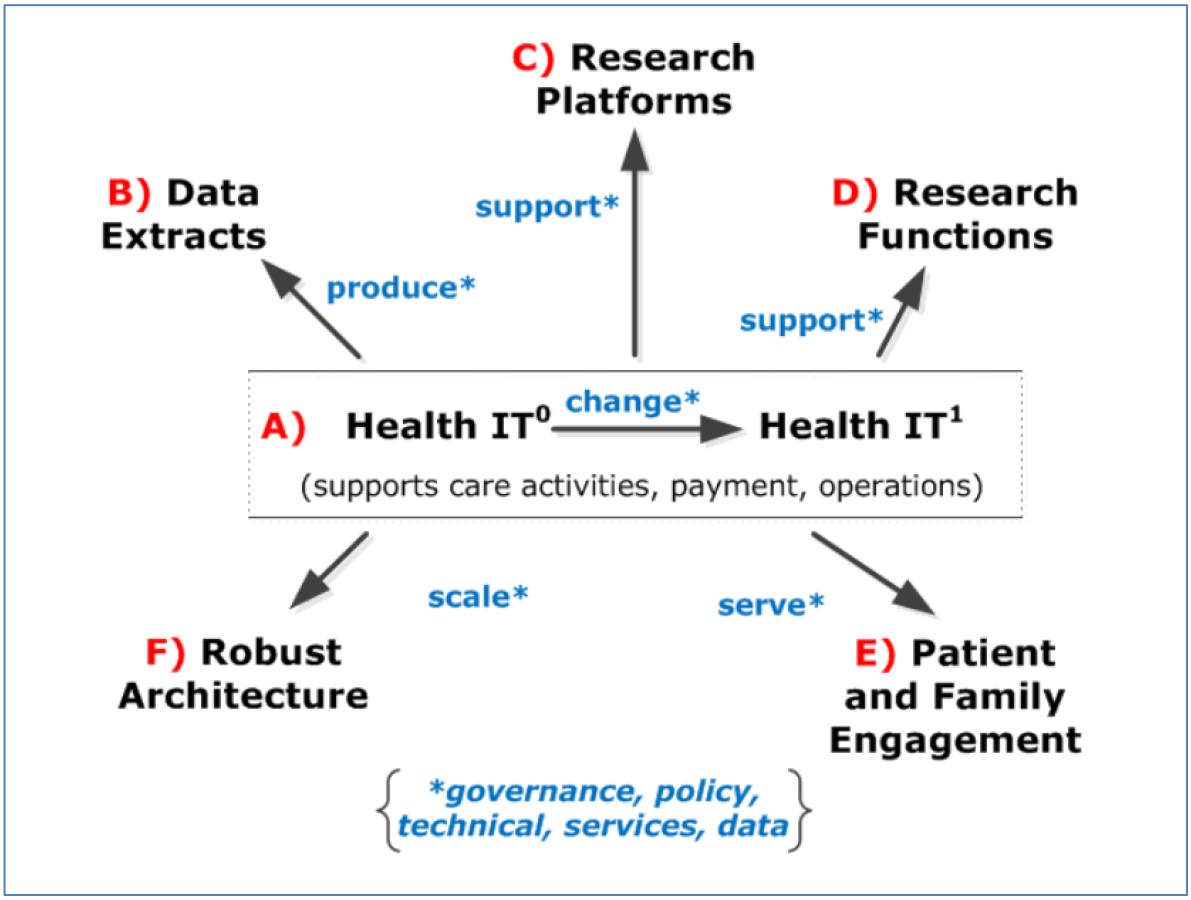


Figure 0-1. Gap Areas

Limited Health IT Support for Research Platforms

Background materials also identified gaps in health IT support for more complex research platforms that might include several health IT instances, multiple organizations, and/or multiple data sources (label C). Although much coordination takes place after data leaves the health IT environment, research challenges such as utilizing multiple data streams that need to be aggregated and related to one another may be mitigated by decisions made before the data leaves its source system, such as common identifiers and useful metadata. It can also be difficult to re-use extraction software on different data stores. Typically, each individual EHR, HIE, Registry, ancillary system (e.g., laboratory, radiology, or pharmacy), claims database, or data warehouse uses a special purpose software tool to extract health data for research. The processes for requesting permission and demonstrating approval also vary widely and must be repeated for each system or organization.

In many cases, several different health IT systems are used together to provide support for research. Varying systems provide metadata, assist with case finding, schedule research activities, and manage research protections. Coordination across different organizations requires strong governance, consistent policies, agreements on methods to link data, and software services that function across a portfolio of systems and organizations participating in the research.

Limited Health IT Support for Research Functions

Specialized functions for researchers, such as locating specific data, searching multiple data sources, indexing data of interest, querying for matching records, and identifying consenting and randomization status are

relevant not only to researchers, but to other stakeholders including providers and patients. In addition, there is a growing need for health IT to support functions that easily incorporate research evidence into practice through decision support triggers and rules, and application programming interfaces with third-party functions.

Some well-established research organizations provide specific information system tools that help researchers search for suitable data among multiple data sources, as well as provide searching, indexing, querying, consenting status, and other functionalities needed to conduct and support research. These functions are relevant not only to researchers, but to other stakeholders including providers and patients exploring data and findings. As the health IT infrastructure support for research becomes a higher priority, more systems and organizations will want these capabilities.

There is also a strong need for health IT to incorporate research evidence into practice. Examples include decision support triggers and rules, expanded indications for test results including genomics, routine use of patient-reported outcome measures, APIs that support clinical trial recruitment, data-driven health screening recommendations, or streaming data from a newly approved medical monitoring device.

Health IT data stewards implementing privacy and security procedures need robust de-identification and re-identification tools to manage risk, and to reassure patients they can be confident in the protections, and in sharing their data for research purposes. Of special relevance in the *All of Us* Research Program is the opportunity to "unlock" key research information for a participant and use health IT to share it with the participant and their providers, if appropriate.

Limited Health IT Support for Patient and Family Engagement in Research

Increasingly, a patient or family decision to engage in some way in the research process will lead them to engage with health IT (label E). Patient and family engagement are needed to accelerate and scale the use of routine health data for research. Patients and families can interact with health IT in a variety of ways to support research.

Table 5-1. Health IT Gap Areas and Specific Gaps to Address

Health IT Infrastructure Gap Areas and Specific Gaps*
A. Limited Health IT Prioritization of Research **Failure to keep health IT current may negatively impact research data or functions, such as:** • Changes to reflect new scientific knowledge • Updated taxonomies and ontologies • Updated dictionaries such as a drug catalogue, lab test catalogue, provider directory, or disease dictionary • Changes in data granularity and metadata captured and reported • Failing to perform robust testing of new system features before full rollout • Lack of support for the use of 3rd party APIs that can strengthen research

Health IT Infrastructure Gap Areas and Specific Gaps*
B. Limited Health IT Production of Research Data **Limited production of research data can occur for a variety of reasons, including:** • Inconsistent access to data and metadata through software tools (e.g., APIs) that perform key functions (e.g., retrieve data, access search and indexing functions, access semantic information, etc.) • Inconsistent or difficult to use software services to extract health IT data • A change record of the taxonomies and ontologies in use at a given point in time • A transparent software architecture that uses (when possible) standardized services, standardized data, and up-to-date taxonomies and ontologies • Useful metadata that conveys provenance, identifiers, context, and permissions • Strong adherence to privacy and security requirements and thorough de-identification when needed • Patient engagement is critical to the routine sharing of data for research
C. Limited Health IT Support for Research Platforms **Conducting research across multiple organizations and data sources is more complex, placing greater demands on health IT:** • Consistent use of ontologies and taxonomies among data sources • Effective ways to unambiguously identify a data source, type of data, medical device, temporal data, person, and other information that supports data aggregation • Permissions management is more complex for multiple organizations and data streams • Health IT decisions may be delayed without strong governance in a multi-organization initiative
D. Limited Health IT Support for Research Functions **When conducting research, there are strong needs for:** • More effective data extraction and curation tools (for unstructured text, for example) that produce granular data • Robust documentation tools that promote extensive data capture, high data granularity, and data curation with reduced effort and cognitive support • Health IT data directories supported by powerful searching and indexing tools • Permission management tools that ease the appropriate sharing of identified and de-identified data for research • Tools within health IT systems to preview data-sharing options and confirm data sharing that has occurred **When disseminating research, there are strong needs for:** • Incorporating new genomic findings • Modifying CDS triggers and rules • Collecting and using patient-reported outcome measures • Software adaptations for authorized use of 3rd party application programming interface • Clinical trial recruitment information • Newly approved biomarkers to be stored using health IT • Data from a newly approved medical device

Health IT Infrastructure Gap Areas and Specific Gaps*
E. Limited Health IT Support for Patient and Family Engagement in Research • Patients that don't directly use health IT may limit the data available for research use • Erroneous medical record information may persist without patient review • Patients may miss opportunities to use health IT to explore relevant medical knowledge • Health IT that does not make use of research findings may result in reduced patient trust in the system • Patients who do not use health IT may have more limited exposure to the benefits of research, such as enrollment in a study
F. Lack of a Robust Health IT Architecture to Support Research • The architecture for a specific health IT system is not easily comparable to a model architecture • Standards used in health IT for both data and services, and exceptions to them, should be transparent and retrievable over time • A known health IT architecture can assist a researcher in understanding data quality • A known health IT architecture can assist a researcher in monitoring its level of data duplication and fragmentation

*Specific gaps may be relevant to more than one gap area. Broad gap areas include gaps related to governance, policy, services, data, and/or standards.

First, a person's use health IT directly for data capture in the form of patient-reported outcomes and consumer or medical device generated data, and provide much EHR observational data through their interactions with clinicians. With increasing frequency, individuals finding errors in their records expect that health IT will enable them to be corrected. Second, people who give permission to use identified data for research purposes may need to do so using health IT systems. Third, many individuals may become very interested in research findings for a specific condition, treatment, test, or lifestyle choice, for themselves or a loved one. Health IT can link patient information to relevant resources to address those needs.

Fourth, patients and families want state-of-the-art care and expect that health IT used in their care is informed by and leverages research evidence for relevant decisions. Fifth, patients and families may wish to share their experience with an innovative treatment, self-monitoring, or personal lifestyle choice to assist others, leveraging health IT to do so. Sixth, patients and families may want to use health IT to participate directly in research, either by making themselves available to researchers looking for participants or data, or by using their own resources coupled with a strong desire to explore an unknown area for themselves. Patients and families are increasingly sharing their health data directly with researchers and taking leadership roles in directing research of relevance to them.

Lack of a Robust Health IT Architecture to Support Research

A robust health IT architecture (label F) is needed to support strong harmonization among system components and data to support health care delivery activities and to accelerate research. A robust health IT architecture is needed to support a new and evolving taxonomy for disease. Scientific advances result from and require advances in information management and computational technology. New knowledge generated through research drives changes in current-day practice. Multidisciplinary informatics and other thought leaders envision health IT capable of routinely capturing and contributing data that is aggregated and analyzed for research. This vision of research informing practice, and practice generating data for research, means that developers must

deploy tools on a large scale that request and manage data, and must utilize a systems architecture that simplifies the process of obtaining data for research, and for incorporating research findings into practice.

Researchers and designers need to understand how to transition toward a new architecture, and its guiding principles. For example, researchers who can see and compare the health IT and data architecture of an existing system to a model architecture, can better understand how to locate data, its context, when it might be duplicated or fragmented, and the standards used to collect and transmit the data in the present and historically.

Table 5-2. Architectural Principles to Advance the Use of Health IT for Research

Principle*	Description and Gaps

1. Principles apply across a wide range of technologies and platforms (G/P; T/S; D/S)	• Principles are technology agnostic with regard to
	  ○ scale;
	  ○ actual locations of the stored data and software systems (desktop, cloud, widely federated system of systems across different organizations); or
	  ○ a specific technology or platform.

2. Open standards and protocols (G/P; T/S; D/S)	• Standards usually are established through a formal process and are endorsed by a standards organization. • Based on published application program interfaces (APIs) and protocols. Protocols dictate the form, content, timing, and order of messages that can be exchanged among cooperating entities, with associated APIs that implement the protocol exchange.
3. Encrypted data and separate key management from data management (G/P; T/S)	• Data must be encrypted at rest and in transit. • Access must be available to control systems that provide identity management, user authentication, and user authorization. • A pre-emptive object-based access control model is necessary for exchange of health information at the national level rather than an audit-based model.
4. Data accompanied by relevant metadata and provenance information (T/S; D/S)	• Provenance—the chain of custody of data from its inception and through its entire history of access, transmission, or modification—is important for understanding research data in context.
5. EHRs and other health IT produce atomic data items and associated metadata (T/S; D/S)	• This promotes maximum flexibility in data handling and security.

Principle*	Description and Gaps
6. The robustness principle is followed (G/P; T/S; D/S)	• Being conservative in sending behavior, and liberal in receiving behavior, fosters interoperability and machine understanding but allows for some flexibility.
7. A migration pathway is provided for legacy health IT systems (G/P; T/S; D/S)	• A new and unifying software architecture leverages data and resources from legacy systems while introducing new components.
8. Privacy and security requirements are honored (G/P; T/S; D/S)	• Sending systems can trust that receiving systems will apply privacy and security restrictions and safeguard data physically and electronically (encryption), without unauthorized copying, modification, or transmission. • These are described in the JASON software architecture model as IAPS: identification, authorization, and privacy services.

Key: G/P = governance/policy; T/S = technology/services; D/S = data/standards

*Adapted from: 2014 JASON report: "A Robust Health Data Infrastructure" (84)

The JASON architecture (84) offers one conceptual framework for describing architectural components relevant to data exchange for both care and research, along with architectural principles (Table 5-2) that future work should address. Using a new systems and data architecture, informatics leaders have an opportunity to decompose complex problems into smaller, more manageable sub-problems, and to describe the intended relationships between key system functions and components.

Finally, the work to identify and address gaps involves tradeoffs. Informatics experts face design choices, balancing the near-term functional and regulatory needs addressed through smaller incremental changes, with more substantial technology changes that better align with the future vision of routinely using clinical care and operational data to accelerate research, and to incorporate research findings routinely into non-research activities. The more substantial changes take longer, require more investment, and could be disruptive, but offer solutions that will advance care, research, and the learning cycles that produce sustained improvement.

Federal and state governments play significant roles in delivering care, sponsoring research, paying for care, and educating stakeholders including patients. The ONC is an important advocate for foundational changes to the health IT infrastructure to support research, and has a significant opportunity to lead and to help coordinate activities across organizations and initiatives.

REFERENCES

1. HealthIT.gov. About ONC. Washington, DC: HealthIT.gov; 2017. Available from: https://www.healthit.gov/newsroom/about-onc.

2. National Committee on Vital and Health Statistics. Personal Health Records and Personal Health Record Systems. Washington, DC: U.S. Department of Health and Human Services; 2006. Available from: https://stacks.cdc.gov/view/cdc/6758.

3. Index for Excerpts from the American Recovery and Reinvestment Act of 2009 (ARRA). 2009. Available from: https://www.healthit.gov/sites/default/files/hitech_act_excerpt_from_arra_with_index.pdf.

4. Office of the National Coordinator for Health Information Technology. 2016 Report to Congress on Health IT Progress: Examining the HITECH Era and the Future of Health IT. 2016. Available from: https://dashboard.healthit.gov/report-to-congress/2016-report-congress-examining-hitech-era-future-health-information-technology.php.

5. U.S. Senate and House of Representatives. 21st Century Cures Act. United States; 2016 December 13. Contract No.: 114-255. Available from: https://www.congress.gov/114/plaws/publ255/PLAW-114publ255.pdf.

6. Darbouze. Structured Data Capture (SDC) & Data Access Framework (DAF). Washington, DC: HealthIT.gov; 2017. Available from: https://www.healthit.gov/sites/default/files/3.CreatingaDataInfrastructureforPatient-COR_FDarbouze.pdf.

7. Office of the National Coordinator for Health Information Technology. Data Access Framework (DAF). 2018. Describes the Data Access Framework (DAF) for disrete data elements and documents from an electronic health record (EHR)]. Available from: https://www.healthit.gov/topic/research-evaluation/data-access-framework-daf.

8. HealthIT.gov. Building Data Infrastructure to Support Patient Centered Outcomes Research (PCOR). Structured Data Capture (SDC). Washington, DC: HealthIT.gov; 2016. Available from: https://www.healthit.gov/topic/research-evaluation/structured-data-capture-sdc.

9. HealthIT.gov. Building Data Infrastructure to Support Patient Centered Outcomes Research (PCOR). Patient-Generated Health Data (PGHD). Washington, DC: HealthIT.gov; 2017. Available from: https://www.healthit.gov/topic/research-evaluation/patient-generated-health-data-pghd.

10. HealthIT.gov. An Overview of the ONC/PCOR Project: Patient Matching, Aggregating, and Linking (PMAL). Washington, DC: HealthIT.gov; 2016. Available from: https://www.healthit.gov/sites/default/files/3.CreatingaDataInfrastructureforPatient-CORDBucci.pdf.

11. Office of the National Coordinator for Health Information Technology. Patient Matching, Aggregating and Linking (PMAL). Washington, D.C.2018. Patient Matching, Aggregating and Linking (PMAL) project - brief description]. Available from: https://www.healthit.gov/topic/research-evaluation/patient-matching-aggregating-and-linking-pmal.

12. HealthIT.gov. Building Data Infrastructure to Support Patient Centered Outcomes Research (PCOR). Privacy and Security Framework for PCOR (PSP). Washington, DC: HealthIT.gov; 2016. Available from: https://www.healthit.gov/topic/research-evaluation/privacy-and-security-framework-pcor-psp.

13. Office of the National Coordinator for Health Information Technology. Patient Choice Technical Project. 2018. Available from: https://oncprojectracking.healthit.gov/wiki/display/PATCH/The+Patient+Choice+Technical+Project+Homepage.

14. MITRE. Electronic Consent Management: Landscape Assessment, Challenges, and Technology. Washington, D.C.; 2014. Available from: https://www.healthit.gov/sites/default/files/privacy-security/ecm_finalreport_forrelease62415.pdf.

15. White PJ, Halamka JD. Letter to Karen DeSalvo, MD. Washington, DC: Health IT Standards Committee; 2015 September 25. Available from: https://www.personalizedmedicinebulletin.com/wp-content/uploads/sites/6/2016/01/Health-IT-Standards-Committee-PMI-Task-force.pdf.

16. HealthITBuzz. NIH and ONC Launch the Sync for Science (S4S) Pilot: Enabling Individual Health Data Access and Donation. Washington, DC: Office of the National Coordinator for Health Information Technology; 2016. Available from: https://www.healthit.gov/buzz-blog/health-innovation/nih-and-onc-launch-the-sync-for-science-pilot/.

17. Alterovitz G, Brown J, Chan M, Daniels J, Haberman E, Johanson E, et al. Sync for Genes. Enabling clinical genomics for precision medicine via FHIR. Washington, DC: U.S. Department of Health and Human Services (HHS) and The Office of the National Coordinator for Health Information Technology (ONC); 2017 June 4. Available from: https://www.healthit.gov/sites/default/files/sync_for_genes_report_november_2017.pdf.

18. Office of the National Coordinator for Health Information Technology. Key privacy and security considerations for healthcare application programming interfaces (APIs). Washington, D.C.; 2017. Available from: https://www.healthit.gov/sites/default/files/privacy-security-api.pdf.

19. NORC. Assessing the SHARP Experience: Final Report. Washington, D.C.; 2014. Available from: https://www.healthit.gov/sites/default/files/sharp_final_report.pdf.

20. NORC. Assessing the SHARP Experience: Final Report Appendices. Washington, D.C.; 2014. Available from: https://www.healthit.gov/sites/default/files/sharp_final_report_appendices.pdf.

21. Smart HealthIT Gallergy. 2018. Available from: https://apps.smarthealthit.org/.

22. Denny J. Overview of EHR phenotyping -- successes, challenges, and examples. Nashville, TN: Vanderbilt University; 2015. Available from: http://informatics.mayo.edu/phema/files/PhEMA_PhenotypingOverview_AMIAjoint_2015.pdf.

23. Doostparast Torshizi A, Petzold LR. Graph-based semi-supervised learning with genomic data integration using condition-responsive genes applied to phenotype classification. J Am Med Inform Assoc. 2018;25(1):99-108.

24. Luo Y, Cheng Y, Uzuner O, Szolovits P, Starren J. Segment convolutional neural networks (Seg-CNNs) for classifying relations in clinical notes. J Am Med Inform Assoc. 2018;25(1):93-8.

25. Martin D. Overview of FDA's active surveillance programs and epidemiologic studies for vaccines. Silver Spring, MD: U.S. Food and Drug Administration; 2012. Available from: https://www.fda.gov/downloads/biologicsbloodvaccines/internationalactivities/ucm464921.pdf.

26. National Institutes of Health. Program FAQ. Bethesda, MD: National Institutes of Health, All of Us Research Program; 2017. Available from: https://allofus.nih.gov/about/program-faq.

27. Department of Energy. 2018. Available from: https://www.va.gov/opa/pressrel/pressrelease.cfm?id=2896.

28. Collins FS. Fiscal Year 2018. Justification of Estimates for Appropriations Committees. National Institutes of Health - Volume 1 Overview. Washington, DC: Department of Health and Human Services; 2017. Available from: https://www.nidcd.nih.gov/about/congressional-justification-2018.

29. National Institutes of Health. All of Us Research Program-Protocol V1. Bethesda, MD: National Institutes of Health; n.d. Available from: https://allofus.nih.gov/sites/default/files/allofus-initialprotocol-v1_0.pdf.

30. Dishman E. 21st Century Cures Act 10-Year Draft Plan for Precision Medicine Funding. Washington, DC: National Institutes of Health, *All of Us* Research Program; 2017. Available from: https://acd.od.nih.gov/documents/presentations/03282017Dishman-PMI.pdf.

31. National Institutes of Health. Precision Medicine Initiative: Data Security Policy Principles and Framework. Washington, D.C.: National Institutes of Health; 2016. Available from: https://allofus.nih.gov/sites/default/files/security-principles-framework.pdf.

32. Precision Medicine Initiative Working Group. The Precision Medicine Initiative Cohort Program – Building a Research Foundation for 21st Century Medicine. Bethesda, MD: National Institutes of Health; 2015. Available from: https://acd.od.nih.gov/documents/reports/DRAFT-PMI-WG-Report-9-11-2015-508.pdf.

33. Hudson K, Lifton R, Patrick-Lake B, Denny J. The precision medicine initiative cohort program—Building a Research Foundation for 21st Century Medicine. Precision Medicine Initiative (PMI) Working Group Report to the Advisory Committee to the Director, ed. 2015.

34. Bhat N. Million Veteran Program. UNTHSC Scholarly Repository; n.d. Available from: https://digitalcommons.hsc.unt.edu/rad/RAD15/GeneralMedicine/8/.

35. Gaziano JM, Concato J, Brophy M, Fiore L, Pyarajan S, Breeling J, et al. Million Veteran Program: A mega-biobank to study genetic influences on health and disease. J Clin Epidemiol. 2016;70:214-23.

36. Office of the National Coordinator for Health Information Technology. Precision Medicine Task Force - Meeting Notes 2016-03-16. Washington, DC: Office of the National Coordinator for Health Information Technology; 2016. Available from: www.healthit.gov/hitac/sites/faca/files/PMTF_Meeting_Slides_2016-03-16.pptx.

37. Kibbe W. NCI-DOE Pilot and Precision Medicine Initiative in Oncology. Bethesda, MD: National Cancer Institute; 2015 September 30. Available from: https://pdfs.semanticscholar.org/presentation/6d1e/7961898ff48559b7591ab5c320faf0b8d719.pdf.

38. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal. 2013;6(269):pl1.

39. Nishimura AA, Tarczy-Hornoch P, Shirts BH. Pragmatic and ethical challenges of incorporating the genome into the electronic health record. Current genetic medicine reports. 2014;2(4):201-11.

40. Sherman RE, Anderson SA, Dal Pan GJ, Gray GW, Gross T, Hunter NL, et al. Real-World Evidence - What Is It and What Can It Tell Us? N Engl J Med. 2016;375(23):2293-7.

41. Eisenberg RS. Shifting institutional roles in biomedical innovation in a learning healthcare system. Journal of Institutional Economics. 2018:1-24.

42. Findlay S. Health Policy Brief: The FDA's Sentinel Initiative. Health Aff (Millwood). 2015.

43. Food and Drug Administration. Digital Health Innovation Action Plan. 2018.

44. Kohane IS. i2b2 - Informatics for Integrating Biology & the Bedside - Overview. Boston, MA2018. Available from: https://www.i2b2.org/about/index.html.

45. Murphy SN, Avillach P, Bellazzi R, Phillips L, Gabetta M, Eran A, et al. Combining clinical and genomics queries using i2b2–Three methods. PLoS One. 2017;12(4):e0172187.

46. Harris PA. Research Electronic Data Capture (REDCap)-planning, collecting and managing data for clinical and translational research. BMC Bioinformatics. 2012;13(12):A15.

47. Research Electronic Data Capture (REDCap) - About Us. 2018. Available from: https://projectredcap.org/about/faq/.

48. Maier C, Lang L, Storf H, Vormstein P, Bieber R, Bernarding J, et al. Towards Implementation of OMOP in a German University Hospital Consortium. Appl Clin Inform. 2018;9(1):54-61.

49. Patient-Centered Outcomes Research Institute. PCORnet: The National Patient-Centered Clinical Research Network. Washington, DC: Patient-Centered Outcomes Research Institute; 2013. Available from: https://www.pcori.org/research-results/pcornet-national-patient-centered-clinical-research-network.

50. National Patient-Centered Clinical Research Network. PCORnet Common Data Model (CDM) Specification, Version 4.1. Washington, D.C.: National Patient-Centered Clinical Research Network; 2018. Available from: http://www.pcornet.org/wp-content/uploads/2018/05/PCORnet-Common-Data-Model-v4-1-2018_05_15.pdf.

51. Lipskiy N. BioSense: Implementation of Data Standards. Washington, D.C.: National Committee on Vital and Health Statistics; 2013. Available from: https://www.ncvhs.hhs.gov/wp-content/uploads/2014/05/131112p17.pdf.

52. The PEW Charitable Trusts. Improving Patient Care Through Safe Health IT. Washington, DC The PEW Charitable Trusts; 2017 December. Available from: http://www.pewtrusts.org/-/media/assets/2017/12/hit_improving_patient_care_through_safe_health_it.pdf

53. Moscovitch B. Pew Letter to the ONC Regarding TEFCA - Feb 20 2018. Washington, DC: The PEW Charitable Trusts; 2018. Available from: http://www.pewtrusts.org/~/media/assets/2018/02/hit_pew_urges_better_patient_matching_data_standards_to_boost_health_it_interoperability.pdf.

54. Medical Device Innovation Consortium. 2015. Available from: http://mdic.org/about-us/overview/.

55. NESTcc - National Evaluation System for health Technology. 2018. Available from: http://mdic.org/cc/about-nestcc/.

56. Amazon Web Services. Genomics in the Cloud. Seattle, WA: Amazon Web Services; 2018. Available from: https://aws.amazon.com/health/genomics/.

57. University of Pittsburgh Schools of the Health Sciences. UPMC Announces $2B Investment to Build 3 Digitally Based Specialty Hospitals Backed by World-Leading Innovative, Translational Science. Pittsburgh, PA: University of Pittsburgh Schools of the Health Sciences; 2017. Available from: http://www.upmc.com/media/NewsReleases/2017/Pages/upmc-innovates-announcement.aspx.

58. IBM. Empowering heroes, transforming health. Armonk, NY: IBM Corporation. Available from: https://www.ibm.com/watson/health/.

59. NVIDIA. NVIDIA Teams Up with GE Healthcare, Nuance to Bring Power of AI to Medical Imaging. Santa Clara, CA NVIDIA Corporate Headquarters; 2018. Available from: https://blogs.nvidia.com/blog/2017/11/26/ai-medical-imaging/.

60. Siemens Healthcare Diagnostics, Inc. White Paper: Medical imaging in the age of artificial intelligence. Tarrytown, NY Siemens Healthcare Diagnostics, Inc; 2017. Available from: https://www.siemens.com/press/pool/de/events/2017/healthineers/2017-11-rsna/white-paper-medical-imaging-in-the-age-of-artificial-intelligence.pdf.

61. Beam AL, Kohane IS. Translating Artificial Intelligence Into Clinical Care. JAMA. 2016;316(22):2368-9.

62. Wallace PJ, Shah ND, Dennen T, Bleicher PA, Crown WH. Optum Labs: building a novel node in the learning health care system. Health Aff (Millwood). 2014;33(7):1187-94.

63. Cerbo M. EUnetHTA: a network for added value of European collaboration on HTA. Rome: European Network for Health Technology Assessment; 2013. Available from: http://www.who.int/medical_devices/Sun_am_HTA_1_CERBO.pdf.

64. ELIXIR. About us. Cambridgeshire, UK: ELIXIR. Available from: https://www.elixir-europe.org/about-us.

65. EUnetHTA. About us. The Netherlands: The EUnetHTA Directorate. Available from: http://www.eunethta.eu/about-us.

66. Global Alliance for Genomics & Health. Enabling genomic data sharing for the benefit of human health. Toronto, Ontario: The Global Alliance for Genomics and Health; 2017. Available from: https://www.ga4gh.org/.

67. European Commission. Horizon 2020. City of Brussels, Belgium: European Commission. Available from: https://ec.europa.eu/programmes/horizon2020/

68. Robinson J. Genomic medicine is going mainstream and pharmacists need to be prepared. London: Pharmaceutical Journal Publications; 2017. Available from: https://www.pharmaceutical-journal.com/news-and-analysis/features/genomic-medicine-is-going-mainstream-and-pharmacists-need-to-be-prepared/20203321.fullarticle?firstPass=false.

69. Nofziger C, Papaluca M, Terzic A, Waldman S, Paulmichl M. Policies to aid the adoption of personalized medicine. Nat Rev Drug Discov. 2014;13(3):159-60.

70. Horizon 2020. Health, demographic change and wellbeing. Brussels, Belgium: European Commission; 2017. Available from: http://ec.europa.eu/research/participants/data/ref/h2020/wp/2018-2020/main/h2020-wp1820-health_en.pdf

71. Kohn LT, Corrigan J, Donaldson M, To E. To Err Is Human: Building a Safer Health System. 2000. Available from: http://www.csen.com/err.pdf.

72. Wolfe A. Institute of Medicine Report: crossing the quality chasm: a new health care system for the 21st century. Policy, Politics, & Nursing Practice. 2001;2(3):233-5.

73. Berwick DM. A user's manual for the IOM's 'Quality Chasm'report. Health Aff (Millwood). 2002;21(3):80-90.

74. Committee on Engaging the Computer Science Research Community in Health Care Informatics. Computational Technology for Effective Health Care The National Academies Press 2009.

75. Desmond-Hellmann S, Sawyers CL, Cox DR, Fraser-Liggett C, Galli SJ, Goldstein DB, et al. Toward Precision Medicine. Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease. Washington, DC: The National Academies Press; 2011.

76. Institute of Medicine. Capturing Social and Behavioral Domains and Measures in Electronic Health Records: Phase 2. Washington, DC: National Academies Press; 2014. Report No.: 0309312426. Available from: https://www.nap.edu/resource/18951/EHRreportbrief.pdf

77. Kasthurirathne SN, Vest JR, Menachemi N, Halverson PK, Grannis SJ. Assessing the capacity of social determinants of health data to augment predictive models identifying patients in need of wraparound social services. J Am Med Inform Assoc. 2018;25(1):47-53.

78. National Academies of Sciences E, and Medicine. Examining the Impact of Real-World Evidence on Medical Product Development: I. Incentives: Proceedings of a Workshop in Brief. Washington, D.C.; 2018. Available from: https://www.ncbi.nlm.nih.gov/books/NBK481618/pdf/Bookshelf_NBK481618.pdf.

79. Hernandez AF, Fleurence RL, Rothman RL. The ADAPTABLE Trial and PCORnet: shining light on a new research paradigm. Ann Intern Med. 2015;163(8):635-6.

80. Elkhenini HF, Davis KJ, Stein ND, New JP, Delderfield MR, Gibson M, et al. Using an electronic medical record (EMR) to conduct clinical trials: Salford Lung Study feasibility. BMC Med Inform Decis Mak. 2015;15(1):8.

81. Woodcock A, Bakerly ND, New JP, Gibson JM, Wu W, Vestbo J, et al. The Salford Lung Study protocol: a pragmatic, randomised phase III real-world effectiveness trial in asthma. BMC Pulm Med. 2015;15(1):160.

82. Office of the Assistant Secretary for Planning and Evaluation. Meeting the ACA Mandate to Build Data Capacity. Washington, DC: Office of the Assistant Secretary for Planning and Evaluation; n.d. Available from: https://aspe.hhs.gov/meeting-aca-mandate-build-data-capacity.

83. Amoozegar J, Blumenfeld B, Brown S, Ortiz A, Renaud J, Richardson J, et al. Building Data Capacity for Patient-Centered Outcomes Research in HHS: A Formative Evaluation of 2012-2016 Projects. Washington, D.C.; 2017. Available from: https://aspe.hhs.gov/system/files/pdf/259016/ASPEPCORTFEvaluation.pdf

84. JASON. A Robust Health Data Infrastructure. Rockville, MD: Prepared by JASON, The MITRE Corporation; 2014 April. Report No.: AHRQ Publication No. 14-0041-EF. Available from: https://www.healthit.gov/sites/default/files/ptp13-700hhs_white.pdf

85. Estiri H, Stephens KA, Klann JG, Murphy SN. Exploring completeness in clinical data research networks with DQe-c. J Am Med Inform Assoc. 2018;25(1):17-24.

86. JASON. Data for Individual Health. McLean, VA: Prepared by JASON, The MITRE Corporation; 2014 November. Report No.: AHRQ Publication No. 15-0006-EF. Available from: https://healthit.ahrq.gov/sites/default/files/docs/publication/2014-jason-data-for-individual-health.pdf.

87. Office of the National Coordinator for Health Information Technology. Connecting Health and Care for the Nation: A 10-Year Vision to Achieve an Interoperable Health IT Infrastructure. Washington, DC: U.S. Department of Health & Human Services. Available from: https://www.healthit.gov/sites/default/files/ONC10yearInteroperabilityConceptPaper.pdf.

88. Mandl KD, Kohane IS. A 21st-Century Health IT System - Creating a Real-World Information Economy. N Engl J Med. 2017;376(20):1905-7.

89. JASON. Artificial Intelligence for Health and Health Care. McLean, VA: The MITRE Corporation; 2017 December 2017. Report No.: JSR-17-Task-002. Available from: https://www.healthit.gov/sites/default/files/jsr-17-task-002_aiforhealthandhealthcare12122017.pdf

90. Chaney KJ, Dymek C, Painter M, Zayas-Cabán T. From Hype to Reality: How Artificial Intelligence (AI) Can Transform Health and Healthcare. Content last reviewed January 2018. . Rockville, MD: Agency for Healthcare Research and Quality; 2018. Available from: http://www.ahrq.gov/news/blog/ahrqviews/hype-reality-artificial-intelligence.html.

91. Trivedi G, Pham P, Chapman WW, Hwa R, Wiebe J, Hochheiser H. NLPReViz: an interactive tool for natural language processing on clinical text. J Am Med Inform Assoc. 2018;25(1):81-7.

92. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3:160018.

93. FORCE11. Guiding Principles for Findable, Accessible, Interoperable and Re-usable Data Publishing version b1.0 | FORCE11. 2016.

94. Lopez MH, Holve E, Sarkar IN, Segal C. Building the informatics infrastructure for comparative effectiveness research (CER): a review of the literature. Med Care. 2012:S38-S48.

95. Kawamoto K, Lobach DF, Willard HF, Ginsburg GS. A national clinical decision support infrastructure to enable the widespread and consistent practice of genomic and personalized medicine. BMC Med Inform Decis Mak. 2009;9(1):17.

96. Khelifi M, Tarczy-Hornoch P, Devine EB, Pratt W. Design Recommendations for Pharmacogenomics Clinical Decision Support Systems. AMIA Summits on Translational Science Proceedings. 2017:237.

97. McCarty CA, Chisholm RL, Chute CG, Kullo IJ, Jarvik GP, Larson EB, et al. The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. BMC Med Genomics. 2011;4(1):13.

98. McDonagh EM, Whirl-Carrillo M, Garten Y, Altman RB, Klein TE. From pharmacogenomic knowledge acquisition to clinical applications: the PharmGKB as a clinical pharmacogenomic biomarker resource. Biomark Med. 2011;5(6):795-806.

99. Green RC, Goddard KA, Jarvik GP, Amendola LM, Appelbaum PS, Berg JS, et al. Clinical sequencing exploratory research consortium: accelerating evidence-based practice of genomic medicine. Am J Hum Genet. 2016;98(6):1051-66.

100. Ogunyemi OI, Meeker D, Kim H-E, Ashish N, Farzaneh S, Boxwala A. Identifying appropriate reference data models for comparative effectiveness research (CER) studies based on data from clinical information systems. Med Care. 2013;51:S45-S52.

101. Shirts BH, Salama JS, Aronson SJ, Chung WK, Gray SW, Hindorff LA, et al. CSER and eMERGE: current and potential state of the display of genetic information in the electronic health record. J Am Med Inform Assoc. 2015;22(6):1231-42.

102. Kho AN, Pacheco JA, Peissig PL, Rasmussen L, Newton KM, Weston N, et al. Electronic medical records for genetic research: results of the eMERGE consortium. Sci Transl Med. 2011;3(79):79re1-re1.

103. McGuire AL, Basford M, Dressler LG, Fullerton SM, Koenig BA, Li R, et al. Ethical and practical challenges of sharing data from genome-wide association studies: the eMERGE Consortium experience. Genome Res. 2011;21(7):1001-7.

104. Robinson JR, Denny JC, Roden DM, Van Driest SL. Genome-wide and phenome-wide approaches to understand variable drug actions in electronic health records. Clin Transl Sci. 2018;11(2):112-22.

105. Williams MS, Buchanan AH, Davis FD, Faucett WA, Hallquist ML, Leader JB, et al. Patient-centered precision health in a learning health care system: Geisinger's genomic medicine experience. Health Aff (Millwood). 2018;37(5):757-64.

106. Blasimme A, Fadda M, Schneider M, Vayena E. Data Sharing For Precision Medicine: Policy Lessons And Future Directions. Health Aff (Millwood). 2018;37(5):702-9.

107. Barnado A, Casey C, Carroll RJ, Wheless L, Denny JC, Crofford LJ. Developing Electronic Health Record Algorithms That Accurately Identify Patients With Systemic Lupus Erythematosus. Arthritis Care Res (Hoboken). 2017;69(5):687-93.

108. Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. IEEE Journal of Biomedical and Health Informatics. 2017.

109. Aronson S, Mahanta L, Ros LL, Clark E, Babb L, Oates M, et al. Information technology support for clinical genetic testing within an Academic Medical Center. Journal of Personalized Medicine. 2016;6(1):4.

110. Danciu I, Cowan JD, Basford M, Wang X, Saip A, Osgood S, et al. Secondary use of clinical data: the Vanderbilt approach. J Biomed Inform. 2014;52:28-35.

111. Winickoff DE, Jamal L, Anderson NR. New modes of engagement for big data research. Journal of Responsible Innovation. 2016;3(2):169-77.

112. Wilbanks J. Portable approaches to informed consent and open data. Privacy, Big Data, and the Public Good: Frameworks for Engagement. 2014;1:234-52.

113. Agniel D, Kohane IS, Weber GM. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. BMJ. 2018;361:k1479.

114. Masys DR, Harris PA, Fearn PA, Kohane IS. Designing a public square for research computing. Sci Transl Med. 2012;4(149):149fs32.

115. Bertier G, Carrot-Zhang J, Ragoussis V, Joly Y. Integrating precision cancer medicine into healthcare-policy, practice, and research challenges. Genome Med. 2016;8(1):108.

116. Nayak L, De RK. Precision Oncology with Electronic Medical Records. Innovations. 2018.

117. Cohen JP. Overcoming regulatory and economic challenges facing pharmacogenomics. N Biotechnol. 2012;29(6):751-6.

118. Addie S, Olson S, Beachy SH. Applying an Implementation Science Approach to Genomic Medicine: Workshop Summary. Washington, DC: The National Academies Press. Report No.: 978-0-309-43776-9. Available from: https://www.ncbi.nlm.nih.gov/books/NBK356505/.

119. IEEE Xplore Digital Library. Open Medical Big Data and Open Consent and Their Impact on Privacy Big Data (BigData Congress), 2017 IEEE International Congress; June 25-30; Honolulu, HI: IEEE; 2017.

120. Giuse NB, Kusnoor SV, Koonce TY, Naylor HM, Chen SC, Blasingame MN, et al. Guiding oncology patients through the maze of precision medicine. J Health Commun. 2016;21(Suppl 1):5-17.

APPENDIX A: GLOSSARY

Term	Definition	Source
Data	see 'health information'	https://www.healthit.gov/sites/default/files/draft-trusted-exchange-framework.pdf
Electronic Health Information (EHI)	Any information that identifies the individual, or with respect to which there is a reasonable basis to believe the information can be used to identify the individual and is transmitted by or maintained in electronic media, as defined in 45 CFR 160.103, that relates to the past, present, or future health or condition of an individual; the provision of health care to an individual; or the past, present, or future payment for the provision of health care to an individual. EHI includes information that is accessed, exchanged, used or maintained in the context of the Trusted Exchange Framework and may be developed for an individual, on behalf of an individual, or provided directly from either an	https://www.healthit.gov/sites/default/files/draft-trusted-exchange-framework.pdf

	individual or from technology that the individual has elected to use. EHI includes but is not limited to ePHI and health information as defined in 45 CFR 160.103. However, unlike ePHI and health information, EHI is not limited to information that is created or received by a health care provider, health plan, public health authority, employer, life insurer, school, university or health care clearinghouse. EHI does not include health information that is de-identified consistent with the requirements of 45 CFR 164.514(b).	
Electronic Health Record (EHR)	A real-time patient health record with access to evidence-based decision support tools that can be used to aid clinicians in decision making. The EHR can automate and streamline a clinician's workflow, ensuring that all clinical information is communicated. It can also prevent delays in response that result in gaps in care. The EHR can also support the collection of data for uses other than clinical care, such as billing, quality management, outcome reporting, and public health disease surveillance and reporting.	https://www.healthit.gov/policy-researchers-implementers/glossary
EHR Certification Program	A voluntary certification program established by the Office of the National Coordinator for Health IT to provide for the certification of health IT standards, implementation specifications and certification criteria adopted by the Secretary. The ONC Health IT Certification Program supports the availability of certified health IT for its encouraged and required use under federal, state and private programs.	https://www.healthit.gov/policy-researchers-implementers/about-onc-health-it-certification-program
EHR Phenotyping	Phenotyping is the practice of developing algorithms designed to identify specific phenomic traits within an individual. A variety of data can be extracted from EHRs including structured and unstructured formats, billing codes, laboratory results, medication data and natural language processing (NLP) which searches text (like doctors' notes and reports) for key words and information. The ability to pull the correct patient records from a phenotype depends on narrowing down the proper search criteria to target the 'true' case.	https://emerge.mc.vanderbilt.edu/phenotyping-cohort-discovery-using-ehr-data/
Fast Healthcare Interoperability Resources	A standard for exchanging health care information electronically.	https://www.hl7.org/fhir/overview.html

(FHIR) Specification		
Health care clinicians	Clinical providers of health care services including physicians, nurses, behavioral health professionals, registered dieticians, chiropractors, and other licensed or certified care providers.	https://www.healthit.gov/sites/default/files/hie-interoperability/Roadmap-Executive%20Summary-100115-4pm.pdf
Health care provider (organizations)	The network of health care service providers that includes hospitals, skilled nursing homes, long term care, and other facilities; pharmacies, lab, and diagnostic facilities reporting test results.	https://www.healthit.gov/sites/default/files/hie-interoperability/Roadmap-Executive%20Summary-100115-4pm.pdf
Health Information	The terms "health information," "health data," and "data" are synonymous in the context of the TEFCA and refer to all electronic health-related data for a patient. Specific references to ePHI refer to the HIPAA definitions of electronic protected health information and protected health information (PHI).	https://www.healthit.gov/sites/default/files/draft-trusted-exchange-framework.pdf
Health Information Technology	The application of information processing involving both computer hardware and software that deals with the storage, retrieval, sharing, and use of health care information, data, and knowledge for communication and decision making.	https://www.healthit.gov/policy-researchers-implementers/glossary
HIPAA Privacy Rule (HIPAA)	The HIPAA Privacy Rule generally requires Covered Entities to take reasonable steps to limit the use or disclosure of, and requests for, protected health information (PHI) to the minimum necessary to accomplish the intended purpose unless an exception applies such as for treatment purposes. In certain circumstances, the HIPAA Privacy Rule permits a Covered Entity to rely on the judgment of the party requesting the disclosure as to the minimum amount of information that is needed. Such reliance must be reasonable under the particular circumstances of the request. This reliance is permitted when the request is made by: a public official or agency who states that the information requested is the minimum necessary for a purpose permitted under 45 C.F.R. §164.512 of the Rule, such as for public health purposes (45 C.F.R. §164.512(b)), another Covered Entity or a professional who is a workforce member or Business Associate of the Covered Entity holding the	https://www.healthit.gov/sites/default/files/draft-trusted-exchange-framework.pdf

	information and who states that the information requested is the minimum necessary for the stated purpose. See generally, 45 C.F.R. §164.502 and 45 C.F.R. §164. 514.	
Interoperable	An interoperable health IT infrastructure is one in which all individuals, their families, and their health care providers have appropriate access to health information that facilitates informed decision-making, supports coordinated health management, allows patients to be active partners in their health and care, and improves the overall health of our population	https://www.healthit.gov/sites/default/files/ONC10yearInteroperabilityConceptPaper.pdf
Medical devices	Instruments, machines and implanted devices monitoring clinical indices, for immediate use as well as for historical purposes.	http://www.hitechanswers.net/wp-content/uploads/2013/05/NAHIT-Definitions2008.pdf
Medical Research	A biomedical, health services, and health policy research infrastructure robust enough to assure continued development of knowledge through primary discovery and in response to clinical and public health insights.	https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/1472-6947-3-1
Outcomes	The metrics by which stakeholders will measure our collective progress on implementing the Roadmap	https://www.healthit.gov/sites/default/files/hie-interoperability/Roadmap-Executive%20Summary-100115-4pm.pdf
Patient-generated health data (PGHD)	Patient-generated health data (PGHD) are health-related data created, recorded, or gathered by or from patients (or family members or other caregivers) to help address a health concern.	https://www.healthit.gov/policy-researchers-implementers/patient-generated-health-data
Patient-generated health information	see 'PGHD'	https://www.healthit.gov/policy-researchers-implementers/patient-generated-health-data
Patient-reported Outcomes (PRO)	Any report of the status of a patient's health condition that comes directly from the patient, without interpretation of the patient's response by a clinician or anyone else	https://www.qualityforum.org/Projects/n-r/Patient-Reported_Outcomes/Patient-Reported_Outcomes.aspx
Policy	Essential policy-related items stakeholders will need to implement in similar or compatible ways to	https://www.healthit.gov/sites/default/files/hie-

	facilitate the development of a health IT infrastructure to support research.	interoperability/Roadmap-Executive%20Summary-100115-4pm.pdf
Public health	Functionally, the network of government health departments, disease surveillance and immunization programs, school-based care providers and social workers, and nongovernmental organizations engaged in health and wellness.	https://www.healthit.gov/sites/default/files/hie-interoperability/Roadmap-Executive%20Summary-100115-4pm.pdf
Social Determinants of Health (SDOH)	The conditions in which people are born, grow, live, work and age. These circumstances are shaped by the distribution of money, power and resources at global, national and local levels.	http://www.who.int/social_determinants/sdh_definition/en/