Building a Holistic View for Patient-Focused Evidence: A Policy Framework for High Quality Source Data Collection, Curation, and Linkage

July 1, 2025

Duke
MARGOLIS INSTITUTE *for*
Health Policy

Authors

Patrick Rodriguez, Policy Analyst

Molly Shields, Policy Research Assistant

Trevan Locke*

Brian Canter, Assistant Research Director

**Former Duke-Margolis researcher*

Acknowledgments

We thank Amy Abernethy, Brad Hirsch, and Nick Honig at the Highlander Health Institute; as well as Kylie Brown, Wenbo Bai, Sabine Sussman, Hannah Graunke, Valerie Lehman, Frank McStay, Christina Silcox, Beena Bhuiyan Khan, Marianne Hamilton Lopez, and Mark McClellan at the Duke-Margolis Institute for Health Policy, for their review. We also thank Hannah Vitiello, Patricia Green, and Laura Hughes for their communications and design support.

About the Institute

The Robert J. Margolis, MD, Institute for Health Policy at Duke University is directed by Mark McClellan, and brings together expertise from the Washington, DC, policy community, Duke University, and Duke Health to address the most pressing issues in health policy. The mission of Duke-Margolis is to improve health, health equity, and the value of health care through practical, innovative, and evidence-based policy solutions. Duke-Margolis catalyzes Duke University's leading capabilities, including interdisciplinary academic research and capacity for education and engagement, to inform policymaking and implementation for better health and health care. For more information, visit healthpolicy.duke.edu.

Funding

This white paper was supported by funding from Highlander Health.

Disclosures

Mark McClellan is an independent director on the boards of Johnson & Johnson, Cigna, Alignment Healthcare, and PrognomIQ and receives fees for serving as an advisor for Arsenal Capital Partners, Blackstone Life Sciences, and MITRE.

Duke
MARGOLIS INSTITUTE *for*
Health Policy

EXECUTIVE SUMMARY

Issue

The United States health care system lacks an integrated, interoperable infrastructure for collecting and linking longitudinal patient data across care settings. Information from electronic health records (EHRs), patient-reported outcomes (PROs), claims, and other real-world data (RWD) sources has the potential to generate real-world evidence (RWE) that supports innovation, clinical decision-making, and value-based payment models.

Opportunity

The Duke-Margolis Institute for Health Policy conducted a review of source data quality, collection, curation, and linkage—informed by an extensive landscape analysis and 41 real-world use cases. The analysis illustrates the promise and current limitations of generating clinical evidence within the U.S. health care system. Its results lay the foundation of a policy framework to enable more routine, reliable use of longitudinal patient data to support learning health care systems.

Key Findings

- The creation of longitudinal data integration is technically feasible but requires operational and policy support.
- Uneven adoption of source data types and data collection methods exist, with wide variability in quality and structure, and few health care organizations able to capture reliable longitudinal data.
- The range of leading use cases is designed to advance the collection, curation, and integration of key longitudinal patient data with potential replicability and scalability.
- Effective policies are needed to drive collaboration between key stakeholders, build on technical advancements, bridge disconnected data siloes, and align incentives.

Policy Recommendations and Applications

1. Policies should support the development of cross-industry pilot projects to test a set of limited data elements leveraging existing initiatives and data repositories. These pilot projects can streamline data collection and support longitudinal evidence generation for clinical decision making, quality improvement, regulatory, and payment needs. Further investment and guidance development can enable data standardization and industry participation in priority RWE research areas through pilot project development. Potential applications are gene therapy outcome-based agreements and long-term monitoring. Pilots that define and standardize key clinical and quality-of-life data elements, e.g., durable response markers, can support these contracts by enabling consistent, auditable RWE collection.

2. Following pilot testing and validation, policies should support the development of demonstration projects to explore real-world research and care in a variety of health care settings. A potential application is Coverage with Evidence Development (CED). CMS linked coverage of transcatheter tricuspid valve replacement to longitudinal patient registry participation demonstrates how public payers can incentivize data infrastructure development by tying reimbursement to evidence generation. Demonstration projects here can test the integration of EHR, claims, and registry data to assess outcomes and inform broader payment policy reforms. Policymakers should expand agency artificial intelligence (AI) governance and curation standards for use in data abstraction, cleaning, and enrichment and exploration of the efficacy and validity of data curation techniques to generate missing or unlinked data. A potential application is gene therapy's long-term monitoring and treatment-specific registries. For conditions like spinal muscular atrophy or inherited retinal disease, curating long-term outcomes from patient notes, imaging, and lab results are key. AI tools, when governed appropriately, can automate data abstraction from varied sources and ensure consistency in monitoring post-treatment trajectories.

3. To foster data interoperability and linkage, policies should support accelerated adoption and integration of standard common data elements, common data models, and application program interfaces. To raise the minimum standard for health information technology (IT) systems, policies should further encourage health data companies and health systems to adhere to the Trusted Exchange Framework and Common Agreement (TEFCA). A potential application is cardiovascular registries. Cardiometabolic interventions, e.g., GLP-1s or structural heart devices, often require outcome tracking across inpatient, outpatient, and pharmacy systems. Promoting standardized data exchange via the Fast Healthcare Interoperability Resources (FHIR) and integrating cardiovascular registries with payer claims through TEFCA-compliant networks can support timely assessments of safety and value.

4. To improve alignment between clinical research, care, and payer priorities to support dynamic learning health systems, policies can create a standardized data infrastructure that seamlessly integrates into routine care visits. Potential applications are diabetes and cardiovascular registries. Standardized data infrastructure can be leveraged to report quality measures, transform data for regulatory submissions and facilitate value-based coverage.

Policies that enhance the availability, quality, and connectivity of longitudinal patient data are not just enablers of better clinical science—they are strategic levers for transforming health care delivery and reimbursement.

By investing in a limited but powerful set of shared data elements and interoperability tools, the U.S. can move toward a holistic patient-focused, learning health system—capable of delivering better outcomes, faster discoveries, and smarter spending.

Conclusion

Policies that enhance the **availability, quality, and connectivity** of longitudinal patient data are not just enablers of better clinical science—they are strategic levers for transforming health care delivery and reimbursement. Real-world applications are starting points to demonstrate how policy frameworks can **catalyze innovation and align stakeholder incentives.** By investing in a limited but powerful set of shared data elements and interoperability tools, the U.S. can move toward a holistic **patient-focused, learning health system—**capable of delivering better outcomes, faster discoveries, and smarter spending.

Background and Status of Current Source Data Landscape

Despite a wealth of source data (see **Appendix A**), the U.S. lacks an integrated health care and data accrual infrastructure, and health systems cannot readily track longitudinal patient data on key health characteristics.[1] Source data and the underlying infrastructure needed to support such an integrated system are siloed and therefore, do not support efficient collection and sharing during routine care. These inefficiencies prevent large-scale reliable data linkage, inhibiting the ability to create longitudinal datasets to support improvements in holistic patient care and patient-focused clinical research.

In an increasingly complex health care ecosystem, expanded evidence-generation beyond the current research paradigm can advance innovation and public health efforts. Innovating on multi-stakeholder data sharing can supplement randomized clinical trials (RCTs), the traditional gold standard approach for evidence to inform patient care and medical product development. RCTs have rigorous protocols and clear endpoints.[2, 3] Unfortunately, because reliable longitudinal data collection is so challenging, RCTs face cost, implementation, and sample size limitations.[4, 5, 6] Additionally, their findings are often not generalizable to broader populations and conditions of care, which leads to limited snapshots of patient health characteristics that can impede insurance coverage decision-making and other clinical needs. Accordingly, a need exists to leverage other sources of evidence available to researchers.

One of these is RWD, considered to be routinely collected source data that covers "usage, or the potential benefits or risks, or a drug derived from sources other than traditional clinical trials."[7] RWD can encapsulate the retrospective, secondary use of already existing data (e.g., medical claims or EHRs) or prospectively generated data according to a pre-specified research protocol (e.g., a disease registry or an expanded access program). Effective analysis of RWD can generate RWE, the "clinical evidence about the usage and potential benefits or risks of a medical product" or clinical care delivery.[8] RWE can assess quality and outcome measures within clinical care settings to enable real time evaluation of interventions towards identifying and scaling best practices (i.g., rapid learning). Additional capabilities of RWE include increases in patient sub-group representation and generalizability, and informing regulatory decisions about the safety and efficacy of medical products.

Longitudinal patient tracking is technically feasible and could significantly benefit efforts to improve both clinical care and research. Synchronized improvements to clinical care and research enable capabilities to collect, curate, and distribute data in a cost-effective manner for stakeholders across the health care landscape. These stakeholders include health systems, health technology companies, payers of clinical care, data and software vendors, pharmaceutical and medical device companies, health care provider professional associations, government agencies,

and international organizations. But our siloed, inefficient system results in fundamental differences in data quality and reliability among source data that result in high barriers and costs that block such research. For example, care delivery and therapeutic approach differences across health care delivery systems and within single-system clinical sub-units create limitations and data quality challenges for EHR data.

The ability to leverage source data, like EHRs, for longitudinal data tracking depends on data quality and stakeholders' ability to document, enhance, and understand its implications for analysis. In other words, the health care system must be restructured to enable merging disparate patient data to create a meaningful picture of individual patients. Improved data linkage centered around patients' movement through the health care system can deliver providers with access to reliable longitudinal health data to improve patient outcomes. Linkage also can support scalable, sustainable infrastructure for rapid learning to improve clinical care that can be re-tooled to address a public emergency or evolving priorities.[21, 22] Governmental leadership has demonstrated interest in real-world, population-based studies and the alignment needed to support multi-stakeholder data sharing.[23, 24]

One impactful method of prospective, rapid learning is to integrate clinical trials where people receive routine care—at the point-of-care. Previous Duke-Margolis work explores how point-of-care trials can better capture post-market performance of medical products in real-world populations and enable continuous health care treatment learning, when data quality and longitudinal completeness are adequate.[25] The COVID-19 pandemic demonstrated the impact point-of-care trials can have, particularly in an interconnected system. Oxford's Randomised Evaluation of COVID-19 Therapy (RECOVERY) trial drew from the United Kingdom's already integrated health system to rapidly enroll and study patients at the point-of-care.[26] Researchers paired routinely collected data with manually accrued data of a limited number of critical variables to quickly and efficiently generate the study dataset. RECOVERY's rapid expansion and associated findings provided timely, salient safety and efficacy information for COVID-19 treatments.[26]

A comparable effort in the U.S., the Accelerating COVID-19 Therapeutic Interventions and Vaccines (ACTIV) studies, demonstrated our system's evidence generation limitations due to systemic infrastructure issues. The studies enrolled fewer patients and had more data linkage challenges despite starting before RECOVERY.[27] While ACTIV was an exciting mobilization of pandemic research efforts, systemic issues with the U.S. evidence generation infrastructure limited the trials' success and highlighted quality and interoperability issues. While both trials required breaking down source data siloes and prioritizing key data for collection, ACTIV showcased the work needed to realize point-of-care trials and the holistic view for patient-focused evidence they engender in the U.S. health care system.

We need rigorous, widely adopted standards and practices that enable researchers to effectively harmonize disparate source data into fit-for-purpose datasets regulators need for decision-making.

To move forward, interventions for addressing significant health problems can be analyzed with advances in technology, coupled with a targeted policy framework. We need rigorous, widely adopted standards and practices that enable researchers to effectively harmonize disparate source data into fit-for-purpose datasets regulators need for decision-making.[28] Addressing these issues across the patient journey could facilitate a scalable clinical research infrastructure that draws on existing health care data and enables learning. Such a system would link currently disconnected evidence generation needs of regulators, payers, health systems, and other stakeholders. In doing so, the existing U.S. research and clinical infrastructure would be further optimized to improve the health of patients, recognizing that within our varied data landscape, the one constant demand driver is the patient. To move toward this goal, Duke-Margolis and Highlander Health partnered to explore source data collection, curation, and linkage; and develop a policy-based framework for achieving an interconnected, continuously learning health care system.[29] Our findings and subsequent recommendations lay out core considerations and functions of an ideal holistic view for patient-focused evidence within learning health systems. Broader benefits of these systems are discussed in a prior Duke-Margolis paper.[30]

Our research identified strategies for integrating quality processes into data collection and curation, methods for linking interoperable data, approaches to building a holistic view for patient-focused evidence, and gaps within each domain. Our findings are informed by 41 real-world examples—or use cases—of source data collection, curation, linkage, and/or interoperability. **Appendix B** discusses our methodology and broad results, **Appendix C** presents a high-level use case overview, and **Appendix D** provides in-depth use case goal and purpose descriptions.

Optimizing High Quality Data Collection and Curation

Data Collection Quality Assurances

Our findings show that data quality and assurance can take place across three dimensions: collection, curation, and at the study level. Quality assurance, paired with the use of limited data elements, at each of these stages is essential for making holistic patient profiles more financially sustainable. The use of limited data elements responds to stakeholder demand signals across three dimensions: easing data mapping challenges for researchers, mitigating provider burden for collection, and maximizing resources while reducing time costs and risks for health systems.

Regarding data collection, our review identified several articles about how health care providers and clinical trial sponsors can improve data quality. Federal guidance offers several standard terminologies that sponsors can incorporate to reduce uneven data collection methods and establish standardized measures from the point of collection. The Asthma RWE and Harmonized Outcome Measures for Asthma Patient use cases relied on International Classification of Disease (ICD) codes to identify patient cohorts and mapped data using ICD codes and Systemized Nomenclature of Medicine (SNOMED) terminology.[31, 32, 33, 34] The U.S. Food and Drug Administration's (FDA) "Data Standards for Drug and Biological Product Submissions Containing Real-World Data" guidance references Clinical Data Interchange Standards Consortium's (CDISC) controlled terminologies and highlights the importance of consistent coding practices, like ICD-9 and ICD-10 codes during data transformation.[35] Presently, CDISC standards are required for electronic data submissions to the FDA, but sponsors relying on RWE have experienced difficulties in transforming data into CDISC formats. The various types of RWD sources make it difficult for sponsors to standardize all RWD in a format that complies with regulatory guidance.[36]

Sponsors, investigators, and providers do not uniformly adopt controlled terminologies and face labor-intensive challenges to map RWD sources to CDISC and ICD terminology that accurately reflect clinical practices.[37] Extensive software tooling is needed to align in-house or system-specific practices with CDISC variables and terminology.[37] Implementation guides for mapping health data to adverse events, vital signs, and other CDISC domains, exist to reduce challenges with mapping and data redundancy.[38] Researchers in the Harmonized Outcome Measures for Asthma Patient Registries and Clinical Practice use case faced challenges aligning RWD source types, like EHRs, with standardized definitions that reflected clinical care practices.[32] Researchers found that even if clinicians collected the same type of routine health data, clinicians in separate health systems coded and classified the data differently. Professional societies should continue to advocate for regulatory agencies to accept common data elements and terminology that better reflect current clinical practices and the use of RWD and RWE in clinical trials.

The use of a limited number of fit-for-purpose data elements can ease administrative burden for providers and help standardize quality evaluations. Prioritizing limited data elements can align stakeholders while also being sufficient to track patient health status across various therapeutic areas and long-term specialty care applications. These discrete data elements can augment existing RWD, e.g., data on avoidable hospital admissions or high-variation procedures. Limited, fit-for-purpose, data elements can achieve the level of completeness required for quality clinical trials and care while reducing provider burden.

Providers are focused on delivering the highest quality care, but they are often constrained by limited resources and supports and face high administrative burdens. Health systems face challenging time and cost burdens while navigating uneven source data collection and varied data elements. The current reimbursement model mostly still pays providers on a fee-for-service basis, although roughly half of traditional Medicare beneficiaries are in accountable care relationships. Fee-for-service reimbursement does not incentivize longitudinal data collection beyond immediate care needs and can lead to upcoding and biased data.[39] Such a system does not incentivize providers to undertake additional data collection or cleaning responsibilities, particularly when they are beyond routine care delivery processes.

In the past several years, however, Accountable Care Organizations (ACOs) comprised of groups of doctors, hospitals, and other health care providers, have offered coordinated care incentivized by health outcomes and quality rather than volume of services.[40] Appropriate regulatory and value-based payment incentives can encourage the adoption of key data elements that support reliable, longitudinal data sufficient for

coverage decisions and medical technology applications. The Center for Medicare and Medicaid Innovation (CMMI) recently announced a strategy that focuses on leveraging technology to empower patients and improve outcomes.[40] Strategic focus areas include increased data access, which can help align individual and payer incentives, and advanced accountable care models, which provide flexibility and enable efficient data usage to advance health outcomes. Other health care payers and purchasers have ongoing reforms to advance value-based payment models and reflect CMMI's strategic vision for technology-enabled and evidence-supported care improvements.[41, 42] Focusing these efforts within value-based frameworks encourages the creation of effective, innovative technologies that are interoperable, patient-centric, and are reimbursed based on outcome rather than utilization. Overall, these announcements are promising and reinforce the need for value-based payment with easy-to-operationalize data elements that lower costs and resources.

Data Curation Quality Assurances

Data curation is the continual processing, cleaning, and digital maintenance of data to enable accessibility, storage, preservation, and dispersion.[43] Curation also can include the addition of metadata to make data more comprehensive. Curation quality assurances can support intermediate steps toward CMMI's goal of evidence-based prevention. More comprehensive, usable data enables individuals to better understand their health status and manage their care. Curation quality assurances also will support complete and accurate data exchange within and across health systems as health leaders look to integrate digital technologies.[40]

A pressing need exists for standards development that includes mathematical and statistical considerations for data curation assessments. Recent technological advancements like AI have made enrichment a more practical solution for improving quality. Many use cases and stakeholders capitalized on the rapid advancement and dispersion of AI and machine learning (ML) algorithms for data abstraction. This effort demonstrates the perceived utility of AI/ML for RWD accrual, curation, and transformation, particularly when that data is unstructured.[44] A more basic example includes a health system IT department modifying an EHR system to account for a discrete data element of interest; however, this approach creates a new clinical workflow and would need to have saliency for patients and care delivery.

Other solutions include care delivery tools where high quality data is a byproduct, such as sensors or AI ambient listening that can automatically augment an EHR.[45] Despite noticeable enthusiasm for AI, in-depth demonstrations of AI validity are not readily nor publicly available. In recognition of this fact, several governmental agencies have released recommendations, guidance, and risk assessment frameworks for the use of AI/ML models throughout the product lifecycle.[46, 47, 48] Additional Duke-Margolis works discuss specific EHR-sourced data AI/ML considerations and general biomedical AI governance.[49, 50]

Operational frameworks can help steer curation quality assurance. Companies develop proprietary tools to solve for data quality because a centralized, regulatory body-endorsed "checklist" does not exist. FDA has previously stated it is not interested in developing a prescriptive data quality checklist, but there is a demonstrated need for good data collection and curation practices. The TransCelerate RWD Audit Readiness initiative assesses data for accuracy, completeness, and adequacy; a recent taskforce expanded the tool to offer sponsors a place to report and track different stages of data curation through a supplemental form.[51] The form helps reduce varied curation quality processes and motivates sponsors to collect the level of data completeness required for regulatory submission. Notably, many industry use cases reveal sponsors and researchers do not universally adopt operational frameworks and instead, create new "in-house" tools to assess quality.[52] Tools' disparate quality thresholds and processes impede reliable curation practices. As priorities shift to a technology-enabled infrastructure, shared data elements and quality assurances are crucial for mitigating uneven collection and curation methods, cutting costs, and supporting reliable, complete longitudinal patient data.

Efforts to standardize quality assurance includes the CMS proposed guidance on study protocols that use RWD.[53] The CMS guidance adapts the original HARPER framework into HARPER+, which includes medical devices and coverage criteria and cross-references the FDA's guidance on using RWE for medical device regulatory decision-making.[54] The CMS proposed guidance outlines how CED for items or services that are likely to benefit the Medicare population may incorporate RWD and highlights the importance of data relevance and reliability. The agencies' shared focus on best practices for RWD regulatory-grade submissions and alignment of coverage and evidence generation marks progress towards accessible longitudinal patient data that informs coverage and care applications.

Study-Level Quality Assurances

The third dimension of data quality and assurance takes place at the study level. Many quality approaches at this level, including the HARPER and HARPER+ frameworks, tend to focus on study methods and assessing whether data is sufficient to address the research question at hand. These tools are necessary but often do not provide accessible and scalable mechanisms for addressing quality beyond the immediate study. Investigators for a Verantos-led FDA demonstration project, Transforming Real-World Evidence with Unstructured and Structured Data to Advance Tailored Therapy (TRUST) study, describe their novel data quality approach that improved data curation, such as completeness, traceability, and accuracy scores by roughly 50 percent compared to traditional approaches.[55, 56] The source data used were medical and pharmacy claims, mortality registry data, and EHRs from 58 hospitals and 1,180 outpatient clinics. The TRUST study, and its documentation of quality assurance at the study level represent a promising step toward improved data quality efforts among industry-government partnerships.

Increased detail and transparency on operationalizing recommendations from projects like the TRUST study present an opportunity to increase the accessibility of generating high-quality, reusable data. Doing so, and providing infrastructural support for newer organizations, can help expand the groups contributing to building a holistic view for patient-focused evidence. The current landscape often includes incomplete or unstandardized data sources, and these quality issues prevent the optimization of high-quality, complete, and error-free data collection in clinical research and care applications. Pilot and demonstration projects, bolstered with aligned governmental data quality guidance, are important steps to identify and test the feasibility of new standards and a limited number of shared data elements that can be fit-for-purpose in various clinical research and care settings.

Recommendations and Action Steps for Optimizing High Quality Data Collection and Curation

Develop pilot projects to identify a set of useful limited data elements where there are existing or emerging drivers for improving data element quality and interoperability. Cross-industry collaborations should focus on developing statistical methods to objectively assess data quality at the curation and aggregation levels. Efforts should prioritize identifying a limited set of data elements with high utility across stakeholders, e.g., providers, payers, health systems, and researchers. This approach could streamline data collection and support longitudinal evidence generation. Standardized collection practices and more efficient evidence generation would support applications, such as long-term gene therapy monitoring and outcome-based agreements. Pilot programs can test how limited, discrete data elements impact payer and regulator priorities, like quality improvement measures. Investment from health data companies is key to building infrastructure that eases provider burden and encourages uptake. These pilots should take advantage of existing or emerging demand signals, often found in specific disease areas or therapies. For instance, CMS coverage lowers out-of-pocket payment expenses while supporting innovative treatments for transcatheter tricuspid valve replacement under CED (cardiovascular care), the Cell and Gene Therapy Access Model (cell and gene therapy), and Enhancing Oncology Model (oncology) using safety- and outcome-based payment models.[57–59] These models can attract industry participation by aligning patient-centered outcome incentives with clear value propositions.

Test the real-world feasibility of identified limited data elements through demonstration projects. Following pilot program identification of existing and readily available performance measures that can reflect routine care priorities and current practices and capture research-relevant datasets, demonstration projects can be established to run proof-of-concept studies in different real-world research and care settings. For example, claims-based measures can assess longitudinal care improvements through avoidable admissions and readmissions rates. Key patient health data recorded in EHRs, like lipid and blood pressure measures, also can support longitudinal chronic disease risk factor management. Measures like these, and other identified ones, should be explored in a variety of health care settings, like large academic medical centers, rural community health systems, and federally qualified health centers. CMS CED, which utilizes a longitudinal patient registry to inform coverage of transcatheter tricuspid valve replacement, is a demonstrative example of using a defined set of data elements to guide quality improvement and coverage assessments. Lessons learned from these projects would help optimize operational workflows, develop generalizable approaches for future expansion, and identify fundamental data and IT infrastructure needs to facilitate high-quality data collection, curation, and transformation needed for longitudinal data tracking.

Expand agency guidance on AI governance and best practices. Expanded AI governance and best practices for sponsor and industry use of AI will help promote standardization. In particular greater clarity on developers' governance scheme for AI applications, such as clinical decision support software and monitoring use cases, would be helpful. These applications do not universally fall under FDA's classification as a clinical device software if it solely analyzes information normally communicated between health care professionals (excluding images) or offers initial recommendations to providers without specific outputs or directives, e.g., probability or risk scores of a disease or condition, but as AI advancements and adoption continues, these practices could theoretically generate evidence used in a product application.[60] Having clarity on these tools, especially when they are used to augment point-of-care data collection, will help with future assessments of data quality. Effective AI can lead to timely care for patients and cost savings for health systems and trial sponsors. Properly governed AI tools can support data abstraction and curation methods for a range of health care applications including long-term monitoring of gene therapies and unstructured registry data. The latter is exemplified in the Pragmatic Asthma Registry use case where AI was employed to abstract unstructured data on patients being treated for severe asthma.[61]

Data Interoperability and Linkage

High-quality data collection, curation, and analysis for longitudinal health data tracking is informed by stakeholders' ability to link interoperable data to one another to form a holistic view for patient-focused evidence. Unlike other health systems, U.S. health data and infrastructure are generally disconnected and, therefore, not readily suited for large-scale linkage. However, some segments within our system are interconnected and may act as a model to expand linkage into our broader network. The Veterans Affairs (VA) Diuretic Comparison Project (DCP) use case is a good example.[62] The project used EHR data from the VA health care system, Medicare claims, and the National Death Index to study non-cancer deaths and major cardiovascular outcomes. Researchers hierarchically organized the three data sources and then validated clinical endpoints across datasets.

The DCP demonstrates that linking, analyzing, and validating health data within a U.S. system is technically feasible. Outlining the stakeholder value in translating technical lessons from this public system into privately operated ones is needed. To achieve this vision, stakeholders must be capable of seamless data transfer and collation between one another. Existing public initiatives and policy mechanisms can be leveraged to improve interoperability and standardization. These include FDA guidance, the 21st Century Cures Act, the TEFCA, and a joint request for information from CMS and the Assistant Secretary for Technology Policy (ASTP) to "strengthen interoperability and...seamless data exchange".[23, 63, 64] Alignment around the technical dimensions of data linkage and interoperability can act as a focal point to demonstrate the utility of an interconnected system to stakeholders.

Dimensions of Data Interoperability and Linkage

Three fundamental dimensions of data interoperability and linkage are needed for a holistic view for patient-focused evidence development: standardized data elements, structures, and transfers. Data elements are discrete "objects that can be collected, used, and/or stored in clinical information systems."[65] Similar to how individual words build full sentences, distinct patient data elements make up larger patient profiles. However, research institutions and health care systems often rely on variable elements. Consequently, the building blocks needed for creating a holistic view for patient-focused evidence are not often interchangeable from one system to another. Common data elements (CDEs) combine a "precisely defined question...with a specified set of responses" and can be standardized across multiple health systems.[66, 67] From a research perspective, CDEs enable better data collation into a cohesive whole; from a clinical care perspective, consensus-developed CDEs can foster rapid evidence generation around novel and persistent health outcome priorities. Additionally, standardized data elements facilitate targeted stakeholder engagement on the usefulness, accuracy, and need for new or existing CDEs.

Common data elements (CDEs) combine a "precisely defined question...with a specified set of responses" and can be standardized across multiple health systems.

CDE repositories, such as the National Institutes of Health's (NIH), support consistent, study-to-study data collection.[68] Another prime example is the United States Core Data for Interoperability (USCDI) and the expanded USCDI+.[69, 70] They present "a standardized set of health data classes and constituent data elements" for national health information exchange.[71] Beneficially, USCDI moves beyond just CDEs and includes standardized data classes too, which aggregate "data elements by a common theme or use case."[72] A key difference between NIH's repository and USCDI is the intended user-group. The NIH CDE repository is geared specifically toward research, whereas USCDI and USCDI+ facilitate the clinical exchange of health information. While USCDI+ improves flexibility to accommodate emerging public health and medical research, further work is needed to actionably harmonize the two CDE repositories.

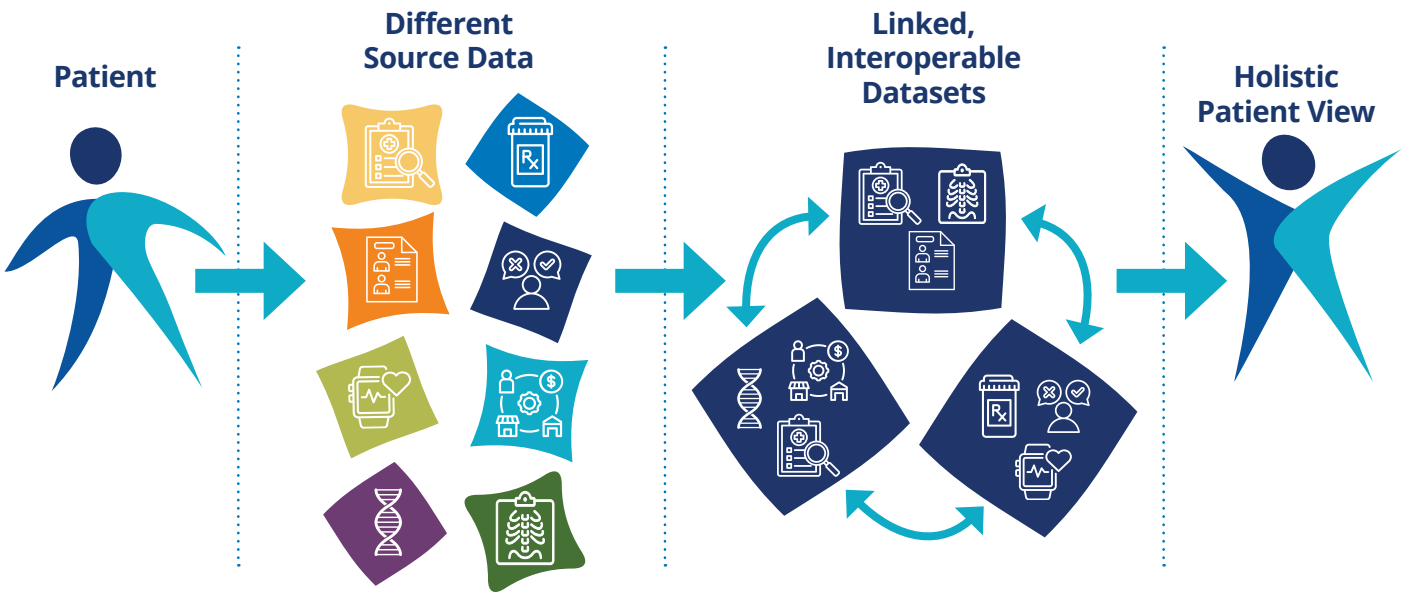
After the CDE building blocks are implemented, the next step is to map and standardize the structure and content of disparate data into a general format, which is done using a common data model (CDM), like PCORnet and the Observational Medical Outcomes Partnership (OMOP).[73, 74] CDMs facilitate interoperability by executing "an identical query…on multiple datasets" to work across different data sources.[9] A holistic view for patient-focused evidence necessarily spans several datasets; CDMs minimize time and resources needed to map diverse data from one another. An infrastructure with built-in, research-ready CDMs fosters rapid development of cross-system patient profiles. An inherent challenge with CDMs is the risk of losing or changing information meaning as data are formatted, due to semantically similar or equivalent data terms that are functionally different between providers and/or health systems.[75] However, if CDEs are present within health systems at the point of data capture, this concern can be partially mitigated.

Working across different datasets necessitates transfer of source data to other programs. Here, two systems are relevant. The first is TEFCA, from ASTP. TEFCA is a nationwide framework that covers the baseline legal and technical requirements required for secure information exchange. Essentially, the framework sets the ground rules for health information sharing across six information exchange purposes—notably, research is not a covered purpose.[76] These exchanges happen between registered Qualified Health Information Networks (QHINs). QHINs, typically large health information networks, manage cross-network data exchange and directly connect with other QHINs. There are also Participants, who contract with QHINs to access data directly from them, and Subparticipants, who are sub-contracted with Participants to use accessible QHIN data.[77] Participants may be large health systems, health IT developers, or public health agencies; and Subparticipants are typically individual physician practices, pharmacies, or laboratories. Each descending role possesses diminished control of data exchange while simultaneously facing fewer infrastructural and workflow compliance expectations. An example of diminishing compliance expectations is that while TEFCA requires certified QHINs' data to minimally conform to USCDI v.1 (and USCDI v.3 by January 1, 2026), Participants do not necessarily need to ensure all their own internal data transfer meets USCDI standards but can still access USCDI-conformed data through contracted QHINs.[77]

The second aspect of data transfer concerns application programming interface (API) tools. These work to streamline how information is transferred across different data holders and users—stakeholders like trial sponsors, health care providers, and regulators. The OneSource Project, a UCSF and FDA collaboration, applied standardized elements including CDISC to transmit UCSF's EHR data to a clinical trial EDC system.[78] The source data capture system was applied in the I-SPY 2 Trial (discussed further in **Appendix D**) and illustrates clinical application of data standardization and linkage to improve clinical trial efficiency. The FHIR is a common API and was relied on by the mCODE Genomics Pilot Project and Cancer Moonshot use cases.[79, 80, 81] To spur adoption of FHIR, seven accelerator programs provide initiation, infrastructure, and standards development aid to stakeholders.[82] These programs support clinical research across different health information-stakeholder domains.[83] For example, the Da Vinci project supports payer-provider data interoperability and the Gravity project focuses on social determinants of health (SDOH) data standards to improve health equity.[84, 85] The ICAREdata Project participated in the oncology-based CodeX accelerator program.[86, 87]

Where FHIR and other APIs represent specific technical processes for data sharing, TEFCA sets overarching expectations for standard data transfer. Combined with standardized data elements and structures, information could move within and across holistic patient profiles (see **Figure 1**). The selection of limited data elements that meaningfully cut across stakeholders supports alignment and streamlines data sharing within and across patient profiles. Efficient and standardized data sharing enables an ecosystem of rapid learning in health care systems and real-time research for a variety of purposes. Examples include improving management of chronic disease conditions, treating a rare genetic disease, or building clinical trial populations for public health emergencies. An important caveat of achieving these dimensions is the significant resources needed to invest in technological infrastructure, which may be prohibitive to stakeholder participation,

Figure 1 | Data Lifecycle in a Holistic View for Patient-Focused Evidence



particularly for small-to-mid-sized organizations. As such, clear demonstrations of long-term utility paired with short-term support is needed. These efforts can be supported via federal policy and initiatives.

Recommendations and Action Steps for Data Interoperability and Linkage

Encourage the adoption and concurrent integration of CDEs, CDMs, and APIs. The ASTP and other government agencies can promote broad adoption of CDEs, CDMs, and APIs to foster data interoperability across stakeholders. ASTP can host convenings to increase trial sponsors' comprehension of how to optimize and integrate CDMs along with corresponding data terminology to minimize adoption of incompatible standardization frameworks. Work is already underway to enhance interoperability between different CDMs, like OMOP and PCORnet.[88] Continued advancement of ASTP's Health IT Alignment policy to harmonize USCDI elements and classes with research-specific CDEs across the NIH, FDA, and CMS is needed.[89] The CMS Merit-based Incentive Payment Systems (MIPS) promotes interoperability through provider incentives; scoring measures take TEFCA adherence—and, consequently, USCDI conformance—into account.[90] FDA is also exploring possibilities for adopting HL7 FHIR for safety and efficacy decision-making related to RWE product submissions.[23] This work should strategically target a limited number of data elements capable of supporting alignment across several clinical care, research, and payment settings. This effort would be an iterative, bidirectional process among stakeholders to minimize upfront time and labor for provider data entry while enhancing long-term efficiency and data sharing capabilities across health systems and therapeutic areas. An application area well-suited for this integration are cardiovascular registries because cardiometabolic interventions, e.g., GLP-1s or structural heart devices, often require outcome tracking across inpatient, outpatient, and pharmacy systems. Promoting data capture and storage within standardized CDEs and CDMs along with standardized data exchange via FHIR can support timely, cross-system safety and value assessments.

Increase industry adherence to TEFCA. A growing number of health information exchanges and EHR vendors, e.g., EPIC and Oracle, have adopted TEFCA, but regulators can further encourage health data companies and health systems to become TEFCA-certified to raise the minimum standard for health IT systems. Adherence to TEFCA will raise patient confidence that their personal health and claims data is protected and supports data-sharing and standardization across health companies and trial sponsors. TEFCA exchange purposes should be evaluated to understand how research capabilities can be folded into existing exchange purposes or expanded to explicitly cover secondary research. Clinical research use cases could have broad applicability across the health care continuum, driving improvements for research, approval, and reimbursement applications. While TEFCA is voluntary, professional associations like the American Health

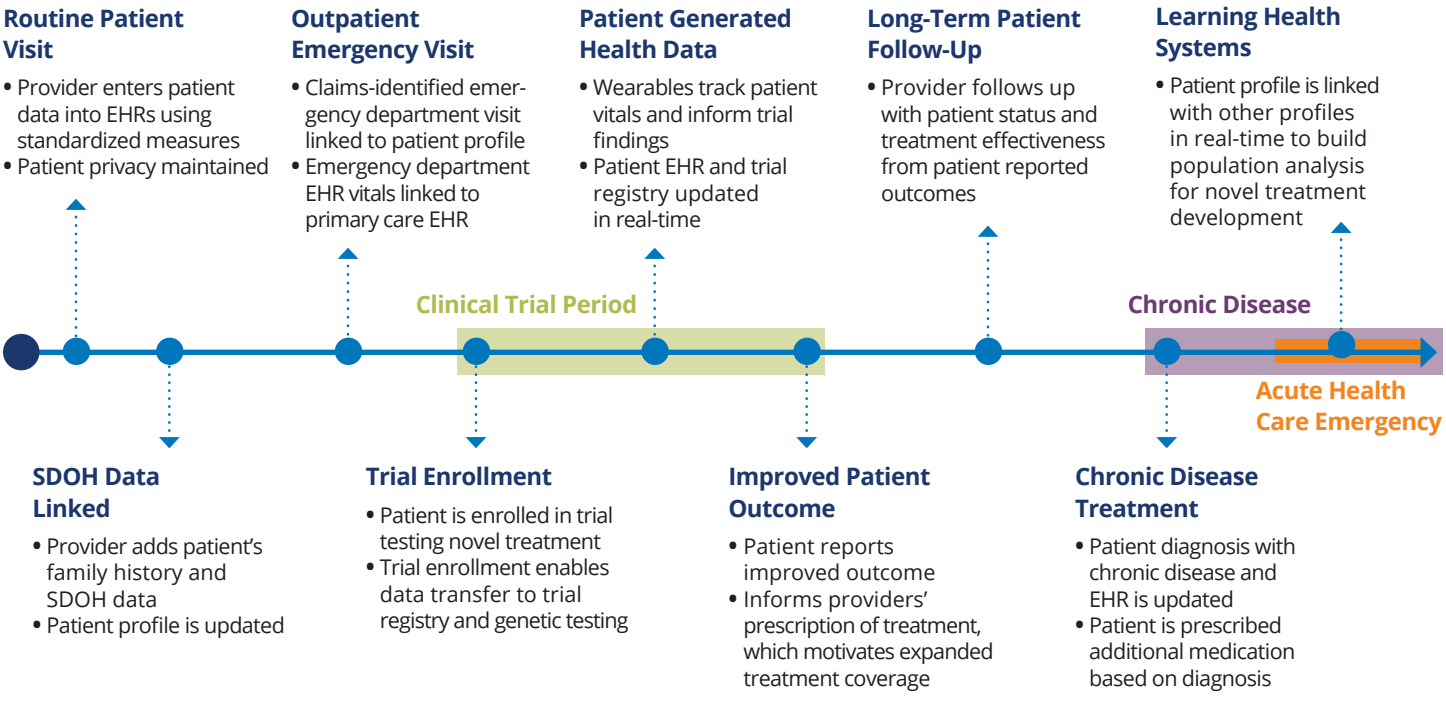
Information Management Association (AHIMA), and trade associations like the American Hospital Association (AHA), could encourage participation.[91, 92] CMS's announcement and commitment to enhance their participation in trusted data exchange reinforces growing recognition of TEFCA's utility to the health care ecosystem. Cardiovascular registries are a prime application area for this initiative—as payer claims through TEFCA-compliant networks can further facilitate rapid safety and value evaluations.

Creating a Holistic View for Patient-Focused Evidence

Our data collection, interoperability, and linkage policy framework illustrates that creating holistic patient profiles is difficult but conceptually feasible. As a patient engages with the health care system (see **Figure 2**) holistic patient profiles would capture, trace, and link health data (e.g., EHRs, genomic data, SDOH, insurance claims, etc.) throughout their health care journey.

A holistic view for patient-focused evidence generation infrastructure would enable reusable multi-purpose source data, expanding clinical research applications and trial efficiency. Our work demonstrates the limitations of currently fragmented and siloed patient data, which make a patient's longitudinal journey through the system difficult to capture effectively. On the other hand, **Figure 2** illustrates the full potential of a holistic view for a patient-focused evidence system through examples of data reuse to inform chronic disease treatment and rapid emergency response. The COVID-19 pandemic illuminated this value. A current demonstration is the clinical trial platform, Cantata by Protas, used in the EASi-KIDNEY trial.[93, 94] Cantata handles study protocols, operating procedures, and other activities in a "single, real-time" interface. Looking forward, opportunities exist to expand applications.[93] Cell and gene therapies are salient examples of long-term, RWD clinical research applications. Cardiometabolic glucagon-like peptide-1s (GLP-1s) are another long-term application that simultaneously addresses chronic conditions like obesity.[95]

Figure 2 | Conceptual Longitudinal Patient Journey



While actors in the health care system strive for improved patient outcomes, demonstrating stakeholder value like cost savings, improved community engagement, or diminished provider burden is necessary to achieve stakeholder investment in data infrastructure improvements. A modernized data accrual infrastructure and learning health system can reduce time-consuming manual data entry and organization of patient records, supporting cost savings for health system leaders.

A modernized data accrual infrastructure and learning health system can reduce time-consuming manual data entry and organization of patient records, supporting cost savings for health system leaders.

Federal agencies can incentivize movement toward sophisticated, longitudinal patient outcome tracking that relies on a limited number of data elements. For example, the CMMI Enhancing Oncology Model is an alternative payment model that uses performance-based payment and recoupment based on quality performance while leveraging data for patient tracking and improving clinical performance.[59] The model includes an implementation guide with a limited number of required data elements and terminologies aligned with HL7 FHIR mCODE.[96] The model's use of a limited number of data elements optimizes data sharing and supports the model's goal of coordinated care, reduced costs, and improved outcomes for Medicare fee-for-service beneficiaries with cancer.[59] Further, while major pharmaceutical companies sponsor RWD clinical research and care, more transparent and cohesive agency regulations can encourage additional participation from smaller, start-up companies. Health system data modernization also creates opportunities for new health technology companies to enter the health data market and collaborate with health systems. As more health system leaders choose to improve their data infrastructure, health data companies may see the value in shifting away from current siloed data practices to accommodate data sharing between health systems. Synergy between clinical research and care priorities will help facilitate the linkage of individual patient care improvements to enable population-based learning and widespread cost savings for health systems. To scale and sustain such a learning health system requires comprehensive stakeholder collaboration.

Recommendations and Action Steps for Creating a Holistic View for Patient-Focused Evidence

Support the development of pilot programs through investment and regulatory guidance to enable data standardization and more industry participation. Research organizations can increase understanding of barriers to entry for new data technology companies. Our findings illustrate that the data ecosystem is largely populated by established pharmaceutical companies, and consideration of infrastructure limitations can help agencies and investors prioritize funding opportunities for start-ups attempting to enter into the RWD/E generation landscape. Several modalities and health care interventions exist that would lend themselves to such pilot programs, including outcomes-based assessment of cell and gene therapy products. These products represent paradigm shifts in provider resources and require the use of longitudinal datasets that are difficult to generate with our current approach to evidence generation. More extensive agency guidance from CMS, FDA, and ASTP, paired with continual infrastructure investment from both public and private sources, can help harmonize data elements and incentivize longitudinal datasets. Several registry use cases struggled to map data, integrate new clinical measures, and update registries accordingly. Uniform adoption of the same data elements and exchange methods among health systems, providers, and researchers, like USCDI+, FHIR, and TEFCA, would improve data mapping efficiency. Stakeholder adoption of the same data exchange methods and increased standardization will also help steer more specific data technology investments.

Continue exploration of the efficacy and validity of data curation techniques to generate missing or unlinked data. The capacity for AI to support holistic data collection is important for comprehensive patient profiles and will help promote more efficient data collection for trial sponsors and researchers alike. Clinical trial sponsors, academic researchers, and health technology companies can drive continued exploration of AI to create complete datasets. Attention also should be given to other data curation techniques, such as manual or workflow efforts to clean, validate, transform, and/or enrich data. Additional transparency and guidance on proper curation process governance from FDA and CMS will help improve the regulatory approval process for agencies and sponsors, such as health technology companies. Promising AI developments include identifying

patients at high-risk for developing a specific disease, collecting complex source data types, like SDOH and genomic data, and collating unstructured notes from a health care visit. These developments can help improve patient outcomes and timely monitoring for various health applications, especially for long-term follow-up of gene therapies. A recent Duke-Margolis publication proposes using natural language processing and third-party software to integrate unstructured data and EHRs into registries.[97] More accurate and efficient care and research can improve patient outcomes and lead to widespread cost savings and waste reduction for health systems, researchers, providers, and payers.

Prioritizing the use of limited data elements across stakeholders can streamline efforts to ensure data measures reflect clinical care practices.

Improve alignment between clinical research, care, and payer priorities to support dynamic learning health systems. Continuous collaboration among providers, payers, trial sponsors, health technology companies, and regulatory agencies can create a standardized data infrastructure that seamlessly integrates into routine care visits. Prioritizing the use of limited data elements across stakeholders can streamline efforts to ensure data measures reflect clinical care practices. This prioritization should be done in partnership with clinicians to maintain their ability to capture clinical nuance and should not be done to the exclusion of unstructured, free-text clinical notes. Ongoing examples include CMS's quality reporting initiatives and encouragements to use APIs for FHIR bulk data and interoperability requirements for Medicare diabetes, cardiovascular, and depression measures.[98, 99] More inter-agency alignment is necessary as FDA requires CDISC for clinical trial data while CMS prioritization of FHIR for health data exchange makes transformation to CDISC standards for regulatory submission highly burdensome. However, a FDA request for comments on using HL7 FHIR standards for RWD shows a promising move towards harmonization.[23] GLP-1s are another future application area for clinical research and care priority alignment, as researching how GLP-1s can improve patient outcomes will help inform providers' understanding of the drug. Research on GLP-1s, informed care, and subsequent patient use of the drug can influence value-based coverage. Cardiovascular registries like TAVR also optimize data collection and Medicare coverage alignment to benefit long-term patient improvements. These examples collectively illustrate the need for collaboration among researchers, providers, and payers to support learning health system improvements that benefit patients.

CONCLUSION

This policy-based framework illustrates how stakeholders can optimize existing RWD and RWE tools and technologies to support holistic patient profiles and longitudinal clinical research and care. Current academic and industry investment in RWD collection and point-of-care trials is encouraging, but opportunities exist to improve existing data collection, data linkage, and interoperability strategies; update and expand existing regulatory guidance; and increase stakeholder engagement. Comprehensive stakeholder engagement can mitigate current siloed practices within the U.S. health care system that prevent longitudinal patient tracking. Providers, payers, trial sponsors, health technology companies, and regulatory agencies are among those who can drive data quality improvement and support the creation of sustainable and scalable holistic patient profiles. Organizations that bridge these stakeholder and infrastructure gaps will be a critical, galvanizing force. Holistic patient profiles are integral to improving quality care, building a clinical research infrastructure, and forming a learning health care system that can develop novel treatments, advance biomedical innovation, and respond to future health care emergencies.

Appendix A | Illustrative Source Data Descriptions, Strengths, and Limitations

Source Data	Brief Description	Strengths	Limitations
Electronic Health Records (EHRs)	Medical records "generated for use in clinical care and...as a basis for billing and for auditing of practice quality measures."[4] They contain structured and unstructured (e.g., free text, images, etc.) data fields recorded by providers and health care professionals with access to patients' records.	EHRs can provide rich detail on clinical decision-making and rationale, as well as behavioral health information.	EHRs are specific to health systems and sub-units within larger systems. Consequently, they may not cover long-term patient movements within the broader health care system.
Medical and Administrative Claims 	Billing records of dispensed medical products and procedures, according to standardized coding, "to support payment for care."[9] Claims data can come from Medicare, Medicaid, or commercial insurance companies.	Claims data can provide a patient's broad, in-network movement throughout the health care system for the duration of their coverage with a given insurance entity.	Claims do not provide in-depth clinical and behavioral health information, may not accurately reflect actual patient conditions, and do not capture out-of-network health care interactions or patients without insurance.
Patient Registry 	"An organized system that uses observational study methods to collect uniform data...to evaluate specified outcomes for a population defined by a particular disease, condition, or exposure."[8] Registries based on a specific disease are 'disease registries'; registries based on the use of a drug, device, or other treatment are 'exposure registries.'	Registries can be valuable for understanding disease progression, particularly for rare disease (with small sample sizes and high disease variability).[10]	Registries can face issues with validation of participant population and standardized data quality measures.[11]
Patient Reported Outcomes (PROs) 	"A measurement based on a report that comes directly from the patient...about the status of the patient's health condition without amendment or interpretation...by a clinician or anyone else."[8] Can be measured through self-report or interviews.	PROs can provide clear, unaltered patient perspectives on clinical care, medical products and/or disease progression.	PRO collection can be limited and fragmented, failing to capture patients without easy access to health care settings and/or those with limited technology.[12]
Patient-Generated Health Data (PGHD) 	"Health-related data created, recorded, or gathered by or from patients, or their family members/other caregivers, to help address a health concern."[13] Data cover medical/health history, treatment history, biometric data, and lifestyle choices.	PGHD methods can gather data not presently available within EHRs (e.g., quality of life, drug adherence, non-serious adverse events, etc.).	PGHD difficulties include reconciling data between different sources and devices, replicating data collection methods, and standardization of quality protocols.
Social Determinants of Health (SDOH) 	"Economic and social conditions that influence the health of people and communities... [including] the social environment, physical environment/total ecology, and health services/medical care."[14] SDOH data can be individual-level (e.g., education level, employment status, housing) or community-level (e.g., environment, neighborhood, socioeconomics).[15]	SDOH data can provide insight on how societal structures impact health outcomes of individual patients and patient sub-groups.	SDOH data can be largely descriptive/passive, be difficult to identify tangible benefits when integrated into health care reforms, and be harder to standardize/structure during the data collection and transformation processes.[16]
Genomic Data 	Information about "all of a person's genes (the genome), including interactions of those genes with each other and the person's environment."[17]	Genomic data can be used to predict the risk of future, adverse health outcomes, how patients may respond to a drug, and can inform preventative care based on a patient's risk profile.	Genomic data on diverse populations is limited, though large research programs like the *All of Us* Research Program are working to increase data on traditionally underrepresented research populations.[18] Additionally, genomic data are large and complex, requiring high storage and analysis capacities.
Imaging Data 	Imaging data are the detailed, high content pictures obtained from imaging tests. Imaging types include radiology images, digital pathology, retinal scans, echocardiography, video, and others.[19]	Imaging data can provide visual evidence of disease progression and treatment effects that may not be available from other sources.	The complex, high detail and/or unstructured nature of imaging data may make uniform extraction and curation difficult and is subject to inconsistencies across health care systems and practices, particularly if machine learning is used.[20]

Appendix B | White Paper Methods and Broad Results

Methods

The Duke-Margolis team carried out a landscape analysis of relevant issues, policies, technologies, methods, and implications for care delivery related to source data collection, processing, and usage for evidence generation. This scoping research broadly characterized the current research landscape and identified key issues. Research was organized around four questions informed by preliminary scoping and stakeholder conversations:

- What gaps or pain points exist for collecting, documenting, and transforming data from different sources?
- What are the barriers to improving the quality of clinical data to support secondary analysis of aggregated data sources?
- What strategies can help raise the minimum quality of clinical data collected and improve data augmentation and curation quality to feed forward into learning health systems and point-of-care approaches?
- How feasible is it to build a holistic view of a patient's health by linking data sources?

To answer these research questions, we identified literature sources addressing the following topic areas: information on source data prevalence, authoritative guidance for working with source data, and real-world examples of source data application. The literature review focused on utilization of different data sources and barriers to working with them. Scoping research was sorted into three categories with the first outlining the general benefits and drawbacks of different source data, including how linked data can be complementary and mutually gap filling. Guidelines from regulatory, governmental, and relevant non-governmental organizations were the second category. These guidelines revolved around source data implementation, operationalization, and/or processing according to different regulatory standards. The final category consisted of real-world examples where organizations applied source data across different domains for research and/or care delivery. These real-world examples—or use cases—were often collaborations from industry, government, and academia.

Broad Use Case and Source Data Findings

We identified 41 use cases across four application areas. Use cases focused on clinical care (n= 12 use cases) involved RWD/E research on best practices so standards of care reflect current knowledge. Use cases with a *regulatory purpose or implication* (n= 12 use cases) used RWE in a regulatory submission (or other activity) to the FDA or other relevant health authority; or presented useful knowledge for regulators. Third, several use cases presented a *RWD/E research framework* (n= 10 use cases), which represents an organization's established method for organizing, processing, or otherwise treating source data to generate RWE. The last use case application, *RWD/E research tool* (n= 7 use cases), involved the specific software or process for organizing, processing, or otherwise treating source data used by a group. While use cases were placed into a single application area for organizational purposes, several may fit into one or more application areas.

Different source data were used more often than others—most common were EHR (n= 26 use cases) and registry data (n= 7 use cases). SDOH (n= 1 use case) and genomic (n= 0 use cases) data, while discussed in the broader literature, were not widely present in identified use cases. Four use cases did not specify target source data. The therapeutic area with the highest number of use cases was respiratory (n= 10 use cases), four of which were focused on infectious diseases. Cardiology (n= 5 use cases) and oncology (n= 5 use cases) were the next two most researched areas; eight use cases were not associated with a specific therapeutic area.

We delineated differences in how use cases publicly discussed their respective data quality strategies based on different stakeholder priorities. For example, use cases centered in academia more frequently focused on data linkage and validity demonstration, likely a consequence of an insulated health care system within a single university. Industry use cases, on the other hand, did not provide public-facing, in-depth descriptions of data quality methods. This may be due to business practices meant to protect innovation and proprietary information. These differences highlight challenges to align stakeholder priorities and underlying needs. Additionally, we found that larger and more established companies currently have a greater capacity to support linkage and interoperability. Smaller start-up organizations appear to face more infrastructure limitations.

Appendix C | High-Level Use Case Descriptions

Use Case	Organization	Source Data	Therapeutic Area
Clinical Care Application			
INFORM Study	AstraZeneca	EHR	Immunology
EPOCH Study	AstraZeneca	Claims	Immunology
CKD-EPI 2021 Equation	Truveta	EHR	Endocrinology
EPIC Cosmos	EPIC	EHR	N/A
Hemophilia A Unmet Needs	PicnicHealth	EHR, PRO	Hematology
Pragmatic Asthma Registry	Verantos and Amgen	Registry	Respiratory
Automated Identification for In-Hospital Clinical Deterioration	Kaiser Permanente, EPIC, and Unlearn.AI	EHR	N/A
Generative AI for Alzheimer's Drug Repurposing	Vanderbilt and Emory	EHR	Neurology
Identifying Elevated Risk for Fabry Disease with Machine Learning	OM1	EHR	Rare Disease
Apple Heart Study	Apple and Stanford University	PGHD	Cardiology
Apple Health Study	Apple and Brigham and Women's Hospital	PGHD	Cardiology, Neurology, and Respiratory
Obesity Healthcare Goals Programme	Eli Lilly and U.K. National Health Service (NHS)	Not Identified	Endocrinology
Regulatory Purpose or Implications			
Cantata by Protas, EASi-KIDNEY Trial	Protas	Not Identified	Nephrology
Blueprint Framework	AstraZeneca	N/A	N/A
Tofactinib Modified Release (MR) Trial	Pfizer	Claims, Registry	Immunology
RECOVERY Trial	UK	EHR, Claims	Respiratory (Infectious Disease)
ACTIV-6	NIH and Industry Partners	PRO	Respiratory (Infectious Disease)
REMAP COVID	UMC Utrecht	EHR	Respiratory (Infectious Disease)
VESALIUS-CV EHR Demonstration Project	Duke Clinical Research Institute (DCRI) and Amgen	EHR	Cardiology
ICAREdata Project	National Clinical Trials Network (NCTN)	EHR	Oncology
Pragmatica Lung	SWOG, Eli Lilly, NIH, and Merck	EHR	Oncology
Salford Lung Study	GSK	EHR	Respiratory
Salford Lung Extension Study	GSK and Ignite Data	EHR	Respiratory
REMAP CAP	UMC Utrecht	EHR	Respiratory (Infectious Disease)
RWD/E Research Framework			
DISCOVER CDK CDM	AstraZeneca	Registry	Endocrinology
PCSK9 Inhibitor	AstraZeneca and ZS Pharma	EHR	Cardiology
COPD and Asthma RWE	SingHealth, Duke, and GSK	EHR	Respiratory
Evidence Powered Operating Framework (EPOF)	Novartis	N/A	N/A
Veradigm Network EHR	Veradigm	EHR, SDOH, Mortality	N/A
GKPTN CAPTIVE	The George Institute	Registry	Endocrinology
UCSF OneSource	UCSF, FDA, and QLHC	EHR	N/A
Substance-exposed Birthing Person-Infant/Child HIE	HHS	EHR	Reproductive Health
Diuretic Comparison Project	VA	EHR, Claims, Mortality	Cardiology
SmartChart: FHIR-based Framework for Syphilis	CDC and Georgia Tech	EHR	Sexually Transmitted Infection
RWD/E Research Tool			
TransFAIR Study	AstraZeneca, Janssen, and Sanofi	EHR	N/A
IDC and EHR-to-EDC	Genentech and Flatiron	EHR	N/A
Parkinson's Disease Population-Wide Registries	EPIC and Verona	Registry	Neurology
mCODE Genomics Pilot Project	Vanderbilt and EPIC	EHR	Oncology
CancerX Data Spring and Cancer Moonshot	ONC, CMS, FDA, NCI, VA, and 30+ Industry Partners	EHR	Oncology
Harmonized Outcome Measures for Use in Non-Small Cell Lung Cancer Patient Registries and Clinical Practice	OM1, AstraZeneca, Flatiron, and 7 Industry Partners	Registry	Oncology
Harmonized Outcome Measures for Asthma Patient Registries and Clinical Practice	OM1 and 10+ Industry Partners	Registry	Respiratory

Appendix D | In-Depth Use Case Descriptions

Clinical Care Application: RWD/E research on clinical care best practices so standards of care reflect current knowledge. Use cases directly impact health care decisions or care delivery.

- **INFORM Study:** Used routine EHR data from primary and secondary care to compare COVID-19 hospitalizations, intensive care unit admissions, and deaths among immunocompromised and general populations. The use case did not include follow-up for interventions with the immunocompromised patient population.

- **EPOCH Study:** Investigators used the Healthcare Integrated Research Database (HIRD) to identify the prevalence and incidence rates of COVID-19 in immunocompromised patients and the resulting health care costs. HIRD is composed of longitudinal medical and pharmaceutical claims, along with social determinants of health data. The use case does not include any interventions to care delivery or change to health care costs.

- **CKD-EPI 2021 Equation:** Investigators used retrospective EHR data to study the effect of the 2021 CKD-EPI Creatinine Equation on the likelihood of patients with chronic kidney disease receiving a diagnosis of hyperkalemia, a potassium-lowering drug prescription, a doubling of serum creatinine, or an arteriovenous graft. While the EHR data revealed disparities in CKD diagnosis, the use case did not include long-term implications for clinical care. Truveta uses EHRs, demographic data, claims, prescriptions, and laboratory results.

- **EPIC Cosmos:** An opinion piece describing how EPIC merges EHR data from various health systems that use EPIC, allowing health care providers to exchange information on rare diseases. EPIC's interoperability network, Care Everywhere, uses standard ontologies, data linkage, and de-duplication. The use case is limited to EPIC systems.

- **Hemophilia A Unmet Needs:** Poster abstract outlining the use of EHR data and patient-reported outcomes to assess unmet needs in patients with hemophilia A. The use case is limited by a relatively small cohort size and does not directly connect study outcomes to changes in clinical care.

- **Pragmatic Asthma Registry:** A registry using RWD from people receiving treatment for severe asthma. AI is used to abstract unstructured data. The use case does not discuss using patient data to change clinical care.

- **Automated Identification for In-Hospital Clinical Deterioration:** Researchers used an automated predictive model to identify patients at high risk for clinical deterioration using Epic's EHR system. Patients tagged with an early-warning system alert from the automated predictive model had shorter hospital stays, a lower incidence of intensive care unit admission, and lower in-hospital mortality rates. Individual clinician responses to an early-warning system alert proved difficult to track, and clinician responses varied. The process-measure analyses did not reveal consistent, significant associations with the interventions, and varied clinician responses may be the true cause of health improvements.

- **Generative AI for Alzheimer's Drug Repurposing:** Sponsors used OpenAI's ChatGPT to generate promising drug repurposing cases to treat Alzheimer's. The AI displayed comprehension of drug repurposing and did not suggest drugs that already had FDA approval. Researchers observed the effects of the AI's three most suggested drugs—metformin, losartan, and minocycline—in large-scale EHRs from VUMC and the *All of Us* datasets that were standardized according to the OMOP CDM. There were modest decreases in the risk of AD among patients using the three drugs and findings suggest ChatGPT can generate quality hypotheses for drug repurposing. However, the high frequency of a drug candidate in ChatGPT queries does not guarantee the drug will lead to positive treatment effects. Incomplete EHRs also posed challenges.

- **Identifying Elevated Risk for Fabry Disease with Machine Learning:** ML predicted patients with a high risk of Fabry Disease (FD) using longitudinal, de-identified health record data. The AI was calibrated using extracted phenotypic patterns from EHR data of approximately 5,000 FD patients. ICD codes and prescription data also helped determine the prevalence of FD in patients.

- **Apple Heart Study:** Apple Watch data identified cardiac arrhythmias using wristwatch-based photoplethysmography. The Apple Heart Study app analyses pulse rate data collected from participants' Apple Watch to identify instances of irregular heart rhythms, including atrial fibrillation and other arrhythmias. Ambulatory electrocardiogram monitoring indicating an irregular rhythm consistent with an app notification and self-reported patient contact with a health care provider within 90 days of an irregular watch notification were secondary outcome measures.

- **Apple Health Study:** The longitudinal study will use participant data from Apple's Research app and other third-party devices to predict, monitor, and manage participants' health across various areas such as cardiovascular, neurologic, and respiratory health.

- **Obesity Healthcare Goals Programme:** A collaboration between NHS and Eli Lilly to improve long-term health outcomes for people living with obesity. The project considers how integrated care systems can support weight management services outside of hospitals.

Regulatory purpose or implications: use of RWE in a regulatory submission (or other activity) to the FDA or other relevant agency, or the use case has regulatory implications more broadly. Use cases directly impact and/or inform decision-making or rationale of regulatory authorities.

- **Cantata and EASi-KIDNEY [trade-marked] Trial:** The world's largest, global study of chronic kidney disease (CKD) intended to analyze whether vicadrostat slows the progression of CKD when it's taken in combination with empagliflozin. Cantata will act as an end-to-end trial management platform including participant enrollment, follow-up assessments, and quality assurance.

- **Blueprint Framework**: Creates a framework to facilitate global-local RWD evidence generation and a voluntary self-assessment tool, including implementation of local Value Teams to improve data collection and data governance. The framework has not been integrated into routine clinical care or the reimbursement process.

- **Tofacitinib Modified Release (MR) Trial**: Linked IBM MarketScan insurance claims data and Corrona Registry (now CorEvitas) for regulatory label expansion.

- **RECOVERY:** Demonstrates a quality by design, point-of-care trial for rapid enrollment of patients with COVID-19. The use case reveals challenges with implementing large, point-of-care trials in the U.S. health care system.

- **ACTIV-6:** Describes a decentralized trial in outpatient settings for people with a positive PCR or antigen test for SARS-CoV-2. ACTIV-6 delivers daily accrual and data quality reports to a clinical coordinating center, identifies missing and inconsistent data, and indicates where supplemental health care data may be needed for verification. Local regulations, institutional review board oversight, and the regulatory landscape presented challenges for multisite research.

- **REMAP COVID:** An international, point-of-care trial to determine the best treatment for intensive care unit patients with community-acquire pneumonia. Investigators found challenges with data reporting.

- **VESALIUS-CV EHR Demonstration Project:** An RWD/E FDA demonstration project to understand operational considerations for EHR-sourced multi-center trial organization. Data was formatted to align with the PCORnet data code and common data model. Limited by site-specific adaptation of broad data policies and operational differences that impacted trial efficiency and standardized data collection.

- **ICAREdata Project:** An RWD/E FDA demonstration project focused on streamlining clinical trial data acquisition from EHRs and eliminating data redundancies by developing new structured EHR fields. Encountered difficulties in clinician uptake due to deviation from routine clinical workflow.

- **Pragmatica Lung:** A phase 3 pragmatic clinical trial using relaxed enrollment eligibility requirements for a new two-drug combination treatment for stage 4 lung cancer. The trial may have larger implications for reducing patient burden and enrollment time for trials using FDA-approved drugs with well-known safety profiles.

- **Salford Lung Study:** Compares the utility of pragmatic RCTs with traditional RCTs for drug repurposing, using approved drugs with clear safety profiles. Best trial performance occurs when relying on a single EHR and pharmacy system within a single community-setting; quantifiable cost effectiveness is not demonstrated.

- **Salford Lung Extension Study:** Proof-of-concept study with retrospectively and prospectively routinely collected health care data with minimal patient and practitioner burden. IgniteData and Graphnet Health extracted data from EHRs and patient questionnaires into a bespoke database. Limitations include maintaining participant continuity over time, impact of legislation on required study documentation, and participant re-identification.

- **REMAP CAP:** Reports on an international randomized point-of-care trial using a randomized embedded multifactorial adaptive platform trial design. A case report form was created from a data dictionary and a single data safety and monitoring board reviewed trial implementation across sites. Limitations around data reporting and clinician-research team relationships across therapeutic areas.

RWD/E research framework: an organization's established method or guideline for organizing RWD or generating RWE. Use cases explain why/how an organization is organizing, processing, or otherwise treating its source data in a particular way.

- **DISCOVER CDK CDM:** Describes the development, validation, and benefits of a study-specific CDM applied to the DISCOVER CDK cohort. Use case required significant up-front labor for data standardization and permissions from several specialists, potentially limiting its scalability.

- **PCSK9 Inhibitor:** Developed an EHR framework to inform clinical trial design. Framework is based on the TriNetX platform, potentially limiting its generalizability.

- **COPD and Asthma RWE:** Designed a research-ready RWD registry to facilitate research and reduce interaction with EHR source data. Patient tracking was only possible at certain points within the hospital system; captured free-text data required additional cleaning.

- **Evidence Powered Operating Framework (EPOF):** Presents a framework for integrating RWE generation strategies into existing pharmaceutical operations to minimize disruptions. States quantifying cost savings requires further research efficiency and productivity. This framework was limited to the TriNetX system.

- **Veradigm Network EHR (VNEHR):** Reports the addition of SDOH and ambulatory EHR mortality data to Veradigm's existing EHR research network. SDOH data is derived from natural language processing, raising questions about validity demonstrations and FDA's fitness-for-use perspective.

- **GKPTN CAPTIVE:** Describes the international GKPTN registry and the modular protocol design of the associated Phase III CAPTIVATE platform trial. Registry relies on local laboratory testing, follows the Recommendations for Interventional Trials (SPIRIT), and adheres to the ICH Guidelines for Good Clinical Practice.

- **UCSF OneSource:** Uses an EDC system to integrate several EHR platforms. Patient follow-up included questionnaires that used Common Terminology Criteria for Adverse Events and HealthMeasures' PROMIS. The system is used in the multi-arm I-SPY COVID TRIAL testing various treatments to improve outcomes for severely ill COVID-19 patients.

- **Substance-Exposed Birthing Person-Infant/Child HIE**: Explored data-sharing capabilities through health information exchange standards to connect data on birthing persons with substance abuse to their infants. The researchers tested, and ultimately recommended, the use of USCDI v4 to connect longitudinal health data. HHS recommended stronger data linkage between EHRs for the birthing person and the infant.

- **Diuretic Comparison Project (DCP)**: Through the DCP, Veterans Affairs (VA) combined Veterans Health Administration EHRs (primary), Medicare claims (secondary), and National Death Index (tertiary verification) data to study non-cancer deaths and major cardiovascular outcomes. Authors present 6 key recommendations: 1) prioritize data sources through a data reliability/trustworthiness hierarchy, 2) develop an analysis plan for specifying event orders with high-validity outcomes, 3) plan for how to manage data changes over time (e.g., data surveillance), 4) determine when final data capture will occur based on validity and data completeness (e.g., timeliness), 5) appropriately select well validated claims-based algorithms, and 6) incorporate novel tools/algorithms (particularly for unstructured data).

- **SmartChart: FHIR-Based Framework for Syphilis:** Overviews the creation of SmartChart, a framework used to support surveillance of syphilis through EHR data. SmartChart retrieves EHR data, integrates the data into a standardized data storage system, and helps diagnosis patients. SmartChart was tested in Atlanta's Grady Health System in 2023. Investigators highlighted the need for TEFCA to support and simplify alignment of responsibilities between health system IT departments and care providers. The researchers used Epic FHIR and noted discrepancies between the Epic FHIR API and standard FHIR protocols.

RWD/E research tool: the specific application and/or process for organizing, processing, or otherwise treating source data employed by an organization.

- **TransFAIR Study:** Developed and carried out a proof-of-concept study with the EHR2EDC tool, which abstracts unstructured EHR data into EDC systems. Investigators used an expanded CONSORT-ROUTINE checklist to ensure transparency. The use case illustrates the need for expanding capabilities across local investigative teams and site readiness best practices.

- **IDC and EHR-to-EDC:** Describes the intentional capture of source data (IDC) method for transferring EHR data into an EDC system during a Phase II open-label trial to reduce data redundancies. EHR variables were mapped to the electronic case report form (eCRF), including free text and standard terminologies (e.g., CDISC). Use case does not include a performance comparison to traditional EDC approaches.

- **Parkinson's Disease Population-Wide Registries:** California's Parkinson's registry and the UCE-PD project funded by the Michael J. Fox Foundation for Parkinson's Research evaluated the registry's completeness and accuracy. Challenges include ICD diagnosis codes not translating to the same diagnosis because PD lacks definitive biomarkers. Additionally, no unified agreement exists on data standards for population-wide PD registries. UCE-PD worked with Epic and Verona to create common PD registry data elements. The case is limited to Parkinson's and may not be generalizable to other diseases and conditions.

- **mCODE Genomics Pilot Project:** The Vanderbilt University Medical Center (VUMC) converted its EHR data into Minimal Common Oncology Data Elements (mCODE) compliant profiles and created a web application that offers cancer risk assessments, demonstrating successful use of mCODE standards throughout a health care system. Barriers to mCODE include insufficient or non-computable genomic EHR data. Challenges included incongruities between mCODE and EHR data standards. Researchers also found FHIR platforms better equipped for data aggregation at the population level rather than the individual patient level.

- **CancerX Data Sprint and Cancer Moonshot:** CancerX, a public-private partnership, initiated the CancerX Data Sprint to collect information from agencies and industry partners to advance cancer-related RWD data standards. Focus was placed on improving CMMI's Enhancing Oncology Model and the USCDI+ oncology extension. Findings from the CancerX Data Sprint Summary include 15 new proposed data elements. Many of these elements include SDOH, drug names, genetic or biomarker testing results, and surgical procedures and outcomes. Current CancerX members called for expanding the scope of data measurements and prioritizing a more holistic data approach, including disease and treatment progression data.

- **Harmonized Outcome Measures for Use in Non-Small Cell Lung Cancer Patient Registries and Clinical Practice:** Developed a minimum set of patient and clinician relevant harmonized outcome measurements for non-small cell lung cancer using 11 lung cancer patient registries and the Agency for Healthcare Research and Quality's Outcome Measurements Framework (OMF).

- **Harmonized Outcome Measures for Asthma Patient Registries and Clinical Practice:** Created a minimum set of patient- and provider-relevant standardized outcome measurements using OMF and 13 asthma registries. Outcome measurements used standardized terminologies, including ICD and SNOMED-CT.

References

[1] Franklin JB, Marra C, Abebe KZ, et al. Modernizing the Data Infrastructure for Clinical Research to Meet Evolving Demands for Evidence. JAMA. 2024;332(16):1378-1385. 10.1001/jama.2024.0268

[1] Hariton E, Locascio JJ. Randomised controlled trials – the gold standard for effectiveness research. *BJOG Int J Obstet Gynaecol.* 2018;125(13):1716-1716. doi:10.1111/1471-0528.15199

[3] Smith JP. Basics of Clinical Trial Design. Presented at: 2024; U.S. Food and Drug Administration. https://www.fda.gov/media/185118/download

[4] Speich B, von Niederhäusern B, Schur N, et al. Systematic review on costs and resource use of randomized clinical trials shows a lack of transparent and comprehensive data. *J Clin Epidemiol.* 2018;96:1-11. doi:10.1016/j.jclinepi.2017.12.018

[5] Pearce W, Raman S, Turner A. Randomised trials in context: practical problems and social aspects of evidence-based medicine and policy. *Trials.* 2015;16(1):394. doi:10.1186/s13063-015-0917-5

[6] Kass NE, Meeker-O'Connell A, Morain SR, Crane MA. Optimizing Informed Consent—A Call to Action. *JAMA Health Forum.* 2025;6(5):e250824. doi:10.1001/jamahealthforum.2025.0824

[7] Food and Drug Administration. Federal Food, Drug, and Cosmetic Act: 21 USC 355g: Utilizing real world evidence. August 18, 2017. Accessed April 30, 2025. https://uscode.house.gov/view.xhtml?req=355g%28b%29&f=treesort&fq=true&num=2&hl=true&edition=prelim&granuleId=USC-prelim-title21-section355g

[8] U.S. Food and Drug Administration. *Framework for FDA's Real-World Evidence Program.* U.S. Food and Drug Administration; 2018:2-40. https://www.fda.gov/media/120060/download

[9] Center for Drug Evaluation and Research, Center for Biologics Evaluation and Research, Oncology Center for Excellence. *Real-world Data: Assessing Electronic Health Records and Medical Claims Data to Support Regulatory Decision□making for Drug and Biological Products.* U.S. Food and Drug Administration; 2022:717-720. Accessed March 20, 2025. https://onlinelibrary.wiley.com/doi/10.1002/pds.5444

[10] Liu F, Panagiotakos D. Real-world data: a brief review of the methods, applications, challenges and opportunities. *BMC Med Res Methodol.* 2022;22(1):287. doi:10.1186/s12874-022-01768-6

[11] Terry S, Runkle D, Wicks P. Challenges and Limitations. In: *Patient- or Participant-Generated Registries: Addendum to Registries for Evaluating Patient Outcomes: A User's Guide, Third Edition [Internet].* Agency for Healthcare Research and Quality (US); 2018. Accessed March 20, 2025. https://www.ncbi.nlm.nih.gov/books/NBK493593/

[12] Maruszczyk K, Aiyegbusi OL, Torlinska B, Collis P, Keeley T, Calvert MJ. Systematic review of guidance for the collection and use of patient-reported outcomes in real-world evidence generation to support regulation, reimbursement and health policy. *J Patient-Rep Outcomes.* 2022;6(1):57. doi:10.1186/s41687-022-00466-7

[13] Emmott N, Nafie M, Shaw N, Hendricks-Sturrup R. *Regulatory Fit-for-Purpose Considerations for Patient-Generated Health Data.* Duke-Margolis Institute for Health Policy; 2024:1-17. Accessed March 20, 2025. https://healthpolicy.duke.edu/publications/regulatory-fit-purpose-considerations-patient-generated-health-data

[14] Centers for Disease Control and Prevention. Frequently Asked Questions | Social Determinants of Health | NCHHSTP | CDC. August 21, 2017. Accessed March 20, 2025. http://medbox.iiab.me/modules/en-cdc/www.cdc.gov/nchhstp/socialdeterminants/faq.html

[15] Cantor MN, Thorpe L. Integrating Data On Social Determinants Of Health Into Electronic Health Records. *Health Aff (Millwood).* 2018;37(4):585-590. doi:10.1377/hlthaff.2017.1252

[16] Frank J, Abel T, Campostrini S, Cook S, Lin VK, McQueen DV. The Social Determinants of Health: Time to Re-Think? *Int J Environ Res Public Health.* 2020;17(16):5856. doi:10.3390/ijerph17165856

[17] National Human Genome Research Institute. Genetics vs. Genomics Fact Sheet. September 7, 2018. Accessed March 20, 2025. https://www.genome.gov/about-genomics/fact-sheets/Genetics-vs-Genomics

[18] All of Us Research Program. All of Us Research Program: Home. 2025. Accessed March 20, 2025. https://allofus.nih.gov/

[19] National Center for Advancing Translational Sciences. Imaging data. Toolkit For Patient-Focused Therapy Development. Accessed April 30, 2025. https://toolkit.ncats.nih.gov/glossary/imaging-data

[20] Yang Y, Zhang H, Gichoya JW, Katabi D, Ghassemi M. The limits of fair medical imaging AI in real-world generalization. *Nat Med.* 2024;30(10):2838-2848. doi:10.1038/s41591-024-03113-4

[21] Goniewicz K, Carlström E, Hertelendy AJ, et al. Integrated Healthcare and the Dilemma of Public Health Emergencies. *Sustainability.* 2021;13(8):4517. doi:10.3390/su13084517

[22] U.S. Department of Health and Human Services. Robert F. Kennedy, Jr. Sworn in as 26th Secretary at HHS, President Trump Signs Executive Order to Make America Healthy Again. February 13, 2025. Accessed March 20, 2025. https://www.hhs.gov/about/news/2025/02/13/robert-kennedy-jr-sworn-26th-secretary-hhs-president-trump-signs-executive-order-make-america-healthy-again.html

[23] Food and Drug Administration. Exploration of Health Level Seven Fast Healthcare Interoperability Resources for Use in Study Data Created From Real-World Data Sources for Submission to the Food and Drug Administration; Establishment of a Public Docket; Request for Comments. Federal Register. April 23, 2025. Accessed April 30, 2025. https://www.federalregister.gov/documents/2025/04/23/2025-06967/exploration-of-health-level-seven-fast-healthcare-interoperability-resources-for-use-in-study-data

[24] Center for Medicare and Medicaid Services, Assistant Secretary for Technology Policy, Department of Health and Human Services. *Request for Information; Health Technology Ecosystem.;* 2025. Accessed May 28, 2025. https://www.federalregister.gov/documents/2025/05/16/2025-08701/request-for-information-health-technology-ecosystem

[25] Propes C, Sheehan S, Hendricks-Sturrup R. *Point-of-Care Clinical Trials: Integrating Research and Care Delivery.* Duke-Margolis Institute for Health Policy; 2022:1-32. Accessed March 20, 2025. https://healthpolicy.duke.edu/publications/point-care-clinical-trials-integrating-research-and-care-delivery

[26] Califf RM, Cavazzoni P, Woodcock J. Benefits of Streamlined Point-of-Care Trial Designs: Lessons Learned From the UK RECOVERY Study. *JAMA Intern Med.* 2022;182(12):1243. doi:10.1001/jamainternmed.2022.4810

[27] Draghia-Akli R, Read SW, Hughes EA. Overview of ACTIV trial-specific lessons learned. *J Clin Transl Sci.* 2024;8(1):e149. doi:10.1017/cts.2023.698

[28] Berger M, Daniel G, Frank K, et al. *A Framework for Regulatory Use of Real-World Evidence.* Duke-Margolis Institute for Health Policy; 2017:1-19. https://healthpolicy.duke.edu/publications/framework-regulatory-use-real-world-evidence

[29] Foley T, Vale L. A framework for understanding, designing, developing and evaluating learning health systems. *Learn Health Syst.* 2023;7(1):e10315. doi:10.1002/lrh2.10315

[30] Locke T, Yankah S, D'Ambrosio M, Hendricks-Sturrup R. *Lessons from Learning Health Care Systems and Recommendations for Successful Implementation.* Duke-Margolis Institute for Health Policy; 2023:1-33. Accessed March 20, 2025. https://healthpolicy.duke.edu/publications/lessons-learning-health-care-systems-and-recommendations-successful-implementation

[31] Lam SSW, Fang AHS, Koh MS, et al. Development of a real-world database for asthma and COPD: The SingHealth-Duke-NUS-GSK COPD and Asthma Real-World Evidence (SDG-CARE) collaboration. *BMC Med Inform Decis Mak.* 2023;23:4. doi:10.1186/s12911-022-02071-6

[32] Gliklich RE, Castro M, Leavy MB, et al. Harmonized outcome measures for use in asthma patient registries and clinical practice. *J Allergy Clin Immunol.* 2019;144(3):671-681.e1. doi:10.1016/j.jaci.2019.02.025

[33] Centers for Disease Control and Prevention. ICD-10. Classification of Diseases, Functioning, and Disability. June 11, 2024. Accessed March 20, 2025. https://www.cdc.gov/nchs/icd/icd-10/index.html

[34] SNOWMED International. What is SNOMED CT. SNOMED International. Accessed March 20, 2025. https://www.snomed.org/what-is-snomed-ct

[35] Center for Drug Evaluation and Research, Center for Biologics Evaluation and Research. *Data Standards for Drug and Biological Product Submissions Containing Real-World Data.* U.S. Food and Drug Administration; 2025:1-13. Accessed March 20, 2025. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/data-standards-drug-and-biological-product-submissions-containing-real-world-data

[36] PHUSE. Submitting RWD: Challenges Faced by Industry While Working on Real-World Evidence Studies. Accessed April 2, 2025. https://phuse-1vge8ug4p-phuse.vercel.app

[37] Facile R, Muhlbradt EE, Gong M, et al. Use of Clinical Data Interchange Standards Consortium (CDISC) Standards for Real-world Data: Expert Perspectives From a Qualitative Delphi Survey. *JMIR Med Inform.* 2022;10(1):e30363. doi:10.2196/30363

[38] Baker RL, Hamidi M, McKenzie L, Denney CK, Edgar TD. The Story Behind the HL7 FHIR to CDISC Mapping Implementation Guide. *J Soc Clin Data Manag.* 2023;4(1). doi:10.47912/jscdm.162

[39] CMS Moves Closer to Accountable Care Goals with 2025 ACO Initiatives | CMS. Accessed June 6, 2025. https://www.cms.gov/newsroom/fact-sheets/cms-moves-closer-accountable-care-goals-2025-aco-initiatives

[40] Sutton A, Center for Medicare and Medicaid Innovation. Making America Healthy Again: Innovation for Healthier Lives | CMS. CMS Newsroom. May 13, 2025. Accessed May 28, 2025. https://www.cms.gov/newsroom/blog/making-america-healthy-again-innovation-healthier-lives

[41] Saunders RS, Bonesteel R, McStay F, McClellan MB. Future Of Accountable Care: Lessons Learned And Potential Paths Forward During A Time Of Transition. doi:10.1377/forefront.20250130.539509

[42] Center for Medicare and Medicaid Services. Innovation Center Strategy Refresh: Driving Health System Transformation – A Strategy for CMS Innovation Center's Second Decade. Published online October 20, 2021:3-32.

[43] National Institutes of Health. Data Curation. NNLM. Accessed April 30, 2025. https://www.nnlm.gov/guides/data-glossary/data-curation

[44] Rubinger L, Gazendam A, Ekhtiari S, Bhandari M. Machine learning and artificial intelligence in research and healthcare. *Injury.* 2023;54:S69-S73. doi:10.1016/j.injury.2022.01.046

[45] Tierney AA, Gayre G, Hoberman B, et al. Ambient Artificial Intelligence Scribes to Alleviate the Burden of Clinical Documentation. NEJM Catal. 2024;5(3):CAT.23.0404. doi:10.1056/CAT.23.0404

[46] National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework (AI RMF 1.0). National Institute of Standards and Technology (U.S.); 2023:NIST AI 100-1. doi:10.6028/NIST.AI.100-1

[47] Center for Devices and Radiological Health. Artificial Intelligence and Machine Learning in Software as a Medical Device. January 6, 2025. Accessed March 20, 2025. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device

[48] Center for Drug Evaluation and Research, Center for Biologics Evaluation and Research, Center for Devices and Radiological Health, et al. *Considerations for the Use of Artificial Intelligence To Support Regulatory Decision-Making for Drug and Biological Products.* U.S. Food and Drug Administration; 2025. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/considerations-use-artificial-intelligence-support-regulatory-decision-making-drug-and-biological

[49] D'Ambrosio M, Rodriguez P, Emmott N, Hendricks-Sturrup R. *Operational Tools and Best Practices to Support Electronic Health Record-Sourced Data Quality, Relevance, and Reliability at the Data Accrual Phase.* Duke-Margolis Institute for Health Policy; 2025:1-20. Accessed March 20, 2025. https://healthpolicy.duke.edu/publications/operational-tools-and-best-practices-support-electronic-health-record-sourced-data

[50] Parker VJ, Economou NJ, Silcox C. *AI Governance in Health Systems: Aligning Innovation, Accountability, and Trust.* Duke-Margolis Institute for Health Policy; 2024:e10315. Accessed March 20, 2025. https://healthpolicy.duke.edu/publications/ai-governance-health-systems-aligning-innovation-accountability-and-trust

[51] Girman C, Mack C, Leahy R, et al. Proposed Contents of a Real-World Data Quality and Relevance Package for Regulatory Submissions. *J Regul Sci.* 2025;13(1). doi:10.21423/JRS.REGSCI.131302

[52] Penev YP, Buchanan TR, Ruppert MM, et al. Electronic Health Record Data Quality and Performance Assessments: Scoping Review. *JMIR Med Inform.* 2024;12(1):e58130. doi:10.2196/58130

[53] Centers for Medicare & Medicaid Services. *Study Protocols That Use Real-World Data.* Centers for Medicare & Medicaid Services; 2025. https://www.cms.gov/medicare-coverage-database/view/medicare-coverage-document.aspx?mcdid=39&docTypeId=-1&status=Open%20for%20Public%20Comment&sortBy=title&bc=-1

[54] Center for Devices and Radiological Health. *Use of Real-World Evidence to Support Regulatory Decision-Making for Medical Devices.* U.S. Food and Drug Administration; 2023:1-24. Accessed March 20, 2025. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/use-real-world-evidence-support-regulatory-decision-making-medical-devices

[55] Center for Drug Evaluation and Research. RWD and RWE-focused Demonstration Projects. FDA. November 16, 2023. Accessed March 20, 2025. https://www.fda.gov/science-research/real-world-evidence/rwd-and-rwe-focused-demonstration-projects

[56] Riskin DJ, Monda KL, Gagne JJ, et al. Implementing Accuracy, Completeness, and Traceability for Data Reliability. *JAMA Netw Open.* 2025;8(3):e250128. doi:10.1001/jamanetworkopen.2025.0128

[57] Center for Medicare and Medicaid Services. *Transcatheter Tricuspid Valve Replacement (TTVR).* Centers for Medicare & Medicaid Services https://www.cms.gov/medicare-coverage-database/view/ncacal-decision-memo.aspx?proposed=N&ncaid=314

[58] Center for Medicare and Medicaid Services. Cell and Gene Therapy (CGT) Access Model | CMS. CMS.gov. Accessed May 28, 2025. https://www.cms.gov/priorities/innovation/innovation-models/cgt

[59] Centers for Medicare & Medicaid Services. Enhancing Oncology Model | CMS. CMS.gov. Accessed March 31, 2025. https://www.cms.gov/priorities/innovation/innovation-models/enhancing-oncology-model

[60] Health C for D and R. Clinical Decision Support Software Frequently Asked Questions (FAQs). *FDA.* Published online January 8, 2025. Accessed June 6, 2025. https://www.fda.gov/medical-devices/software-medical-device-samd/clinical-decision-support-software-frequently-asked-questions-faqs

[61] Verantos. Asthma Pragmatic Registry Data Sheet. Respiratory Pragmatic Registry. 2024. Accessed March 20, 2025. https://go.verantos.com/rs/077-MLC-183/images/Asthma%20Pragmatic%20Registry%20Data%20Sheet.pdf?version=1

[62] Hau C, Woods PA, Guski AS, et al. Strategies for secondary use of real-world clinical and administrative data for outcome ascertainment in pragmatic clinical trials. *J Biomed Inform.* 2024;150:104587. doi:10.1016/j.jbi.2024.104587

[63] 21st Century Cures Act: Interoperability, Information Blocking, and the ONC Health IT Certification Program.

[64] Assistant Secretary of Technology Policy. TEFCA | HealthIT.gov. Accessed March 20, 2025. https://www.healthit.gov/topic/interoperability/policy/trusted-exchange-framework-and-common-agreement-tefca

[65] Institute of Medicine (US) Committee on Data Standards for Patient Safety. Health Care Data Standards. In: Aspden P, Corrigan JM, Wolcott J, Erickson SM, eds. *Patient Safety: Achieving a New Standard for Care.* National Academies Press (US); 2004. Accessed March 20, 2025. https://www.ncbi.nlm.nih.gov/books/NBK216088/

[66] Sheehan J, Hirschfeld S, Foster E, et al. Improving the value of clinical research through the use of Common Data Elements. *Clin Trials.* 2016;13(6):671-676. doi:10.1177/1740774516653238

[67] Ward SL, Flori HR, Bennett TD, et al. Design and Rationale for Common Data Elements for Clinical Research in Pediatric Critical Care Medicine. *Pediatr Crit Care Med.* 2020;21(11):e1038. doi:10.1097/PCC.0000000000002455

[68] National Institutes of Health. NIH Common Data Elements (CDE) Repository. Accessed March 20, 2025. https://cde.nlm.nih.gov/home

[69] Assistant Secretary for Technology Policy. United States Core Data for Interoperability (USCDI). March 5, 2025. Accessed March 20, 2025. http://www.healthit.gov/isp/united-states-core-data-interoperability-uscdi

[70] Assistant Secretary for Technology Policy. USCDI+ | HealthIT.gov. Accessed March 20, 2025. https://www.healthit.gov/topic/interoperability/uscdi-plus

[71] USCDI+ - USCDI+ Service Portal. Accessed April 1, 2025. https://uscdiplus.healthit.gov/uscdiplus

[72] National Committee for Quality Assurance. Introduction to USCDI. NCQA. Accessed March 20, 2025. https://www.ncqa.org/resources/introduction-to-uscdi/

[73] The National Patient-Centered Clinical Research Network. PCORnet: Home. The National Patient-Centered Clinical Research Network. Accessed March 20, 2025. https://pcornet.org/

[74] Observational Health Data Science and Informatics. Data Standardization – OHDSI. Accessed March 20, 2025. https://www.ohdsi.org/data-standardization/

[75] Kent S, Burn E, Dawoud D, et al. Common Problems, Common Data Model Solutions: Evidence Generation for Health Technology Assessment. *PharmacoEconomics.* 2021;39(3):275-285. doi:10.1007/s40273-020-00981-9

[76] The Sequoia Project. *Standard Operating Procedure (SOP): Exchange Purposes (XPs).* Assistant Secretary for Technology Policy; 2024:1-7.

[77] Sequoia Project. How Can I Participate in TEFCA? ONC TEFCA Recognized Coordinating Entity. Accessed May 28, 2025. https://rce.sequoiaproject.org/participate/

[78] Rocca M, Asare A, Esserman L, Dubman S, Gordon G. *Source Data Capture from EHRs: Using Standardized Clinical Research Data.* U.S. Food and Drug Administration; 2019:1-50.

[79] Electronic Clinical Quality Improvement Resource Center. FHIR® - Fast Healthcare Interoperability Resources® | eCQI Resource Center. March 11, 2025. Accessed March 20, 2025. https://ecqi.healthit.gov/fhir?qt-tabs_fhir=about

[80] Li Y, Ye J, Huang Y, et al. Minimal Common Oncology Data Elements Genomics Pilot Project: Enhancing Oncology Research Through Electronic Health Record Interoperability at Vanderbilt University Medical Center. *JCO Clin Cancer Inform.* 2024;(8):e2300249. doi:10.1200/CCI.23.00249

[81] Patel S, Goldsack JC, Mirsky DE, Konya S, Wood WA. CancerX Data Sprint: Advancing national oncology data standards through a unique public-private partnership. J Clin Oncol. 2024;42(16_suppl):e23283-e23283. doi:10.1200/JCO.2024.42.16_suppl.e23283

[82] Health Level Seven International. HL7 FHIR ACCELERATOR™ Program | HL7 International. Accessed March 20, 2025. https://www.hl7.org/about/fhir-accelerator/

[83] Duda SN, Kennedy N, Conway D, et al. HL7 FHIR-based tools and initiatives to support clinical research: a scoping review. *J Am Med Inform Assoc.* 2022;29(9):1642-1653. doi:10.1093/jamia/ocac105

[84] Health Level Seven International. Da Vinci Project | HL7 International. Accessed March 20, 2025. https://www.hl7.org/about/davinci/

[85] Health Level Seven International. Gravity Project | HL7 International. Accessed March 20, 2025. https://www.hl7.org/gravity/

[86] George S, Campbell N, Hillman SL, et al. Feasibility of structuring electronic health record data to facilitate real-world data research: ICAREdata methods applied to multicenter cancer clinical trials. *Cancer.* 2025;131(1):e35528. doi:10.1002/cncr.35528

[87] Health Level Seven International. CodeX | HL7 International. Accessed March 20, 2025. https://www.hl7.org/codex/

[88] Yu Y, Zong N, Wen A, et al. Developing an ETL tool for converting the PCORnet CDM into the OMOP CDM to facilitate the COVID-19 data integration. *J Biomed Inform.* 2022;127:104002. doi:10.1016/j.jbi.2022.104002

[89] Assistant Secretary for Technology Policy. HHS Health IT Alignment Program | HealthIT.gov. September 30, 2025. Accessed April 30, 2025. https://www.healthit.gov/topic/hhs-health-it-alignment-program

[90] Explore Measures & Activities - QPP. Accessed June 6, 2025. https://qpp.cms.gov/mips/explore-measures?tab=advancingCareInformation&py=2025

[91] The American Health Information Management Association (AHIMA). AHIMA Home. The American Health Information Management Association (AHIMA). Accessed March 20, 2025. https://www.ahima.org/

[92] American Hospital Association. American Hospital Association Homepage | Hospitals USA | AHA. March 19, 2025. Accessed March 20, 2025. https://www.aha.org/

[93] Protas. Protas launches novel unified clinical trial management platform. Protas. January 27, 2025. Accessed March 20, 2025. https://protas.co.uk/protas-launches-novel-unified-clinical-trial-management-platform-cantata/

[94] Oxford Population Health. EASi-KIDNEY. Accessed March 20, 2025. https://www.ctsu.ox.ac.uk/research/easi-kidney

[95] Palanca A, Ampudia-Blasco FJ, Calderón JM, et al. Real-World Evaluation of GLP-1 Receptor Agonist Therapy Persistence, Adherence and Therapeutic Inertia Among Obese Adults with Type 2 Diabetes. *Diabetes Ther.* 2023;14(4):723-736. doi:10.1007/s13300-023-01382-9

[96] Centers for Medicare & Medicaid Services, Center for Medicare and Medicaid Innovation. EOM Clinical Data Elements Guide V2.2. Published online November 22, 2024. Accessed April 30, 2025. https://www.cms.gov/priorities/innovation/media/document/eom-clinical-data-elements-guide

[97] Khan BB, Graunke H, McClellan M. *Margolis Tricuspid Valve Intervention.* Duke-Margolis Institute for Health Policy; 2024:1-21. Accessed March 31, 2025. https://healthpolicy.duke.edu/sites/default/files/2024-02/Margolis%20Tricuspid%20Valve%20Intervention.pdf

[98] Centers for Medicare and Medicaid Services. Aligning Quality Measures Across CMS - the Universal Foundation | CMS. February 26, 2025. Accessed April 30, 2025. https://www.cms.gov/medicare/quality/cms-national-quality-strategy/aligning-quality-measures-across-cms-universal-foundation

[99] Centers for Medicare and Medicaid Services. Meaningful Measures 2.0: Moving from Measure Reduction to Modernization | CMS. April 23, 2025. Accessed April 30, 2025. https://www.cms.gov/medicare/quality/cms-national-quality-strategy/meaningful-measures-20-moving-measure-reduction-modernization