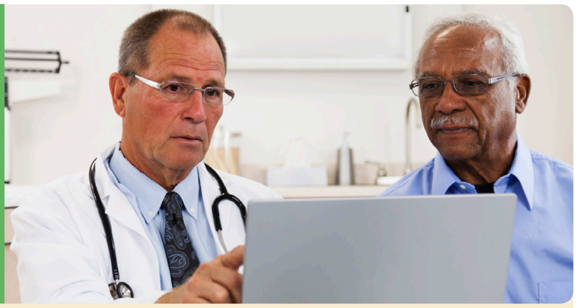
AcademyHealth

Finding Value in Volume: An Exploration of Data Access and Quality Challenges



The Health IT for Actionable Knowledge project examines the experiences of six large health care systems that have used data from electronic health records and other information technology to conduct research and analysis related to health care delivery. This document is one of five reporting the results of this AcademyHealth initiative. Each report draws on examples from these early-adopting health systems to explore a range of issues relevant to the conduct of health services and other research using electronic clinical data. The six health system partners in this effort are Denver Health, Geisinger Health System, Kaiser Permanente, the New York City Department of Health and Mental Hygiene's Primary Care Information Project, the Palo Alto Medical Foundation Research Institute, and the Veterans Health Administration. AcademyHealth gratefully acknowledges the generous support of the California HealthCare Foundation in funding this project, and the U.S. Agency for Healthcare Research and Quality (AHRQ) for providing seed funding.

Summary

Aware of the potential benefits of health information technology (health IT), many early adopting health systems have leveraged electronic health records (EHRs) for far more than point-of-care documentation. Many rely on these electronic data to conduct quality assurance, quality improvement (QI) and reporting; surveillance; operations research; and clinical and health services research. However, as these health systems have generated more electronic data—and from an increasingly diverse array of IT systems and sources—they have encountered challenges associated with accurately recording, reconciling, and contextualizing these data in a way that supports a multitude of potential uses. Through AcademyHealth's Health IT for Actionable Knowledge project, six participating health system partners—Denver Health, Geisinger Health System, Kaiser Permanente, the New York City Department of Health's Primary Care Information Project (PCIP), the Palo Alto Medical Foundation Research Institute (PAMFRI), and the Veterans Health Administration (VHA)—raised a number of issues and challenges associated with using health IT to support other critical functions, focusing primarily on quality improvement and health services research. This report describes some of the key challenges to capturing and then extracting "research grade" data from health IT systems, and elaborates on the experiences of these six health system partners that, recognizing the value of using EHR data to support research, have devised approaches to mitigate these challenges. Their collective experience offers lessons from which others can learn.

Introduction

The proliferation of EHRs, which has been significantly fortified by recent federal investments in health IT, presents an unprecedented opportunity to collect more—and more useful—information in the service of individual patients, and the health care sector as a whole. For some providers, the impetus for EHR adoption stems from the desire to modernize documentation, improve clinical care, and more effectively manage their patients and practices. The presence of financial incentives (and the imminence of disincentives) established in the Health Information Technology for Economic and Clinical Health (HITECH) Act, likely has had an effect as well. Though these are all important and sufficient reasons for adoption, the focus has turned to other potential "downstream" uses of the data captured through these systems and their potential to help to transform care at the institution level, as well as inform broader national improvement and efficiency efforts. National policymakers in particular have called for the creation of "a learning health care system," which essentially relies upon the use of information initially collected via EHRs as part of the clinical care process to improve health care quality and effectiveness, and constrain the growth of costs.[1]

Many early adopting health systems, already sold on the potential benefits of health IT, are modeling the ideals of a learning health system on a smaller scale. These systems have leveraged EHRs for far more than point-of-care documentation. Many rely on these electronic data to conduct quality assurance, quality improvement and reporting; surveillance; operations research; and clinical and health services research. However, as these health systems have generated more electronic data from an increasingly diverse array of IT systems and sources, they have encountered challenges associated with accurately recording, reconciling, and contextualizing these data in a way that supports a multitude of potential uses.

Through AcademyHealth's Health IT for Actionable Knowledge project, six participating health system partners raised a number of issues and challenges associated with using health IT to support these other critical functions, focusing primarily on internal improvement (which includes QI) and health services research. These challenges include:

- Capturing and then integrating EHR data with data from other sources;
- Establishing technical and procedural mechanisms to facilitate access and appropriate use by (often many) internal audiences (e.g., operations staff, clinicians); and
- Assessing and (where possible) improving data quality to support research efforts.

This report will describe some of the key challenges to capturing and then extracting electronic clinical data for research purposes, and elaborate on the experiences of these six health system partners that, recognizing the value of using EHR data to support research, have devised approaches to mitigate these challenges that could prove useful to others. The report is based largely on site visits, interviews, meetings, and a review of documents provided by the systems themselves. In these interactions, many of the researchers involved stressed that while care providers also rely on the information in these systems to care for patients, the providers (in most cases) have the benefit of temporal proximity, context, and other information of immediate relevance to the situation. Researchers noted that one particular challenge is that research involves post-facto investigation of only a subset of possible data elements, and is therefore subject to systematic omission of potentially important information.

The six health system partners in the Health IT for Actionable Knowledge project include:

- Denver Health
- Geisinger Health System
- Kaiser Permanente
- New York City Department of Health's Primary Care Information Project
- Palo Alto Medical Foundation Research Institute
- Veterans Health Administration

This report is organized in two parts: 1) a high-level overview of key challenges associated with accurately and appropriately *capturing* data through the use of EHRs, and 2) a review of promising strategies employed by the six partner health systems for *extracting* useful, quality data from these systems.

Capturing Quality Data

Nearly all physician practices—whether they are single-provider entities or large practice associations—have struggled to capture reliable, accurate data in their EHR systems. The task becomes even more complicated when the intention is to use these EHR data (sometimes integrated with other sources of information) to support research, given the added requirements of reproducibility and generalizability.

Although there are market indications that the EHR vendor community is trying to respond to these challenges, one significant impediment is the fact that EHR systems are not generally structured in a manner that allows users to extract the full value of the data.[2] In other words, the principle of "collect once and use many times" is much easier said than done with existing technologies.[3] In fact, it has been suggested that the very features of most EHR systems that make them attractive to clinical users actually contribute to their lack of utility as efficiently designed data management systems. In order to meet the needs and conform to the dominant workflow patterns of providers, most EHRs resemble digital versions of paper records. Somewhat ironically, this attempt at familiarity often makes it far more challenging for providers (and other potential users) to subsequently locate and then use the information they need. One oft-cited example of this is the common use of the free text "notes" field, which resembles the process of taking paper-based notes, but does an equally poor job of organizing and categorizing the content.

In fact, most EHRs are set up and used in a manner that often defaults to recording content in an unstructured format. When providers cannot quickly locate the appropriate field for a particular piece of information, or if there is no discrete or structured field that's been built into the system to capture said element, the typical response is to record the information in a free-text field. This does not necessarily affect the clinician's ability to locate information needed to care for the patient, but does make it nearly invisible to most researchers. Without standardized dialogue boxes, drop-down lists, or other pre-defined menu options, this means that there is no consistency in the content captured in free-text fields. As such, comparisons of entries across patients, providers, or other systems are exceedingly challenging and require the use of sophisticated coding tools and analytics to interpret the underlying content.[4]

One approach to harvesting meaningful information from free-text fields is the use of natural language processing (NLP) software tools. Though still in early stages of development, some institutions are actively testing the ability of NLP software to cull useful data from free text fields. The VA, in particular, has been engaged in the development of science and computerized solutions for NLP. The Consortium for Healthcare Informatics Research is a project grant with emphasis on information retrieval, information extraction, and de-identification. It focuses on advancing the science to improve NLP methods. Its partner, VA Informatics and Computing Infrastructure (VINCI), works to create a robust platform and user interface for NLP. Recent successes in NLP include extraction of information on left ventricular ejection fraction. VINCI is now in the process of indexing all concepts in more than two billion notes.

PAMFRI is also exploring opportunities for using NLP as part of its research efforts. In one study, they are trying to identify indicators of depression that may not be encoded in the formal "problem list" of its EHR. In another (under review), NLP will be used to find indications in the progress notes of pediatrician's concerns about autism that appear significantly before a formal diagnosis is made.

These explorations aside, even in circumstances where data are captured in structured fields, a number of issues can impact the extent to which the data are accurate, and of a high enough quality to be used for research. These include:

- Inconsistent or differential data entry by providers;
- Recording errors;
- Missing data;
- Lack of data standards and consistent coding practices; and
- Reconciliation of data from multiple sources.

As individuals, providers differ considerably in their use of EHRs. Thus, the application of structured data capture across providers—and provider institutions—can be highly inconsistent. Sometimes this inconsistency results from organizational policy or processes. For example, given the huge array of data, institutions prioritize the standardized capture and use of certain data elements over others, but this prioritization likely will not be consistent across organizations.

Similarly, some EHRs offer users the ability to establish defaults from previous entries into successive fields, and/or permit for the "cloning" of previous entries (either within or between patient records). A clinician treating a patient with a stabilized chronic condition, for example, may wish to allow for entries from past encounters to default to the last value (e.g., HbA1c = 8). Though such structured default entries may represent the "truth," their use may increase the likelihood of capturing inaccurate or inappropriate data (i.e., the patient's normally stable value may have changed, but without notice of the clinician).[5]

Other sources of inconsistency have less to do with organizational priorities or practices, and more to do with individual level "error." One such source is the practice of recording a data element in the appropriate field, but using incorrect or inconsistent terminology. A common example is blood pressure, a routinely measured and important indicator of health that is often recorded in different ways within the same practice or organization. A blood pressure of 120/80, for example, can be entered as 120 80; 120/80; 120/80 sitting; 120/80 left arm; etc. In essence, the number of permutations for a non-standardized data element is limited only by the establishment of clear policies, consistently followed, at the provider and practice levels. Despite efforts and progress, this is still fairly rare.

Another common occurrence is the correct recording of a data element, but in an unusual location. This poses a challenge in that anyone hoping to use the data element for research or other purposes downstream will look only in the usual or appropriate field and (unless there is reason to suspect or anticipate the problem) will assume that it is missing. The issue of missing data also poses challenges, as downstream users often do not have adequate information as to whether the absence of data in a field is intentional or the result a recording (or some other) error. In other words, is the patient's recent allergic reaction to a new medication recorded in the "allergies" section of the record, or only mentioned in the free-text notes?

Some data capture challenges stem from technical as opposed to user-driven issues. It is widely acknowledged, for example, that there is very little consistency in the use of clinical vocabularies/terminologies across EHR systems, and also across provider institutions.[6] Sometimes the actual clinical information is recorded differently, whereas in other cases the unit of measurement is different. Either way, given this lack of standardization, the interpretation and analysis of even the "same" data element collected from more than one EHR system or organization is not straightforward. It is likely that the introduction of ICD-10, a medical classification list for the coding of diseases, signs and symptoms, abnormal findings, complaints, social circumstances, and external causes of injury or diseases,[7] will further complicate this issue. Another possible permutation of this is that, though recorded in the right field, data elements are not always mapped to the appropriate data model.[8]

Many of these same data issues surface in situations where the data derive from sources other than EHRs (e.g., registries), and are compounded when there is a need to aggregate data across multiple sites and sources. A challenge unique to the act of collecting data from multiple sources is that of data reconciliation (i.e., determining which version of the same data element represents the "truth" or best defines the concept of interest). When multiple versions of the same data element are identified, the tasks of deciding which to use, establishing a record of that decision, and then developing a process for subsequent determinations, is critical. Partners involved in the Health IT for Actionable Knowledge project pointed out another complicating factor, which is that such determinations may change depending on the specific research question being posed. This is sometimes also an issue in multi-site research efforts, but a more common data quality challenge in aggregating data from multiple sites is that of distinguishing between "real" versus "artifactual" variations in data (i.e., what variation in data elements is real versus due to differences in data collection).[9]

Promising Practices for Extracting Quality Data

All health systems participating in the Health IT for Actionable Knowledge project indicated a keen awareness of the issues outlined above, and acknowledged that data cleaning required an inordinate amount of time, energy, and organizational resources that was unlikely to have a substantial benefit for core clinical care functions. Despite those hurdles, each partner organization expressed a motivation to work with its data to ensure that it would be of high enough quality to support various research needs. This motivation derives from the appreciation of the potential value of the data, and the lessons that can be learned from its effective use – not just in terms of treating individual patients, but also in terms of contributing to the overall performance of the institution, and the generalizable knowledge base. As one partner noted during a site visit, "We pitched research as part of the cake (not just the icing); it's part of our business model. It's the pervasive value system at every level of the organization."

In describing their various institutional efforts to generate high-value, "research-grade" data from EHR systems, partners involved in the Health IT for Actionable Knowledge project identified several important themes and features of "successful" endeavors. These include:

- Creation of duplicate databases to support research (i.e., data redundancy strategy);
- Investments in IT infrastructure and staff support to facilitate data access;
- Creation of automated data quality checks;
- A commitment to collaboration with colleagues (particularly those in QI and operations); and
- External policy pressures (e.g., reporting requirements, pay for performance).

Some of these features, and other characteristics are presented in more detail in Figure A. One characteristic of nearly all partners is that their respective organizations—recognizing the value of research—have invested in IT and data structures to support the cultivation of "research grade" data. In fact, all but one of the partner institutions participating in the Health IT for Actionable Knowledge project have established an IT infrastructure that allows the research departments to set up and manage separate working copies of the clinical database on which to perform research studies.[10] The creation of these separate data warehouses was deemed by partners as critical because it allowed for research staff to clean and work with "real" clinical data (typically refreshed in a timely manner) without posing any risk to or compromising the integrity of the "live" data streams used to support clinical care.

These data resources also provide opportunities to delve into the data and identify its strengths and limitations. Partners noted that, in some instances, this scrutiny has served to simply highlight different perspectives within the organization. Several partners confirmed that researchers often require far more rigorous examination of certain data fields than those who need the information for transactional purposes (e.g., to pay a claim), and this may have implications for how data are

Investigators at PAMFRI realized that the routinely collected data on race, ethnicity, and language were not only inadequate for research, but were not optimal for patient care. They developed data collection tools based on the U.S. Census, with two fields for race, ethnicity and ancestry as self-defined by the patient. They also initiated the capture of preferred language and preferences with respect to access for translation services. These tools were rolled out across the organization, subjected to an audit period, and now inform both day-to-day practice as well as a wide range of research studies.

managed and for how long they are stored. For example, researchers might want to see all versions of a record for mailing address in order to support longitudinal research efforts. This has led some partners to suggest a rule of thumb that no data ever get deleted or overwritten, so that there is always a mechanism for recovery. In others instances, this level of collaborative data exploration has developed into an opportunity for organizational improvement.

A number of partners also referenced the importance of IT support, both in gaining access to data for research purposes, but also in the development, integration, and implementation of any changes required to the IT infrastructure (including EHRs) based on data quality assessment efforts. All emphasized the necessity to establish the health system's research unit as an important IT client. Despite this view, partners described mixed levels of success in making this happen.

The VA, for example, indicated that, because its IT group is separate, a lot of paperwork and process effort is required to implement changes to the EHR based on data quality issues "discovered" by research. Another partner confirmed that, if the relationship between research and IT is not strong, the ensuing experience will be one of frustration. In other words, all acknowledged the benefits of having dedicated IT staff who understand research data needs and uses, and who perceive value in facilitating data access to research units.

Regarding data access, Geisinger noted that the development of its enterprise data warehouse, the Clinical Decision Intelligence System (CDIS), is what has allowed researchers (and other data users) to avoid significant bottlenecks. Specifically, de-identified, analytic databases are created from CDIS as needed for research, and can be used (by approved individuals) outside of the firewall.

Partners also noted the importance of data validation mechanisms and logic checks (i.e., checks to ensure that the data make sense) to evaluate data quality. For example, if a record indicates that a patient is deceased, subsequent entries about health care use should trigger an investigation as to which data are in error. Or, perhaps a less dramatic example is the need to ensure that lab and diagnostic values fall within an appropriate range for the condition or measure of interest (e.g., blood pressure cannot be in the 1000s). PCIP acknowledged that because its health department receives data from so many different types of provider organizations,the team is creating a system of automated logic or content checks to validate or assess data quality. This process has also generated significant volumes of meta-data (i.e., data about the content and context of the data) which will help them determine which fields or data elements and/or sources are likely to be problematic. Such systems can help determine the reliability of a given source (e.g., a provider) during intervals of interest. All partners confirmed that the establishment of such processes and meta-data is critical to the continuous improvement of their data resources, and can help support the research function.

Health services researchers from the health systems participating in AcademyHealth's project —wanting to take full advantage of electronic health data available—also emphasized that their role includes figuring out how to improve the quality of routinely collected clinical information to meet research standards (i.e., getting "research grade" data into the system) without overburdening clinicians, and while ensuring the buy-in (and where possible, participation) of operations personnel and other critical institutional partners. As one partner expressed,"I *never* assume that operations will put data in for me. Everyone is way overworked. So, how can we use and interpret what's already there?" This belief in the notion that "what gets studied gets improved" has led nearly all of them to seek out opportunities for collaboration with their operations and quality improvement colleagues.

Sometimes those opportunities are identified because—based on examination of the data—a researcher uncovers a limitation or problem. For example, noting the inconsistent location within the EHR of documentation of patient advanced directive information, the research department within PAMFRI worked with QI staff to assess provider workflow and documentation practices, which resulted in a restructuring of the EHR flow for entering these data. In other instances, the motivating force behind such change may not be research; rather, it reflects the priorities of either the operations or QI department.

Several partners noted that this type of collaboration—based on institutional priorities—is often a high-value/high-yield approach, as it tends to involve investigation of the most "clinically relevant" questions and data, or an issue of immediate concern (as opposed to a longer-term care delivery issue).

Despite examples of success, partners also cautioned that it was the researcher's role to point out where there are limitations in the data available that could or should preclude its use for (at least) certain research questions. Some partners simultaneously observed that EHR data will always have ambiguities because the things they are trying to measure are inherently ambiguous. At some point, trying to improve data quality becomes the "quest for the unicorn," so standard research practice should involve the conduct of multiple analyses to assess sensitivity of results to data definitions, missing data, and other realities of research based on electronic clinical data.

Finally, all partners referenced the power of external policy pressures to help fuel the research imperative within their respective organizations. They noted that as payers and providers experiment with alternatives to pure fee-for-service payment models, health systems face an ever-expanding array of new financial incentives and reporting requirements, setting the stage for a new cultural paradigm in which continuous data collection, review, and assessment (often in the form of research) is the norm. And, though many partners were already well positioned to take full advantage of these opportunities, they acknowledged the impact of meaningful use measures (required to receive payment for EHR adoption) as a strong force for greater data standardization among providers. They also cited the emergence of new organizational formations like accountable care organizations which, in order to be sustainable, will need to figure out how to effectively and efficiently care for diverse populations.

Conclusions

The experiences of the six institutions examined as part of AcademyHealth's Health IT for Actionable Knowledge project confirms that there are a number of issues and challenges associated with using electronic clinical data to support critical functions such as quality improvement and HSR. These include challenges of:

- Capturing and then integrating EHR data with data from other sources;
- Establishing technical and procedural mechanisms to facilitate access and appropriate use by (often many) internal audiences (e.g., operations staff, clinicians); and
- Assessing and (where possible) improving data quality to support desired functions, including research.

However, recognition of the potential value of EHR data, supported by examples from those who are pioneering its use by demonstrating value and generating new and much-needed evidence, is likely to compel further exploration of these issues. The themes and features of "successful endeavors" that these innovators have identified could prove useful to other health systems considering significant health IT investments.

As the value proposition of these investments is realized, it is hoped that even further progress can be made to resolve (or at least evolve) many of the challenges associated with capturing data and then making it available for research purposes. Those engaged in this transformation process recognize that it will require commitment and significant investment of time and resources, particularly given the desire to expand from having high-quality data within one health system, to attaining some level of comparability across multiple institutions (something that will be required for the rigorous conduct of multi-site research).

Those health systems with aspirations for broader use of EHR data might want to consider the promising practices reflected in the above examples from the Health IT for Actionable Knowledge project partners (e.g., establishment of data redundancy strategies, investments in IT infrastructure and staff support, commitment to collaboration with QI and operations colleagues). Furthermore, all might benefit from seeking opportunities for cross-system communication and—to the extent possible—coordination and collaboration, so that their common interests can be more effectively addressed.

About The Author:

Alison Rein, M.S., is the director of health innovation and information infrastructure at AcademyHealth. Bryan Kelley, research assistant at AcademyHealth, provided research support for this report.

Acknowledgements:

AcademyHealth gratefully acknowledges the time and expertise provided by the researchers, clinicians, IT professionals, and executives at the health systems participating in AcademyHealth's HIT for Actionable Knowledge project—Denver Health, Geisinger Health System, Kaiser Permanente, the New York City Department of Health's Primary Care Information Project, the Palo Alto Medical Foundation Research Institute, and the Veterans Health Administration. Any errors are AcademyHealth's.

Endnotes

[1] National Research Council. 2007. *The Learning Healthcare System: Workshop Summary (IOM Roundtable on Evidence-Based Medicine).* Washington, DC: The National Academies Press.

[2] President's Council of Advisors on Science and Technology. (December 2010). Report to the President: Realizing the Full Potential of Health Information Technology to Improve Healthcare for Americans: The Path Forward. Retrieved from http://www.whitehouse.gov/sites/default/files/microsites/ostp/pcast-health-it-report.pdf, accessed on February 23, 2012.

[3] Many respected health services researchers are skeptical that – even given the ultimate EHR tool – their research data needs could ever be adequately fulfilled by relying on data captured via EHRs at the "front end" of the care process. This is in part due to expectations of provider and systems variability, and the importance of considering context (i.e., why the patient had the encounter and why certain information was (or was not) collected).

[4] It should be noted that, though the use of free text fields presents challenges to use of the data for research, documentation in this manner is often necessary to fully capture and describe the specifics of individual patients.

[5] American Health Information Management Association. "Quality Data and Documentation for EHRs in Physician Practice." *Journal of American Health Information Management Association,* Vol. 79, No. 8, August 2008, pp. 43-48.

[6] Defined by the Computer-Based Patient Record Institute as "standardized terms and their synonyms, which (allow one to) record patient findings, circumstances, events and interventions with sufficient detail to support clinical care, decision support, outcomes research and quality improvement; and can be efficiently mapped to broader classifications for administrative, regulatory, oversight, and fiscal requirements." Sue Bowman. "Coordination of SNOWMED-CT and ICD-10: Getting the Most Out of Electronic Health Record Systems," *Perspectives in Health Information Management,* Spring 2005. Retrieved from http://library.ahima.org/xpedio/groups/public/documents/ahima/bok1_027171.pdf, accessed on February 23, 2012.

[7] World Health Organization. "International Classification of Diseases (ICD)". Retrieved from http://www.who.int/classifications/icd/en/, Accessed on February 23, 2012.

[8] A data model explicitly determines the structure of data or structured data. Typical applications of data models include database models, design of information systems, and enabling exchange of data. Usually data models are specified in a data modeling language.

[9] Kahn, M. et al. "A Pragmatic Framework for Single-Site and Multi-Site Data Quality Assessment in Electronic Health Record-Based Clinical Research," publication forthcoming in *Medical Care.*

[10] PCIP has not created such a database because it does not collect patient-level data (only aggregated "count" data) and is not directly part of an institution that provides patient care.

Figure A: **Key Features of Six Health IT for Actionable Knowledge Partners**

Health IT for Actionable Knowledge Partner	Research Data Storage and Management Strategies	Types of Data Typically Available for Research	Data Access Processes and Strategies	Location of IT Support for Research Endeavors
Denver Health	Centralized storage of (nearly) all data, with data warehouse that extracts information from patient records for clinical decision support.	From the EHR and other sources, core clinical and ancillary services including: • Inpatient • Pharmacy • Imaging • Laboratory • Surgery • Blood bank • Emergency department Access to patient demographics and payer information also available. Clinical notes captured electronically from inpatient nursing, but not yet implemented for physicians.	Data in central warehouse and from other sources are "cloned" for research purposes so as not to interfere with clinical care and operations functions.	IT support is not housed within research unit; requests for research data can be made electronically, but general Denver Health IT staff fulfill such requests based on priority. Research staff have also worked with IT staff to learn how to perform data extractions themselves. Actual data extraction and associated technical support (e.g., to define appropriate samples) have generally become bottlenecks that impede access to data for research purposes.
Geisinger Health System	Enterprise data warehouse (EDW) functions as an "operational data store" and is a centralized database that connects different data sets but does not combine them into a single database. EDW is used by administrators, business analysts, health services and clinical researchers, and members of the Innovations team.	• Encounters • Orders for lab tests, medications, imaging and procedures • Appointments • Digital imaging • Clinical notes and results summaries (e.g., lab and pathology) • Billing and claims data, and additional administrative data for patients en — rolled in Geisinger Health Plan • Researchers also have access to patient demographics and vitals (e.g., problem lists, personal and family histories) and patient satisfaction data. • Expansion efforts are underway to provide access to additional data source systems (e.g., oncology) in stages.	A stand alone, de-identified "shadow copy" of the EDW resides on a separate server and caters to researcher-specific needs (and other "high-end" users within the hospital system). A Web interface allows those with potential research questions to access de-identified databases created from the EDW of the system in advance of pursuing IRB approval. Use is audited, and data and analyses remain behind firewall.	IT support of EDW for research centers and Clinical Innovations resides within the clinical Innovations team.
Kaiser Permanente Each of the eight regional KP health care organizations supports research activities. Administrative support is provided at the KP Program (i.e., national) level by the Kaiser Foundation Research Institute.	All eight regional KP research organizations participate in the HMO Research Network (HMORN), including the development and maintenance of HMORN's Virtual Data Warehouse (VDW). The common VDW data model is populated from KP's EHR and other clinical and administrative data systems. In general, each KP region stores VDW data and manages the update process. Some regions (e.g., Northwest and Hawaii) collaborate on these activities.	VDW includes data on: • Enrollment • Demographics • Pharmacy • Utilization (e.g., diagnoses and procedures) • Vitals • Facility census • Lab (including results) • Cause of death (if applicable) • VDW also contains specialized data for oncology, rehabilitation services, and implantable medical device tracking.	In each KP region, VDW data are readily available to researchers within that region. KP regions make VDW data available for multi-region research activities using a "distributed query" approach modeled on HMORN protocols and processes. The recently created Program-level Center for Effectiveness and Safety Research (CESR) is streamlining these multi-region processes for rapid response activities (e.g., the mini-sentinel network).	Some KP regions (e.g., Northern California) maintain separate IT staff and infrastructure to support the regional VDW. Other (generally smaller) regions depend on program-level IT staff and infrastructure.

Figure A: **Key Features of Six Health IT for Actionable Knowledge Partners** (continued)

Health IT for Actionable Knowledge Partner	Research Data Storage and Management Strategies	Types of Data Typically Available for Research	Data Access Processes and Strategies	Location of IT Support for Research Endeavors
New York City Primary Care Information Project	Data from participating provider EHRs remain at the practice office, so there is no sharing of patient-level data. Aggregated and de-identified "count" data are collected through nightly queries and maintained on a Department of Health and Mental Hygiene (DOHMH) server.	PCIP maintains a limited set of clinical and other data that reflect the clinical priorities of the DOHMH: • Practice demographics • Patient and provider satisfaction measures • Aggregated practice management data by CPT code and provider • Aggregated clinical information including effectiveness of care and syndromic surveillance measures • Statistics on EHR use	DOHMH researchers have access to aggregated, de-identified data gathered through queries of practices. Clinical care providers have access only to the data of their patients, as well as city-wide benchmark levels for standardized quality measures.	EHR system IT support is provided on-site for all participating practices. PCIP has EHR IT support for the query system architecture and internal IT support for analysis and storage of data.
Palo Alto Medical Foundation Research Institute	Data are extracted from an enterprise data warehouse and used to populate a separate "clone" database within the PAMF firewall that includes identifiable information, such as progress notes and images. Data without identifiable personal health information are transferred to a secure PAMFRI server. As new variables are created for various projects, these are stored in the PAMFRI database for use by other studies.	• Billing • Scheduling • Care management • Vitals • Lab orders and results • Medication orders • Pathology • Oncology	Projects require IRB review before project-specific data files can be made available for analysis. Many projects can rely on data that have already been de-identified, and thus require only an expedited review to determine their exempt status. These data, however, are still protected as if they were limited data sets, affording extra security.	Has a dedicated IT staff for the research function – responsible for creating project data files and providing necessary IT support to researchers.
Veterans Health Administration	The Veterans Affairs (VA) Informatics and Computing Infrastructure (VINCI) is a virtualized computing environment that serves most VHA clinical data back to 2000 in a rationalized database. It creates incentives for researchers to keep data in a central repository—a practice designed to minimize data loss.	VINCI has national data with nightly updates on: consults, preventative health, clinical guidelines, immunizations, laboratory results, microbiology results, primary care panels, inpatient and outpatient pharmacy, vitals, appointments, mental health assessments, administrative data, orders, inpatient movement, staff, billing, radiology, patient demographics. Other types of data are updated on a less frequent basis: notes, beneficiary travel, non-VA-filled medications, secure messaging transactional data (not content). Registry data for cancer, surgery, HIV, etc., is available through partnerships with data stewards.	VINCI manages authorizations for data access through the Data Access Request Tool (DART), which was developed in collaboration with VA Information Resource Center (VIReC) and coordinates the processing of requests through various VHA offices. VINCI controls access to data, so only authorized users can access data for specific research projects under an active IRB protocol. This practice is designed to prevent researchers from accessing data for one project and then reusing data for multiple other projects without IRB approval. VINCI caches and randomly audits outbound data transfers to verify that patient data are not inappropriately transferred out of VINCI.	VINCI provides support through dedicated and escalated staff. IT support exists at each VHA medical center. VIReC provides assistance on data quality and meaning issues.