10-14-2014

The Community Health Applied Research Network (CHARN) Data Warehouse: a Resource for Patient-Centered Outcomes Research and Quality Improvement in Underserved, Safety Net Populations

Reesa Laws
Kaiser Permanente Northwest, Center for Health Research, Reesa.Laws@kpchr.org

Suzanne Gillespie
Kaiser Permanente Northwest, Center for Health Research, suzanne.e.gillespie@kpchr.org

Jon Puro
OCHIN, Inc., puroj@ochin.org

Stephan Van Rompaey
Center for AIDS Research, University of Washington, kelpie@u.washington.edu

See next pages for additional authors

Recommended Citation

Laws, Reesa; Gillespie, Suzanne; Puro, Jon; Van Rompaey, Stephan; Quach, Thu; Carroll, Joseph; Chang Weir, Rosy; Crawford, Phil; Grasso, Chris; Kaleba, Erin; and McBurnie, Mary Ann (2014) "The Community Health Applied Research Network (CHARN) Data Warehouse: a Resource for Patient-Centered Outcomes Research and Quality Improvement in Underserved, Safety Net Populations," *eGEMs (Generating Evidence & Methods to improve patient outcomes)*: Vol. 2: Iss. 3, Article 11.
DOI: http://dx.doi.org/10.13063/2327-9214.1097
Available at: http://repository.edm-forum.org/egems/vol2/iss3/11

The Electronic Data Methods (EDM) Forum is supported by the Agency for Healthcare Research and Quality (AHRQ), Grant 1U18HS022789-01. eGEMs publications do not reflect the official views of AHRQ or the United States Department of Health and Human Services.

The Community Health Applied Research Network (CHARN) Data Warehouse: a Resource for Patient-Centered Outcomes Research and Quality Improvement in Underserved, Safety Net Populations

Abstract

Background: The Community Health Applied Research Network, funded by the Health Resources and Services Administration, is a research network comprising 18 Community Health Centers organized into four Research Nodes (each including an academic partner) and a data coordinating center. The network represents more than 500,000 diverse safety net patients across 11 states.

Objective: The primary objective of this paper is to describe the development and implementation process of the CHARN data warehouse.

Methods: The methods involved regulatory and governance development and approval, development of content and structure of the warehouse and processes for extracting the data locally, performing validation, and finally submitting data to the data coordinating center.

Progress to Date: Version 1 of the warehouse has been developed. Tables have been added, the population and the years of electronic health records (EHR) have been expanded for Version 2.

Conclusions: It is feasible to create a national, centralized data warehouse with multiple Community Health Center partners using different EHR systems. It is essential to allow sufficient time: (1) to develop collaborative, trusting relationships among new partners with varied technology, backgrounds, expertise, and interests; (2) to complete institutional, business, and regulatory review processes; (3) to identify and address technical challenges associated with diverse data environments, practices, and resources; and (4) to provide continuing data quality assessments to ensure data accuracy.

Acknowledgements

This work was funded by Health Resources and Services Administration (HRSA). The authors gratefully acknowledge the CHARN community health centers (CHCs), research node centers and the data coordinating center for contributing the time and data used in preparation of this article. Association of Asian Pacific Community Health Organizations (AAPCHO) CHARN affiliates include four CHCs: Asian Health Services Community Health Center, Charles B. Wang Community Health Center, Waianae Coast Comprehensive Health Center, Waimanalo Health Center, and their academic affiliate; University of California, Los Angeles. Alliance of Chicago Community Health Services (Alliance) CHARN affiliates include eight CHCs: Erie Family Health Center, Inc., Glide Health Services, Heartland Health Outreach, Howard Brown Health Center, Near North Community Health Center, North Country HealthCare, PCC Community Wellness, Children's Hospital, NU, and their academic affiliate; Northwestern University. Fenway Health CHARN affiliates include three CHCs: Chase-Brexton Health Services, Beaufort-Jasper-Hampton Comprehensive Health Services, Fenway Health, and their academic affiliate; University of Washington. OCHIN, Inc. CHARN affiliates include four CHCs: Open Door Community Health Centers, Virginia Garcia Memorial Health Centers, Multnomah County Health Department, OHSU Richmond Family Medicine Clinic, and their academic affiliate; Oregon Health & Science University. The CHARN Data Coordinating Center (DCC) is located at the Kaiser Permanente Center for Health Research. COMPETING INTERESTS: There are no competing interests. FUNDING: This work was funded by Health Resources and Services Administration (HRSA). FIGURE LEGENDS: Figure 1: Sample IRB Process

Keywords
Health Information Technology, Data Use Agreements, Research Networks, Data Warehouse, PCOR, Safety-net, Community Health Centers, Electronic Health Records

Disciplines
Clinical Epidemiology | Community Health and Preventive Medicine | Environmental Public Health | Epidemiology | Health Information Technology | Health Services Administration | Health Services Research

Authors
Reesa Laws, *Kaiser Permanente Northwest, Center for Health Research*; Suzanne Gillespie, *Kaiser Permanente Northwest, Center for Health Research*; Jon Puro, *OCHIN, Inc.*; Stephan Van Rompaey, *Center for AIDS Research, University of Washington*; Thu Quach, *Asian Health Services*; Joseph Carroll, *Open Door Community Health Centers*; Rosy Chang Weir, *Association of Asian Pacific Community Health Organizations (AAPCHO)*; Phil Crawford, *Kaiser Permanente Northwest, Center for Health Research*; Chris Grasso, *The Fenway Institute, Fenway Health*; Erin Kaleba, *Alliance of Chicago Community Health Services*; Mary Ann McBurnie, *Kaiser Permanente Northwest, Center for Health Research*.

The Community Health Applied Research Network (CHARN) Data Warehouse: a Resource for Patient-Centered Outcomes Research and Quality Improvement in Underserved, Safety Net Populations

Reesa Laws, BS;[i] Suzanne Gillespie, MA, MS;[i] Jon Puro, MPA-HA;[ii] Stephan Van Rompaey, PhD;[iii] Thu Quach, PhD, MPH;[iv] Joseph Carroll, MD, PhD;[v] Rosy Chang Weir, PhD;[vi] Phil Crawford, MS;[i] Chris Grasso, MPH;[vii] Erin Kaleba, MPH;[viii] Mary Ann McBurnie, PhD[i]

Abstract

Background: The Community Health Applied Research Network, funded by the Health Resources and Services Administration, is a research network comprising 18 Community Health Centers organized into four Research Nodes (each including an academic partner) and a data coordinating center. The network represents more than 500,000 diverse safety net patients across 11 states.

Objective: The primary objective of this paper is to describe the development and implementation process of the CHARN data warehouse.

Methods: The methods involved regulatory and governance development and approval, development of content and structure of the warehouse and processes for extracting the data locally, performing validation, and finally submitting data to the data coordinating center.

Progress to Date: Version 1 of the warehouse has been developed. Tables have been added, the population and the years of electronic health records (EHR) have been expanded for Version 2.

Conclusions: It is feasible to create a national, centralized data warehouse with multiple Community Health Center partners using different EHR systems. It is essential to allow sufficient time: (1) to develop collaborative, trusting relationships among new partners with varied technology, backgrounds, expertise, and interests; (2) to complete institutional, business, and regulatory review processes; (3) to identify and address technical challenges associated with diverse data environments, practices, and resources; and (4) to provide continuing data quality assessments to ensure data accuracy.

Introduction

Federally funded Community Health Centers (CHCs) provide comprehensive primary care health services to "safety net" patients, i.e., patients who are medically underserved and who experience substantial barriers to accessing health care. CHCs serve low income patients; individuals with racial- and ethnic-minority status; nonconforming sexual orientation or gender identity; limited English proficiency; housing, food, or employment insecurity; and complex comorbid health conditions, among others. There are 7,000 CHCs in the United States and territories, serving approximately 22 million people. Safety net patient populations have historically been underrepresented in clinical trials, industry-financed studies, and studies using large data warehouses due to explicit exclusion criteria and lack of access to their medical information. Thus the results of these research studies are not directly generalizable to this underserved population.[1]

Including safety net populations in large-scale electronic health record (EHR) databases could provide valuable insights by describing population demographics, common conditions and comorbidities, current care practices, and, to some extent health care utilization. Such databases have the potential to engage more researchers, advance the evidence base for issues affecting these populations, and influence health care delivery.[2] A data warehouse for safety net populations can be developed by moving EHR data from multiple locations to one or more investigative centers, where the data are aggregated and analyzed for a specific purpose. Large samples can be selected and used to draw powerful inferences about populations.[3] Classical clinical trials that collect data through structured interviews are much more limited in scope, whereas a large data warehouse that captures a broader range of clinical data associated with disease management at the point-of-care allows researchers to look at long-term outcomes and complications.[4]

[i] Kaiser Permanente Northwest, Center for Health Research, [ii] OCHIN, Inc., [iii] Center for AIDS Research, University of Washington, [iv] Asian Health Services, [v] Open Door Community Health Centers, [vi] Association of Asian Pacific Community Health Organizations (AAPCHO), [vii] The Fenway Institute, Fenway Health, [viii] Alliance of Chicago Community Health Services

However, the design, development, and implementation of a data warehouse for health care delivery system data is a complex, sociotechnical challenge.[5] Large, organized health systems have been building EHR databases (or Learning Health Systems) to better answer research and clinical questions: the Department of Veterans Affairs (VA), with more than 8 million enrollees; Kaiser Permanente, also with more than 8 million enrollees; and Geisinger Health System, with more than 2 million enrollees. Other large federally funded networks have also built EHR databases, including the National Cancer Institute (NCI), the Vaccine Safety Datalink at the Centers for Disease Control and Prevention, and the American Medical Group Association.[2,6-8] One limitation of large, organized health system databases is the exclusion of safety net patients due to unemployment, lack of insurance, or other barriers. Additional research infrastructure is needed in order to identify and address specific health-care needs of this underrepresented population.

The Community Health Applied Research Network (CHARN) data warehouse was created to address this gap. CHARN was funded by the Health Resources and Services Administration (HRSA) to establish research capacity and infrastructure for carrying out patient-centered outcomes research (PCOR) in a network of CHCs serving safety net patients. Aims of the CHARN were the following: (1) to describe CHCs' medically vulnerable, diverse safety net populations, (2) to establish a multisite, multidisciplinary collaborative infrastructure to advance PCOR, (3) to inform design of PCOR studies, and (4) to identify scientific questions that can be addressed by a large-scale combined CHC warehouse. Unique aspects of this funding include the exclusive focus on safety net patients and the intention to build a CHC- and safety net-based research agenda and research capacity within the CHCs, including developing staff and resources and providing access to colleagues and researchers at a national level. A specific objective of CHARN was to create a centralized, robust data warehouse comprising data extracted primarily from the CHCs' EHRs. CHARN partners include 18 CHCs organized into four Research Nodes (each including an academic partner) and a Data Coordinating Center (DCC).

The CHARN data warehouse pools data from multiple CHCs to address important questions relevant to safety net populations in primary care practice under real-world conditions. The data warehouse was envisioned to serve multiple purposes, including the following: characterizing the safety net population; and supporting data-only studies, research proposals, and quality improvement (QI) efforts. This paper describes the experience of the CHARN network in defining and implementing a "proof-of-concept" data warehouse—developing procedures and tools for defining and standardizing data content, extracting and uploading the data, and carrying out basic data quality and validation checks. In addition, this paper describes the design and governance of the CHARN, while highlighting lessons learned and challenges to inform future development of research networks and research focused in CHCs.

Development and Implementation

Developing Collaborative Relationships

The four Research nodes included in the CHARN include the Association of Asian Pacific Community Health Organizations (AAPCHO), Alliance of Chicago Fenway Health, and OCHIN, Inc. Each of CHARN's Nodes coordinate at least three CHCs and an academic partner. Names and locations of participating CHCs and academic partners are listed in Table 1 by Node. The Nodes are responsible for overseeing the activities of their CHCs and serving as liaisons with the DCC. Academic partners are responsible for supporting the development of CHC research priorities and the DCC is responsible for coordinating the activities of the national network, developing and maintaining the data warehouse infrastructure, and providing analytic and technical support to the network.

Table 1. CHARN Partners

Partner Name, Location	Participating Community Health Centers and Academic Affiliates,* Location
Kaiser Permanente Center for Health Research (DCC), Portland, OR	N/A
Association of Asian Pacific Community Health Organizations (AAPCHO), Oakland, CA	• Asian Health Services (AHS), Oakland, CA • Charles B. Wang (CBW), New York, NY • Waianae Coast Comprehensive Health Center (WCCHC), Oahu, HI • Waimanalo Health Center (WHC), Oahu, HI • NYU Center for the Study of Asian American Health, New York, NY*
Alliance of Chicago Community Health Services (Alliance) Chicago, IL	• Alliance of Chicago Community Health Services, Chicago, IL • Erie Family Health Center, Chicago, IL • Heartland Health Outreach, Chicago, IL • Howard Brown Health Center, Chicago, IL • Near North, Chicago, IL • North Country Healthcare, Flagstaff, AZ • PCC Community Wellness Center, Oak Park, IL • Feinberg School of Medicine Northwestern University, Chicago, IL*
Fenway Health (Fenway), Boston MA	• Beaufort-Jasper-Hampton Comprehensive Health Services, Ridgeland, SC • Chase Brexton Health Services, Baltimore, MD • Fenway Health, Boston, MA • University of Washington, Seattle, WA*
OCHIN, Inc. (OCHIN), Portland, OR	• Multnomah County Health Department, Portland, OR • Open Door Community Health Center, McKinleyville, CA • OHSU Richmond Clinic, Portland, OR • Virginia Garcia Memorial Health Center, Portland, OR • Oregon Health and Sciences University, Portland, OR*

* Each node had an academic affiliate to support research in their CHCs.

With so many participating organizations and wide diversity in research and clinical skills and experience, it was essential for the data warehouse development to establish and adhere to procedures and policies to assure CHCs that their EHR data would be protected and that ownership would remain with them—not with external researchers. CHARN is governed by a Steering Committee and four subcommittees: Research Planning, Data, Institutional Review Board Regulatory (IRB), and Communications. A representative from each of the four Research Nodes is required on each committee, and Nodal representatives are responsible for ensuring that their respective CHCs are informed and approve of the scope of proposed projects. Monthly conference calls and biannual face-to-face meetings were held that helped to develop collaborative relationships across and within the committees. Because the development of the data warehouse was an integral part of the network, data representation was needed on each committee. For the data warehouse, guidelines were defined to assure that CHCs reviewed and approved protocols for identifying data elements, for accessing and sharing data (within Nodes and with the DCC), and for publication. Version 1 (V1) of the data warehouse was intended as a "proof-of-concept" to test the operationalization of technical and governance procedures.

Data Use Agreements and Institutional Review Board (IRB) Approval

The first step in creating a CHARN data warehouse was to develop and implement multilevel data use agreements (DUAs) between all participating CHCs and their representative Nodes, and between the Nodes and the DCC. The DCC, Nodes, and CHCs worked together to develop a shared understanding of the purpose, the process for extraction, the content, and the proposed use of the warehouse. CHC involvement in the approval of all data sharing activities, including manuscript review prior to publication, was a critical component of the DUAs. DUAs were executed at each point where data were exchanged. Fully executed DUAs were completed within nine months. DUAs between the CHCs and the Nodes were executed within one to six months, DUAs between the Nodes and the DCC were executed within one to nine months.

In addition to creating DUAs for data sharing, each CHC sought IRB approval for the data warehouse protocol according to their specific needs and requirements. Thus, the IRB approval process varied, as did the time required to obtain approval. For example, some CHCs ceded to a Node-level IRB while others had their own local IRB. One CHC has a research committee that includes community members and that reviews all proposed research projects prior to IRB submission. This process requires extra time, but engages local stakeholders directly. Overall, the mean time to obtain IRB approval was 5 weeks and ranged from 1 to 25 weeks across the 18 CHCs. A sample IRB review process is included as Figure 1. Key issues for most IRBs included clarity regarding data sharing, data ownership, and the processes for approving projects and manuscripts based on the data warehouse. Particular vigilance was accorded to the vulnerable and unique populations served within each CHC (e.g., HIV and AIDs, the homeless, racial and ethnic minorities). For example, only coarsely aggregated race information was approved for V1 due to the potential for identifying patients falling into racial categories occurring with low frequency.

Data Warehouse Population and Data Content

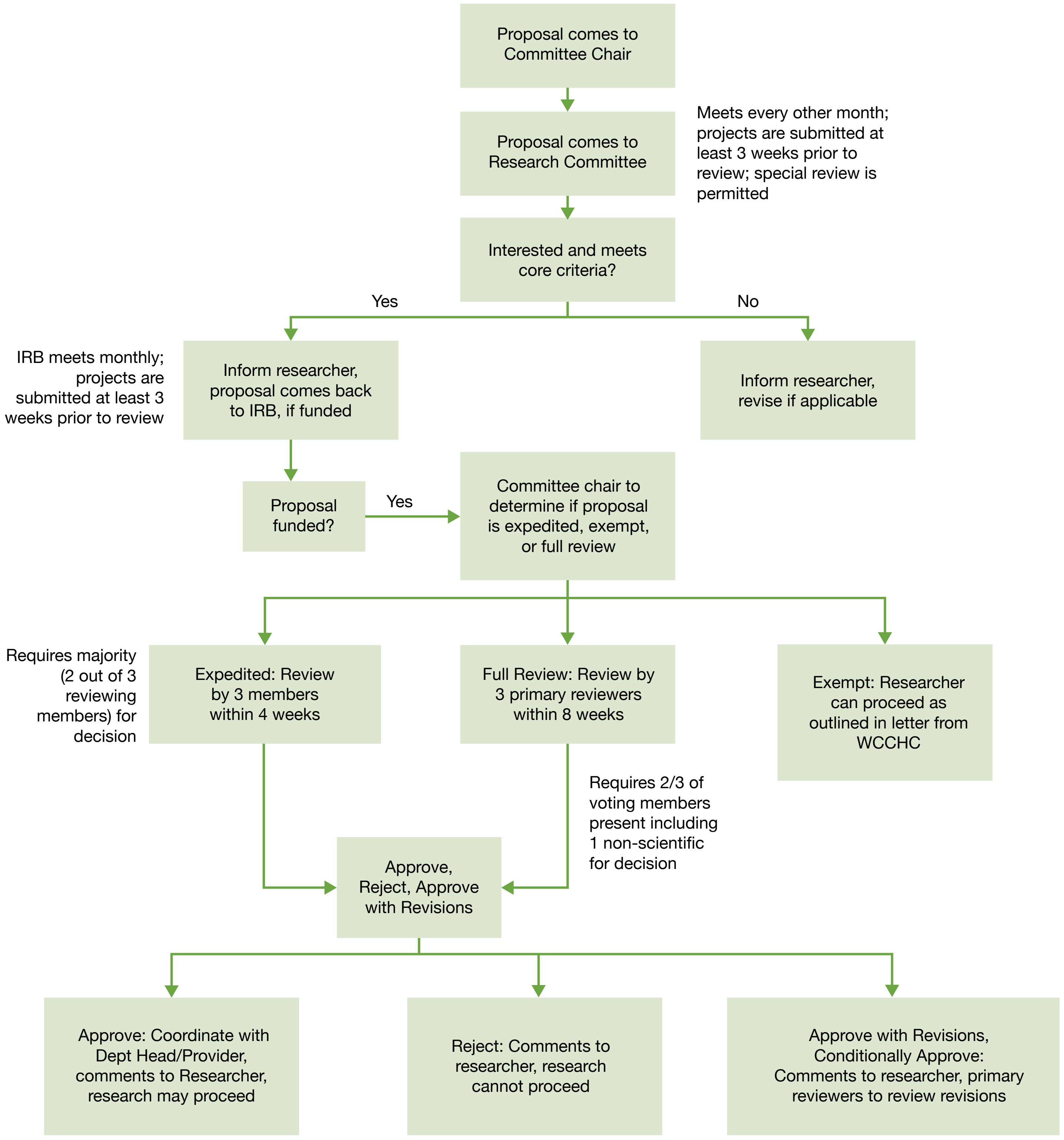
A multidisciplinary team of community clinicians, academic researchers, and data programmers defined the CHARN patient population and data elements needed to support CHARN's near-term goals for carrying out PCOR (Table 2). According to CHARN governance and oversight processes, the Research Planning Subcommittee guided the concept development and overall scope of V1 of the data warehouse. The CHARN Data Subcommittee (DS), comprising representatives with Node- and CHC-specific data expertise, was integral to the content development process. The DS made recommendations based on their experience with local data regarding ease of extraction, and completeness and accuracy of the data. An interactive, iterative process was developed between researchers, clinicians, and data programmers to finalize V1 content. The Steering Committee approved the final content and protocol.

Table 2. CHARN V1 Data Warehouse Tables

Subject Area	Types of Data Included
Patient Demographics	Birth and death dates, birth and current gender, transgender status, race, ethnicity, primary language, and CHC enrollment date
Encounter Data	Encounter start and end dates, encounter type, department, location, provider, insurance for encounter, and smoking status
Diagnosis Data	Diagnosis code, coding system, diagnosis descriptive name, and where data was collected from (patient reported, provider coded, billing data)
Laboratory Results	Lab code, coding system, lab descriptive name, collection and results dates, lab result, reference units (including high and low values), interpretation
Medications Ordered	Medication names (generic and brand names), medication code and coding system, form, total dose, units, route, frequency, start and end dates, stop reason, diagnosis associated with medication

Goals for the V1 data warehouse included (1) establishing protocols for sharing data among the four participating Nodes, (2) creating the technical processes for extracting and merging common data sets, and (3) establishing a repository of data to inform the first set of CHARN-approved research studies. V1 content focused on seven specific disease cohorts: diabetes, HIV- and AIDS-related conditions, hepatitis B, hepatitis C, cardiovascular disease, hypertension, and dyslipidemia. These conditions were chosen because they impose heavy burdens on patients and health care systems and because patients with these conditions could be identified using data elements commonly available and existing in discrete fields in the EHR: ICD-9 codes, medication prescriptions, and lab results. Patients were included in V1 if they had at least

Figure 1. Sample IRB Process


Proposal comes to Committee Chair
Proposal comes to Research Committee
Meets every other month; projects are submitted at least 3 weeks prior to review; special review is permitted
Interested and meets core criteria?
Yes
No
IRB meets monthly; projects are submitted at least 3 weeks prior to review
Inform researcher, proposal comes back to IRB, if funded
Inform researcher, revise if applicable
Proposal funded?
Yes
Committee chair to determine if proposal is expedited, exempt, or full review
Requires majority (2 out of 3 reviewing members) for decision
Expedited: Review by 3 members within 4 weeks
Full Review: Review by 3 primary reviewers within 8 weeks
Exempt: Researcher can proceed as outlined in letter from WCCHC
Requires 2/3 of voting members present including 1 non-scientific for decision
Approve, Reject, Approve with Revisions
Approve: Coordinate with Dept Head/Provider, comments to Researcher, research may proceed
Reject: Comments to researcher, research cannot proceed
Approve with Revisions, Conditionally Approve: Comments to researcher, primary reviewers to review revisions

one primary care encounter between January 1, 2008, and December 31, 2010. The 2008 start date was chosen because most CHCs had implemented an EHR by 2006 and had transitioned patients' medical records to the EHR by 2008 (one CHC did not have an EHR system in place during V1 work). Demographic data were requested for all CHARN patients meeting this criterion while more comprehensive data were obtained on the specified cohorts, including lab results, medication orders, and encounter data. We requested data elements we anticipated would be needed to create computable phenotypes, but the work of developing and validating them is ongoing and too lengthy to include in this paper.

Warehouse Design and Implementation

The CHARN DS was established to provide feedback to researchers on the content of the warehouse, to develop a database schema and methods to facilitate the extraction of data from CHCs, as well as creating the processes for transferring those data to the DCC. In order to implement V1, the DS created a standardized data dictionary to define requested data elements, along with a data submissions process to specify procedures for compiling, querying, and transmitting the data. After review and discussion by the DS, Microsoft SQL Server was chosen as the platform to manage the data at the CHC, Node, and DCC.

Microsoft's SQL server was chosen because the platform leveraged existing skills and tools across the network. Additional considerations included the following: MS SQL was available and in use at all participating CHCs; it is a robust database management system and offers significant flexibility over existing platforms—e.g., Informatics for Integrating Biology and the Bedside (i2b2)—for adapting the data warehouse as the needs of researchers and uses of the data warehouse evolve; furthermore, similar data infrastructure had previously been successfully developed and implemented in MS SQL by the DCC, and existing programs and procedures could be easily modified—greatly reducing start-up and development time.

In order to ensure accurate extraction of data and translation of the data elements into our CHARN data warehouse, each Node created a mapping document that cross-referenced data values in their respective systems with the data values in the CHARN data dictionary, then submitted that mapping document to the DCC. Although much has been accomplished in standardizing data entry into EHRs (due in part to federal incentives to encourage health systems to adopt Meaningful Use standards), more work is needed to standardize formats for data storage and data extraction for purposes outside of routine care delivery, such as research. Addressing how researchers would access and use EHR data for analysis and QI with CHCs necessitates standardization to allow for quicker, easier access. An example would be to apply a standard code to lab tests that are differentially labeled across the CHCs so they could be more easily aggregated, because not all EHR vendors used a standardized coding system—i.e., Logical Observation Identifiers Names and Codes (LOINC). Finally, HIPAA-limited data sets were created with patient identifiers removed, as defined by HIPAA privacy rules.[9]

The DCC set up a Secure File Transfer (SFT) site for transferring data and data schema between the Nodes and the DCC. The SFT site secures Protected Health Information (PHI) via encryption. PHI comprises potentially identifying patient information—e.g., names, addresses, dates of birth, etc.—and its use in research is restricted to protect patients' privacy. For limited data sets, like the CHARN data warehouse, the only allowable PHI fields are dates. Each Nodal data manager is assigned a unique client certificate, otherwise known as a "Digital ID," which is the digital equivalent of a handwritten signature. This is used to verify that the person is a registered user of the SFT site.

The DCC developed and posted SQL scripts to the SFT site for each Node. Scripts were customized to the appropriate version of SQL at the Node. Data managers downloaded these files and created their Node-level data warehouses by running designated scripts, which created a set of empty data warehouse tables and set up standardized Node-level variables.

Data Validation

Data provided by the Nodes' CHCs were subsequently uploaded to the warehouse, and additional scripts ran standardized data queries to validate key data characteristics before transfer to the DCC. These queries contained two types of data checks. The first level addressed conformity to the data structure and primary keys and therefore had to be successfully "passed" (i.e., run without errors) in order to load data into the tables. The second level of data checks focused on formatting, completeness of data, accuracy of data, etc., which were important checks from a data quality perspective but were not a requirement for loading the data into the tables (see Table 3).

Table 3. Data Queries

Level 1	Confirming that all data conformed to the defined SQL server field data types
	All records loaded into the tables conformed to the primary key constraints
Level 2	Data format (field level data conformity)
	Required fields (no missing data in required fields)
	Foreign key (data values exist in other tables where an explicit relationship exists)
	Valid code (values conform to a list of predefined codes), valid range (values conform to a predefined range)
	Orphan records (every record in a "nonpatient" table links to a record in the "patient" table)

After completion of level 1 and 2 queries, data were uploaded to the DCC SFT. Initial submissions for V1 were completed by all Nodes within seven months of finalizing V1 of the CHARN data dictionary. The DCC produced additional centralized data queries and cross-Node reports to carry out additional quality assessment.

CHARN Working Groups began submitting concept papers for review and approval by the Research Planning Subcommittee for proposed data-only studies and prep-to-research for research proposals within weeks of the data load and the first phase of data validation was complete. Data validation efforts continued throughout the remainder of the project funding as data issues were uncovered during new variable creation and analysis. Leaders of CHARN working groups are collaborating with the DCC to develop standardized coding and categorizations across the Nodes' data that are relevant to the conditions of interest—for example, diabetes medications categories, lab types, and results. V1 of the CHARN data warehouse represents more than 500,000 safety net patients: 58 percent female; 28 percent under age 18; 42 percent white, 25 percent Asian, and 18 percent black (African-American) (Table 4).

Multiple data complexities were discovered during the development and implementation of the CHARN data warehouse, including the following:

- A lack of standardized data classification systems for labs and medications, which made it more difficult to clearly define our populations for analysis based on either labs or medications;
- A high proportion of missing data for certain data elements;
- Complications in limiting data extracts to the seven disease cohorts;
- Implementation of an EHR system not being done at the time of the data warehouse development (One participating CHC had its data entered into an internal database structure and was able to provide most of the V1 data, but this required additional CHC effort to extract the data into the CHARN data model.);

Table 4. CHARN Patient Characteristics by Node and Overall

	AAPCHO	%	Alliance	%	Fenway	%	OCHIN	%	Total	%
Total Number of Patients	126,353	—	159,300	—	59,900	—	156,848	—	502,401	—
Percent of CHARN Total Patients	25.1%	—	31.7%	—	11.9%	—	31.2%	—	100.0%	—
Gender										
Male	50,998	40.4%	63,317	39.7%	28,417	47.4%	67,473	43.0%	197,770	41.8%
Female	75,355	59.6%	95,904	60.2%	30,498	50.9%	89,320	56.9%	267,048	57.9%
Transgender	0	0.0%	0	0.0%	958	1.6%	55	0.0%	1,013	0.2%
Unknown or missing	0	0.0%	79	0.0%	27	0.0%	0	0.0%	106	0.0%
Age										
Under 18	35,484	28.1%	43,816	27.5%	6,981	11.7%	52,552	33.5%	130,744	27.6%
18-25	18,523	14.7%	25,648	16.1%	10,744	17.9%	19,376	12.4%	68,276	14.8%
26-39	23,466	18.6%	37,129	23.3%	18,064	30.2%	35,516	22.6%	105,102	22.7%
40-64	37,382	29.6%	44,814	28.1%	20,974	35.0%	41,403	26.4%	132,937	28.8%
65-79	9,032	7.1%	6,708	4.2%	2,538	4.2%	6,574	4.2%	23,544	4.9%
80 and older	2,466	2.0%	1,180	0.7%	599	1.0%	1,426	0.9%	5,328	1.1%
Missing/bad birth date	0	0.0%	5	0.0%	0	0.0%	1	0.0%	6	0.0%
Race										
White	6,928	5.5%	74,398	46.7%	23,729	39.6%	105,016	67.0%	209,269	41.8%
Black	2,129	1.7%	56,605	35.5%	20,223	33.8%	9,249	5.9%	86,489	17.6%
American Indian	143	0.1%	3,188	2.0%	189	0.3%	1,243	0.8%	4,217	0.9%
Asian/Native Hawaiian and other Pacific Islander*	111,198	88.0%	4,021	2.5%	3,212	5.4%	6,906	4.4%	92,734	24.9%
Multiracial	859	0.7%	924	0.6%	940	1.6%	972	0.6%	5,250	0.7%
Other	5,096	4.0%	20,164	12.7%	8,434	14.1%	0	0.0%	31,342	6.7%
No Race Indicated	0	0.0%	0	0.0%	3,173	5.3%	33,462	21.3%	36,636	7.3%
Hispanic										
Hispanic or Latino	597	0.5%	45,829	28.8%	1,734	2.9%	53,823	34.3%	101,386	20.3%
Not Hispanic or Latino	0	0.0%	66,312	41.6%	4,237	7.1%	75,896	48.4%	146,445	29.1%
Missing (reported unknown)	0	0.0%	47,159	29.6%	53,929	90.0%	27,129	17.3%	128,217	25.5%
Missing (left blank)	125,756	99.5%	0	0.0%	0	0.0%	0	0.0%	89,889	25.0%

*Although the CHARN network advocates for disaggregated Asian and Native Hawaiian and other Pacific Islander race data due to the vast differences in health within these subgroups, the data were aggregated to protect patient identifiable information.

- Impossibility of linking encounters (visits) to medication orders and lab orders for some CHCs because of differing EHR structures and workflows; and
- A need for additional time to compile all required data because of multiple data sources at the CHC level.

In addition, although the V1 warehouse was intended as a "proof-of-concept" exercise and focused only on the selected conditions of interest, this restriction actually resulted in a more labor- and time-intensive effort at the CHC/Node level than if all patients had been included in the local data extractions. Also, CHARN working groups investigating other conditions or patterns (e.g., undiagnosed and untreated hypertension) were unable to use V1 to conduct their research. Finally, since patient identifiers were removed to conform to HIPPA, the DCC was not able to check for duplicate records in the central data warehouse and Nodes were responsible for identifying deleting duplicate records. Metadata were captured on the procedures used at the Nodes for this process.

Discussion

The three-year funding period for this infrastructure development project supported the development of a robust data warehouse. Critically, the project led to relationships among partners and the development of the following: (1) governance and data use policies, (2) a data model that would be efficient and effective across all 18 CHCs, and (3) basic data integrity and validation procedures.

Building collaborative, trust-based relationships among the CHCs, Nodes, and DCC requires transparent processes and assurance that these will be applied. These processes require adequate time for input, review, and revision from network participants at all levels, as well as a commitment by all participants to continue to cultivate and maintain trust, respect, and transparency in the use and further development of the data warehouse. Input from CHCs, Nodes, DCC, and academic partners also helped to foster commitment and ownership of this data warehouse. Sharing CHC patient data required trust, respect, and clear governance policies. Policies were created that outlined the use and sharing of the data that were vetted by each of the Nodes and their CHCs. For example, keeping the CHC name blinded in the warehouse was a critical factor for allowing the transfer of data to the DCC. CHCs did not want to be singled out in the analysis and reporting process. Most importantly, CHCs needed to be part of the design process to see value in this effort, in contrast to prior, imbalanced research relationships with academic partners.

The collaborative model for development of a centralized data warehouse enabled CHCs with less experience in warehouse development to learn from more experienced members of the CHARN network. Queries that the DCC built into the SQL database were run on the CHC/Node level data, and both brought up unknown issues with Nodes' local data, resulting in a better understanding of the complexities of their local EHR data. As a result, additional data queries and processes were created at the CHC level to have cleaner data extracted from their EHRs, which proved beneficial to the individual CHC as well as the CHARN network. Local CHCs and Nodes now have a standardized data warehouse and can use data queries already created for local data cleaning to investigate local QI or research questions using this shared data structure. Another benefit to the CHCs is that having access to a local copy of the data warehouse made it possible for the CHCs to more easily answer questions about their local data and to conduct their own analysis on their specific populations.[10]

It was important to set up a multiple-layer query process for data validation to ensure better data quality by identifying and resolving problems along a continuum of local and central checks.[11]

Impact of the CHARN Data Warehouse

The CHC experiences gained through this project have far-reaching impacts beyond CHARN. The validation queries and processes developed among the network to facilitate the goal of standardized data that can be used for research are critical to conducting research within and across CHCs. Partners within a data network need to come to a shared understanding of the context of the CHC data. This will enable them to create and implement appropriate data validation processes before using and publishing CHC data.

A benefit of the data warehouse from the CHC perspective was that clinics could participate in research without having their primary functions disrupted. The main deterrent to clinic participation in practice-based research is the time that it takes away from patient care. As a result, our project paved the way for low impact but active clinic engagement in research. Other benefits of the CHARN warehouse for CHCs include the following: providing CHARN clinicians with the opportunity to use the warehouse to ask questions about the data relevant to their population's needs and interests; exposing clinicians who are not particularly interested in research to findings that are relevant to their practices and that can improve their capacity to deliver high quality care to the patients they serve; and providing clinicians with a deeper understanding of the value and importance of collecting and recording data in a standardized way.

The centralized data warehouse currently supports CHARN working groups, research proposals, and prep-to-research requests. Examples include investigations of the impact of insurance transitions among diabetics, screening for HIV, viral hepatitis and tuberculosis, and providing preliminary data for proposals to study the impact of a team-care model in safety net diabetes patients and the impact of Meaningful Use on documentation and delivery of smoking cessation counseling.

Future Opportunities Created by the CHARN Data Warehouse

Additional possibilities for using the CHARN data warehouse are numerous. For example, we could combine census data on socioeconomic status with our warehouse data to study patient mobility and migration systematically.[12] Adding data from other existing sources (e.g., Medicare, in-patient data), new sources (e.g., primary data collection of patient-reported outcomes), and CHCs (e.g., from the National Association of Community Health Centers, which currently has over 1,000 CHCs), would enrich our warehouse and expand its potential uses. It will also be important to add new Meaningful Use data as standardized data fields are rolled out across EHRs. There is also the possibility of adding web-based query tools to facilitate easier access to the information contained in the warehouse and to help researchers create and prepare new research studies. A query tool would also enable CHC clinicians to conduct QI projects.

Limitations and Challenges

In particular, concerns about data quality and the suitability of EHR data for research have significant potential to have an impact on research because clinical data are not captured with the same rigor as are research data.[9] For example, as data are being analyzed, we are discovering differences in clinical practice recording methods. For example, some lab results are being recorded in discrete fields, while others are recorded in difficult-to-capture open-text fields. Also, some fields had a large percentage of missing data due to either the complexity of extracting and formatting the individual elements or to the data not being collected or codified at some of the CHCs. Another limitation was the inability to link different but related events (e.g., diagnosis, medication order, lab result) to a single encounter episode (e.g., diabetes diagnosis) via an encounter ID—likely due to differing workflows and EHR structures across CHCs. These issues point to the complexities of using EHR data for research.

As the movement to conduct more research using EHR data continues, improvements in methodology to increase the quality of EHR data, similar to those used for classical clinical trials, can be implemented. For example, the use of more standardized intake forms—requiring field completion, standardizing medications, labs tests, and diagnosis lists—as well as standardizing clinician documentation, can improve the accuracy and completeness of these data.[4] Conversely, federal reporting requirements and initiatives have enforced standardization of demographic data collection across EHR vendors. These types of requirements could have similar impacts on data collection of other clinical measures.

The selection of specific cohorts for conditions of interest resulted in a more labor-intensive effort at the CHCs/Nodes than would have been the case if all patients and their encounters had been included; extracting and processing EHR data on all patients takes considerably less time than does selecting multiple populations based on differing diagnosis codes, medication prescriptions, and lab results criteria. This specific approach, combined with the differences in CHC data structures and clinical workflows, limited researchers' ability to fully characterize and compare patient populations and cohorts against the backdrop of the full patient population. Furthermore, the potential uses of the data warehouse were limited regarding conditions that were not explicitly defined. Finally, even relatively complete EHR data only provide a partial picture of the health and health care experience in a safety net population—emergency visits and hospitalization are typically not documented and the most vulnerable patients often change locations (and health providers) frequently.

Progress in Version 2.0 of the Data Warehouse

As the focus of the warehouse expanded through the development of Version 2 (V2), we developed additional procedures to address some of the limitations encountered with V1. V2 has been implemented and includes patients seen through December 31, 2012, as well as additional clinical data; for example, we have expanded the warehouse to accommodate data on vital signs, procedures, tobacco use, problem list, enabling services, and referrals for other clinical services.

A key change to the data warehouse design was to include the full population of patients from each CHC. This change substantially simplified the extraction processes at the CHC/Node level and provides greater flexibility with regard to studying patients and cohorts, conditions, comorbidities, health care utilization, and care practices. It also facilitates interpretation of results and understanding regarding generalizability of those results to a diverse safety net population. Another improvement for V2 was to have the DCC manage the standardization of coding for lab results and medications rather than having separate efforts at each Node. In addition, in order to minimize and characterize missing data we clarified "optional" versus "nonoptional" fields in the data dictionaries and added codes to indicate types of missingness (e.g., the field exists in the EHR and was data entered as "unknown/missing," or the field exists but is not populated, or the field does not exist). Furthermore, CHARN working groups outline specific queries and targeted chart reviews relevant to their subject in order to understand the quality and appropriate uses of the data, e.g., for defining disease cohorts or care quality measures. With regard to the variable encounter structures, linking events via dates rather through a common encounter ID is a feasible, standardized alternative.

Summary

CHARN plans to refresh V2 of the data warehouse EHR data through 2013 and include additional data elements. Working groups are currently using (or planning to use) the data warehouse for the following: to develop and validate computable phenotypes for several conditions, including diabetes, cardiovascular disease, substance abuse, and HIV; to investigate the impact of insurance coverage transitions on outcomes; to identify patients at highest risk for poor outcomes in general; and to characterize care delivery with respect to cholesterol management, treatment of ischemic valve disease, and identification of undiagnosed hypertension. In addition, de-identified data sets will be released to the Nodes and their CHCs, enabling them to carry out additional local- or network-wide projects. The data warehouse will also be used in prep-to-research activities to develop new research proposals.

It is feasible, but time intensive, to create a centralized data warehouse with multiple CHC partners using different EHR systems. It is essential to allow sufficient time (1) to develop collaborative, trusting relationships among new partners with varied backgrounds, experiences, and interests; (2) to complete institutional and regulatory review processes; (3) to identify and address technical challenges associated with diverse data environments, practices and resources; and (4) to provide continuing data quality assessments to ensure data accuracy.

Acknowledgements

This work was funded by Health Resources and Services Administration (HRSA). The authors gratefully acknowledge the CHARN community health centers (CHCs), research node centers and the data coordinating center for contributing the time and data used in preparation of this article. Association of Asian Pacific Community Health Organizations (AAPCHO) CHARN affiliates include four CHCs: Asian Health Services Community Health Center, Charles B. Wang Community Health Center, Waianae Coast Comprehensive Health Center, Waimanalo Health Center, and their academic affiliate; University of California, Los Angeles. Alliance of Chicago Community Health Services (Alliance) CHARN affiliates include eight CHCs: Erie Family Health Center, Inc., Glide Health Services, Heartland Health Outreach, Howard Brown Health Center, Near North Community Health Center, North Country HealthCare, PCC Community Wellness, Children's Hospital, NU, and their academic affiliate; Northwestern University. Fenway Health CHARN affiliates include three CHCs: Chase-Brexton Health Services, Beaufort-Jasper-Hampton Comprehensive Health Services, Fenway Health, and their academic affiliate; University of Washington. OCHIN, Inc. CHARN affiliates include four CHCs: Open Door Community Health Centers, Virginia Garcia Memorial Health Centers, Multnomah County Health Department, OHSU Richmond Family Medicine Clinic, and their academic affiliate; Oregon Health & Science University. The CHARN Data Coordinating Center (DCC) is located at the Kaiser Permanente Center for Health Research. COMPETING INTERESTS: There are no competing interests. FUNDING: This work was funded by Health Resources and Services Administration (HRSA). FIGURE LEGENDS: Figure 1: Sample IRB Process

References

1. Cleland CL, Tully MA, Kee F, Cupples ME. The effectiveness of physical activity interventions in socio-economically disadvantaged communities: a systematic review. Prev Med 2012 Jun;54(6):371-80.
2. Etheredge LM. A rapid-learning health system. Health Aff (Millwood) 2007 Mar;26(2):w107-w118.
3. Friedman CP, Wong AK, Blumenthal D. Achieving a nationwide learning health system. Sci Transl Med 2010 Nov 10;2(57):57cm29.
4. Kitahata MM, Rodriguez B, Haubrich R, Boswell S, Mathews WC, Lederman MM, et al. Cohort profile: the Centers for AIDS Research Network of Integrated Clinical Systems. Int J Epidemiol 2008 Oct;37(5):948-55.
5. Sittig DF, Hazlehurst BL, Brown J, Murphy S, Rosenman M, Tarczy-Hornoch P, et al. A survey of informatics platforms that enable distributed comparative effectiveness research using multi-institutional heterogenous clinical data. Med Care 2012 Jul;50 Suppl:S49-S59.
6. Wagner EH, Greene SM, Hart G, Field TS, Fletcher S, Geiger AM, et al. Building a research consortium of large health systems: the Cancer Research Network. J Natl Cancer Inst Monogr 2005;(35):3-11.
7. Centers for Disease Control and Prevention. Vaccine Safety Datalink Project (VSD). http://www. cdc gov/od/science/iso/research_activities/vsdp htm 2006 October 23 [cited 2006 Nov 7];
8. American Medical Group Association. Anceta/AMGA Collaborative Data Warehouse: Background. http://www.amga org/QMR/Anceta/bg anceta asp 2006 December [cited 2006 Dec 12];
9. United States Department of Health and Human Services. OCR Privacy Rule Summary, 1996, revised 2003. 2003.
10. Li V, Weir R, Quach T, Gillespie S, McBurnie M, Oster A, et al. Building a Community Health Center Data Warehouse to Promote Patient Centered Research in the Asian American, Native Hawaiian, and Pacific Islanders Population. AAPI Nexus. In review 2014.
11. Bauck A, Gillespie S. Integrating layered data validation for multi-site clinical trials. SOCRA Source 2010 Aug;65:20-4.
12. Roos LL, Nicol JP. A research registry: uses, development, and accuracy. J Clin Epidemiol 1999 Jan;52(1):39-47.