Published in final edited form as:
Med Care. 2012 July ; 50(Suppl): S49–S59. doi:10.1097/MLR.0b013e318259c02b.

A survey of informatics platforms that enable distributed comparative effectiveness research using multi-institutional heterogeneous clinical data

Dean F. Sittig, PhD[1], **Brian L. Hazlehurst, PhD**[2], **Jeffrey Brown, PhD**[3], **Shawn Murphy, MD, PhD**[4], **Marc Rosenman, MD**[5], **Peter Tarczy-Hornoch, MD**[6], and **Adam B. Wilcox, Ph.D**[7]

[1]School of Biomedical Informatics, University of Texas Health Science Center, Houston, TX

[2]Kaiser Permanente Center for Health Research, Portland, OR

[3]Department of Population Medicine, Harvard Pilgrim Health Care Institute and Harvard Medical School, Boston, MA

[4]Partners Healthcare System and Harvard Medical School, Boston, MA

[5]Regenstrief Institute and Indiana University School of Medicine, Indianapolis, IN

[6]Division of Biomedical and Health Informatics, University of Washington, Seattle, WA

[7]Department of Biomedical Informatics Columbia University, NY

Abstract

Comparative Effectiveness Research (CER) has the potential to transform the current healthcare delivery system by identifying the most effective medical and surgical treatments, diagnostic tests, disease prevention methods and ways to deliver care for specific clinical conditions. To be successful, such research requires the identification, capture, aggregation, integration, and analysis of disparate data sources held by different institutions with diverse representations of the relevant clinical events. In an effort to address these diverse demands, there have been multiple new designs and implementations of informatics platforms that provide access to electronic clinical data and the governance infrastructure required for inter-institutional CER. The goal of this manuscript is to help investigators understand why these informatics platforms are required and to compare and contrast six, large-scale, recently funded, CER-focused informatics platform development efforts. We utilized an 8-dimension, socio-technical model of health information technology use to help guide our work. We identified six generic steps that are necessary in any distributed, multi-institutional CER project: data identification, extraction, modeling, aggregation, analysis, and dissemination. We expect that over the next several years these projects will provide answers to many important, and heretofore unanswerable, clinical research questions.

Keywords

Methods; Comparative Effectiveness Research; Organization and Administration; Medical Informatics; Methods

Address for correspondence: Dean F. Sittig, PhD UT-Memorial Herman Center for Healthcare Quality & Safety 6410 Fannin St, UTPB 1100.43 Houston, TX 77030 W: 713-500-7977 Dean.F.Sittig@uth.tmc.edu.

Introduction

The American Recovery and Reinvestment Act of 2009 provided $1.1 billion for Comparative Effectiveness Research (CER)[1]. The goal of CER is to generate new evidence on the potential effectiveness, benefits, and harms of different treatments, diagnostics, preventions, and care models under "real world" conditions. Widespread adoption of CER has potential to radically change healthcare. CER also places enormous demands on existing informatics research infrastructure[2], as it requires aggregation and analysis of disparate data held by different institutions, each with its own representation of relevant events and accountabilities for protecting data as a matter of patient confidentiality and business operations.

Currently, most data manipulations are performed using non-coordinated applications (e.g., data collection forms, electronic health records [EHRs], research databases, condition-specific registries, and statistical analyses) with disjointed institutional control. In an effort to address these demands, there have been new designs and implementations of informatics platforms that provide access to electronic clinical data and the governance required for inter-institutional comparative effectiveness research[3,4,5,6]. Briefly, a "platform" is a suite of interconnected, coordinated applications, together with the operational environment that hosts those applications.

The goal of this manuscript is to compare and contrast six large-scale, projects that are either developing or extending existing informatics platforms for CER. Rather than compare the informatics platforms at an abstract level, we focus on specific CER projects that provide implementations of informatics platforms and highlight design requirements and solutions.

The following sections provide an overview of the projects surveyed.

Washington Heights/Inwood Informatics Infrastructure for Comparative Effectiveness Research (WICER)

WICER is creating infrastructure to facilitate patient-centered outcomes research in Washington Heights, NY. The project facilitates comprehensive understanding of populations by leveraging data from existing EHRs, and combining data from institutions representing various healthcare processes. For example, it includes data from hospitals, clinics, specialists, homecare agencies and long-term care facilities. It also includes survey data from community residents with assessments on socioeconomic status, vital statistics, support networks, health and illness perceptions, quality of life, and health literacy. Data from multiple sources are merged in a data warehouse, where deeper analysis is performed by clinical and public health researchers. WICER investigators are using the infrastructure and methods on three clinical trials in hypertension care around diagnosis, adherence to therapy, and care management.

Scalable PArtnering Network for Comparative Effectiveness Research: Across Lifespan, Conditions, and Settings (SPAN)

The HMO Research Network (HMORN) is a consortium of 19 Health Plans with formal, research capabilities[7]. SPAN, a project within the HMORN, uses its Virtual Data Warehouse (VDW) to provide a standardized, federated data system across 11 partners, to address CER in ADHD and obesity[8]. The VDW consists of commonly-defined linked tables within each health plan that capture medical care utilization, clinical data, health plan enrollment information, demographics, detailed inpatient and outpatient encounter information, outpatient pharmacy dispensing data, laboratory test results and vital signs[9]. The VDW is augmented with State and local cancer registry information on date and cause of death for health plan members. Each plan maintains control of individual VDW data files and does

not have access to files held by other HMORN sites. All HMORN participants must be capable of running - without modification - SAS programs distributed by other sites to execute against their local VDW. SPAN is pioneering use of a new platform – PopMedNet™ – that facilitates creation, operation, and governance of multi-site, distributed health data networks[10].

Enhancing Clinical Effectiveness Research with Natural Language Processing of EHR Data – CER-HUB

The CER-HUB is an Internet-based platform for conducting CER. A central function of CER-HUB is facilitating (through online, interactive tools) development of a shared, data processor library that can be downloaded by registered researchers to provide uniform, standardized coding of both free-text and structured clinical data. This shared library permits researchers to assess data on clinical effectiveness in multiple healthcare areas and gain access to information locked in freetext notes. Using CER-HUB, researchers collaboratively build software applications (MediClass applications[11]) that will process EHR data within their respective healthcare organizations, creating standardized datasets that can be pooled to address specific CER protocols. Participating researchers contribute IRB-approved, limited data sets to a centralized coordinating center to be pooled with data similarly processed from other healthcare organizations to answer CER questions. The CER-HUB is being used to conduct 2 CER studies addressing effectiveness of medication for controlling asthma and of smoking cessation counseling services, across 6 geographically-distributed and demographically-diverse health systems. Researchers and data providers for these initial studies come from 3 Kaiser health plans (Northwest, Hawaii, and Georgia regions), one consortium of Federally Qualified Health Centers located primarily along the west coast (OCHIN, Inc), one Veterans Administration service region (Puget Sound VA in Washington), and an integrated network of hospitals and physicians in the greater Dallas/ Fort Worth area (Baylor Health Care System).

The Partners Research Patient Data Registry (RPDR)

The RPDR is an enterprise data warehouse combined with a multi-faceted user interface that enables clinical research and CER across Partners Healthcare in Boston, MA. The RPDR is used to recruit patients for clinical trials, and to perform active surveillance. It amasses data from billing, decision support, and EHRs in the Partners' system. Data are available to researchers through a drag-and-drop web Query Tool[12] allowing users to construct exploratory, ad hoc, queries for hypothesis generation from structured data, and to get aggregate totals and graphs of age, race, gender and vitals. A utility exists for finding matched controls for patients. Requests can be made for detailed data on patients identified through the query tool with proper IRB authorization through an automated wizard. The RPDR has proven useful for gathering clinical trial cohorts, and for CER. This strategy was later adopted as the core of “Informatics for Integrating Biology and the Bedside” (i2b2)[13]. The RPDR was first released in December, 1999 and has been in production at multiple sites since March, 2002.

The Indiana Network for Patient Care (INPC) Comparative Effectiveness Research Trial of Alzheimer's Disease Drugs (COMET-AD)

INPC was begun in 1994 as an experiment in community-wide health information exchange serving five major hospitals in Indianapolis, IN. Today, it includes data from hospitals and payers statewide[14,15,16]. Entities participating in INPC submit patient registration records, laboratory test results, diagnoses, procedure codes, and other data for various types of healthcare encounters. Data are also obtained from health departments and a pharmacy benefit manager consortium. Data are standardized (e.g., laboratory test results are mapped to LOINC[17] with common units of measure) to the extent possible, prior to storage in a

central repository. Data for a patient with visits to multiple INPC institutions can be linked using a patient matching algorithm. The COMET-AD project is using data from INPC to monitor healthcare processes and outcomes and to build systems to monitor patients for adverse drug events. The project also involves building infrastructure and workflows to support integration of biospecimen results with clinical data from the INPC.

The Surgical Care Outcomes Assessment Program Comparative Effectiveness Research Translation Network (SCOAP-CERTN)

The goal is to assess how well an existing statewide quality assurance and quality improvement registry (i.e., the Surgical Care Outcomes Assessment Program) can be leveraged to perform CER. The SCOAP-CERTN leverages relationships built collaboratively in SCOAP to improve surgical care and outcomes and aims to build infrastructure for streamlined, electronic data abstraction from EHRs, patient reported outcomes, and healthcare payments across hospitals. Through a partnership with Microsoft Health Solutions Group (Redmond, WA), SCOAP-CERTN is identifying ways to maximize automatic capture of data from EHRs, to:

- Allow longitudinal clinical data capture across healthcare encounter types (i.e., surgical, interventional);
- Reduce clinical workflow and staffing burdens for maintenance of the SCOAP registry at participating hospitals;
- Provide capacity and interoperability to incorporate outpatient care delivery into SCOAP.

In addition, SCOAP developers plan to add functions to capture patient reported outcomes for research and quality improvement evaluation. The primary informatics goal is to assess how, and to what degree, the collection of SCOAP-CERTN measures can be automated across sites.

Conceptual model for CER platform evaluation

Designing, developing, implementing, and using health information technology (HIT) within healthcare delivery systems is a complex, socio-technical challenge. To provide a theoretical basis for our comparison of six CER informatics platforms we adapted an 8-dimension, socio-technical model of safe and effective HIT use[18]. This model prescribes attention to: (1) appropriate ***hardware/software***, (2) a spectrum of ***clinical content*** ranging from case narrative, to standard vocabularies, to algorithms representing best practices, (3) ***human-computer interfaces*** enabling productive interactions with technology, (4) ***personnel*** who develop systems and how systems meet the needs of users in their social contexts, (5) ***workflow and communications*** (both between people and technology components) required to accomplish tasks using the technology, (6) organizational policies, procedures, culture, and environment that prescribe and govern how and where things happen and who is responsible, (7) ***external rules, regulations, and pressures*** which shape these organizational constraints, and (8) system ***measurement and monitoring*** which ensures adequate performance for primary intended use cases, i.e., the conduct of CER.

These eight constructs[18] are used to investigate and evaluate aspects of CER platform design and implementation by ensuring that both the social as well as the technical aspects are considered. Failure to consider who will use the applications, how they will use them, and why they are necessary often leads to sub-optimal technology design and utilization.

Methods

Data sources

We developed a written survey and sent it to informatics experts representing six large CER projects focusing on the design, development, and use of multi-institutional informatics platforms. Projects were selected by convenience, yet they are representative of vastly different approaches researchers have taken to address numerous CER challenges.

Survey instrument

We (DFS, BLH) developed a 2-page, open-ended survey that highlighted project-specific similarities and differences. We created 2 – 8 questions within each of the 8 dimensions to ensure that all important aspects were captured[18]. For example, within **Workflow/ communication** we asked, "How do data get into your warehouse?" and "What stages do the data go through?" Similarly, within the **Hardware/Software** dimension we asked, "What computing infrastructure is required to run your system?"

Data collection and analysis

Completed surveys were returned by e-mail and checked for completeness. DFS and BLH read through the 6–10 page responses from each of the co-authors looking for key concepts highlighting project similarities and differences. After review and discussion, it became clear that the following 4 dimensions of the 8-dimension model were the key differentiators: content or data (Table 2); workflow/communication regarding how data moved from sources to analysis (Table 3); people (investigators, data programmers, research analysts, managers) involved in the projects (Table 4); and organizational policies, procedures, and culture (Table 5). We extracted data items to fill-in the tables from surveys. In addition to survey items, two authors (DFS, BLH) gathered information regarding project descriptions and funding from websites and journal articles (Table 1). Drafts of completed tables were sent to co-authors for review.

Results

All projects implement six generic data processing steps necessary for distributed, multi-institutional CER projects:

- Identification of applicable data within health care transaction systems,
- Extraction to a local data warehouse for staging,
- Modeling data to enable common representations across multiple health systems,
- Aggregation of data according to this common data model,
- Analysis of data to address research questions,
- Dissemination of study results.

All projects performed these activities, although there were variations in how (real-time aggregation of HL-7 transactions vs. nightly or as-needed extraction, transformation, and loading), where (local site vs. coordinating center), and with what tools (web-based query interfaces for researchers vs. tools to develop Natural Language Processing (NLP) modules).

Table 2 compares data sources, types, models, and handling of duplicate patients. All projects collected data from multiple sources (i.e., hospitals, clinics, billing, long-term care) and included different data types (eg, numeric test results, ICD-9-encoded problems, and free text progress notes). Only three projects used a "master patient index" that enabled them

to combine data from patients who received treatment at different organizations. All projects used different, and sometimes multiple, data storage and manipulation formats ranging from SAS tables to XML-based documents to relational databases.

Table 3 provides a comparison of data flow and transformation, from local EHRs to aggregated analyses. The most important differences highlighted in Table 3 pertain to when patient-identifiable data leave local sites. In two projects, this occurs immediately following extraction from the local transaction-based clinical or administrative systems. In SPAN and CER-HUB, transfer of “raw” patient-identifiable data never occurs (i.e., all data are processed at the local site by data analysis programs that are distributed from the central site, and only data conforming to protocol-specific Limited Data Sets are shared). Only three sites had any form of natural language processing capability; the other sites relied solely on numeric or coded data elements.

Also of interest in Table 3 is the state of data analysis tools offered by projects. All projects are working on “user-friendly” tools to facilitate researchers' direct access to data via ad hoc queries, while concurrently meeting multi-institutional requirements for protecting patient data and corporate business interests. To date, only the RPDR has a working version.

Table 4 describes key personnel. The most important difference is that some projects either have or are working on Internet-based interfaces that allow non-technical investigators to perform a limited set of data queries and analyses on the combined data set. For example, the SPAN project currently requires all queries be coded as SAS programs and sent to the local site where they are executed and the results returned after manual review; SPAN is beta-testing an internet-based approach using the PopMedNet architecture to allow non-technical users to issue queries.

Table 5 provides a comparison of project governance and internal organizational policies and procedures. All projects have an oversight committee; most consisting of representatives from all sites involved in the project. Often this committee is responsible for governing all aspects of data ownership and sharing, project membership, and publication rights and responsibilities.

Discussion

We compared six large CER projects and described how they employ informatics platforms to provide data aggregation, analysis, and research management capabilities. Many of these platforms were originally designed and developed to address widely different healthcare, organizational and research objectives; only after significant amounts of work had been completed were they transitioned to focus on CER. For example, the RPDR was originally designed to answer the question, “How many patients with a specific set of characteristics have we treated within our integrated delivery network?” On the other hand, INPC and WICER started as a means of improving the quality and efficiency of care in large metropolitan areas by creating centrally-managed health information exchanges (HIEs). Similarly, SCOAP-CERTN started as a registry to improve surgical outcomes and efficiency. SPAN (and to a lesser extent CER-HUB) build upon existing research networks comprised of similarly organized and managed, large, integrated health plans.

CER requires comprehensive data on patients

Different data types are required to create complete, patient-centered views of patient's medical history. The surveyed projects demonstrate that creating a useful CER platform requires enormous amounts, and a large variety, of data. To access these data, CER investigators need to collect them from as many different sources within their participating

organizations as possible. Therefore, we see researchers collecting data from inpatient and outpatient EHRs (including the text narrative of clinical encounters), from billing and ancillary systems such as laboratory, pharmacy, and radiology. In addition, it is important to collect data that document that patients actually received the care that was ordered, so we see organizations collecting pharmacy dispensing and patient-reported data when available. This vast array of data, while large, is nearly always incomplete (i.e., they generate sparse representations in a large-dimensional space of patient care facts in the real world) and methods which use these data must be appropriate to the task of measuring health status and care events with available data.

CER requires data on populations from multiple organizations

Researchers need to aggregate data from multiple organizations to have enough information to identify small differences, address bias, perform subgroup analyses, improve generalizability, allow evaluation of demographic and geographic variation, and identify rare events. Therefore, CER informatics platforms must be able to extract and collect data from many different organizations to compile as complete a view of conditions, treatments, and individuals as possible. Towards that end we see investigators working to include data from multiple organizations, pursuing non-traditional research data sources, such as long-term care facilities, home and public health agencies, and attempting to reliably ascertain patients' socioeconomic status on a widespread basis.

A key requirement for data collection across healthcare provider organizations located in the same geographic region is the need to merge data from the same patient who has received healthcare services and had clinical data captured at multiple institutions. Such efforts require a community-wide master patient index that identifies patients based on multiple demographic data (e.g., first name, last name, date of birth, gender, social security or telephone numbers) and keeps track of all patient identifiers used by various participating organizations to create a single, master patient identifier[19]. To date, only the CER projects that were built on top of existing health information exchange platforms designed for patient care have tackled this extraordinarily difficult problem[20, 21], but in the future patient matching capabilities will be a critical success factor.

CER requires data extraction, modeling, aggregation and analysis methods and tools

Researchers must be able to extract required data from various electronic data systems, map data types to standardized clinical representations, and analyze it. Design and development of these “mapping” applications is one of the biggest challenges in any multi-institutional research project, because it is often the case that different organizations refer to the same activity, condition, or even procedure by different names, and the same names can refer to different things across institutions. Further, even with accurate mapping it is difficult for researchers to fully appreciate local idiosyncratic data issues (e.g., non-random incomplete data capture) without active engagement of local data experts.

Furthermore, conducting CER is a complex undertaking requiring people with widely different skills, often in different locations and subject to different organizational policies and practices. In an attempt to reduce potential for misunderstanding in collaboration processes, platform developers are working to create powerful, user-friendly tools for data extraction, manipulation, and analysis. These tools are being designed so CER project staff, who often have little informatics training, can perform their tasks more efficiently. In addition, several projects are developing tools to help researchers make sense of highly variable and clinically-rich free-text notes documenting patient care.

CER must conform to local organization's internal governance and Institutional Review Board's (IRB) rules and local and federal legislation

The social, legal, ethical, and political challenges involved in setting up and conducting large, multi-institutional CER projects must not be underestimated. Friedman et al. stated that "organizations are understandably reluctant to move data beyond their own boundaries absent a clear and specific need to do so, and patients will be less likely to consent to allow this to happen."[22] Therefore, in addition to providing the technical infrastructure required to collect, standardize, normalize, and analyze disparate data, informatics platforms must conform to local organization's internal governance and IRB's rules and regulations as well as existing state and federal guidelines. One design to address use of protected health information is to retain physical control of raw data while providing for their aggregation as limited data sets to answer specific questions. Other ways in which projects have accommodated inter-institutional governance issues include standardizing data models across the project; limiting access to authorized personnel while facilitating remote access; restricting the types of queries that can be executed and masking patient-specific, identifiable data; and logging all data transactions and access activities. As rules, regulations, and guidelines evolve (eg, proposed Common Rule revision[23]) CER platforms and governance processes must evolve accordingly.

Summary and Conclusion

CER stands to transform the current healthcare delivery system by identifying which therapies, procedures, preventive tests, and healthcare processes are most effective from the standpoints of cost, quality, and safety. State-of-the-art informatics platforms are necessary to carry out this type of research across organizations with disparate patient populations, health information systems, data types, and local governance structures.

We used an 8-dimension, socio-technical model to develop a survey enabling us to compare and contrast informatics platforms that are under development or in use in six large CER efforts. Based on the data we collected, we identified six generic steps necessary in any distributed, multi-institutional CER project: data identification, extraction, modeling, aggregation, analysis, and dissemination.

We conclude that all of the informatics platforms for CER studied are on their way to creating the socio-technical infrastructure required to enable researchers from multiple institutions to conduct high-quality, cost-effective CER. We expect that over the next several years, these projects will provide answers to many important CER questions that in the past were virtually inaccessible. In addition, we expect many more CER-focused informatics research platforms to be designed, developed, and tested as the fields of informatics and CER continue to evolve.

Acknowledgments

Dr. Sittig is supported in part by a grant from the National Library of Medicine R01- LM006942 and by a SHARP contract from the Office of the National Coordinator for Health Information Technology (ONC #10510592).

Dr. Hazlehurst's work is supported in part by grants from the National Library of Medicine (R21LM009728), and the Agency for Healthcare Research and Quality (AHRQ) (R01HS019828, R18HS18157).

Dr. Brown is supported in part by the AHRQ grant 1R01HS019912

Dr. Murphy is supported by grants from the National Institute of Health U54LM008748, UL1RR025758, U24RR025736 and by a grant from the Office of the National Coordinator 90TR0001/01.

In this work Dr. Rosenman was supported by a grant from AHRQ (R01HS019818)

Dr. Tarczy-Hornoch is supported in part by AHRQ 1 R01 HS 20025-01 “Surgical Care and Outcomes Assessment Program (SCOAP) Comparative Effectiveness Research Translation Network (CERTN)” and by NIH NCRR 1 UL1 RR 025014 “Institute of Translational Health Sciences”.

Dr. Wilcox is supported in part by AHRQ grant R01 HS019853-01, Washington Heights/Inwood Informatics Infrastructure for Community-Centered Comparative Effectiveness Research (WICER)

We also thank Andrea Bradford, PhD for editorial assistance.

References

1. The American Recovery and Reinvestment Act of 2009. Public Law 111-5-February 17, 2009. Available at: http://www.gpo.gov/fdsys/pkg/PLAW-111publ5/pdf/PLAW-111publ5.pdf
2. VanLare JM, Conway PH, Rowe JW. Building academic health centers' capacity to shape and respond to comparative effectiveness research policy. Acad Med. Jun; 2011 86(6):689–94. [PubMed: 21512371]
3. Stang PE, Ryan PB, Racoosin JA, Overhage JM, Hartzema AG, Reich C, Welebob E, Scarnecchia T, Woodcock J. Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership. Ann Intern Med. Nov 2; 2010 153(9):600–6. [PubMed: 21041580]
4. Ohno-Machado L, Bafna V, Boxwala AA, Chapman BE, Chapman WW, Chaudhuri K, Day ME, Farcas C, Heintzman ND, Jiang X, Kim H, Kim J, Matheny ME, Resnic FS, Vinterbo SA, iDASH team. iDASH: integrating data for analysis, anonymization, and sharing. J Am Med Inform Assoc. Nov 10.2011 Epub ahead of print.
5. Behrman RE, Benner JS, Brown JS, McClellan M, Woodcock J, Platt R. Developing the Sentinel System--a national resource for evidence development. N Engl J Med. Feb 10; 2011 364(6):498–9. [PubMed: 21226658]
6. Payne P, Ervin D, Dhaval R, Borlawsky T, Lai A. TRIAD: The Translational Research Informatics and Data management grid. Appl Clin Inf. 2011; 2:331–344. http://dx.doi.org/10.4338/ACI-2011-02-RA-0014.
7. Greene SM, Hart G, Wagner EH. Measuring and improving performance in multicenter research consortia. J Natl Cancer Inst Monogr. 2005; (35):26–32. [PubMed: 16287882]
8. Toh S, Platt R, Steiner JF, Brown JS. Comparative-effectiveness research in distributed health data networks. Clin Pharmacol Ther. Dec; 2011 90(6):883–7. doi: 10.1038/clpt.2011.236. [PubMed: 22030567]
9. Hornbrook MC, Hart G, Ellis JL, Bachman DJ, Ansell G, Greene SM, Wagner EH, Pardee R, Schmidt MM, Geiger A, Butani AL, Field T, Fouayzi H, Miroshnik I, Liu L, Diseker R, Wells K, Krajenta R, Lamerato L, Neslund Dudas C. Building a virtual cancer research organization. J Natl Cancer Inst Monogr. 2005; (35):12–25. [PubMed: 16287881]
10. Brown JS, Holmes JH, Shah K, Hall K, Lazarus R, Platt R. Distributed Health Data Networks: A Practical and Preferred Approach to Multi-Institutional Evaluations of Comparative Effectiveness, Safety, and Quality of Care. Med Care. Jun; 2010 48((6) Suppl 1):S45–51. [PubMed: 20473204]
11. Hazlehurst B, Frost HR, Sittig DF, Stevens VJ. MediClass: A system for detecting and classifying encounter-based clinical events in any electronic medical record. J Am Med Inform Assoc. Sep-Oct;2005 12(5):517–29. [PubMed: 15905485]
12. Murphy SN, Gainer VS, Chueh H. A Visual Interface Designed for Novice Users to find Research Patient Cohorts in a Large Biomedical Database. Journal of the American Medical Informatics Association, Symposium Supplement. 2003:489–493. PMID: 14728221.
13. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc. 2010; 17(2):124–30. [PubMed: 20190053]
14. Overhage JM, Tierney WM, McDonald CJ. Design and implementation of the Indianapolis Network for Patient Care and Research. Bull Med Libr Assoc. 1995; 83:48–56. [PubMed: 7703939]
15. McDonald CJ, Overhage JM, Barnes M, Schadow G, Blevins L, Dexter PR, Mamlin B, INPC Management Committee. The Indiana network for patient care: a working local health information

infrastructure. An example of a working infrastructure collaboration that links data from five health systems and hundreds of millions of entries. Health Aff (Millwood). 2005; 24:1214–20. [PubMed: 16162565]

16. Zhu VJ, Tu W, Rosenman MB, Overhage JM. Facilitating Clinical Research through the Health Information Exchange: Lipid Control as an Example. AMIA AnnuSymp Proc. Nov 13.2010 :947–51. 2010.
17. McDonald CJ, Huff SM, Suico JG, Hill G, Leavelle D, Aller R, Forrey A, Mercer K, DeMoor G, Hook J, Williams W, Case J, Maloney P. LOINC, a universal standard for identifying laboratory observations: a 5-year update. Clin Chem. Apr; 2003 49(4):624–33. [PubMed: 12651816]
18. Sittig DF, Singh H. A New Socio-technical Model for Studying Health Information Technology in Complex Adaptive Healthcare Systems. Quality & Safety in Healthcare. Oct; 2010 19(Suppl 3):i68–74. doi:10.1136/qshc.2010.042085.
19. AHIMA. Fundamentals for Building a Master Patient Index/Enterprise Master Patient Index (Updated). *Journal of AHIMA* (Updated September 2010). Available at: http://library.ahima.org/xpedio/groups/public/documents/ahima/bok1_048389.hcsp?dDocName=bok1 _048389.
20. Adragna L. Implementing the enterprise master patient index. J AHIMA. Oct; 1998 69(9):46–8. 50, 52. [PubMed: 10187470]
21. McDonald CJ, Overhage JM, Tierney WM, Dexter PR, Martin DK, Suico JG, Zafar A, Schadow G, Blevins L, Glazener T, Meeks-Johnson J, Lemmon L, Warvel J, Porterfield B, Warvel J, Cassidy P, Lindbergh D, Belsito A, Tucker M, Williams B, Wodniak C. The Regenstrief Medical Record System: a quarter century experience. Int J Med Inform. Jun; 1999 54(3):225–53. [PubMed: 10405881]
22. Friedman CP, Wong AK, Blumenthal D. Achieving a nationwide learning health system. Sci Transl Med. Nov 10.2010 2(57):57cm29.
23. [Accessed 9/13/11] Regulatory Changes in Advanced Notice of Proposed Rulemaking: Comparison of Existing Rules with Some of the Changes Being Considered. at: http://www.hhs.gov/ohrp/humansubjects/anprmchangetable.html
24. Health & Human Services Research Awards: Use of Recovery Act and Patient Protection and Affordable Care Act Funds for Comparative Effectiveness Research. U.S. Government Accountability Office; Washington, D.C.: Jun 14. 2011 Available at: http://www.gao.gov/new.items/d11712r.pdf
25. Tatonetti NP, Denny JC, Murphy SN, Fernald GH, Krishnan G, Castro V, Yue P, Tsao PS, Kohane I, Roden DM, Altman RB. Detecting drug interactions from adverse-event reports: interaction between paroxetine and pravastatin increases blood glucose levels. Clin Pharmacol Ther. Jul; 2011 90(1):133–42. doi: 10.1038/clpt.2011.83. Epub 2011 May 25. [PubMed: 21613990]
26. Kurreeman F, Liao K, Chibnik L, Hickey B, Stahl E, Gainer V, Li G, Bry L, Mahan S, Ardlie K, Thomson B, Szolovits P, Churchill S, Murphy SN, Cai T, Raychaudhuri S, Kohane I, Karlson E, Plenge RM. Genetic basis of autoantibody positive and negative rheumatoid arthritis risk in a multi-ethnic cohort derived from electronic health records. Am J Hum Genet. Jan 7; 2011 88(1): 57–69. [PubMed: 21211616]
27. CHAPTER 4: VIRTUAL DATA WAREHOUSE (VDW) In Collaboration Toolkit: A guide to multicenter research in the HMO Research Network. 2011. Available at: www.hmoresearchnetwork.org/resources/toolkit/HMORN_CollaborationToolkit.pdf#4
28. Friedman C, Shagina L, Lussier Y, Hripcsak G. Automated encoding of clinical documents based on natural language processing. J Am Med Inform Assoc. Sep-Oct;2004 11(5):392–402. [PubMed: 15187068]
29. Baorto D, Li L, Cimino JJ. Practical experience with the maintenance and auditing of a large medical ontology. J Biomed Inform. Jun; 2009 42(3):494–503. [PubMed: 19285569]
30. Zeng QT, Goryachev S, Weiss S, Sordo M, Murphy SN, Lazarus R. Extracting principal diagnosis, comorbidity and smoking status for asthma research: evaluation of a natural language processing system. BMC Med Inform Decis Mak. Jul 26.2006 6:30. [PubMed: 16872495]
31. Perlis RH, Iosifescu DV, Castro VM, Murphy SN, Gainer VS, Minnier J, Cai T, Goryachev S, Zeng Q, Gallagher PJ, Fava M, Weilburg JB, Churchill SE, Kohane IS, Smoller JW. Using electronic medical records to enable large-scale studies in psychiatry: treatment resistant

depression as a model. Psychol Med. Jan; 2012 42(1):41–50. Epub 2011 Jun 20. [PubMed: 21682950]

32. Friedlin J, McDonald CJ. Using a natural language processing system to extract and code family history data from admission reports. AMIA Annu Symp Proc. 2006:925. [PubMed: 17238544]
33. HMO Research Network Collaboration Toolkit. Available at: http://www.hmoresearchnetwork.org/resources/collab_toolkit.htm
34. CERHUB Mission and Membership Overview. Available at: http://www.cerhub.org/mission.html

Table 1

Overview of the six informatics platform-based comparative effectiveness research projects analyzed in this manuscript.

CER Project Name	Project PI/Representative	Project-specific Funding[24]	Geographic diversity	Project Website	Areas of study (# patients involved)	Number of Participating Institutions	Stage of the infrastructure and project
SPAN	Daley/ Brown	$8,272,272	USA-based Hawaii - Georgia	popmednet.org (software platform); Hmoresearch network.org has information on the VDW	Attention-deficit hyperactivity disorder (~240,000 adolescents); obesity (~2.4million)	11 US-based health systems	**Infrastructure:** VDW in production 10 yrs, SPAN is adding new variables; developing query enhancements to the PopMedNet platform **Project:** Developing research protocols, initial data assessments
WICER	Wilcox/Wilcox	$8,855,607	Limited to Washington Heights / Inwood community (New York City, NY)	Wicer.org	Population in research data warehouse (~200,000); population cohort analysis using community surveys (12,000); care management, new diagnostic algorithms, 3 clinical trials of (600 patients total)	4 local organizations	**Infrastructure:**Under development; **Project:** In planning stages
CER-HUB	Hazlehurst/Hazlehurst	$8,735,759	USA-based Hawaii - Georgia	cerhub.org	Asthma (200,000); smoking cessation (500,000)	6 large health systems – 3 HMOs, 1 VA, 1 consortium of FQHCs, 1 large regional physician/ hos pital network	**Infrastructure:**Under development; **Project:** Two studies underway
RPDR	Murphy/ Murphy	~$18M over past 12 years	Limited to Boston, MA area	rc.partners.org/rpdr, I2b2.org	Supports about 400 studies/year; 4 studies have been completed,[25,26,30,31]	7 local hospitals	**Infrastructure:**System in operation since 1999 **Project:** Hundreds of studies underway
INPC COMET-AD	Dexter and Boustani/Rosenman	$8,422,410	Indianapolis, IN metropolitan area	www.regenstrief.org/medinformatics/inpc	Drug treatment of Alzheimer's disease (300)	5 major hospital systems in Indianapolis area	**Infrastructure:** HIE used routinely for patient care and research **Project:** Beginning patient enrollment
SCOAP-CERTN	Flum/Tarczy-Hornoch	$11,690,974	Washington state	www.scoap.org	Primary: peripheral artery disease; Secondary: under development	10 institutions planned (3 agreed)	**Infrastructure:** WA statewide QA/QI registry in use at 50+ sites for 4 + years

CER Project Name	Project PI/Representative	Project-specific Funding[24]	Geographic diversity	Project Website	Areas of study (# patients involved)	Number of Participating Institutions	Stage of the infrastructure and project
							Project: Recruiting 7 additional sites (3 agreed)

Abbreviations: Health Maintenance Organization (HMO); Virtual Data Warehouse (VDW); Quality Assurance/Quality Improvement (QA/QI); Federally-Qualified Health Care facilities (FQHCs)

Table 2

Comparison of data sources, types, models, and handling of duplicate patients.

CER Project Name	Data sources	Data types	Standard data model(s) used	Duplicate patients identified across organizations?
SPAN	Health plan enrollment, pharmacy dispensing, utilization data, billing data, vitals, lab results, tumor registry, death info	Local codes Standard codes No unstructured text	Expanded version of the HMO Research Network Virtual Data Warehouse Version 3 (13 tables linked by a unique identifier)[27]	No. Two organizations unlikely to have information for the same patient during a defined enrollment period.
WICER	Patient surveys, vital statistics, health literacy, socioeconomic status, in-patient, ambulatory clinics, long term care facilities, home care agencies	Local codes Standard codes Processed Free text	Early version of the HL-7 Reference Information Model	Yes, many patients are participants in New York Care Connect HIE
CER-HUB	Ambulatory EHR, In-patient discharge summaries, billing, pharmacy dispensing, lab results; all are extracted based on project need via standard extraction mechanism.	Local codes Standard codes Processed Free text	Implementation of HL-7 Clinical Document Architecture that extends the CCD (Continuity of Care Document)	No. Unlikely for sites currently involved to have overlap in patient populations. One site operates a single instance EHR for its multiple consortium member FQHC organizations.
RPDR	Demographics and labs data loaded nightly; EHR, billing and decision support systems data (including vitals and inpatient and ambulatory clinics data), death info and pharmacy data loaded monthly; text clinical notes available on project-need basis.	Local codes Standard codes Processed Free text	Star schema data model codes clinical events as “facts” in relational database structure with radiating tables that further define facts, along with metadata tables	Yes, Enterprise Master Patient Index
INPC COMET-AD	Multiple hospital systems, healthcare payers, practice organizations (eg, primary care group practice, radiology practice, sports medicine practice), laboratory organizations	Local codes Standard codes Processed Free text and unstructured text	Identifier, timestamp, “question” term, and “answer” term - where answer term is numeric, coded, date, person (e.g., patient or clinician), or free text value. Also some “compound” results.	Yes, patients are linked across, institutions in the Indiana Network for Patient Care via the global person ID service
SCOAP-CERTN	ADT/Registration, Laboratory, Medications, text Reports (e.g. Doppler Ultrasound report), text Notes (e.g. Operative Note)	Local codes Standard codes Processed Free text Unstructured text	HL7-v3 in warehouse augmented by data elements from the SCOAP data collection forms	No

Abbreviations: Health Information Exchange (HIE); Federally-Qualified Health Care facilities (FQHCs); Health Level 7 – Version 3 (HL7-v3)

Table 3

Comparison of data flow and transformation from EHR to aggregated analysis. Grey cells indicate when data first leaves control of local site.

CER Project Name	Raw data	Natural Language Processing	Data Normalization	Data aggregation	Data Analysis Tools
SPAN	Data from local EHR, billing, and other sources are accessed in their native form, or extracted and stored in research data warehouse	No	Each site transforms local data to the common data model (HMORN VDW) using native coding systems (NDC, ICD9, HCPCS, LOINC); transformations checked centrally for consistency	Centrally-developed application (analytic programs, queries) runs at local site, results sent to Data Coordinating Center or study lead site	Menu-driven query tools for encounter-level VDW datasets are under development; tools available querying of aggregate data
WICER	Extracted from local EHRs, sent to central data warehouse where it is combined, by patient, with data from multiple organizations (HIE warehouse – model)	Yes, using general purpose MedLee[28]	Central DB uses Columbia's Medical Entities Dictionary (MED)[29] to map to LOINC, ICD, SnoMed	Central site	Tool under development to allow end-users to identify patients with characteristics of interest, specify query constraints, and combine data elements
CER-HUB	Data extracted from Local EHR, stored in local data warehouse	Yes, using project-specific, knowledge-based MediClass applications[11]	Local site uses centrally-developed data processor – MediClass application–to populate Clinical Research Documents using Unified Medical Language System as standard knowledge base.	Centrally-developed processors are downloaded to create study-specific clinical events that are transmitted to central site for aggregation and analysis	Tools available for researchers to create new and review existing NLP query modules and test them against new data sets. Tools in development to enable investigators to analyze and review final aggregated results
RPDR	Data extracted from local systems (some daily, some monthly), aggregated centrally in data warehouse.	Yes, using project-specific pattern recognition models[30,31]	Data mapped to ICD-9-CM and COSTAR (for diseases), NDC (for medications), LOINC (for labs), CPT and HCPCS (for procedures) data stored centrally in local coding systems and dynamically mapped in query tool.	All aggregate data is derived from queries generated through query tool accessing central data repository, detailed data on patients are gathered from central repository and through enterprise web services.	Drag-and-drop web Query Tool allows users to construct ad hoc, Boolean queries for hypothesis generation from structured data, to get aggregate totals and to graph age, race, gender and vitals
INPC COMET-AD	Extracted from local EHR (or payer), sent to central data warehouse, stored distinctly but can be combined at patient-level across multiple organizations; HIE-model	Yes, using project-specific pattern recognition models[32].	Data are normalized and coded using standard vocabularies (LOINC, ICD9, CPT, NDC, etc.) when/where appropriate	Specially developed data aggregation routines developed and run centrally	All SAS and other statistical data analysis done at the central site; new data analysis tools are under development
SCOAP-CERTN	Data are extracted from source systems and transmitted via VPN channels to the central data warehouse servers; message stream is queued then parsed;	Text mining tools operate on the data within the central data warehouse; use hybrid text mining (rules base and statistical) to extract and tag the text and derive data elements needed for QA/QI and CER	Parsed data stream is either directly mapped onto the HL-7 data model or into a staging table from which it is mapped to the HL-7 derived and augmented data model.	Data extracted from local systems and sent to secure remote server cluster running Amalga	Sites and SCOAP-CERTN personnel will use secure web sites and interfaces with appropriate authentication, authorization and logging to query the Amalga data warehouse

Table 4

Key personnel involved in various stages of project.

CER Project Name	Data extraction, transformation, loading	Data queries	Data analysis	Total personnel accessing system
SPAN	Local site programmers create HMORN VDW (multi-purpose research warehouse)/ expand VDW on study-by-study basis; distributed queries create study-specific analytic file extracts from the VDW.	SAS analysts at the lead study site create SAS programs to run against the VDW. Non-technical researchers use the SPAN user interface to query summary count tables; user interface enhancements will allow menu-driven querying of individual-level data derived from the VDW.	Statisticians, investigators, analysts at lead study site	Planned 20–40 investigators (2–4/site); 25 analytic/administrative staff to respond to queries (~2 per site)
WICER	Local site programmers work with central programmers to create extracts – real-time & monthly	Designed to be completed by researchers	Statisticians and investigators at local or central sites	Platform still under development
CER-HUB	Local site programmers create XML-based clinical research document	Study staff create and validate standardized data processors—a MediClass application—for each study.	Statisticians at central data coordinating center with guidance from study investigators	Per study (two currently). Local site programmers: 1–3 per study site (currently 12 total). Research staff: 5–10 investigators, RA's, and PM's per (currently 15 total)
RPDR	Local site programmers create real-time and periodic data extracts	Clinical investigators run queries using structured user interface	Preliminary statistics automatically generated by query tool. Investigators at study sites perform more detailed analysis.	>2,500 across the Partners' organization
INPC COMET-AD	Central staff extracts data from the INPC; additional NLP is planned; data are also collected by project staff	SQL queries on central DB are written by central staff in collaboration with the investigator team	Analysis is by statisticians at the central location using extracted data	For the pilot study – one project manager and staff, one INPC data analyst, one NLP expert programmer, six investigators. The more general platform is under development.
SCOAP-CERTN	Central staff work with local site programmers to create site-specific data extracts that are sent to centrally-developed Amalga message processing architecture	Users at participating sites, centralized SCOAP-CERTN staff (site associated data coordinators, centralized data coordinators and scientists doing data analysis)	By statisticians at central location using extracted data	System is still under development. When ready, expect users across the state of Washington

Abbreviations: Natural Language Processing (NLP); Structure Query Language (SQL); Virtual Data Warehouse (VDW)

Table 5

Comparison of Project Governance and Internal Organizational Policies and Procedures.

CER Project Name	Project Governance	Project Membership	Data Ownership & Access control	Data use agreements	IRB oversight	Study Design Oversight	Publication Rights
SPAN	Project Governance Core, SPAN PI and executive team[33]	Inclusion in the project grant submission; additional membership possible via governance committee	Local sites control access to all data; opt-in on a study-by-study basis	Between local sites and coordinating center and study lead site(s); BAA agreement between sites and IT vendor that hosts the SPAN networking infrastructure	Each project overseen by local IRBs; most sites have ceded IRB oversight to the study lead site	All projects result from external contracts or grants. Each site must agree to participate.	Maintained by the study team in accordance with grant and contract language.
WICER	Project Governance committee	All members were part of New York Care Connect HIE; new members via governance committee (community surveys are a new group, that were not represented by NYCC HIE)	PHI removed from data after matching patient IDs; all researchers have access to data via login ID	Local sites have DUA with central site. Each data source has a steward	Local sites and central coordinating center all have IRBs	Researchers control study design; access to data via remote UI	Data steward defines right of access and first publication
CER-HUB	Steering committee (site Pis) authorize each project (currently two projects) but projects are autonomous teams addressing approved study protocol.[34]	Based on projects run by a lead investigator; approved by Steering Committee	**Raw data:** Stays at sites; **Study data:** Project-specific Limited Data Sets shared (using DUA) with data coordination center; **Results:** Project team controls with sites represented by an investigator.	Between each site and the central data coordinating center	Sites each require an IRB submission to address local concerns and circumstance. One of the five secondary sites has ceded IRB oversight to the lead site.	Approved by steering committee through acceptance of project. Managed by study lead investigator.	Maintained by study team in accord with contract language and publications policy
RPDR	Project governance committee	Based on lead investigator always being a faculty member at the organization	**Raw data:** Local sites; **Study data:** Stored centrally, controlled by query tool; **Results:** Local sites control, requires login managed by organizations involved	Established along with initial data transfer to central DB	Central IRB, enforced by query tool and requirement of digital signatures to obtain data extracts	Researchers control study design, user interface limits search strategy and patient detail that may be viewed without explicit IRB approval.	No oversight
INPC COMET-AD	INPC management committee (and/or its research subcommittee) oversees all	All INPC members may participate (optional)	**Raw data:** Local site owns data, but Rl is data	Established between the INPC and each	All research projects have separate IRB	Rl/university investigators overseen by the IRB, with initial data	Controlled by the Rl/Indiana University (or

CER Project Name	Project Governance	Project Membership	Data Ownership & Access control	Data use agreements	IRB oversight	Study Design Oversight	Publication Rights
	data use; PI and co-investigators lead the project; data and safety monitoring board for specific trials		custodian; **Study data:** Stored centrally; **Results:** Managed by RI/ lndiana University (or Purdue University or other) investigators and data analysts	local hospital, or payer, etc.	approvals, most with the Indiana University IRB	use approval from the INPC management/ research committee/sub committee	Purdue University or other) project investigator
SCOAP-CERTN	Overall External Advisory Boards including community representation (SCOAP and SCOAP-CERTN), also Health Information Technology specific external advisory board	SCOAP-CERTN executive leadership identifies potential partners who are invited to participate	Sites retain control over their data; can elect to stop participation and withdraw their data from the central repository up until extraction for CER occurs	DUAs and BAAs being developed with IRB, compliance, Regulatory and Legal offices of the University of Washington and participating sites	Each sub-study has an IRB protocol;	Each CER project will oversee work;	SCOAP-CERTN project publication guidelines overseen by SCOAP-CERTN executive committee

Abbreviations: Health Maintenance Organization Research Network (HMORN); Indiana Network for Patient Care (INPC); Regenstrief Institute (RI) Data Use Agreements (DUA); Business Associate Agreements (BAA)