4-16-2014

How the Provenance of Electronic Health Record Data Matters for Research: A Case Example Using System Mapping

Karin E. Johnson
Group Health Research Institute, johnson.ke@ghc.org

Aruna Kamineni

Sharon Fuller

Danielle Olmstead

See next pages for additional authors

Recommended Citation

Johnson, Karin E.; Kamineni, Aruna; Fuller, Sharon; Olmstead, Danielle; and Wernli, Karen J. (2014) "How the Provenance of Electronic Health Record Data Matters for Research: A Case Example Using System Mapping," *eGEMs (Generating Evidence & Methods to improve patient outcomes)*: Vol. 2: Iss. 1, Article 4.
DOI: http://dx.doi.org/10.13063/2327-9214.1058
Available at: http://repository.edm-forum.org/egems/vol2/iss1/4

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.

How the Provenance of Electronic Health Record Data Matters for Research: A Case Example Using System Mapping

Abstract

Introduction: The use of electronic health records (EHRs) for research is proceeding rapidly, driven by computational power, analytical techniques, and policy. However, EHR-based research is limited by the complexity of EHR data and a lack of understanding about data provenance, meaning the context under which the data were collected. This paper presents system flow mapping as a method to help researchers more fully understand the provenance of their EHR data as it relates to local workflow. We provide two specific examples of how this method can improve data identification, documentation, and processing.

Background: EHRs store clinical and administrative data, often in unstructured fields. Each clinical system has a unique and dynamic workflow and EHR, which may be influenced by broader context such as documentation required for billing.

Methods: We present a case study with two examples of using system flow mapping to characterize EHR data for a local colorectal cancer screening process.

Findings: System flow mapping demonstrated that information entered into the EHR during clinical practice required interpretation and transformation before it could be accurately applied to research. We illustrate how system flow mapping shaped our knowledge of the quality and completeness of data in two examples: 1) determining colonoscopy indication as recorded in the EHR, and 2) discovering a specific EHR form that captured family history.

Discussion: Researchers who do not consider data provenance risk compiling data that are systematically incomplete or incorrect. For example, researchers who are not familiar with the clinical workflow under which data were entered might miss or misunderstand patient information or procedure and diagnostic codes.

Conclusions/Next steps: Data provenance is a fundamental characteristic of research data from EHRs. Given the diversity of EHR platforms and system workflows, researchers need tools for evaluating and reporting data availability, quality, and transformations. Our case study illustrates how system mapping can inform researchers about the provenance of their data as it pertains to local workflows.

Acknowledgements

We would like to thank John Steiner for his guidance on the vision for this manuscript and Chris Tachibana for editing the manuscript. This work was funded by the National Cancer Institute under award number U54CA163261 and the Agency for Healthcare Research and Quality under award number U13HS019564. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Cancer Institute or the Agency for Healthcare Research and Quality.

Keywords

Context, Data Use and Quality, Health Information Technology

Disciplines

Health Services Research

Authors

Karin E Johnson, *Group Health Research Institute*; Aruna Kamineni; Sharon Fuller; Danielle Olmstead; Karen J Wernli.

How the Provenance of Electronic Health Record Data Matters for Research: A Case Example Using System Mapping

Karin E. Johnson, PhD; Aruna Kamineni, PhD, MPH; Sharon Fuller, BA; Danielle Olmstead, BS; Karen J. Wernli, PhD, MS[i]

Abstract

Introduction: The use of electronic health records (EHRs) for research is proceeding rapidly, driven by computational power, analytical techniques, and policy. However, EHR-based research is limited by the complexity of EHR data and a lack of understanding about data provenance, meaning the context under which the data were collected. This paper presents system flow mapping as a method to help researchers more fully understand the provenance of their EHR data as it relates to local workflow. We provide two specific examples of how this method can improve data identification, documentation, and processing.

Background: EHRs store clinical and administrative data, often in unstructured fields. Each clinical system has a unique and dynamic workflow, as well as an EHR customized for local use. The EHR customization may be influenced by a broader context such as documentation required for billing.

Methods: We present a case study with two examples of using system flow mapping to characterize EHR data for a local colorectal cancer screening process.

Findings: System flow mapping demonstrated that information entered into the EHR during clinical practice required interpretation and transformation before it could be accurately applied to research. We illustrate how system flow mapping shaped our knowledge of the quality and completeness of data in two examples: (1) determining colonoscopy indication as recorded in the EHR, and (2) discovering a specific EHR form that captured family history.

Discussion: Researchers who do not consider data provenance risk compiling data that are systematically incomplete or incorrect. For example, researchers who are not familiar with the clinical workflow under which data were entered might miss or misunderstand patient information or procedure and diagnostic codes.

Conclusions: Data provenance is a fundamental characteristic of research data from EHRs. Given the diversity of EHR platforms and system workflows, researchers need tools for evaluating and reporting data availability, quality, and transformations. Our case study illustrates how system mapping can inform researchers about the provenance of their data as it pertains to local workflows.

Introduction

Electronic health record (EHR) implementation is proceeding rapidly in both ambulatory and hospital settings.[1-3] Adoption of EHR systems in the United States has accelerated since 2009 under the Health Information Technology for Economic and Clinical Health Act, the American Recovery and Reinvestment Act, and more recently through Affordable Care Act incentives.[4]

The increasing use of EHRs offers great potential for research, especially with growing computational power and methodological developments such as natural language processing to capture unstructured data.[5-7] EHR data reflect real-world, real-time health care information that can answer a range of research and quality improvement questions.

While EHR data are useful because they make previously paper-based information electronically accessible, their full potential to support research remains nascent for several reasons. First, EHRs are complex. Their data are often unstructured and difficult to extract into measures useful for research.[8] The second reason is the often unrecognized but critical role of data provenance—the original context under which the information was collected. Specifically, the data in EHRs reflect the nuances of clinical system activities, including variation in the provision and documentation of care.[9, 10] Researchers planning to use EHR data for health research or quality improvement must understand the original intended use—including drivers and workflow—to ensure the validity of data elements and to assess variations in data quality and completeness. The third factor limiting the utility of EHR data is the lack of

[i] Group Health Research Institute

published guidance on approaches and best practices for accessing and using EHR data for research.[11-13] Researchers typically rely on their tacit or assumed knowledge about EHR data. Therefore, the next step in using EHR data for research is applying approaches that enable researchers to recognize, understand, and communicate the limitations and the potential of their data, including information about data provenance.

In this paper, we present a colorectal cancer (CRC) screening case study to illustrate how site-specific clinical workflow and EHR setup affected the availability and completeness of data for research. We introduce system flow mapping as a technique to help researchers achieve two key advantages of investigating data provenance: (1) understanding and documenting original intended use of electronic health information, and (2) uncovering new information and new data sources.

Case Study

The SuCCESS Study

The goal of the Studying Colorectal Cancer: Effectiveness of Screening Strategies (SuCCESS) project is to develop evidence to inform personalized CRC screening recommendations. SuCCESS is part of the Population-based Research Optimizing Screening through Personalized Regimens (PROSPR) program funded by the National Cancer Institute. The aims of SuCCESS are examining the comparative effectiveness of CRC screening as currently practiced, evaluating the potential for personalizing screening and surveillance recommendations, and modeling the long-term comparative effectiveness of CRC screening.

This case study is from SuCCESS, which is based at Group Health Research Institute, the independent research arm of Group Health, a large integrated health care system in Washington State that implemented its EHR system in 2005. Our case highlights how attention to data provenance during study planning identified access points for CRC-specific EHR data, as well as workflow and system-specific contexts that needed to be considered when using the data for research.

The SuCCESS research data originate from multiple clinical and administrative Group Health data sources. Laboratory and claims databases identify CRC tests. Pathology databases and tumor registries identify test outcomes, including diagnosis of preinvasive lesions and cancers. Information collected in the EHR includes patient risk factors such as age, sex, race, family history, and comorbidities. Several of these variables, for example, Group Health patient demographics, are available to SuCCESS researchers through the HMO Research Network (HMORN) Virtual Data Warehouse (VDW).[14] The HMORN is a consortium of delivery systems with public domain research programs. The HMORN VDW supports consortium members, including Group Health, with ongoing processes to make data ready for research. First, variables are standardized into formats and values typically used for research. Second, data undergo quality control (e.g., programmers investigate unusual amounts of missing data to determine if they are truly missing or absent because of a coding error). However, the VDW contains only a small portion of the vast clinical information in the EHR. For example, several CRC screening-related variables, including family history and pathology results, are not yet available in the VDW. As a result, SuCCESS researchers can access some research-ready patient data from the HMORN VDW, but must identify and transform many other data elements directly from the Group Health EHR without the benefit of HMORN VDW support.

System Flow Mapping

During SuCCESS study planning, the research team realized a need for a fuller understanding of the CRC screening data we were collecting from the Group Health EHR. To better understand the CRC screening process at the system- and provider levels, we used system flow mapping, a process widely employed in the analysis of manufacturing facilities. System flow mapping is most broadly known in association with Lean Methodology or the Toyota Production System, adapted from methodology and management principles originally used at the Toyota manufacturing plant in Japan.[15] In addition to identifying opportunities to improve a process, system flow mapping can describe the context for data generated from a system and identify access points for those data.

System Flow Mapping Methods in the SuCCESS Study

To characterize the CRC screening process at Group Health from a system flow perspective, a study team member followed these steps:

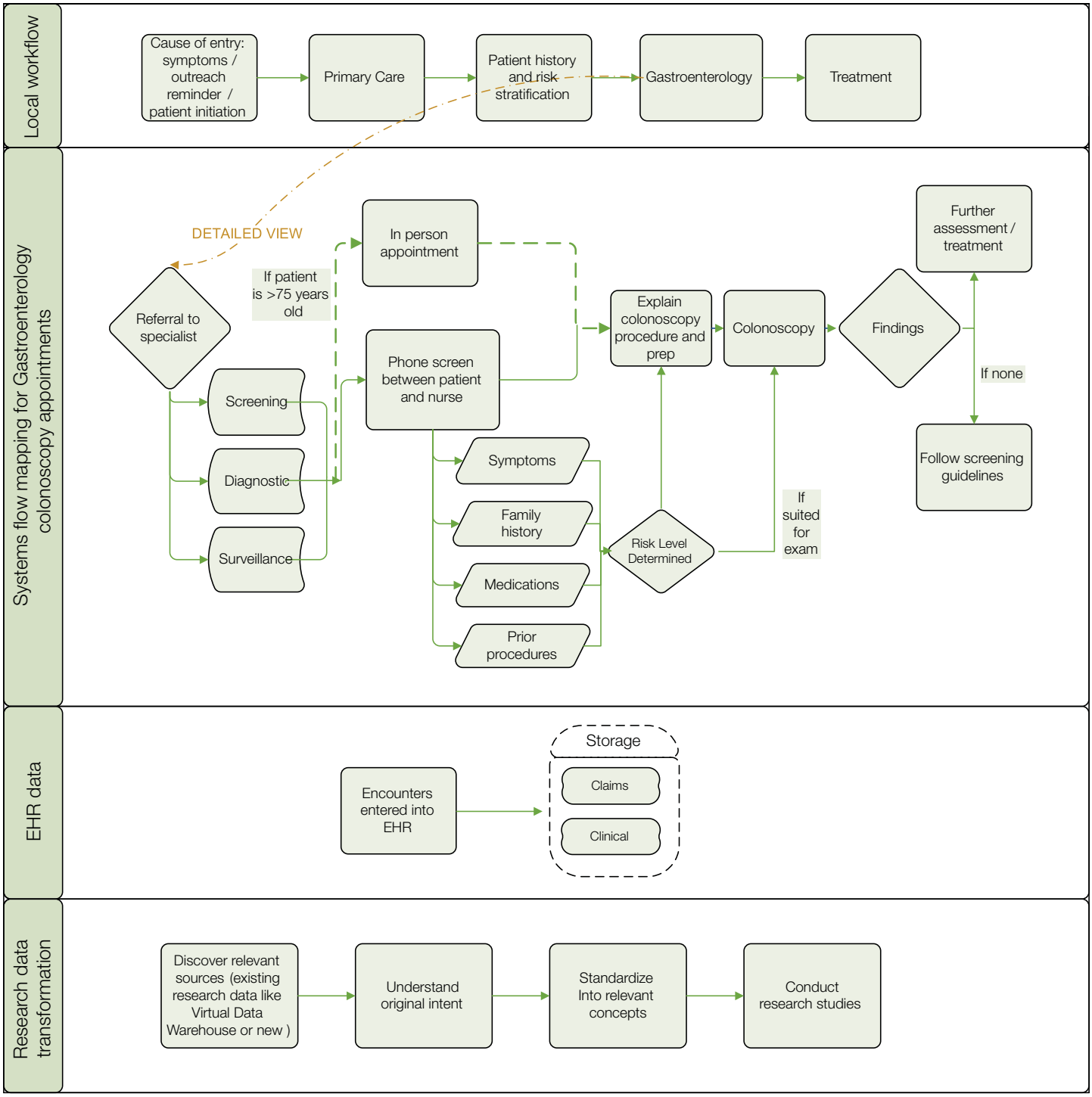
1. Gathered background information on CRC epidemiology, screening, and treatment to understand disease-specific context;
2. Identified key Group Health delivery system members with first-hand knowledge of CRC screening outreach and CRC clinical management, including nurse practitioners, gastroenterology specialists, pathologists, primary care providers, medical oncologists, and screening and outreach coordinators;
3. Drafted lists of questions for each delivery system group about the CRC screening process;
4. Conducted in-person interviews with delivery system staff and providers, using initial contacts to get referrals to other interviewees with additional knowledge;
5. Generated a system flow diagram using Microsoft Visio software;
6. Traced access points where researchers might obtain clinical and administrative data entered at each stage of the CRC screening process.

Institutional Review Board review and approval was not required for the system flow mapping because no individual-level data were used.

Figure 1 presents a high-level summary of the CRC screening process at Group Health. The top tier summarizes the overall local workflow. The process begins when patients are contacted for CRC screening at a clinic visit or through outreach (e.g., mailed screening kits), initiate CRC screening themselves, or are

referred for diagnostic testing because of symptoms or positive screening results. From the entry point, patients follow a number of paths for screening or follow-up involving both primary care and specialty providers. The second tier highlights one part of the workflow: the colonoscopy intake process that occurs upon referral to gastroenterology specialists. The third tier shows that information on all aspects of this process— scheduling, phone screening, and results—is entered into the EHR and stored in an underlying database. Standard coding systems (e.g., International Classification of Diseases (ICD(diagnoses, Current Procedural Terminology (CPT) procedures, or visit type service codes) enable use for care and billing purposes. The fourth tier shows that data are also available for research use but require interpretation and transformation. The next section of the paper contains two examples of how system flow mapping helped the SuCCESS research team document colonoscopy indications and CRC family history.

Figure 1. From Local Workflow to Research Data: An Example of System Flow Mapping of EHR Data and Original Intended Use



Example One: Colonoscopy Indication — Implications of Intended Use

Studies on screening efficacy or effectiveness must be able to distinguish screening tests of asymptomatic individuals from diagnostic tests. Including tests of symptomatic individuals in the analysis can introduce bias: since symptomatic individuals likely have a higher risk of cancer than asymptomatic individuals, including their test results can lead to a false observation of screening benefit.

In the SuCCESS project, system flow mapping helped us understand that the terms "screening" and "diagnostic" in EHR documentation might not correspond to our research definitions. For example, EHR documentation that listed a scheduled colonoscopy as a "screening" procedure might not be updated to "diagnostic" for a patient who reported CRC-related symptoms during colonoscopy intake on the day of the procedure. Mandates for screening could lead to the coding of more exams as screening, illustrating the interface of local workflows and broader context in shaping data provenance.

Thus, the full meaning of a procedure indication in the EHR is not always apparent to a downstream data user. Reconciliation of EHR and research definitions for procedure indications is not possible without investigating the original intended use of the information and its clinical context for individual patients. For the SuCCESS project, system flow mapping highlighted that the most accurate way to determine asymptomatic patients in our EHR data was to use both the procedure code for colonoscopy indication *and* preprocedure information on symptoms. This resolution arose from the system flow map, which showed that clinical workflow in our delivery system included multiple points when symptoms may be assessed and entered into the EHR. This workflow observation was also critical for the data discovery described in Example Two.

Example Two: Family History Ascertainment – Data Discovery

In the process of creating the system flow diagram, we learned that a phone screen is completed by a nurse practitioner for every patient who is scheduled for a gastroenterology procedure. During the gastroenterology phone screen, patients are asked about their family history of CRC. From the nurse practitioners who conducted the phone screens, we obtained the questions and information about available responses (i.e., text-based answers or yes/no boxes). With this information, SuCCESS programmers focused their data search and located the relevant data source in the EHR reporting database. As a result, this family history is now available for evaluation and use in the SuCCESS study. We previously determined that family history can be ascertained from a structured EHR data field if recorded at Group Health primary care or other visits. However, the newly discovered source of family history information may be preferred if information at the time of a specific colonoscopy procedure is desired. In addition, it may fill gaps in other sources of family history.

This example, like the previous example on assessing colonoscopy indication, demonstrates how researchers can use a system flow map to understand the availability of particular data points in the context of clinical workflow. Discovering additional sources of information and understanding their clinical use can greatly enhance the validity and completeness of EHR data for research.

Discussion

This case study highlights how system flow mapping can help researchers understand sources and nuances of clinical data that might not otherwise be apparent, given the complex workflows that generate EHR data. System flow mapping can facilitate successful research by elucidating components of data provenance, locating relevant information, and guiding data processing by identifying critical data features (e.g., available as structured fields, suitable for extraction using natural language processing, or requiring manual chart review).

In rapidly developing efforts to make the most of EHRs for research, data provenance has received little attention as a fundamental characteristic of research data. This case study highlights the fact that local workflows shape the data available for research. This occurs in the broader context of the many uses of EHR data, which include patient care, billing and, increasingly, system-level legal requirements and reimbursement incentives such as the EHR Meaningful Use Standards from the United States Centers for Medicare and Medicaid Services and the Healthcare Effectiveness Data and Information Set (HEDIS) performance measures. Other influences on EHR data are periodic additions and revisions to data sources, for example, patient-entered information from patient portals or mobile health applications and the ICD revision planned in the United States. Given that each health care organization has a unique data system that reflects its particular administrative processes, data collection, staff roles and practice culture,[16, 17] researchers who do not consider these aspects of provenance risk working with data that may be systematically incomplete or incorrect. However, provenance is often invisible to researchers, which limits the utility of EHR-based research findings to guide evidence-based care.

We propose that EHR data should be understood as being the product of a host of clinical, technical and policy factors. The quality of research using these data depends on recognizing and evaluating data provenance.[18] As part of efforts to ensure that EHR data are ready for research use, we recommend these steps:

1. Apply system flow mapping to understand and communicate how available data are shaped by local workflows and EHR systems, as well as the implications for research.
2. Support background work by research analysts such as training in the health system's EHR. This upfront investment will pay off in more accurate data and a fuller understanding of the potential and limitations of the data.

3. Ensure effective communication between data providers and researchers. This step is essential for both single- and multisite research[19] and is facilitated by a research team that invests time and resources (e.g., in system mapping) to understand the systems that generate the data. For example, research teams who want to understand trends in coding and recognize opportunities for linking data must coordinate extensively with both information technology staff responsible for data systems and clinical providers who enter data. However, these staff can have competing priorities[8] and limited availability so advance preparation by the research team can ease these collaborations. Researchers can also point out how collaborative investigation of data systems and location of key variables benefits clinical and administrative users.
4. Generate processes for transforming clinical concepts into consistent, research-specific ontologies that take data origin and consequent variability into account.[20] The HMORN VDW processes are an excellent model of efficiency in constructing standardized data.
5. Ensure that data provenance is a component of data quality assessment and documentation procedures for aggregate EHR data.[21,22] Initial work in this area is occurring through initiatives such as Observational Medical Outcomes Partnership (OMOP), a public-private partnership using EHRs for drug safety surveillance.[23]

Conclusions and Next Steps

With abundant data that capture a comprehensive picture of clinical care, EHRs open the possibility of research that would previously have been prohibitively expensive or time-consuming. However, appropriate application of these data requires that researchers consider provenance. Given the diversity of EHR platforms and system workflows, researchers need tools for evaluating and reporting data availability, quality, and transformations. Our case study illustrates how system mapping can inform researchers about the provenance of their data as it pertains to local workflows. We encourage the continued exchange of lessons learned and best practices about working with EHRs, including discussing the utility of system flow mapping and ways to design data systems to support both high-quality care and research. These rapidly evolving approaches advance the potential of EHR data to promote learning health care system activities ranging from developing patient-centered risk screening to implementing point-of-care decision-making, to conducting comparative effectiveness research.

References

1. Desroches CM, Charles D, Furukawa MF, Joshi MS, Kralovec P, Mostashari F, et al. Adoption Of Electronic Health Records Grows Rapidly, But Fewer Than Half Of US Hospitals Had At Least A Basic System In 2012. Health Aff (Millwood). 2013;32(8):1478-85. Epub 2013/07/11.
2. Hsiao CJ, Jha AK, King J, Patel V, Furukawa MF, Mostashari F. Office-based physicians are responding to incentives and assistance by adopting and using electronic health records. Health Aff (Millwood). 2013;32(8):1470-7. Epub 2013/07/11.
3. Xierali IM, Hsiao CJ, Puffer JC, Green LA, Rinaldo JC, Bazemore AW, et al. The rise of electronic health record adoption among family physicians. Ann Fam Med. 2013;11(1):14-9. Epub 2013/01/16.
4. The Office of the National Coordinator for Health Information Technology. Health IT Legislation. 2013 [cited 2013 11/15]; Available from: http://www.healthit.gov/policy-researchers-implementers/health-it-legislation.
5. Pittman P. Health Services Research in 2020: data and methods needs for the future. Health Serv Res. 2010;45(5 Pt 2):1431-41. Epub 2010/08/11.
6. Lopez MH, Holve E, Sarkar IN, Segal C. Building the informatics infrastructure for comparative effectiveness research (CER): a review of the literature. Med Care. 2012;50 Suppl:S38-48. Epub 2012/06/22.
7. McCoy AB, Wright A, Eysenbach G, Malin BA, Patterson ES, Xu H, et al. State of the art in clinical informatics: evidence and examples. Yearbook of medical informatics. 2013;8(1):13-9. Epub 2013/08/27.
8. Marsolo K. Informatics and operations--let's get integrated. J Am Med Inform Assoc. 2013;20(1):122-4. Epub 2012/09/04.
9. Spence D. Data, data everywhere. BMJ. 2013;346:f725. Epub 2013/02/06.
10. Hersh WR, Weiner MG, Embi PJ, Logan JR, Payne PR, Bernstam EV, et al. Caveats for the use of operational electronic health record data in comparative effectiveness research. Med Care. 2013;51(8 Suppl 3):S30-7. Epub 2013/06/19.
11. Platt R, Andrade SE, Davis RL, Destefano F, Finkelstein JA, Goodman MJ, et al. Pharmacovigilance in the HMO Research Network. Pharmacovigilance: John Wiley & Sons, Ltd; 2002. p. 391-8.
12. Bradley CJ, Penberthy L, Devers KJ, Holden DJ. Health services research and data linkages: issues, methods, and directions for the future. Health Services Research. 2010;45(5 Pt 2):1468-88.
13. Glasgow RE. Commentary: Electronic health records for comparative effectiveness research. Med Care. 2012;50 Suppl:S19-20. Epub 2012/06/22.
14. Ross TR, Ng D, Brown JS, Pardee R, Hornbrook MC, Hart G, et al. The HMO Research Network Virtual Data Warehouse: A Public Data Model to Support Collaboration. eGEMs (Generating Evidence & Methods to improve patient outcomes). 2014;2(1).
15. Liker J. The Toyota Way: 14 Management Principles from the World's Greatest Manufacturer. New York, NY: McGraw-Hill; 2004.
16. Fernald DH, Wearner R, Dickinson WP. The journey of primary care practices to meaningful use: a colorado beacon consortium study. J Am Board Fam Med. 2013;26(5):603-11. Epub 2013/09/06.
17. Saleem JJ, Flanagan ME, Russ AL, McMullen CK, Elli L, Russell SA, et al. You and me and the computer makes three: variations in exam room use of the electronic health record. J Am Med Inform Assoc. 2013. Epub 2013/09/05.

18. Hersh WR, Cimino JJ, Payne PR, Embi PJ, Logan JR, Weiner MG, et al. Recommendations for the Use of Operational Electronic Health Record Data in Comparative Effectiveness Research. eGEMs (Generating Evidence & Methods to improve patient outcomes). 2013;1(1).
19. Kahn MG, Raebel MA, Glanz JM, Riedlinger K, Steiner JF. A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research. Med Care. 2012;50 Suppl:S21-9. Epub 2012/06/22.
20. Rea S, Pathak J, Savova G, Oniki TA, Westberg L, Beebe CE, et al. Building a robust, scalable and standards-driven infrastructure for secondary use of EHR data: the SHARPn project. Journal of biomedical informatics. 2012;45(4):763-71. Epub 2012/02/14.
21. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc. 2013;20(1):144-51. Epub 2012/06/27.
22. Richesson RL, Hammond WE, Nahm M, Wixted D, Simon GE, Robinson JG, et al. Electronic health records based phenotyping in next-generation clinical trials: a perspective from the NIH Health Care Systems Collaboratory. J Am Med Inform Assoc. 2013. Epub 2013/08/21.
23. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. J Am Med Inform Assoc. 2012;19(1):54-60. Epub 2011/11/01.