Clinical Research Informatics and Electronic Health Record Data

R. L. Richesson[1], M. M. Horvath[2], S. A. Rusincovitch[3]
[1] Duke University School of Nursing, Durham, NC, USA
[2] Health Intelligence and Research Services, Duke Health Technology Solutions, Durham, NC, USA
[3] Duke Translational Medicine Institute, Duke University, Durham, NC, USA

Summary

Objectives: The goal of this survey is to discuss the impact of the growing availability of electronic health record (EHR) data on the evolving field of Clinical Research Informatics (CRI), which is the union of biomedical research and informatics.

Results: Major challenges for the use of EHR-derived data for research include the lack of standard methods for ensuring that data quality, completeness, and provenance are sufficient to assess the appropriateness of its use for research. Areas that need continued emphasis include methods for integrating data from heterogeneous sources, guidelines (including explicit phenotype definitions) for using these data in both pragmatic clinical trials and observational investigations, strong data governance to better understand and control quality of enterprise data, and promotion of national standards for representing and using clinical data.

Conclusions: The use of EHR data has become a priority in CRI. Awareness of underlying clinical data collection processes will be essential in order to leverage these data for clinical research and patient care, and will require multi-disciplinary teams representing clinical research, informatics, and healthcare operations. Considerations for the use of EHR data provide a starting point for practical applications and a CRI research agenda, which will be facilitated by CRI's key role in the infrastructure of a learning healthcare system.

Keywords

Biomedical research, electronic health records, data collection, research design

Yearb Med Inform 2014:215-23
http://dx.doi.org/10.15265/IY-2014-0009
Published online August 15, 2014

Introduction

The use of data derived from electronic health records (EHRs) for research and discovery is a growing area of investigation in clinical research informatics (CRI), defined as the intersection of research and biomedical informatics [1]. CRI has matured in recent years to be a prominent and active informatics sub-discipline [1, 2]. CRI develops tools and methods to support researchers in study design, recruitment, and data collection, acquisition (including from EHR sources), and analysis [1]. To complement the "Big Data" theme of the IMIA 2014 Yearbook, this summary explores the impact of increasing volumes of EHR data on the field of CRI.

There is tremendous potential for leveraging electronic clinical data to solve complex problems in medicine [3]. The impact on the CRI domain is exemplified by a growing number of publications related to the use of EHRs, including medical record systems, algorithms and methods [4]. The analysis of existing clinical, environmental, and genomic data for predicting diseases and health outcomes is growing [5-7]. The regulatory and ethical challenges for using EHR data for research – though complex – are being addressed [8, 9]. Research use of EHR data is inherent to the vision of the learning healthcare system [10]. In this context, CRI will play a central role bridging different perspectives from research and healthcare operations, particularly as they relate to new demonstrations of interventional clinical trials embedded within healthcare systems [11]. The more immediate uses of EHR data are for observational research (i.e., investigations that observe and explore patient phenomena related to the "natural" – rather than researcher controlled – assignment of interventions), because these designs have less inherent risk and disruption to clinical workflows than do interventional trials.

Definitions

Clinical research is the science that supports the evaluation of safety and effectiveness of therapeutics (medications and devices), diagnostic tools, and treatment regimens. Clinical research includes a variety of study designs and methods to support patient-oriented research (i.e., conducted with human subjects or their biospecimens), clinical trials, outcomes research, epidemiologic and behavioral studies, and health services research [12]. Clinical research informatics, then, is the branch of informatics that supports all these research activities, particularly the collection, management, and analysis of data for varied types of studies. Research approaches can be characterized broadly as either interventional (or experimental trials, where the researcher assigns treatments) or observational (where treatments are not assigned by the researcher). To date, CRI has focused largely on the support of interventional trials, but there is momentum around observational research and clinical data mining [6], both of which are particularly relevant to this IMIA Yearbook theme of "Big Data". We defer to other issue authors for precise definitions of the term "Big Data," but premise this discussion on the assumption that the large amounts of clinical and administrative data from institutional repositories and EHR sys-

tems qualify as Big Data. This summary and discussion, therefore, focus on informatics activities and trends related to the use of data collected from clinical practice for purposes of research and discovery.

Interventional Research

In interventional studies, researchers control the assignment of the intervention or treatment under investigation. In randomized controlled trials (RCTs) – the gold standard for evidence generation – researchers assign the participant to an intervention using randomization. The widespread availability of EHR systems in clinical practice are enhancing the potential for *pragmatic clinical trials* (PCTs), randomized controlled trials designed for broad generalizability, typically using multiple clinical sites and broader eligibility criteria. In contrast to explanatory trials, for which the goal is to detect the effects of new treatments, PCTs evaluate interventions in "real-world" practice conditions [13]. The routine implementation of PCTs is an important component of a learning health system [10, 14]. Pragmatic trials require EHR data to identify research cohorts based on patient features and "clinical profiles", including co-morbidities, severity, and health outcomes [14]. Current informatics challenges for PCTs include developing ethical and regulatory standards and public trust [8, 9], integrating data from multiple databases, identifying appropriate study populations, unambiguously identifying procedures and treatments, and consistently and explicitly characterizing diseases in terms of progression, severity, and patient impact [14].

Observational Research

Observational research is non-experimental research encompassing different research designs (e.g., cross sectional, cohort, and case control) and directional components (prospective, retrospective, or non-longitudinal) [15]. The distinguishing factor is that there is no researcher-controlled assignment of treatment or intervention. In observational studies, the treatment occurs "naturally:" that is, as a result of patient factors and decisions made as part of routine healthcare delivery. In quasi-experimental design, the criteria used for treatment might be unknown, or determined using non-random methods (e.g., a summary score) outside the control of the researcher. A control group component in some observational study designs facilitates the evaluation treatment-outcome associations, making observational studies an appealing complement to RCTs [16-18]. Observational research principles underlie the growing use of patient registries for research [19-21] and management of chronic disease [22], quality measurement and improvement [23-37] activities, and comparative effectiveness research (CER) [28-30]. CER is the examination of available therapies relative to a broad range of health outcomes – or "what works best" in healthcare [16]. Because the goal of CER is to evaluate and compare real world treatments in large and diverse populations, the use of EHR data and observational research methods are essential [31-33].

Data mining is the exploratory and computationally-guided process of discovering patterns and signals in large data sets, which can then be used for hypothesis generation, predictive modeling, and other analytic activities. Data mining methods are counter to traditional hypothesis-based research, and instead developed in response to Big Data challenges in the business sector. Nonetheless, data mining has been embraced by some biostatisticians, and is gaining respect in the research community [6]. Data mining supports very large data sets obtained from legacy databases or data warehouses [34], and deals with the secondary analysis of clinical data, meaning the data are collected as a byproduct of routine clinical care and not purposely collected for research [6].

Research Fundamentals

The general process of research investigation includes formulating a research question, identifying relevant concepts and measures (variables), and collecting, analyzing, and interpreting data. A variety of statistical techniques can be used to demonstrate associations between patient features (e.g., laboratory value, genetic marker), experience (e.g., treatment), or events (e.g., onset of disease, hospitalization, death); these associations can sometimes be due to chance, bias, or confounding [35]. Bias is any systematic error that affects the estimates of the association under study, and can emerge from the identification of subjects (i.e., selection bias) or their evaluation (i.e., observation bias). Confounding results from the contamination or oversaturation of measured effects, influenced from related factors external to the study [35]. The strength behind RCTs is the belief that randomization eliminates confounding by 'randomly distributing' these factors – both known and unknown – across comparison groups. Both bias and confounding are major issues for observational studies [36, 37] and CER in particular [16, 37].

General research considerations for all research studies are the somewhat competing notions of validity and generalizability. Validity refers to confidence in the observed associations, and is increased when chance, bias, and confounding are well addressed. Bias and confounding can be minimized with strict eligibility criteria to limit the differences between comparison groups, but at the cost of making study populations 'different' from (or less generalizable to) the greater population.

Methods

Drawing from the literature of both the informatics and clinical research communities, we isolated important themes related to the use of electronic clinical data for research, including the heterogeneity and quality of EHR data, integrating data from multiple organizations, identifying research cohorts based on explicit clinical profiles, and the role of informatics in a learning health system. Emergent from these themes, a set of considerations for the use of EHR data is presented as a tool for coping with these challenges in the present and for guiding improvements for the future.

Current Themes Related to the Use of Electronic Clinical Data for Research

Important areas of informatics activity and recent advances are summarized below.

Heterogeneity of Data from EHRs

The definition and content of EHRs vary greatly [38, 39]. However, reimbursement requirements and common healthcare delivery processes do result in broad areas of similar data collection across many health care providers. Common subject areas shared between most EHRs include patient demographics, healthcare encounters, diagnoses, procedures, laboratory findings, and vital signs. While there is commonality in subject areas, there is variation in how these concepts are operationalized into variables and codes [40]. What is notably missing from typical EHR data are standardized data related to disease severity, including disease-specific and general measures of patient functioning that are necessary for health outcomes research [41]. Estabrooks et al convened consensus groups of experts, patients, and stakeholders and identified critical patient-reported elements that should be systematically incorporated into EHRs for standard and continuous assessment, which include health behaviors (e.g., exercise), psychosocial issues (e.g., distress), and patient factors (e.g., demographics) [42].

In addition, there are multiple sources for some concepts that need to be well defined for meaningful analyses – within or across organizations. For example, medication usage can be identified using electronic orders, pharmacy fulfillment, administration records, or medication reconciliation. EHR data are inherently subject to the institution's workflows and data-generating activities [43-45]. For research to be reproducible and for results to be valid, the source and limitations of different types of data within an organization must be clearly defined. Data provenance is the understanding of definitive or authoritative sources for particular data and any transformation of the data from their original state. This understanding is critical both for the valid use of these data in the present, and to drive future improvement in the quality of data from clinical systems [46]. Curcin et al have constructed a set of recommendations for modeling data provenance that includes the formal representation of relevant domain knowledge and business processes [47]. Hersh et al (2013) provide an illustrative model of data provenance, as part of their comprehensive description of caveats for the use of operational EHR data in research contexts [46]. Other caveats identified include the prevalence of missing and inaccurate data, the fragmentation of patient data across providers, operational features that introduce bias, inconsistencies in free text clinical notes, and differences in data granularity [46]. These aspects of EHR data are not generally reported, but likely have important implications for most research designs.

Data Quality

The notion of data quality is complex and context dependent [48, 49]. Weiskopf presents a multi-dimensional (completeness, correctness, concordance, plausibility, and currency) model of data quality, as well as common data quality assessment methods that include comparison with gold standards, data source agreement, distribution comparison, and validity checks [50]. Accuracy and completeness [51] are the dimensions of quality that are most commonly assessed in both observational and interventional trials [23, 52]. These dimensions closely indicate the capability of the data to support research conclusions, and have been prioritized in the high-profile Healthcare Systems Collaboratory, an NIH-funded exploratory collaboration to advance PCTs, cooperative research partnerships, and evidence-based healthcare) [52].

Challenges for Studies Involving Multiple Healthcare Organizations

Multi-site PCTs, safety surveillance, and observational research projects that identify patients from heterogeneous practice settings pose challenges for reconciling the variation in healthcare operations, widely disparate information systems, and differences in data capture fidelity. The impact of the selection of healthcare system and database on results of previously conducted studies is illustrated by a sobering study recently published in the American Journal of Epidemiology [53]. Using administrative claims data, Madigan et al systematically explored differences of relative risk and standard error estimates across a network of 10 health system data bases (130 million patients combined) for 53 drug-outcome test cases. They demonstrated variant results on studies in different clinical networks, despite identical sampling and statistical methods, in some cases reversing the drug-outcome associations detected. Authors concluded that 20% to 40% of observational database studies can swing from statistically significant in one direction to statistically significant in the opposite direction, depending on the choice of database [53]. The specific causes for this variance is unknown, but a growing number of methods reports are addressing approaches for using EHR data in observational research, including methods related to patient sampling and data quality [32, 54, 55].

Research studies are mandated to report patient characteristics for each study site as part of Consolidated Standards Of Reporting Trials (CONSORT) [56] and STrengthening the Reporting of OBservational studies in Epidemiology (STROBE) [57] guidelines. Data from different healthcare organizations represent different patient populations, treatment patterns, and operational factors related to the collection, coding, and reporting of clinical and administrative data. Brown et al provided a set of recommendations for data quality checks and trend monitoring in distributed data networks for CER, using experiences from their multi-site networks [58]. Moving forward, features related to the EHR system will be a critical factor, and at times an unknown confounder, in research conducted with healthcare systems and electronic clinical data. There is growing appreciation that EHR structure and features might someday be reported as important qualifying data [59] and that it is important to have processes in place to monitor and characterize potential issues.

EHR Phenotypes

The heterogeneity of EHR data creates challenges for identifying the presence of any specific clinical condition – such as diabetes, heart disease, or obesity – in patient populations, and greatly limits opportunities to observe or learn from regional variation of disease patterns or trends over time [60]. For example, a recent report on trends in diabetes lists several conditions (including hypoglycemia, neuropathy, chronic kidney disease, peripheral vascular disease, cognitive decline) whose national prevalence could not be calculated due to lack of consistency of EHR documentation and definitions across the United States [61]. The methods for defining a clinical condition and applying it to EHR data are encompassed by the concept of EHR-driven computational phenotyping. Currently there are no standardized or well-established EHR-based definitions (or "phenotypes") for most conditions, although many different operational definitions are used in different contexts. Within the past few years, a growing number of publications have emerged to describe methods related to EHR-driven phenotype definitions and resources for accessing and evaluating validated definitions [60, 62-65] An important original work by Shivade et al reviews the development of phenotyping approaches using EHR data for identifying cohorts of patients [66]. Their assessment creates a useful framework of classification, but is focused specifically on cohort identification. In addition, there is a scarcity of information about performance (e.g., specificity, sensitivity, positive and negative predictive values) of these various conditions definitions used with EHR data. The estimation of validity, using any performance measure, requires a "gold standard," defined as the best classification available for assessing the true or actual disease status. This requirement poses feasibility challenges because a "gold standard" does not often exist and must be constructed in order to be used for evaluation. Many EHR-based definition developers have conducted validation studies [65, 67, 68] but there is no standard approach or uniformly operationalized clinical "gold standard". Further, the performance indications of phenotype definitions are not typically included in the reports of many research studies published in scientific journals. Although greatly needed, a standard process for validation of these definitions (including statistical considerations and procedures and qualifications for expert reviewers) does not yet exist. Standardized methods in this area could support measuring the impact of the different versions of the International Classification of Disease coding systems, a transition that will have broad impact across healthcare operations in the U.S. As discussed by Boyd et al, the sometimes-convoluted semantic relationships between mapping the same disease conditions from ICD-9-CM and ICD-10-CM will create complex logistics, with predicted repercussions to accuracy [69].

Although used predominantly to identify positive "cases" (or negative cases or controls) for research, an important application of phenotypes is in the definition of health outcomes. Here, the methods already developed with administrative and claims data should be a foundation for application to EHR data, in particular the well-constructed and mature work of the U.S. Food and Drug Administration's Mini-Sentinel program [70]. Also utilizing claims data, Fox et al have described methods for expert review and consensus to define health outcome measurements in claims data [71].

The eMERGE consortium [65] and phenotype knowledge base [72] has lead the vanguard effort in many areas, including representation of phenotype definitions in characterizing eligibility criteria for clinical trials [64], metadata mapping [73], and a targeted evaluation of the National Quality Forum quality measure model [74]. This work is framed within its core objective of genetic research; in another context, Richesson et al have described a broader set of use cases for clinical research [14]. The SHARPn project [75, 76] describes development of a data normalization infrastructure for phenotyping activity [77] and its new library [78] represents an important repository for implementation efforts, particularly for Meaningful Use application. Institutional infrastructure solutions [79], machine learning platforms [62], and user interfaces [79,80] are also a significant area of development.

Phenotype definitions based upon EHR data are inexorably tied to their health services context. Disparate processes are reflected within these data, including measurement of patient biological status, diagnostic progression, treatment decisions, and management of disease. As discussed by Hripcsak, the true patient state is interpreted through the health care model processes, which in turn informs and creates the basis for the phenotype itself [81]. The logic and parameters of each phenotype definition may lead to significantly different yields [82].

Due to the relatively recent development of phenotyping methods, most applications have been at single institution or with a relatively small group of collaborators. In order to achieve uniform implementation and reproducibility of results, especially among heterogeneous data sources for multi-site research and national networks, more expansion is needed for logical and consistent representation and dissemination across sites [83]. The development of robust and standardized validation methods is an important area for future development, and will ensure that individual phenotype definitions are widely generalizable. Further development of phenotyping methods and applicability within a variety of settings will become increasingly important for a broad set of applications in observational and interventional research settings.

Observational Data and Learning Health Systems

The vision of a Learning Healthcare System (LHS) has been described in both the EU and the US. The paradigm depends upon operationalizing proven insights from research into the health care delivery system [84]. In this environment, quality improvement and research studies increasingly use the same data sources, cohorts, and personnels. The core concept is a circular flow of data and knowledge between patients, clinicians, health provider organizations, and communities so that data related to the healthcare experience inform research, and research builds evidence, which in turn informs

healthcare practices. Achieving this vision will require new ethical frameworks [8, 9], robust methods for managing and analyzing observational data, and effective partnerships among healthcare systems [85]. Future work around the themes presented here will be essential to the vision of LHS, and CRI will play a key role. The growing appreciation for the generalizability and convenience of observation studies has increased their prominence on the evidence hierarchy, and observational research is gaining respect as critical part of the LHS infrastructure.

Both interventional and observational research methods are important components of the core vision of the learning healthcare system (IOM), as shown in Figure 1, inspired by interpretations of the learning healthcare system as a continuous communication of data, information and knowledge between healthcare systems, researchers, and public.

There have been fruitful collaborations for integrating clinical data across large organizations, including models for networked research [86] and national pharmacovigilance [87]. We look forward to continued demonstration and dissemination of knowledge from the above, in particular their common challenge to convincingly demonstrate that they can overcome issues related to data quality, validity, and reproducibility, to which observational research and secondary use of clinical data are inherently vulnerable.

The ability to overcome these issues will be critical to combining data, applying guidelines, or comparing results across different settings. Essentially, generalized evidence synthesis (e.g., comparative effectiveness research or meta-analyses) of any kind will be dependent upon shared semantics and data quality [46, 88-91]. This will require increased collaboration between the researcher and health system enterprises and commitment to quantify data quality, lineage, and provenance as part of an ongoing health informatics strategy. Informed by the experience of research and clinical networks that have successfully leveraged electronic clinical data for research insight and pharmacovigilance, we articulate the guidelines in the next section.

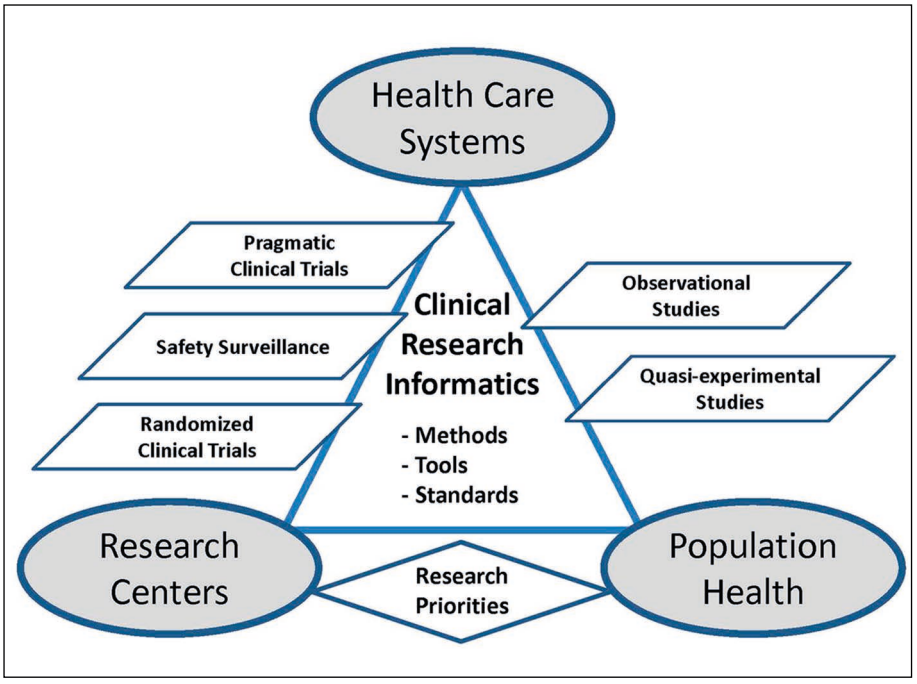

Fig. 1 Central role of CRI in a learning healthcare system. Adapted from: http://ehealth.johnwsharp.com/2012/12/14/the-learning-healthcare-system-and-order-sets/

Considerations for Using EHR Data in Research

Caveats and Guidelines

Regardless of the study design, there are some important issues that must be considered for research investigations involving EHR data. The following principles address risks to research integrity and points of attention for using EHR data, and also to describe outstanding challenges for clinical research informatics and informaticians in general.

- ***Data on a given cohort's outcomes will be limited.*** Only a fraction of a patient's lifetime care will be housed within any EHR [39], particularly in the U.S. Inconsistencies in defining a study population can affect the validity and generalizability of results. For example, readmissions can only be identified within those care sites where EHR data are available and patients can be linked across different locations and care delivery settings. Many research studies examine all-cause mortality as an endpoint; for deaths not directly observed within the inpatient setting, this is a very incomplete data point in most EHRs without supplementation by either the social security death index data, which has certain limitations, or the National Death Index (NDI), which can be costly to acquire [92].
- ***EHR adoption is continually evolving in healthcare environments.*** Observational studies require longitudinal data, but use of an EHR does not imply its data is consistent over time. A recent Black Book Ratings report has noted that 17% percent of surveyed health organizations planned to switch their current EHR to a new vendor by the end of 2013 in order to comply with growing government meaningful use requirements [93]. For researchers, this means that collected data may span multiple systems (i.e. different EHRs) and thus require separate data dictionaries as well as design and documentation of a strategy for spanning one or more timelines. Major upgrades to existing EHRs can also change the production data tables as well.
- ***Data quality will be an ongoing issue.*** Depending on the subject domain, some portion of data for a given field may be

nonsensical and should be omitted. For example, it is possible that 'dummy' or 'test' patients may be left within a production EHR which could be discovered by scrutinizing patient keys that seem to have an unusual volume of encounters. At times, this may impact between 10%-15% of the EHR data examined. The decision to omit or restrict data will depend upon the project needs.

- ***Trustworthy data dictionaries are essential.*** A project-specific data dictionary should be created and include the following for each data element: completeness in the EHR, range of possible values, data types (e.g. integer, character), and definitions.
- ***The use of EHR data must be accompanied by an understanding of health care workflow.*** EHRs have plentiful timestamps that could be used to understand the process of care, but care must be taken to understand exactly what triggers the recording of those dates and times. For example, a timestamp recorded for when a clinic visit starts can potentially precede a patient's appointment if the provider opens the encounter early to get a head start on documentation.
- ***EHR data will reflect billing needs and not just clinical care.*** There may be concepts that have significance for billing but not care provision or disease progression. Researchers should be cautioned that diagnosis and procedure codes that were designed with billing and reimbursement uses in mind may not reflect a clinical state to the resolution needed for research purposes.
- ***Observational research using EHR data must be a team activity.*** Success requires partnerships between clinical staff, informaticians, and researchers. Given data complexity, a new discipline of health data science is emerging [94]. As of this writing, the National Institutes of Health announced the appointment of the Associate Director for Data Science, a new senior scientific position, charged to lead a series of NIH-wide strategic initiatives that collectively aim to capitalize on the exponential growth of biomedical research data such as from genomics, imaging, and electronic health records [95].
- ***Advocate for a research voice in the creation of organizational data governance.*** Quality EHR data is ultimately dependent upon a clear organizational data governance structure that enforces process, definitions, and standards. Researchers should have a voice within the governance structure to ensure that their needs are met and that they can communicate any data quality issues identified over the course of their investigations.

CRI includes a growing collection of methods that offer methodological and technical solutions for processing clinical data, but guidelines for the valid assessment and analyses of these data are needed. Current clinical research data management practices are tailored to the rigor and regulatory requirements of clinical trials [96]. The secondary use of data generated from clinical care activities has created new challenges for data management and analysis, and the amount of data available greatly exceeds the number of analysts capable of making sense of it [97]. Interpretation and application of study findings will require teams of dedicated informatics professionals working collaboratively with researchers and practitioners. These multi-disciplinary teams should appreciate the fundamental research design principles, as well as organizational and workflow factors that affect the completeness and appropriateness of data for different research needs.

The above issues have emerged from the lack of representational standards for EHR data. Clinical researchers and informaticians have developed complex strategies to deal with the resulting EHR heterogeneity on a per-study basis; however, the identification of strategies and solutions will continue to be a major activity area for CRI and researchers alike.

Future Directions and Challenges

EHR technologies are evolving to permit not just management of patients at care sites, but also telemedicine and the management of population health. Vendors are exploring how to allow integration of the data generated by platforms for mobile technologies, and wearable devices [98-100]. These data streams bring new challenges as patients will use these resources to different depths; there is a large potential for missing data from patients on the less engaged side of the digital divide, especially due to lack of technical acumen or barriers to access [101, 102].

EHRs contain information primarily generated during routine clinical care. As technology has evolved, there is a great deal of data generated outside the traditional healthcare system (e.g., wearable devices, social networks) with volume expected to increase exponentially; these data represent an important opportunity to understand the complete context of patient health and behavior, but will require integrated solutions for analytic use. Similarly, increased focus on socio-economic status will likely driven broader inclusion of different data types, including geospatial and patient-reported data, areas addressed by recommendations of the committee formed by the IOM to identify domains and measures that capture the social determinants of health to inform the development of recommendations for meaningful use of EHRs [102]. Images and video data represent other areas of largely untapped potential where challenges for interoperability have precluded large-scale analytic application [103].

As more data sources are available, there are also significant challenges associated with person-level disambiguation and linkage across sources. The U.S. lacks a unique person identifier used for healthcare settings, a strategy that has been adopted in other countries [104]. Multiple techniques exist to perform entity resolution and create an integrated view of all data points associated with the patient, but accuracy and performance of these methods is variable [105].

Emerging computational methods have arisen to address the demands of molecular analyses conducted upon large volumes of genetic data, including cloud-based solutions and the massively parallel processing supported by Hadoop, MapReduce, and other platforms [105, 106]. As the availability and volume of clinical data increases, extending these technologies beyond the translational sciences offers great potential for overcoming some limitations of traditional relational

databases management systems [107]. These emerging tools hold potential to support better prediction models, with the goal of supporting learning healthcare to achieve better outcomes of patient health. The future role of informatics will be to build upon successful clinical and research networks and collaborations to create a data infrastructure that will support a new generation of continuous learning [108]. This includes understanding the limitations of EHR data, operationalizing appropriate research questions for the data, and proactively devising approaches improving data quality for more robust uses. There are also opportunities for CRI professionals to formulate informatics research questions as well as provide leadership in building better approaches to data standards and exchange to support varied uses.

Conclusions

There is a great amount of activity surrounding the use of EHRs for observational research. Strategies are being designed by the CRI community to grapple with the complexities of observational and interventional study designs, data governance, technical integration issues, and query models. This work will continue to grow and inform the design and conduct of research as well as the eventual application of evidence based practice to complete the learning health system cycle.

We provide a set of principles, grounded in research design, to cope with the problems of leveraging EHR data for various research and learning objectives, and highlight outstanding and important areas for future research and attention. EHR data is complex and intertwined with business and operational aspects of an organization, which may be unfamiliar to researchers and statisticians accustomed to data from clinical trials and registries. Using EHR data points means that one must attempt to understand the workflow that created them, particularly if a strong program of data governance is not in place. Despite this, there is growing appreciation for the inherent limitations of these data as well as momentum to improve its content and quality for research. Successful strategies will address fundamentals of research design (including confounding, sampling, and measurement bias) while embracing data quality limitations.

References

1. Embi PJ, Payne PR Clinical research informatics: challenges, opportunities and definition for an emerging domain. J Am Med Inform Assoc 2009;16(3):316-27.
2. Embi PJ.Clinical research informatics: survey of recent advances and trends in a maturing field. Yearb Med Inform 2013;8(1):178-84.
3. Murdoch TB, Detsky AS. The inevitable application of big data to health care. JAMA 2013;309(13): 1351-2.
4. Jiang X, Tse K, Wang S, Doan S, Kim H, Ohno-Machado L. Recent trends in biomedical informatics: a study based on JAMIA articles. J Am Med Inform Assoc 2013;20(e2):e198-205.
5. Shah NH, Tenenbaum JD. The coming age of data-driven medicine: translational bioinformatics' next frontier. J Am Med Inform Assoc 2012;19(e1):e2-4.
6. Bellazzi R, Zupan B. Predictive data mining in clinical medicine: current issues and guidelines. Int J Med Inform 2008;77(2):81-97.
7. Denny JC, Bastarache L, Ritchie MD, Carroll RJ, Zink R, Mosley JD, et al, Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. Nat Biotechnol 2013;31(12):1102-11.
8. Kass NE., Faden RR, Goodman SN, P. Pronovost P, Tunis S, Beauchamp TL. The research-treatment distinction: a problematic approach for determining which activities should have ethical oversight. Hastings Cent Rep 2013;Spec No:S4-S15.
9. Faden RR, Kass NE, Goodman SN, Pronovost P, Tunis S, Beauchamp TL. An ethics framework for a learning health care system: a departure from traditional research ethics and clinical ethics. Hastings Cent Rep 2013;Spec No S16-27.
10. Grossmann C, Powers B, McGinnis JM, editors. IOM, in Digital Infrastructure for the Learning Health System: The Foundation for Continuous Improvement in Health and Health Care: Workshop Series Summary, 2011; Washington (DC); 2011.
11. D'Avolio L, Ferguson R, Goryachev S, Woods P, Sabin T, O'Neil J, et al. Implementation of the Department of Veterans Affairs' first point-of-care clinical trial. J Am Med Inform Assoc 2012;19(e1):e170-6.
12. NIH. The NIH Director's Panel on Clinical Research Report to the Advisory Committee to the NIH Director, December, 1997. 1997 [cited 2011 May 15]; Available from: http://www.oenb.at/de/img/executive_summary--nih_directors_panel_on_clinical_research_report_12_97_tcm14-48582.pdf.
13. Patsopoulos NA. A pragmatic view on pragmatic trials. Dialogues Clin Neurosci 2011;13(2):217-24.
14. 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;20(e2):e226-31.
15. Hennekens CH, Buring JE. Mayrent, SL, editor. Epidemiology in Medicine. Boston: Little, Brown, and Company; 1987
16. Concato J, Lawler EV, Lew RA, Gaziano JM, Aslan M, Huang GD. Observational methods in comparative effectiveness research. Am J Med 2010;123(12 Suppl 1):e16-23.
17. Johnson ML, Crown W, Martin BC, Dormuth CR, Siebert U. Good research practices for comparative effectiveness research: analytic methods to improve causal inference from nonrandomized studies of treatment effects using secondary data sources: the ISPOR Good Research Practices for Retrospective Database Analysis Task Force Report--Part III. Value Health 2009;12(8):1062-73.
18. Rawlins M. De testimonio: on the evidence for decisions about the use of therapeutic interventions. Lancet 2008;372(9656):2152-61.
19. Gliklich RE, Dreyer NA, editors. AHRQ, Registries for Evaluating Patient Outcomes: A User's Guide. Agency for Healthcare Research and Quality: Rockville, MD; 2010.
20. Guilloud-Bataille M, De Crozes D, Rault G, Degioanni A, Feingold J. Cystic fibrosis mutations: report from the French Registry. The Clinical Centers of the CF. Hum Hered 2000;50(2):142-5.
21. Richesson R, Vehik K. Patient registries: utility, validity and inference. Adv Exp Med Biol 2010;686:87-104.
22. Metzger J. Using Computerized Registries in Chronic Disease Care. California HealthCare Foundation; 2004
23. Arts DG, De Keizer NF, Scheffer GJ. Defining and improving data quality in medical registries: a literature review, case study, and generic framework. J Am Med Inform Assoc 2002;9(6):600-11.
24. Paxton EW, Kiley ML, Love R, Barber TC, Funahashi TT, Inacio MC. Kaiser Permanente implant registries benefit patient safety, quality improvement, cost-effectiveness. Jt Comm J Qual Patient Saf 2013;39(6):246-52.
25. Molina-Ortiz EI, Vega AC, Calman NS. Patient registries in primary care: essential element for quality improvement. Mt Sinai J Med 2012;79(4):475-80.
26. McNeil JJ, Evans SM, Johnson NP, Cameron PA. Clinical-quality registries: their role in quality improvement. Med J Aust, 2010;192(5):244-5.
27. Butler J, Kalogeropoulos A. Registries and health care quality improvement. J Am Coll Cardiol 2009;54(14):1290-2.
28. Franklin PD, Allison JJ, Ayers DC. Beyond joint implant registries: a patient-centered research consortium for comparative effectiveness in total joint replacement. JAMA 2012;308(12):1217-8.
29. Aghayev E, Henning J, Munting E, Diel P, Moulin P, Röder C, et al; SWISSspine and Spine Tango Registry groups. Comparative effectiveness research across two spine registries. Eur Spine J 2012;21(8):1640-7.
30. Shah BR, Drozda J, Peterson ED. Leveraging observational registries to inform comparative effectiveness research. Am Heart J 2010;160(1):8-15.
31. Psaty BM, Larson EB. Investments in infrastruc-

ture for diverse research resources and the health of the public. JAMA 2013;309(18):1895-6.
32. Methodology Committee of the Patient-Centered Outcomes Research (PCORI). Methodological standards and patient-centeredness in comparative effectiveness research: the PCORI perspective. JAMA 2012;307(15):1636-40.
33. Hersh WR, Weiner MG, Embi PJ, Logan JR, Payne PR, Bernstam EV, et al. Recommendations for the Use of Operational Electronic Health Record Data in Comparative Effectiveness Research eGEMs (Generating Evidence & Methods to improve patient outcomes), Med Care 2013 Aug;51(8 Suppl 3):S30-7.
34. Hand DJ. Data mining: statistics and more? The American Statistician 1998;52(2).
35. Bacchieri A, della Cioppa G. Observational Studies, in Fundamentals of Clinical Research. Bridging Medicine, Statistics, and Operations. Springer-Verlag: Milano; 2007. p. 28-56.
36. Vandenbroucke JP. When are observational studies as credible as randomised trials? Lancet 2004;363(9422):1728-31.
37. Green SB, Byar DP. Using observational data from registries to compare treatments: the fallacy of omnimetrics. Stat Med 1984;3(4):361-73.
38. Häyrinen K, Saranto K, Nykänen P. Definition, structure, content, use and impacts of electronic health records. Int J Med Inform 2008 May;77(5):291-304.
39. Köpcke F, Trinczek B, Majeed RW, Schreiweis B, Wenk J, Leusch T, et al. Evaluation of data completeness in the electronic health record for the purpose of patient recruitment into clinical trials: a retrospective analysis of element presence. BMC Med Inform Decis Mak 2013;13: p. 37.
40. Hammond WE, McCourt B. Making sense of standards. J AHIMA 2007;78(60–61).
41. Kane RL. Understanding Health Care Outcomes Research. Gaithersburg: Aspen; 1997.
42. Estabrooks PA, Boyle M, Emmons KM, Glasgow RE, Hesse BW, Kaplan RM, et al. Harmonized patient-reported data elements in the electronic health record: supporting meaningful use by primary care action on health behaviors and key psychosocial factors. J Am Med Inform Assoc 2012 Jul-Aug;19(4):575-8.
43. Hripcsak G, Albers DJ. Correlating electronic health record concepts with healthcare process events. J Am Med Inform Assoc 2013;20(e2):e311-8.
44. Kokkonen EW, Davis SA, Lin HC, Dabade TS, Feldman SR, Fleischer AB Jr. Use of electronic medical records differs by specialty and office settings. J Am Med Inform Assoc 2013;20(e1):e33-8.
45. Embi PJ, Weir C, Efthimiadis EN, Thielke SM, Hedeen AN, Hammond KW. Computerized provider documentation: findings and implications of a multisite study of clinicians and administrators. J Am Med Inform Assoc 2013;20(4):718-26.
46. 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.
47. Curcin, V., S. Miles, R. Danger, Y. Chen, Bache R, Taweel A. Implementing interoperable provenance in biomedical research. Future Generation Computer Systems-the International Journal of Grid Computing and Escience 2014;34:1-16.
48. I.O.f. Standardization, editor. ISO, ISO-8000-2:2012(E) Data Quality - Part 2: Vocabulary; 2012.
49. 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.
50. 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.
51. Weiskopf NG, Hripcsak G, Swaminathan S, Weng C. Defining and measuring completeness of electronic health records for secondary use. J Biomed Inform 2013;46(5):830-6.
52. Zozus MN. Assessing Data Quality for Healthcare Systems Data Used in Clinical Research. An HSC Research Collaboratory Core White Paper; 2013.
53. Madigan D, Ryan PB, Schuemie M, Stang PE, Overhage JM, Hartzema AG, et al. Evaluating the impact of database heterogeneity on observational study results. Am J Epidemiol 2013;178(4):645-51.
54. Schmiemann G. [The preliminary draft of the methodology report by the Patient-Centered Outcomes Research Institute]. Z Evid Fortbild Qual Gesundhwes 2012;106(7):496-9.
55. PCORI. PCORI Methodology Standards; 2012.
56. Moher D. CONSORT: an evolving tool to help improve the quality of reports of randomized controlled trials. Consolidated Standards of Reporting Trials. JAMA 1998;279(18):1489-91.
57. Vandenbroucke JP, von Elm E, Altman DG, Gøtzsche PC, Mulrow CD, Pocock SJ, et al., STROBE Initiative. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. Epidemiology 2007;18(6):805-35.
58. Brown JS, Kahn M, Toh S. Data quality assessment for comparative effectiveness research in distributed data networks. Med Care 2013;51(8 Suppl 3):S22-9.
59. Holve E, Kahn M, Nahm M, Ryan P, Weiskopf N. A comprehensive framework for data quality assessment in CER. AMIA Summits Transl Sci Proc 2013;2013:86-8.
60. Pathak J, Kho AN, Denny JC. Electronic health records-driven phenotyping: challenges, recent advances, and perspectives. J Am Med Inform Assoc 2013;20(e2):e206-11.
61. Gregg EW, Li Y, Wang J, Burrows NR, Ali MK, Rolka D, Williams DE, Geiss L. Changes in diabetes-related complications in the United States, 1990-2010. N Engl J Med 2014;370(16):1514-23.
62. Chen Y, Carroll RJ, Hinz ER, Shah A, Eyler AE, Denny JC, et al. Applying active learning to high-throughput phenotyping algorithms for electronic health records data. J Am Med Inform Assoc, 2013. 20(e2): p. e253-9.
63. Cobb JN, Declerck G, Greenberg A, Clark R, McCouch S. Next-generation phenotyping: requirements and strategies for enhancing our understanding of genotype-phenotype relationships and its relevance to crop improvement. Theor Appl Genet 2013;126(4):867-87.
64. Conway M, Berg RL, Carrell D, Denny JC, Kho AN, Kullo IJ, et al. Analyzing the heterogeneity and complexity of Electronic Health Record oriented phenotyping algorithms. AMIA Annu Symp Proc 2011;2011:274-83.
65. Newton KM, Peissig PL, Kho AN, Bielinski SJ, Berg RL, Choudhary V, et al. Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. J Am Med Inform Assoc 2013;Jun;20(e1):e147-54
66. Shivade C, Raghavan P, Fosler-Lussier E, Embi PJ, Elhadad N, Johnson SB, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. J Am Med Inform Assoc 2014 Mar-Apr;21(2):221-30.
67. Rosenman M, He J, Martin J, Nutakki K, Eckert G, Lane K, et al. Database queries for hospitalizations for acute congestive heart failure: flexible methods and validation based on set theory. J Am Med Inform Assoc 2014;21(2):345-52.
68. Peissig PL, Rasmussen LV, Berg RL, Linneman JG, McCarty CA, Waudby C, et al. Importance of multi-modal approaches to effectively identify cataract cases from electronic health records. J Am Med Inform Assoc 2012;19(2):225-34.
69. Boyd AD, Li JJ, Burton MD, Jonen M, Gardeux V, Achour I, et al. The discriminatory cost of ICD-10-CM transition between clinical specialties: metrics, case study, and mitigating tools. J Am Med Inform Assoc 2013;20(4):708-17.
70. Carnahan RM, Moores KG. Mini-Sentinel's systematic reviews of validated methods for identifying health outcomes using administrative and claims data: methods and lessons learned. Pharmacoepidemiol Drug Saf 2012;21 Suppl 1:82-9.
71. Fox BI, Hollingsworth JC, Gray MD, Hollingsworth ML, Gao J, Hansen RA. Developing an expert panel process to refine health outcome definitions in observational data. J Biomed Inform 2013;46(5):795-804.
72. eMERGE. What is the Phenotype KnowledgeBase? 2012 [cited 2013 March 25]; Available from: http://www.phekb.org/.
73. Pathak J, Wang J, Kashyap S, Basford M, Li R, Masys DR, et al. Mapping clinical phenotype data elements to standardized metadata repositories and controlled terminologies: the eMERGE Network experience. J Am Med Inform Assoc 2011;18(4):376-86.
74. Thompson WK, Rasmussen LV, Pacheco JA, Peissig PL, Denny JC, Kho AN, et al. An evaluation of the NQF Quality Data Model for representing Electronic Health Record driven phenotyping algorithms. AMIA Annu Symp Proc 2012;2012:911-20.
75. Chute CG, Pathak J, Savova GK, Bailey KR, Schor MI, Hart LA, et al. The SHARPn project on secondary use of Electronic Medical Record data: progress, plans, and possibilities. AMIA Annu Symp Proc 2011;2011:248-56.
76. 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. J Biomed Inform 2012;45(4):763-71.
77. Pathak J, Bailey KR, Beebe CE, Bethard S, Carrell DC, Chen PJ, et al. Normalization and standardization of electronic health records for high-throughput phenotyping: the SHARPn consortium. J Am

Med Inform Assoc 2013;20(e2):e341-8.
78. Pathak J, Endle C, Suesse D, Peterson K, Stancl C, Li D, et al. PhenotypePortal: An Open-Source Library and Platform for Authoring, Executing and Visualization of Electronic Health Records Driven Phenotyping Algorithms. AMIA Summits Transl Sci Proc 2013.
79. Post AR, Kurc T, Cholleti S, Gao J, Lin X, Bornstein W, et al. The Analytic Information Warehouse (AIW): a platform for analytics using electronic health record data. J Biomed Inform 2013;46(3):410-24.
80. Post AR, Kure T, Willard R, Rathod H, et al. Clinical Phenotyping with the Analytic Information Warehouse. in AMIA Summits Transl Sci Proc 2013.
81. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. J Am Med Inform Assoc 2013;20(1):117-21.
82. Richesson RL, Rusincovitch SA, Wixted D, Batch BC, Feinglos MN, Miranda ML, et al. A comparison of phenotype definitions for diabetes mellitus. J Am Med Inform Assoc 2013;20(e2):e319-26.
83. Richesson RL, Rusincovitch SA, Smerek MM, Pathak J. Standardized Representation for Electronic Health Record-Driven Phenotypes, in Accepted for presentation at: AMIA Joint Summits for Translational Research, Summit on Clinical Research Informatics. San Francisco: AMIA; 2014
84. Olsen L, Aisner D, McGinnis JM, editors. The Learning Healthcare System: Workshop Summary (IOM Roundtable on Evidence-Based Medicine). The National Academies Press; 2007.
85. Eapen ZJ, Vavalle JP, Granger CB, Harrington RA, Peterson ED, Califf RM. Rescuing clinical trials in the United States and beyond: a call for action. Am Heart J 2013;165(6):837-47.
86. 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 2010;48(6 Suppl):S45-51.
87. Platt R, Carnahan RM, Brown JS, Chrischilles E, Curtis LH, Hennessy S, et al. The U.S. Food and Drug Administration's Mini-Sentinel program: status and direction. Pharmacoepidemiol Drug Saf 2012;21 Suppl 1:1-8.
88. Bhandari M, Giannoudis PV. Evidence-based medicine: what it is and what it is not. Injury 2006;37(4):302-6.
89. Manchikanti L, Datta S, Smith HS, Hirsch JA. Evidence-based medicine, systematic reviews, and guidelines in interventional pain management: part 6. Systematic reviews and meta-analyses of observational studies. Pain Physician 2009;12(5):819-50.
90. Manchikanti L, Benyamin RM, Helm S, Hirsch JA. Evidence-based medicine, systematic reviews, and guidelines in interventional pain management: part 3: systematic reviews and meta-analyses of randomized trials. Pain Physician 2009;12(1):35-72.
91. Yuan Y, Hunt RH. Systematic reviews: the good, the bad, and the ugly. Am J Gastroenterol 2009;104(5):1086-92.
92. Wells BJ, Nowacki A, Chagin SK, Kattan MW. Strategies for Handling Missing Data in Electronic Health Record Derived Data. eGEMs (Generating Evidence & Methods to improve patient outcomes), 2013. 1: Iss. 3, Article 7.
93. McCann E. EHR users unhappy, many switching. Vendors often failing to meet provider needs, report finds. Healthcare IT News 2013.
94. Ohno-Machado L. Data science and informatics: when it comes to biomedical data, is there a real distinction? J Am Med Inform Assoc 2013;20(6):1009.
95. NIH. NIH Names Dr. Philip E. Bourne First Associate Director for Data Science. 2013 [cited 2013 December 19]; Available from: http://www.nih.gov/news/health/dec2013/od-09.htm.
96. ICH. Clinical Safety Data Management and Definitions and Standards for Expedited Reporting E2A 1994. International Conference on Harmonisation; 1994.
97. Califf, R.M., Attribution of causality in clinical research: an important element for learning health systems. Am Heart J, 2013. 165(3): p. 252-3.
98. Peeples MM, Iyer AK, Cohen JL. Integration of a mobile-integrated therapy with electronic health records: lessons learned. J Diabetes Sci Technol 2013;7(3):602-11.
99. Eggleston EM, Weitzman ER. Innovative uses of electronic health records and social media for public health surveillance. Curr Diab Rep 2014;14(3):468.
100. Ansermino JM. Universal access to essential vital signs monitoring. Anesth Analg 2013;117(4):883-90.
101. Fortney JC, Burgess JF Jr, Bosworth HB, Booth BM, Kaboli PJ. A re-conceptualization of access for 21st century healthcare. J Gen Intern Med 2011;26 Suppl 2:639-47.
102. IOM. Capturing Social and Behavioral Domains in Electronic Health Records: Phase 1. The National Academies Press; 2014.
103. Rojo MG, Castro AM, Goncalves L. COST Action "EuroTelepath": digital pathology integration in electronic health record, including primary care centres. Diagn Pathol 2011;6 Suppl 1:S6.
104. Hammond WE, Bailey C, Boucher P, Spohr M, Whitaker P. Connecting information to improve health. Health Aff (Millwood) 2010;29(2):284-8.
105. Joffe E, Byrne MJ, Reeder P, Herskovic JR, Johnson CW, McCoy AB, et al. A benchmark comparison of deterministic and probabilistic methods for defining manual review datasets in duplicate records reconciliation. J Am Med Inform Assoc 2014;21(1):97-104.
106. Dong X, Bahroos N, Sadhu E, Jackson T, Chukhman M, Johnson R, et al. Leverage hadoop framework for large scale clinical informatics applications. AMIA Summits Transl Sci Proc 2013;2013:53.
107. Russom P. Integrating Hadoop into Business Intelligence and Data Warehousing. TDWI Best Practices Report; 2013.
108. Embi PJ, Payne PR. Evidence generating medicine: redefining the research-practice relationship to complete the evidence cycle. Med Care 2013;51(8 Suppl 3):S87-91.

Correspondence to:
Rachel Richesson, PhD, MPH
Duke University School of Nursing
2007 Pearson Bldg, 311 Trent Drive
Durham, NC, 27710
USA
Tel: +1 (919) 681-0825
E-mai: rachel.richesson@duke.edu