# Clinical Research Informatics for Big Data and Precision Medicine

C. Weng[1], M. G. Kahn[2]

[1] Department of Biomedical Informatics, Columbia University, New York, NY 10032 USA

[2] Department of Pediatrics, University of Colorado, Denver, CO 80045 USA

## Summary

**Objectives**: To reflect on the notable events and significant developments in Clinical Research Informatics (CRI) in the year of 2015 and discuss near-term trends impacting CRI.

**Methods**: We selected key publications that highlight not only important recent advances in CRI but also notable events likely to have significant impact on CRI activities over the next few years or longer, and consulted the discussions in relevant scientific communities and an online living textbook for modern clinical trials. We also related the new concepts with old problems to improve the continuity of CRI research.

**Results**: The highlights in CRI in 2015 include the growing adoption of electronic health records (EHR), the rapid development of regional, national, and global clinical data research networks for using EHR data to integrate scalable clinical research with clinical care and generate robust medical evidence. Data quality, integration, and fusion, data access by researchers, study transparency, results reproducibility, and infrastructure sustainability are persistent challenges.

**Conclusion**: The advances in Big Data Analytics and Internet technologies together with the engagement of citizens in sciences are shaping the global clinical research enterprise, which is getting more open and increasingly stakeholder-centered, where stakeholders include patients, clinicians, researchers, and sponsors.

## Keywords

Evidence-based medicine; learning health system; individualized medicine; patient selection; electronic health records; data bank; data quality

**Yearb Med Inform 2016:211-8**
**http://dx.doi.org/10.15265/IY-2016-019**
**Published online November 10, 2016**

## Introduction

Clinical Research Informatics (CRI), a recently defined subfield of biomedical informatics that focuses on informatics support for medical evidence generation [1], has continued to enlarge its scope and importance in supporting the broadening agendas in clinical and translational sciences [2]. Over the past decade, accelerating at an explosive pace, biomedical research has moved into the era of massive-scale digitalization of data and computationally-intensive quantitative analytics spanning molecular, clinical, and population-level data and including measuring events from picoseconds to decades-long time scales. Novel digital devices, from high-throughput next generation deep sequencing machines to continuous real-time bio-sensing tattoos [3], continue to challenge the CRI community to develop new infrastructure capacities in addition to data and knowledge discovery tools that can handle petabyte-size data stores. The "Information Commons" highlighted in the IOM report on Precision Medicine is envisioned to integrate vast amounts of data with constantly evolving biomedical knowledge [4]. The informatics underpinnings that enable and accelerate this transition to large-scale integrated data and knowledge systems has to respond with innovations across the CRI spectrum. At the same time, research and discovery at the scale that is technically possible presents new challenges, not only to CRI but also to data sharing and privacy policies, and the regulatory bodies that must respond to this rapidly changing data-driven agenda [5].

In 2012, we presented a conceptual model intended to capture the CRI landscape of activities and challenges to contextualize eighteen new publications in a special supplement of the Journal of the American Medical Informatics Association focused exclusively on CRI research results and innovations [6]. The central thesis of our model was CRI's unique role in enabling "informatics-enabled clinical research workflow" and the methods and tools needed to support early-stage translational discoveries and later-stage evidence generation and synthesis, personalized evidence application and populations surveillance. We ended that publication with the following predictions:

> "We expect the CRI research agenda will continue to evolve to become more precise, predictive, preemptive, and participatory, in parallel with the development of P4 medicine. We anticipate more patient-centered research decision support and innovative consent programs to strengthen patient participation, including specifying how an individual's research data will be used and by whom. We also expect more CRI research that is informed by and responsive to patient or population needs."

We revisited our 2012 conceptual model and examined current advances in CRI against that model, adding new elements where needed and modifying those that have evolved. We evaluated our predictions from four years ago and updated them to reflect both the anticipated and unanticipated shifts in the translational research landscape and their impact on CRI in 2016 and into the immediate future.

## Methods

We did not conduct an exhaustive formal literature review. Instead, we selected notable publications and events based on our personal weighing of their importance and by referring to the public expert opinions on the Internet, such as Dr. Peter Embi's "CRI Year in Review" (http://www.embi.net/cri.html) and "Rethinking Clinical Trials" provided by Duke University (http://sites.duke.edu/rethinkingclinicaltrials/), and the discussions within the AMIA CTSA community. We summarized our understanding of the state of the art and the recent trends in CRI and described them below.

# Findings

Figure 1 illustrates our updated conceptual model of the state-of-the-art CRI methods and issues. Comparing this model to the one that we previously presented in year 2012 [6], changes have occurred in the overall workflow, the underlying data sources, and the CRI foundational components. New workflow components include the addition of Big Data Sciences as a source of new research questions and the expansion of evidence generation and synthesis by including evidence appraisal. Evidence appraisal involves critical and systematic review of medical evidence to judge its trustworthiness, value and applicability in a particular context. For example, it uncovers potential biases in clinical research participant selection and examines factors such as internal validity, generalizability and relevance [7-10]. It is particularly relevant given the rapid growth of new high-throughput data analytics and hypothesis generation methods that give rise to more controversial findings than ever [11-13]. New data sources are acknowledged in our conceptual framework with the addition of wearable devices, patient-reported outcomes, social media, and environmental sensors. We anticipate that this list of electronic data sources will continue to grow as portable, wearable, and always connected devices become more widely used. The largest changes are reflected in the core informatics foundational components that now include Big Data Analytics, Data Fusion, Workflow Support, and Phenotyping using electronic patient data. In addition, record linkage has been generalized to data linkage, information extraction now includes natural language processing and text mining, and knowledge management has been expanded to knowledge engineering. All these additions are preparing the CRI community to better advance the Precision Medicine [4] and Learning Health System [14, 15] agendas.

## 1 The Arrival of Big Data

The National Academy of Medicine (the former Institute of Medicine) predicted back in 2003 that the wide adoption of electronic health record (EHR) systems would eventually enable the collection and aggregation of large amounts of electronic patient data to facilitate clinical decision support and accelerate evidence generation [16]. With the continued widening adoption of EHR systems globally, this prediction has been partially realized. According to the latest statistics, 75% of the hospitals in the United States have adopted at least one basic EHR system and the adoption rate is still steadily rising [17]. Less successful has been the broad adoption of real-time clinical decision support and rapid-cycle evidence generation as envisioned by Embi [18].

The Big Data acquired by EHRs enables us to pursue a long-sought vision, a rapid learning healthcare system that integrates clinical research and clinical care, where clinical data are a basic staple of health learning [14, 15, 19, 20], and enables large-scale observational studies and large pragmatic trials for rapid evidence generation and validation using EHR data [21, 22]. Citizen engagement is central to the success of this learning health system [23]. Regional or national learning health systems, such as PaTH and PEDSnet, have been developed based on enterprise data warehouses or clinical data research networks [24-28]. These working examples help researchers continue to make the functionalities of learning health systems more specific and concrete [29-31].

Clearly, Big Data has been recognized as the foundation of a learning health system and a catalyst for optimizing clinical research design [32]. In order to harness Big Clinical Data increasingly made available by EHRs, numerous clinical data research networks, loosely coupled or tightly coupled, have been developed across the world. In the United States, the National Center for Advancing Translational Sciences (NCATS) has established the Accrual to Clinical Trials (CTSA ACT) network (https://ncats.nih.gov/pubs/features/ctsa-act). The Patient Centered Outcomes Research Institute (PCORI) has launched the PCORnet [33], which includes thirteen clinical data research networks and nineteen patient-powered research networks that cover most of the states in the United States (USA) to conduct both randomized trials and observational comparative effectiveness studies using EHR data. These networks expand existing large-scale data sharing networks such as the CDC-sponsored Vaccine DataLink [34], Health Care Systems Research Network (formerly called the HMORN) [35], FDA-sponsored Mini-Sentinel drug surveillance network [36] and AHRQ-sponsored large-scale platforms that support multi-institutional comparative effectiveness research [37, 38]. In September 2015, the Electronic Medical Records and Genomics (eMERGE) network sponsored by the National Human Genomics Research Institute [39] also embarked on its third phase of research with a particular focus on returning actionable pathogenic genetic variants to patients and families via genomic decision support in clinical care settings using EHRs or personal health records across 9 participating sites in the USA.

In Europe, the large-scale EHR4CR project has entered its 5th year as the flagship project for using EHR data for accelerating clinical trials [40]. A recent cost-effectiveness study of using EHR data for clinical trials based on the EHR4CR project has suggested that optimizing clinical trial design and execution with the EHR4CR platform would generate substantial added value for pharmaceutical industry, as the main sponsors of clinical trials in Europe, and beyond [41]. Similarly, the EU-ADR project has established a multi-national drug safety surveillance system [42].

Internationally, a global collaborative network called The Observational Health Data Sciences and Informatics (OHDSI) has enabled large-scale evidence aggregation using more than 680 million patients' electronic data [43] and helped shape the emerging networked science for biomedical research based on interoperable data [44].

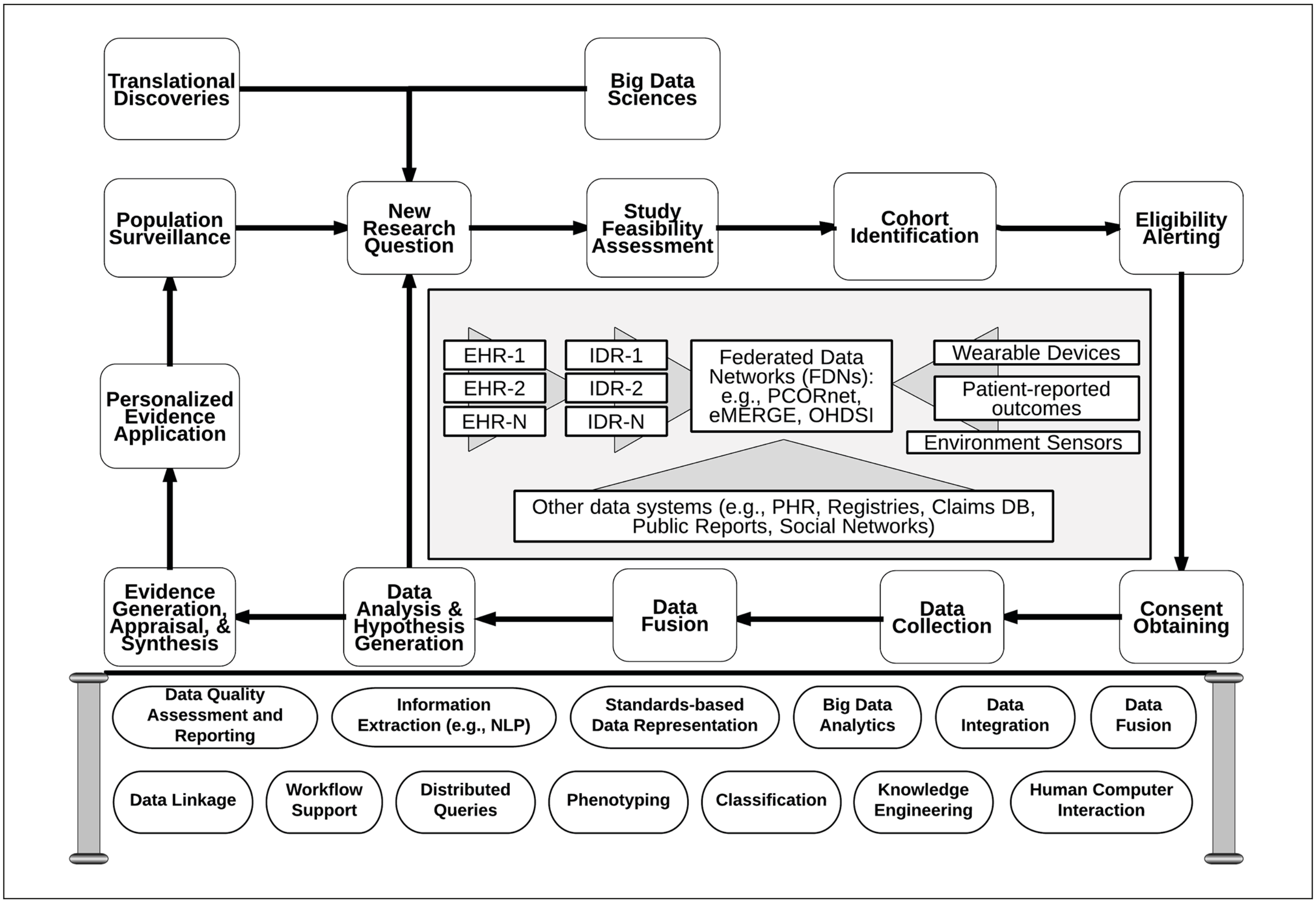
Translational Discoveries
Big Data Sciences
Population Surveillance
New Research Question
Study Feasibility Assessment
Cohort Identification
Eligibility Alerting
Personalized Evidence Application
EHR-1
EHR-2
EHR-N
IDR-1
IDR-2
IDR-N
Federated Data Networks (FDNs): e.g., PCORnet, eMERGE, OHDSI
Wearable Devices
Patient-reported outcomes
Environment Sensors
Other data systems (e.g., PHR, Registries, Claims DB, Public Reports, Social Networks)
Evidence Generation, Appraisal, & Synthesis
Data Analysis & Hypothesis Generation
Data Fusion
Data Collection
Consent Obtaining
Data Quality Assessment and Reporting
Information Extraction (e.g., NLP)
Standards-based Data Representation
Big Data Analytics
Data Integration
Data Fusion
Data Linkage
Workflow Support
Distributed Queries
Phenotyping
Classification
Knowledge Engineering
Human Computer Interaction

**Fig. 1** Our conceptual framework of the field of clinical research informatics updated/expanded from [6].

Although the proliferation and success of such networks are exciting, we should not forget the lessons learned from previously heavily investigated but later abandoned or terminated large-scale research networks due to troubles from impractical project goals and study designs, ineffective management, and failed oversight, such the National Children's Study network [45, 46] and The Cancer Bioinformatics Grid (caBIG) [47] program in the USA. The sustainability of large-scale data infrastructures remains a largely unresolved issue and a primary concern of the CRI community, who is heavily involved in the development and maintenance of such large and complex systems [48, 49]. Unlike most tightly-coupled networks that operate by external funding support, as a loosely-coupled network, OHDSI shows unusual sustainability promise in that only the very early experimental sites received funding and nearly all of the existing data partners continue to be active in the OHDSI Collaborative without depending solely upon designated external funding, illustrating the resilience of open community-based collaborations rather than the brittleness of top-down centralized collaborations. This phenomenon has been described by others who have achieved sustained adoption of widely-deployed CRI tools [50, 51].

The growing availability of Big Clinical Data promises to accelerate drug discovery [52]. It has also accelerated the knowledge engineering of reproducible and portable computerized clinical phenotypes [53] in the hope of using standards-based algorithms for achieving interoperability in genome-wide association studies and phenome-wide association studies of determinants of disease risks and for clinical study cohort identification and recruitment. Big Data comes from not only EHRs, but also sensors, wearable devices, and consumer-generated Big Data in social media [54] such as Facebook [55] and Twitter [56]. Weber et al. highlights that EHR data reflects only a small portion of data relevant to understanding the full context of health and disease [57]. A recently published JAMA article pointed out that Twitter streams can be very effective for public health surveillance [56]. To support Big Data sciences, the National Institutes of Health also launched the Big Data to Knowledge (BD2K) Initiative [58, 59] and funded a series of Centers of Excellence for using Big Data Analytics as well as a range of training and curriculum grants to train the next generation data scientists [59].

## 2 Advances in Big Data Analytics

The era in which the integration, fusion, or linkage of data across the biological, clinical, patient, and environmental spectrums is increasingly common has arrived. Ma et al. linked public clinical trial summaries with a medical encyclopedia to identify questionable exclusion criteria [7]. Lorgelly et al. linked cancer data with commonwealth reimbursement data to infer which patient, disease, genomic and treatment characteristics explain variation in health expenditure [60]. Data integration is also called data aggregation. It is a process where data of the same type from multiple sources, such as EHR data from multiple institutions are integrated in a shared central data warehouse [61]. It is used often to improve sample size for clinical research. In contrast, data fusion emphasizes arriving at improved understanding using different but complementary data about the same object [62, 63]. For example, Wu et al. developed a multi-omic data fusion approach to map the crosstalk between metabolic phenotype and microRNA data to understand the systemic consequences Roux-en-Y gastric bypass surgery [64]. The major challenge for data integration is semantic interoperability, whereas the primary challenges facing data fusion include integration of data and knowledge representation from multiple different yet complementary perspectives and with different granularities and resolutions.

One of the most striking developments in recent years has been the massive expansion in data storage capabilities, driven by cloud-based technologies developed to support the enormous data needs of search and e-commerce vendors, such as Google (now Alphabet) and Amazon, and rapidly adopted in biomedicine and health sciences [65-67]. With near limitless storage, data previously difficult to access, such as geographic, climate, economic and social media data sets, can now be linked to enable population-based analytics that have never before been possible. For life sciences research and CRI professionals, these new data storage and data retrieval architectures, coupled with on-demand, scalable computational resources, has enable petabyte-size data sets to be stored and shared for worldwide access and analysis. For example, Amazon and Google provide access to The Cancer Genome Atlas, The International Cancer Genomic Consortium, 1000 Genomes Project, and 3000 Rice Genome data sets on their cloud services (http://aws.amazon.com/public-data-sets/). Google has a similar library of large published data sets that can be accessed worldwide (http://google-genomics.readthedocs.org/en/latest/use_cases/discover_public_data/genomic_data_toc.html). In this respect, the public sharing of clinical data remains far behind the sharing of genomic data, due to widespread concerns about the growing ability to re-identify individuals [68]. An increasing body of literature suggests re-identification risks also exist with genomic data [69].

With massive data sets that are far too large or too complex to analyze using traditional local computational methods, new approaches for performing analytics using distributed techniques that “bring analytics to the data” rather than “submit data to the analytics” are a new area of active CRI research [70-74]. These have also been adapted to enable distributed analytics of sensitive clinical data without requiring data partners to release any patient-level data, offering a new approach for reducing concerns about patient re-identification. One set of tools that continue to evolve slower than anticipated are systems that automate semantic harmonization and annotation for data and knowledge integration [75]. Current methods remain difficult to use, mostly relying on human annotation. The promise of semantic web technologies has not materialized for general use although some striking examples show the potential of these methods [76, 77].

## 3 Reproducibility, Generalizability, and Ethical Implications of Big Data

Research based on reuse of clinical data is frequently questioned for reproducibility [78]. Publishers and scientists have increasingly recognized the importance of sharing data for improving reproducibility [79]. The Scientific Data (http://www.nature.com/sdata/) journal was launched this year in response to the rising need to help scientists permanently archive, share, and disseminate valuable research data. It is foreseeable that more journals will start accommodating data archiving needs for future publications. Related to archiving data to support the principles of Open Science and Reproducible Research is a newly funded BD2K effort, called bioCADDIE (https://biocaddie.org), to develop a comprehensive set of descriptors of data sets to support the search and discovery of available sharable data resources.

Safran recently summarized the value of reuse of clinical data made available by EHRs, the potential problems with large aggregations of these data that do not necessarily have consistent meanings, the policy frameworks that have been formulated, and the major challenges in the coming years [80]. More recently, Hersh and colleagues expanded these concerns in the context of more recent large scale comparative effectiveness research networks [81].

Understanding the ethnical implications of Big Data lags behind [82] and the existing regulatory framework falls short to meet the needs of the evolving data capabilities. Mittelstadt et al. identified five key areas of concerns [82]: 1) informed consent, (2) privacy, (3) ownership, (4) epistemology and objectivity, and (5) ‘Big Data Divides’ created between those who have or lack the necessary resources to analyze increasingly large datasets. Data breach is still a significant threat to organizations and individuals curating, using, and sharing these data. The imperative for protecting patient privacy and data confidentiality requires advanced network security safeguards and enhanced patient privacy and data confidentiality protection. The conversation about privacy has shifted away from ensuring privacy to assessing risk instead. It is no longer possible to guarantee privacy. It is only possible to estimate and manage risk.

Six additional areas of concern were suggested to require much closer scrutiny in the immediate future: (6) the dangers of ignoring group-level ethical harms; (7) the importance of epistemology in assessing the ethics of Big Data; (8) the changing nature of fiduciary relationships that become increasingly data saturated; (9) the need to distinguish between ‘academic’ and ‘commercial’ Big Data practices in terms of potential harm to data subjects; (10) future problems with ownership of intellectual property generated from analysis of aggregated datasets; and

(11) the difficulty of providing meaningful access rights to individual data subjects that lack necessary resources. For this last theme, data access by non-technical stakeholders such as clinical researchers, studies have found that data query mediation is a laborious and error-prone process and has not received adequate attention but can negatively affect research reliability for studies based on these data [83-85]. As Mittelstadt et al. pointed out, "these themes provide a thorough critical framework to guide ethical assessment and governance of emerging Big Data practices." New studies have shed light on borrowing ideas from library and information sciences or dialogue system research to improve query mediations for biomedical Big Data [86, 87].

Meanwhile, continued progress on data interoperability and clinical research regulations has reached a new milestone this year. Richesson and Chute published a special issue for JAMIA on data interoperability standards and concluded that "data standards are finally down to business for enabling emerging interoperability" [88]. The Notice of Proposed Rulemaking (NPRM) for regulating clinical research was released this September to enhance protection of research participants while streamlining IRB review efficiency [89].

The above progress together with the technology readiness well prepare the CRI community to engage and lead in the newly launched initiative for advancing Precision Medicine in the USA, which has an urgent need for Big Data and large representative population samples. Genetic variants discovered from small and unrepresentative population samples may mislead the public (http://www.theatlantic.com/science/archive/2015/09/genome-big-data-disease-genes/404356/). In order to help verify genetic discoveries, libraries of genetic variants have been developed. ClinVar was created to address the need for transparency in genetic evidence [90, 91]. Food and Drug Administration has also launched OpenFDA (https://open.fda.gov/) and PrecisionFDA (https://precision.fda.gov) to help improve the transparency and collaboration in safety data around drugs and devices.

The newly released strategic plan of the National Institutes of Health of the United States [92] also highlights the imperative for developing the "science of science" and "evaluating steps to enhance rigor and reproducibility". It is foreseeable even future funding decisions will be based on data-driven evidence of research quality and impact.

## 4 Data Quality Challenges

Unlike "traditional" prospective clinical trials that utilize detailed data collection tools and procedures and rely on trained data collection personnel, EHR and PHR databases contain data collected during routine clinical care by practitioners focused on patient care or by patients focused on capturing their health care experiences rather than research. Differences in clinical workflows, practice standards, patient populations, available technologies, and referral resources impact what data are collected and how they are documented. Numerous studies have highlighted significant concerns about the quality of data in EHRs [34, 93-100]. CER studies seek to exploit real-world diversity in order to detect and understand determinants impacting outcome variation. Data quality and completeness problems, however, may affect the validity of CER findings [101, 102]. The importance of good quality data in clinical research is well accepted [103, 104]. There are substantial efforts to develop robust analytic methods for extracting valid knowledge from observational data, but there are no formal data quality assessment guidelines, analytic methods, or reporting requirements. Methods for categorizing, analyzing, and reporting on data quality, however, are poorly developed. Most approaches to data quality (DQ) assessment are ad hoc, developed based on an intuitive understanding of data quality challenges, and focused on specific research questions [105-108]. Few systematic approaches to DQ assessment for the secondary use of clinically-obtained data have been proposed. Current methods do not emphasize the need to improve the reporting of DQ results [109].

# Discussion

In the four years since the publication of our initial conceptual model, clinical research informatics continues to evolve and expand. Our previous work emphasized tools that support clinical research workflow and new clinical research data networks. A similar review of the CRI landscape by Embi and Payne described six core CRI activities: (1) data capture, collection and re-use, (2) standards, (3) tensions with regulatory and ethical issues, (4) research networks and team science, (5) improved user experiences, and (6) integration of clinical research and practice [110]. While these efforts have continued over the intervening period, a clear shift toward a more data-centric perspective permeates this update. A widening array of data sources, data sharing methods, and Big Data architectures, tools and analytics are dominating the current CRI agenda. Tied closely to this shift is the rapid development of large-scale data sharing networks and new distributed query and analytics infrastructures, including the appearance of a new common data model from PCORI [38, 111, 112]. New infrastructure and methods for record linkage, data fusion, natural language processing (NLP), and standardized phenotyping have enabled new data discovery opportunities that were being discussed but not widely implemented during our previous CRI overview [6].

As CRI investigators implement these expansive data resources and develop new tools for linking, exploring, visualizing and analyzing complex data sets, how will these data be used to accelerate translational research and new discoveries? "Traditional" uses include retrospective clinical research, study feasibility, and cohort selection or patient recruitment. New data sources also enable new capabilities, including the development of "deep clinical phenotypes" that include the use of biomarkers, imaging results, and NLP to extract clinical features not available from typical databases based on "coded" data elements [113-116]. Data linkages that combine clinical and billing data allow analysis of longitudinal outcomes; linkages with environmental exposures adds new dimensions to determining disease risks across broad patient populations [113, 116]. The inclusion of diverse clinical practices allows assessment of the relationship between health system features on disease diagnosis, treatment patterns, and outcomes [117]. While these types of studies have been performed by investigators for many years, the new data infrastructures hold the promise of dramatically reducing the cost and effort required to do similar studies at population sizes and in diverse practice settings that were not previously available or affordable [49, 118].

New opportunities also bring new challenges. We have noted the lack of clarity around the ethical use of large-scale, linked data, the growing gap between the regulatory restrictions and the ability to maintain patient privacy, the need to promote patient engagement in complex data sciences programs, the need to better understand the impact of data quality and biases across various data sources, and the lack of competent infrastructure to fully support the principles of Open Science / Reproducible Research. We have raised concerns about the need to improve transparency in the use of large-scale data sets and the analytical discoveries derived from them, especially in validating disease risks and predicted outcomes for both highly refined populations and individuals. Evidence of the profound negative consequences of not doing this well is beginning to appear as publications of false positives in genomic discoveries or chilling anecdotes in the misuse of genetic risk information [11-13, 119-122]. Guidelines for developing robust risk models do exist and should be adapted and incorporated into the analytics platforms that CRI investigators create [123]. Furthermore, the long-term financial sustainability of large-scale data networks and the associated administrative, regulatory, and technical infrastructure costs has yet to be demonstrated. While not entirely under the control of CRI investigators, the CRI community must continue to seek novel value-based approaches to developing tools and infrastructures that have high, recognized value to organizations that would be willing to contribute to the financial stability of these significant investments. Each of these challenges represents a new area for CRI investigators to both lead and contribute novel methodologies and tools to support evolving data governance and regulatory frameworks.

# Conclusion

Four years ago, we published a conceptual model for Clinical Research Informatics that highlighted the importance of data sources, research workflows, and underlying core technologies. Our current update highlights the growth of the diversity and size of data resources and expands the underlying core technologies to include more data-sciences centered activities. As the predictive capabilities of Big Data Analytics becomes more precise, CRI, in partnership with colleagues in biostatistics, research ethics, patient empowerment, and community engagement, will need to include patients and policy makers in difficult conversations about validating and communicating the findings of these predictive models. Also high on our priority list is a significant investment in developing new incentives and methods for promoting data sharing while protecting privacy and confidentiality, including analytic methods to create a true reproducible research / open science culture. With the rise of the “citizen scientist” [124], “quantified self” [125], and engaged patients as research partners and co-investigators, the timing is right for engaging and empowering all these stakeholders and communities in establishing how best to leverage these new opportunities to generate robust medical evidence faster than ever.

## Conflict of Interest Notification

CW declares no conflict of interest. MGK is a member of the external advisory board for TriNetX Corporation who provides data tools for clinical trial design and recruitment.

## Acknowledgements

This study was funded by National Library of Medicine grant R01LM009886 (PI: Weng) and Patient Centered Outcomes Research Institute award ME-1303-5581 (PI: Kahn).

# 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. Zerhouni EA. Translational and clinical science--time for a new vision. N Engl J Med, 2005. 353(15): p. 1621-3.
3. Bandodkar AJ, Jia W, Yardimci C, Wang CX, Ramirez J, Wang J. Tattoo-based noninvasive glucose monitoring: a proof-of-concept study. Anal Chem 2015;87(1):394-8.
4. IOM, in Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease. Washington (DC): IOM; 2011.
5. Kohane IS. Health Care Policy. Ten things we have to do to achieve precision medicine. Science 2015;349(6243):37-8.
6. Kahn MG, Weng C. Clinical research informatics: a conceptual perspective. J Am Med Inform Assoc 2012;19(e1):e36-42.
7. Ma H, Weng C. Identification of Questionable Exclusion Criteria in Mental Disorder Clinical Trials Using a Medical Encyclopedia. Pac Symp Biocomput 2016;21:219-30.
8. He Z, Wang S, Borhanian E, Weng C. Assessing the Collective Population Representativeness of Related Type 2 Diabetes Trials by Combining Public Data from ClinicalTrials.gov and NHANES. Stud Health Technol Inform 2015;216:569-73.
9. He Z, Carini S, Sim I, Weng C. Visual aggregate analysis of eligibility features of clinical trials. J Biomed Inform 2015;54:241-55.
10. Weng C, Li Y, Ryan P, Zhang Y, Liu F, Gao J, et al. A distribution-based method for assessing the differences between clinical trial target populations and patient populations in electronic health records. Appl Clin Inform 2014;5(2):463-79.
11. Dixon S, Shackley P, Bonham J, Ibbotson R. Putting a value on the avoidance of false positive results when screening for inherited metabolic disease in the newborn. J Inherit Metab Dis 2012;35(1):169-76.
12. Macarthur D. Methods: Face up to false positives. Nature 2012;487(7408):427-8.
13. Moyer AM. Handling false positives in the genomic era. Clin Chem 2012;58(11):1605-6.
14. IOM, in Clinical Data as the Basic Staple of Health Learning: Creating and Protecting a Public Good: Workshop Summary. Washington (DC): IOM; 2010.
15. IOM, in Digital Infrastructure for the Learning Health System: The Foundation for Continuous Improvement in Health and Health Care: Workshop Series Summary, C. Grossmann C, Powers B, McGinnis JM, editors. Washington (DC): IOM; 2011.
16. IOM, in Key Capabilities of an Electronic Health Record System: Letter Report. Washington (DC); 2003.
17. Adler-Milstein J, DesRoches CM, Kralovec P, Foster G, Worzala C, Charles D, et al. Electronic Health Record Adoption In US Hospitals: Progress Continues, But Challenges Persist. Health Aff (Millwood) 2015;34(12):2174-80.
18. 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.
19. Etheredge LM. A rapid-learning health system. Health Aff (Millwood) 2007;26(2):w107-18.
20. Rubin JC, Friedman CP. Weaving together a healthcare improvement tapestry. Learning health system brings together health data stakeholders to share knowledge and improve health. J AHIMA 2014;85(5):38-43.
21. IOM, in Observational Studies in a Learning Health System: Workshop Summary. Washington (DC); 2013.
22. IOM, in Large Simple Trials and Knowledge Generation in a Learning Health System: Workshop Summary. Washington (DC); 2013.

23. IOM, in Patients Charting the Course: Citizen Engagement and the Learning Health System: Workshop Summary. Olsen LA, Saunders RS, McGinnis JM, editors. Washington (DC); 2011.
24. Amin W, Tsui FR, Borromeo C, Chuang CH, Espino JU, Ford D, et al, PaTH: towards a learning health system in the Mid-Atlantic region. J Am Med Inform Assoc 2014;21(4):633-6.
25. Forrest CB, Margolis PA, Bailey LC, Marsolo K, Del Beccaro MA, Finkelstein JA, et al. PEDSnet: a National Pediatric Learning Health System. J Am Med Inform Assoc 2014;21(4):602-6.
26. Krumholz HM. Big data and new knowledge in medicine: the thinking, training, and tools needed for a learning health system. Health Aff (Millwood) 2014;33(7):1163-70.
27. McGlynn EA, Lieu TA, Durham ML, Bauck A, Laws R, Go AS, Jet al. Developing a data infrastructure for a learning health system: the PORTAL network. J Am Med Inform Assoc 2014;21(4):596-601.
28. Starren JB, Winter AQ, Lloyd-Jones DM. Enabling a Learning Health System through a Unified Enterprise Data Warehouse: The Experience of the Northwestern University Clinical and Translational Sciences (NUCATS) Institute. Clin Transl Sci 2015;8(4):269-71.
29. Delaney BC, Curcin V, Andreasson A, Arvanitis TN, Bastiaens H, Corrigan D, et al. Translational Medicine and Patient Safety in Europe: TRANSFoRm-Architecture for the Learning Health System in Europe. Biomed Res Int 2015;2015:961526.
30. Friedman C, Rubin J, Brown J, Buntin M, Corn M, Etheredge L, et al. Toward a science of learning systems: a research agenda for the high-functioning Learning Health System. J Am Med Inform Assoc 2015;22(1):43-50.
31. Psek WA, Stametz RA, Bailey-Davis LD, Davis D, J. Darer J, Faucett WA, et al. Operationalizing the learning health care system in an integrated delivery system. EGEMS (Wash DC) 2015;3(1):1122.
32. Weng C. Optimizing Clinical Research Participant Selection with Informatics. Trends Pharmacol Sci 2015;36(11):706-9.
33. Fleurence RL, Curtis LH, Califf RM, Platt R, Selby JV, Brown JS. Launching PCORnet, a national patient-centered clinical research network. J Am Med Inform Assoc 2014;21(4):578-82.
34. Chen, R.T., J.W. Glasser, P.H. Rhodes, R.L. Davis, W.E. Barlow, R.S. Thompson, Mullooly JP, Black SB, Shinefield HR, Vadheim CM, Marcy SM, Ward JI, et al. Vaccine Safety Datalink project: a new tool for improving vaccine safety monitoring in the United States. The Vaccine Safety Datalink Team. Pediatrics 1997;99(6):765-73.
35. Ross, T.R., D. Ng, J.S. Brown, R. Pardee, M.C. Hornbrook, G. Hart, and J.F. Steiner, The HMO Research Network Virtual Data Warehouse: A Public Data Model to Support Collaboration. EGEMS (Wash DC), 2014. 2(1): p. 1049.
36. Platt, R., R.M. Carnahan, J.S. Brown, E. Chrischilles, L.H. Curtis, S. Hennessy, J.C. Nelson, J.A. Racoosin, M. Robb, S. Schneeweiss, S. Toh, and M.G. Weiner, The U.S. Food and Drug Administration's Mini-Sentinel program: status and direction. Pharmacoepidemiol Drug Saf, 2012. 21 Suppl 1: p. 1-8.
37. Sittig DF, Hazlehurst BL, Brown J, Murphy S, Rosenman M, Tarczy-Hornoch P, et al. A survey of informatics platforms that enable distributed comparative effectiveness research using multi-institutional heterogenous clinical data. Med Care 2012 Jul;50 Suppl:S49-59
38. Randhawa GS. Building electronic data infrastructure for comparative effectiveness research: accomplishments, lessons learned and future steps. J Comp Eff Res 2014;3(6):567-72.
39. Gottesman O, Kuivaniemi H, Tromp G, Faucett WA, Li R, Manolio TA, et al. The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. Genet Med 2013;15(10):761-71.
40. De Moor G, Sundgren M, Kalra D, Schmidt A, Dugas M, Claerhout B, et al. Using electronic health records for clinical research: the case of the EHR-4CR project. J Biomed Inform 2015;53:162-73.
41. Beresniak A, Schmidt A, Proeve J, Bolanos E, Patel N, Ammour N, et al. Cost-benefit assessment of using electronic health records data for clinical research versus current practices: Contribution of the Electronic Health Records for Clinical Research (EHR4CR) European Project. Contemp Clin Trials, 2015. 46: p. 85-91.
42. Avillach P, Coloma PM, Gini R, Schuemie M, Mougin F, Dufour JC, et al; E.-A. consortium. Harmonization process for the identification of medical events in eight European healthcare databases: the experience from the EU-ADR project. J Am Med Inform Assoc 2013;20(1):184-92.
43. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. Stud Health Technol Inform 2015;216: 574-8.
44. Nielsen M. Reinventing Discovery: The New Era of Networked Science. Princeton: Univ. Press; 2011.
45. Perrin JM, Batlivala SP, Cheng TL. In the Aftermath of the National Children's Study. JAMA Pediatr 2015;169(6):519-20.
46. Landrigan PJ, Baker DB. The National Children's Study--end or new beginning? N Engl J Med 2015;372(16):1486-7.
47. caBIG Strategic Planning Workspace. The Cancer Biomedical Informatics Grid (caBIG): infrastructure and applications for a worldwide research community. Stud Health Technol Inform 2007;129(Pt 1):330-4.
48. Wilcox A, Randhawa G, Embi P, Cao H, Kuperman GJ. Sustainability considerations for health research and analytic data infrastructures. EGEMS (Wash DC) 2014;2(2):1113.
49. Randhawa GS, Slutsky JR. Building sustainable multi-functional prospective electronic clinical data systems. Med Care 2012;50 Suppl:S3-6.
50. Masys DR, Harris PA, Fearn PA, Kohane IS. Designing a public square for research computing. Sci Transl Med, 2012. 4(149):149fs32.
51. Brailer DJ. From Santa Barbara to Washington: a person's and a nation's journey toward portable health information. Health Aff (Millwood) 2007;26(5):w581-8.
52. Chen B, Butte AJ. Leveraging Big Data to Transform Target Selection and Drug Discovery. Clin Pharmacol Ther 2016 Mar;99(3):285-97.
53. Yu S, Liao KP, Shaw SY, Gainer VS, Churchill SE, Szolovits P, et al. Toward high-throughput phenotyping: unbiased automated feature extraction and selection from knowledge sources. J Am Med Inform Assoc 2015;22(5):993-1000.
54. Hansen MM, Miron-Shatz T, Lau AY, Paton C. Big Data in Science and Healthcare: A Review of Recent Literature and Perspectives. Contribution of the IMIA Social Media Working Group. Yearb Med Inform 2014;9:21-6.
55. Gittelman S, Lange V, Gotway Crawford CA, Okoro CA, Lieb E, Dhingra SS, et al. A new source of data for public health surveillance: Facebook likes. J Med Internet Res 2015;17(4):e98.
56. Kuehn BM. Twitter Streams Fuel Big Data Approaches to Health Forecasting. JAMA 2015;314(19):2010-2.
57. Weber GM, Mandl KD, Kohane IS. Finding the missing link for big biomedical data. JAMA 2014;311(24):2479-80.
58. Margolis R, Derr L, Dunn M, Huerta M, Larkin J, Sheehan J, et al. The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data. J Am Med Inform Assoc 2014;21(6):957-8.
59. Bourne PE, Bonazzi V, Dunn M, Green ED, Guyer M, Komatsoulis G, et al. The NIH Big Data to Knowledge (BD2K) initiative. J Am Med Inform Assoc 2015;22(6):1114.
60. Lorgelly PK, Doble B, Knott RJ, Cancer I. Realising the Value of Linked Data to Health Economic Analyses of Cancer Care: A Case Study of Cancer 2015. Pharmacoeconomics 2016 Feb;34(2):139-54.
61. Kaushal R, G. Hripcsak G, Ascheim DD, Bloom T, Campion TR, Jr., Caplan AL, et al. Changing the research landscape: the New York City Clinical Data Research Network. J Am Med Inform Assoc 2014;21(4):587-90.
62. Deng B, Fradkin M, Rouet JM, Moore RH, Kopans DB, Boas DA, M. Lundqvist, and Q. Fanget al. Characterizing breast lesions through robust multimodal data fusion using independent diffuse optical and x-ray breast imaging. J Biomed Opt 2015;20(8):80502.
63. Blanchet L, Smolinska A. Data Fusion in Metabolomics and Proteomics for Biomarker Discovery. Methods Mol Biol 2016;1362:209-23.
64. Wu Q, Li JV, Seyfried F, le Roux CW, Ashrafian H, Athanasiou T, et al. Metabolic phenotype-microRNA data fusion analysis of the systemic consequences of Roux-en-Y gastric bypass surgery. Int J Obes (Lond) 2015;39(7):1126-34.
65. Heath AP, Greenway M, Powell R, Spring J, Suarez R, Hanley D, et al. Bionimbus: a cloud for managing, analyzing and sharing large genomics datasets. J Am Med Inform Assoc 2014;21(6):969-75.
66. Hsieh PJ. Healthcare professionals' use of health clouds: Integrating technology acceptance and status quo bias perspectives. Int J Med Inform 2015;84(7):512-23.
67. Ohmann C, Canham S, Danielyan E, Robertshaw S, Legre Y, Clivio L, et al. 'Cloud computing' and clinical trials: report from an ECRIN workshop. Trials 2015;16:318.
68. Ohm P. Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization. UCLA Law Review 2010;57:1701–77.
69. Naveed M, Aydayn E, Clayton EW, Fellay J, Gunter CA, Hubaux JP, et al. Privacy in the Genomic Era. ACM Comput Surv 2015;48(1).
70. Li Y, Jiang X, Wang S, Xiong H, Ohno-Machado L. VERTIcal Grid lOgistic regression (VERTIGO).

J Am Med Inform Assoc 2016;23(3):570-9.
71. Lu CL, Wang S, Ji Z, Wu Y, Xiong L, Jiang X, et al. WebDISCO: a web service for distributed cox model learning without patient-level data sharing. J Am Med Inform Assoc 2015;22(6):1212-9.
72. Wu Y, Jiang X, Wang S, Jiang W, Li P, Ohno-Machado L. Grid multi-category response logistic models. BMC Med Inform Decis Mak 2015;15:10.
73. Wu Y, Jiang X, Kim J, Ohno-Machado L. Grid Binary LOgistic REgression (GLORE): building shared models without sharing data. J Am Med Inform Assoc 2012;19(5):758-64.
74. Wang S, Jiang X, Wu Y, Cui L, Cheng S, Ohno-Machado L. EXpectation Propagation LOgistic REgRession (EXPLORER): distributed privacy-preserving online model learning. J Biomed Inform 2013;46(3):480-96.
75. Bodenreider O. Biomedical ontologies in action: role in knowledge management, data integration and decision support. Yearb Med Inform 2008;67-79.
76. Musen MA, Bean CA, Cheung KH, Dumontier M, Durante KA, Gevaert O, et al. Rocca-Serra, S.A. Sansone, J.A. Wiser; CEDAR team. The center for expanded data annotation and retrieval. J Am Med Inform Assoc 2015;22(6):1148-52.
77. Oellrich A, Collier N, Groza T, Rebholz-Schuhmann D, Shah N, Bodenreider O, et al. The digital revolution in phenotyping. Brief Bioinform 2015.
78. Anderson WP. Reproducibility: Stamp out shabby research conduct. Nature 2015;519(7542):158.
79. Ohno-Machado L. A journal's role in resource sharing and reproducibility. J Am Med Inform Assoc 2015;22(3):491.
80. Safran C. Reuse of clinical data. Yearb Med Inform 2014;9:52-4.
81. 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.
82. Mittelstadt BD, Floridi L. The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts. Sci Eng Ethics 2016 Apr;22(2):303-41.
83. Hruby GW, Boland MR, Cimino JJ, Gao J, Wilcox AB, Hirschberg J, et al. Characterization of the biomedical query mediation process. AMIA Jt Summits Transl Sci Proc 2013;2013:89-93.
84. Hruby GW, Cimino JJ, Patel V, Weng C. Toward a cognitive task analysis for biomedical query mediation. AMIA Jt Summits Transl Sci Proc 2014;2014:218-22.
85. Hruby GW, Ancker J, Weng C. Use of Self-Service Query Tools Varies by Experience and Research Knowledge. Stud Health Technol Inform 2015;216:1023.
86. Hruby GW, Matsoukas K, Cimino JJ, Weng C. Facilitating biomedical researchers' interrogation of electronic health record data: Ideas from outside of biomedical informatics. J Biomed Inform 2016;60:376-84.
87. Hoxha J, Chandar P, He Z, Cimino J, Hanauer D, Weng C. DREAM: Classification scheme for dialog acts in clinical research query mediation. J Biomed Inform 2016;59:89-101.
88. Richesson RL, Chute CG. Health information technology data standards get down to business: maturation within domains and the emergence of interoperability. J Am Med Inform Assoc 2015;22(3):492-4.
89. Emanuel EJ. Reform of Clinical Research Regulations, Finally. N Engl J Med 2015;373(24):2296-9.
90. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res 2014;42(Database issue):D980-5.
91. Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, et al. ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res 2016;44(D1):D862-8.
92. NIH. NIH Strategic Plan for 2016-2020; 2015. Available from: http:///www.nih.gov/sites/default/files/about-nih/strategic-plan-fy2016-2020-508.pdf.
93. Hohnloser JH, Fischer MR, Konig A, Emmerich B. Data quality in computerized patient records. Analysis of a haematology biopsy report database. Int J Clin Monit Comput 1994;11(4):233-40.
94. Hogan WR, Wagner MM. Accuracy of data in computer-based patient records. J Am Med Inform Assoc 1997;4(5):342-55.
95. Aronsky D, Haug PJ. Assessing the quality of clinical data in a computer-based record for calculating the pneumonia severity index. J Am Med Inform Assoc 2000;7(1):55-65.
96. Thiru K, Hassey A, Sullivan F. Systematic review of scope and quality of electronic patient record data in primary care. BMJ 2003;326(7398):1070.
97. de Lusignan S, Valentin T, Chan T, Hague N, Wood O, van Vlymen J, et al. Problems with primary care data quality: osteoporosis as an exemplar. Inform Prim Care 2004;12(3):147-56.
98. Berner ES, Kasiraman RK, Yu F, Ray MN, Houston TK. Data quality in the outpatient setting: impact on clinical decision support systems. AMIA Annu Symp Proc 2005:41-5.
99. Botsis T, Hartvigsen G, Chen F, Weng C. Secondary Use of EHR: Data Quality Issues and Informatics Opportunities. AMIA Jt Summits Transl Sci Proc 2010;2010:1-5.
100. Chan KS, Fowles JB, Weiner JP. Review: electronic health records and the reliability and validity of quality measures: a review of the literature. Med Care Res Rev 2010;67(5):503-27.
101. Nahm M. Data quality in clinical research. Clinical Research Informatics. London: Springer-Verlag; 2012. p. 175–201.
102. 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.
103. Nahm ML, Pieper CF, Cunningham MM. Quantifying data quality for clinical trials using electronic data capture. PLoS One 2008;3(8):e3049.
104. Weiner MG, Embi PJ. Toward reuse of clinical data for research and quality improvement: the end of the beginning? Ann Intern Med 2009;151(5):359-60.
105. Van den Broeck J, Cunningham SA, Eeckels R, Herbst K. Data cleaning: detecting, diagnosing, and editing data abnormalities. PLoS Med 2005;2(10):e267.
106. Hayes P. The ethics of cleaning data. Clin Nurs Res 2004;13(2):95-7.
107. Pipino LL, Lee YW, Wang RY. Data quality assessment. Commun. ACM 2002;45(4):211-8.
108. Wang RY, Strong DM. Beyond accuracy: what data quality means to data consumers. J Manage Inf Syst 1996;12(4):5-33.
109. Kahn MG, Brown JS, Chun AT, Davidson BN, Meeker D, Ryan PB, et al. Transparent reporting of data quality in distributed data networks. EGEMS (Wash DC) 2015;3(1):1052.
110. Embi PJ, Payne PR. Advancing methodologies in Clinical Research Informatics (CRI): foundational work for a maturing field. J Biomed Inform 2014;52:1-3.
111. Curtis LH, Brown J, Platt R. Four health data networks illustrate the potential for a shared national multipurpose big-data network. Health Aff (Millwood) 2014;33(7): 1178-86.
112. Brown JS, Rusincovitch SA, Kho AN, Marsolo K, Curtis L. Development of a national distributed research network data infrastructure: Design of the PCORnet common data model. In: Proceedings of the 2015 American Medical Informatics Association. San Francisco, CA; 2015. p. 302.
113. Delude CM. Deep phenotyping: The details of disease. Nature 2015;527(7576): S14-5.
114. Frey LJ, Lenert L, Lopez-Campos G. EHR Big Data Deep Phenotyping. Contribution of the IMIA Genomic Medicine Working Group. Yearb Med Inform 2014;9:206-11.
115. Robinson PN. Deep phenotyping for precision medicine. Hum Mutat 2012;33(5):777-80.
116. Stepniak B, Papiol S, Hammer C, Ramin A, Everts S, Hennig L, et al. Accumulated environmental risk determining age at schizophrenia onset: a deep phenotyping-based study. Lancet Psychiatry 2014;1(6):444-53.
117. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. J Am Med Inform Assoc 2013;20(1):117-21.
118. Collins FS, Hudson KL, Briggs JP, Lauer MS. PCORnet: turning a dream into reality. J Am Med Inform Assoc 2014;21(4):576-7.
119. Error prone. Nature 2012;487(7408):406.
120. Kingsmore SF. Incidental swimming with millstones. Sci Transl Med 2013;5(194):194ed10.
121. Tse H. Publishing: Curb temptation to skip quality control. Nature 2012;488(7413):591.
122. Clinical Genetics Has a Big Problem That's Affecting People's Lives. [December 29, 2015] Available from: http://www.theatlantic.com/science/archive/2015/12/why-human-genetics-research-is-full-of-costly-mistakes/420693/.
123. Pepe MS, Etzioni R, Feng Z, Potter JD, Thompson ML, Thornquist M, et al. Phases of biomarker development for early detection of cancer. J Natl Cancer Inst 2001;93(14):1054-61.
124. Bonney R, Phillips TB, Ballard HL, Enck JW. Can citizen science enhance public understanding of science? Public Underst Sci 2016;25(1):2-16.
125. Bottles K. Will the quantified self movement take off in health care? Physician Exec 2012;38(5):74-5.

**Correspondence to:**
Chunhua Weng, PhD, FACMI
Department of Biomedical Informatics
Columbia University
622 W 168 Street, PH-20
New York, NY 10032, USA
E-mail: chunhua@columbia.edu