Opportunities and Challenges related to the use of Electronic Health Records data for research

Rex L. Chisholm, Ph.D., Northwestern University (Co-chair); Joshua Denny, M.D., M.S., Vanderbilt University; Doug Fridsma, M.D., Ph.D., American Medical Informatics Association; Sachin Kheterpal M.D., University of Michigan Medical School; Daniel Masys, M.D., University of Washington (Co-chair); Lucila Ohno-Machado, M.D., Ph.D, University of California, San Diego.

Publication note: This document is an invited "white paper" created as background information for participants at a workshop held on February 11-12, 2015. It represents the opinions of the authors, and is intended to help inform the planning process for the Precision Medicine Initiative research cohort.

The correlation of genotype and phenotype across very large numbers of persons offers the promise of improved understanding of the molecular basis of human health and disease, and an evidence base that can help guide health-related decisions made by care providers and by patients. Electronic Health Records (EHRs) are an essential data resource for such correlations, with two characteristics that are rarely found in traditional cohort research. First, they contain longitudinal data representing significant health events that may extend over decades, and increasingly will represent an individual's entire lifetime as well as that of family members. In addition, to the extent that each individual is a unique "experiment of Nature", EHRs document real world clinical manifestations of both common and rare personal molecular variation. These potentially powerful scientific opportunities are balanced by challenges related to technology, policy, organizational motivations, and concerns over data privacy and security. This white paper outlines some of the most important issues related to research use of EHR-derived clinical data in a national cohort of a million or more, as viewed by the members of the EHR working group.

The term "Electronic Medical Records" has commonly been used as a synonym for computerized clinical systems. It represents the historical evolution of care delivery records by health professionals from paper-based handwritten formats to electronic formats. As such, it carries an unfortunate connotation of systems that are primarily designed to be passive *after-the-fact* archives of the observations and actions initiated by clinicians on behalf of individual patients in the course of providing medical care. A more general and contemporary term is "Electronic Health Records" (EHRs), which expands the scope of systems to include data relevant to both health and disease, and expands the sources of data beyond the records created by health professionals, to include the patient's own observations, and family history data generated by relatives. The data can consist of standardized codes, text, or other electronic forms such as video and audio, and may include electronic communications among these multiple 'stakeholders'.

The effective use of EHR-derived health data from a national cohort, whose participants can be expected to receive healthcare services from hundreds or thousands of organizations that maintain an EHR system, will depend critically upon the ability to address the following high priority issues:

1. Human and business factors: Individual and institutional motivation to participate, and integrated consent management systems

The first and perhaps most difficult challenge to be addressed in a Precision Medicine (PM)[1] research program is creating and articulating a value proposition for project participation that is compelling to all data providers, including patients and healthcare organizations. The issues of technology are moot unless organizations and individual participants are motivated to join the effort and give permission for use of the data that are under their control.[2] The EHR working group considered three complementary pathways by which clinical data could become available.

The first is the union of existing research cohorts in which both clinical data and biosamples (or the data derived from biosamples) already exist and for which observational studies of the correlation of genotype and phenotype are currently underway. Existing NIH research consortia provide examples of both scientific motivation and cost efficiency.[3] Building the national cohort in this way would supplement current scientific agendas and research capabilities without replacing or diminishing work already underway.

The second potential approach is to acquire clinical data *ab initio* via new organizational partnerships. These range in scale, from community-based hospitals and clinics to established individual academic health centers (particularly those that have strengthened recruitment, research data management, and biospecimen capabilities via the CTSA program), to health systems such as Kaiser Permanente and Geisinger, and consortia of heterogeneous health systems such as the HMO Research Network and PCORnet, the PCORI-supported consortium of Clinical Data Research Networks (CDRNs) and Patient-Powered Research Networks (PPRNs).[4] Dispensing data from pharmacies and pharmacy benefits management companies, and claims data from various organizations could potentially add substantial value to EHR data. Crafting a value proposition and business case based on *quid pro quo* models of returning PM research data (particularly genetic data[a]) that help advance other interests of those organizations (e.g., quality improvement, new capabilities to enable them as a learning healthcare organization) should be a component of developing a value proposition and increasing the motivation to participate.

The third and most novel pathway for access to clinical data is via the rights granted to each individual by HIPAA[5]/HITECH legislation to obtain electronic copies of their EHR data. This access and download capability has been termed the "Blue Button" by the Office of the National Coordinator, and it is the focus of a technology and data standards issues described later in Section 3 of this paper. Potentially, each study participant could become the active agent and channel to allow access to their own EHR data by a PM coordinating center. This scenario also begins with articulation of a personal value proposition to every potential study participant.

The EHR in this model can serve as a foundation that provides the opportunity for participants to enrich their data with patient-reported observations, including quality of life information and other related data that could be entered in the system through mHealth apps and patient portals. Robust engagement of patients in such a PM cohort also aligns well with the on-going

[a] It should be noted that some careful thought needs to be given to the potential consequences of providing genetic data to some of these organizations. The Genetic Information Non-discrimination Act should be helpful in minimizing some of these concerns.

discussions about creating learning healthcare systems and sharing health management and research decision-making with the patients. Thus, an important motivation for participation is the opportunity for these data to be used not only for discovery and implementation of PM, but also to improve the quality of data in 'upstream' EHR systems and consequently improve the information available to patients and providers. Sustainability of PM efforts could result in part by leveraging this type of return on investment.

The three pathways described are complementary and do not need to rely on the assumption that every patient consents to every use of his or her data. For example, in the first approach, merging of existing cohorts may require that patients be contacted to update the terms of their consent. In the second case, institutions may be located in different states, and therefore under different regulatory constraints in establishing partnerships.[6] In the third case, patient decisions regarding which entities will have access to the data do not need to be done on a one-by-one basis, and should be modifiable at any time. A consent management system will thus be an important component of the systems infrastructure for the project.

2. Technical issues in integrating and analyzing data from heterogeneous systems

Once motivation to participate is obtained and permission to access clinical data is in place, the technical issues relevant to research use of EHR data become the focus of attention. The experience of NIH-supported consortia such as eMERGE (electronic MEdical Records and Genomics)[7] and the nascent experience of PCORnet is both relevant and encouraging, since they have demonstrated the ability to normalize and analyze data on "virtual cohorts" of individuals defined by the EHR. The creation of "research grade" selection logic for specific disease phenotypes has been based on both structured data, such as ICD and CPT billing codes, clinical laboratory values, and medication codes, and importantly also on use of natural language processing (NLP) techniques applied to unstructured narrative text contained in discharge summaries, operative notes and procedure summaries, radiology and pathology reports, history and physicals, progress notes, etc.[8] An important finding of the eMERGE consortium, which began with five institutions with five different EHR systems, was that selection logic based on codes, labs, meds, and NLP was transportable across those systems with preservation of its positive predictive value, and relatively little need to adapt it to work in EHR systems different from the one(s) used to develop the "computable phenotype".[9–11] However, identification of potential subjects is only the beginning: in order to follow the health of these subjects over time it is important to keep track of their encounters with the healthcare system, regardless of where they were treated. Completeness of heath information will need to be addressed by placing the patient at the center of the information hub with initiatives such as the one that utilizes information from the 'blue button', described later in Section 3 of this paper.

Another challenge relates to the lack of a practical mechanism to uniquely identify participants. The integrity and value of EHR data depends fundamentally on being able to unambiguously link it to one and only one individual whose healthcare it represents. This is sometimes achieved by matching records based on several characteristics (probabilistic matching) instead of using a unique identifier, which circumvents but does not eliminate privacy concerns. Unintentional duplicate records, name ambiguities, and utilization of another person's identity to obtain care all contribute complexity to the seemingly straightforward task of linking data to the correct

individual. This problem grows in proportion to the number of individuals who participate and the number of different care organizations that generate data about those individuals. On the flip side, unique identification also can confer a higher risk of privacy breach. Several techniques to decrease the risk of re-identification have been proposed,[12–16] but they all come with a cost in terms of the usefulness of "de-identified" data. While it is easy to show that HIPAA-compliant de-identification does not prevent re-identification, it is not easy to develop a solution that preserves data utility.

Finally, existing consortia that extract, transform, "clean" and analyze clinical data for research purposes have extensively depended upon the local expertise of informatics teams at each of the data contributing sites.[17] Such expertise is currently available in a relatively small number of healthcare institutions (typically in large academic medical centers) and the paucity of experts constitutes a bottleneck in the development of a national PM cohort that includes all care settings.

3. Patients in control: Enhancing "Blue Button" functionality for research

As part of its mandate to promote adoption of EHRs through HITECH and the meaningful use program[18] the federal Office of the National Coordinator (ONC) has supported data standards for certification of EHR systems and has extended HIPAA to include electronic access to healthcare information. This has included a campaign for patient to access their own clinical and administrative data via the "Blue Button."[19] The Blue Button campaign has worked to raise visibility among patients that they have a right to access their information electronically, and encouraged data producers to release their data in electronic formats. The Veterans Administration was an early implementer of the Blue Button campaign and had over 3.6 million veterans access their information electronically as of April 2013.[20] Over time, this access has included not only free text electronic documents, but also the same standardized summary format that is used by providers in exchanging care summaries. Additional formats have been added over time, and include not only clinical care summaries, but also administrative data that includes explanations for patients of insurance benefits that have been paid.

Once a patient has downloaded this information, they are free to do with it as they wish – upload it to a personal healthcare record, share it with their provider, or provide it to researchers or other third parties. For purposes of this workshop the EHR working group visualized this as a "Synch for Science" (S4S) button that updates databases as new data becomes available in the EHRs maintained by the various care organization an individual chooses to use. Access could be extended to third parties if the patient wanted, including the possibility of granting research organizations the permission to act on behalf of the patient as needed, serving as a broker for their health data. Furthermore, this could be accomplished by an access control system that could obviate the need to transmit data, but allowed their secure consultation via the network. A distributed model for analyzing data "in situ" has already been selected for use by several CTSA awardee institutions in the Accrual for Clinical Trials project and PCORnet clinical data research networks, and hence the main challenge would be to integrate EHR Blue-Button functionality into such models.

Currently, the technical specifications that support Blue Button functionality are not complete,

and have only a few types of documents: A clinical care summary format (based on the consolidated Clinical Document Architecture[21] standard), an explanation of benefits format, and three mechanisms to transit and share the information (secure download/upload, secure email, and web-based publication/subscription models). While these document formats have sections that have higher degrees of structure and require the use of standardized terminologies, (e.g., problem lists, labs, meds, and allergies) in addition to free text, they are not yet sufficiently standardized or detailed for research-quality cohort identification and case selection.

Therefore, to make progress toward a reliable research data resource based on individuals accessing their records as study participants, we would need to 1) define a common structure for granular research-related data,(common data element standard[22]) 2) have agreed-upon definitions (semantics) for key research data, and 3) develop an extensible "comprehensive electronic data format" that will accommodate both structured and unstructured data. In the clinical informatics community and at NIH, relevant standards development work is currently underway to develop common data elements and their definitions.

In addition, there is a need for development of a simple "complete medical record" electronic format for extraction of the entire collection of structured and unstructured information in an individual's electronic health record. While such a data standard might not have full functionality at the outset, it will set the stage for a S4S data standard to provide automatic updates of EHR information to linked research systems. Ultimately, this approach could become the basis of a second-generation "Secure Health Applications Programming Interface (API)" that would allow for more data-driven (rather than document-based) access, by allowing patients to authorize secure programmatic access and to update the EHR data that they wish to donate to PM research. One attractive quality control aspect of this process model is that patients could view the data of interest and verify that it indeed belongs to them prior to sending it to, or enabling access by, a research data center. This would help overcome the challenge of being sure that the clinical data are associated with the correct individual.

4. Industry engagement

Clinical data availability for research from both healthcare organizations and individuals will depend critically upon engagement and support of the wide range of healthcare and consumer industry vendors involved in EHRs, health information exchanges, reimbursement, home healthcare, and personal computing. However, clinical system functionality directly supporting PM has historically been limited. Moreover, even within the research community, research of the type envisioned for the PM cohort has not been a prominent "use case" until recently. Industry engagement must be based upon four tenets: a) practical, specific, and certifiable functionality; b) compatibility with existing government-supported requirements; c) recognition that keeping clinical systems operational will have priority over formal certifications and d) technology agnostic engagement.

a) Practical, specific, and certifiable functionality

   As the ONC Meaningful Use program has demonstrated, the EHR industry is capable of delivering specific functionality when the target is well established and testable. In order to

increase adoption of any standards related to the PM data collection process, EHR vendors must know the "target" of their efforts. Transforming the realities of real-world EHR data into interoperable, standardized data structures requires significant vendor-to-vendor and site-to-site interpretation. EHR vendors and healthcare providers must be given guidance and testing capabilities regarding how to configure systems for these new requirements.

b) Leverage existing government-supported EHR requirements

Healthcare industry vendors are already overwhelmed with a range of software and data formats from various agencies and organizations involved with informatics (e.g., i2b2, PopMedNet, CDA), regulation (e.g., HIPAA, Meaningful Use), and reimbursement (Physician quality reporting systems, Value based purchasing). Identifying which existing standards and requirements most closely meet the needs of the PM Initiative will be essential. Next, the PM project leadership will need to perform ongoing assessments of adoption by not only vendors, but also providers. The financial pressure induced by CMS standards for value based purchasing and physician quality reporting systems are potent drivers of adoption and execution. Integrating research and reimbursement driven standards will balance the realities of competing industry requests with optimal functionality needed for PM.

c) Recognizing operational imperatives

Although the academic research and vendor communities share a foundation of innovation defining success, industry requires practical, achievable, fixed endpoints that can be delivered and measured in quarters, not years. Shareholders and investors demand return on investment, and the return is not always measured in dollars, but also in lives affected or patients enrolled. As a result, the PM Initiative must publicly recognize and reward not only vendor participation and certification, but also objective measures of operational performance. This will help create a culture of transparency, accountability, and competition that fosters adoption.

d) Technology agnostic engagement

Interoperability should be emphasized: Favoring a particular EHR technology for PM implementation may jeopardize widespread industry engagement. Any specific technical platform requirements must be rooted in existing standards that are already experiencing high adoption, since new technical requirements will be met by resistance due to the wide variety of practice settings, customer resources, and vendor installed base.

5. Cybersecurity

A large national cohort study project will depend critically upon the use of the Internet and cell phone networks, and will involve transmission, storage and analysis of sensitive personal data, some of which will have been originally created as Protected Health Information (PHI) within HIPAA covered entities. It will inherit the security vulnerabilities of a variety of computing devices, including data servers, workstations, and smartphones. Thus data and communications security will be an important issue, and the effort will need to employ effective means for both

administrative and technical safeguarding of confidential data. Cyberthreats are real, pervasive, and will continue to evolve over the duration of any longitudinal study, so data and communications security will be an essential ongoing component of such an activity. The EHR workgroup does not believe that any new or unique cybersecurity measures will need to be developed specifically for the project, but adoption of industry-wide best practices and technologies, and ongoing vigilance will be essential to establishing and maintaining the trust of study participants and that of collaborating organizations.

Summary

The scientific utility of EHR-derived "routine" clinical data generated as a byproduct of care delivery has been previously demonstrated, and there are no insurmountable technical challenges or barriers to achieving the effective use of this type of data for a national cohort of millions of individuals. Work will need to be done in the areas of policy and data standards, and extensive software engineering will be needed to develop the specialized applications that will support the types of data sources described here. The critically important first step will be to understand the potential motivations and goals of organizations and individuals related to participating in the project, and design systems and flows of information and data that operationally support those goals and motivations, while advancing the science of genome-phenome correlation and contributing to a national learning healthcare system.

References

1. Collins, F. S. & Varmus, H. A New Initiative on Precision Medicine. *N. Engl. J. Med.* **0,** null (2015).
2. Kim, K. K. *et al.* Data governance requirements for distributed clinical research networks: triangulating perspectives of diverse stakeholders. *J. Am. Med. Inform. Assoc. JAMIA* **21,** 714–719 (2014).
3. Bowton, E. *et al.* Biobanks and electronic medical records: enabling cost-effective research. *Sci. Transl. Med.* **6,** 234cm3 (2014).
4. Collins, F. S., Hudson, K. L., Briggs, J. P. & Lauer, M. S. PCORnet: turning a dream into reality. *J. Am. Med. Inform. Assoc. JAMIA* **21,** 576–577 (2014).
5. Health Information Privacy. at <http://www.hhs.gov/ocr/privacy/>

6. Kim, K. K., McGraw, D., Mamo, L. & Ohno-Machado, L. Development of a privacy and security policy framework for a multistate comparative effectiveness research network. *Med. Care* **51,** S66–72 (2013).
7. Gottesman, O. *et al.* The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. *Genet. Med. Off. J. Am. Coll. Med. Genet.* (2013). doi:10.1038/gim.2013.72
8. Kho, A. N. *et al.* Electronic Medical Records for Genetic Research: Results of the eMERGE Consortium. *Sci. Transl. Med.* **3,** 79re1 (2011).
9. Denny, J. C. *et al.* Variants Near FOXE1 Are Associated with Hypothyroidism and Other Thyroid Conditions: Using Electronic Medical Records for Genome- and Phenome-wide Studies. *Am. J. Hum. Genet.* **89,** 529–542 (2011).
10. Peissig, P. L. *et al.* Importance of multi-modal approaches to effectively identify cataract cases from electronic health records. *J. Am. Med. Inform. Assoc. JAMIA* **19,** 225–234 (2012).
11. Kho, A. N. *et al.* Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. *J. Am. Med. Inform. Assoc. JAMIA* **19,** 212–218 (2012).
12. Loukides, G., Denny, J. C. & Malin, B. The disclosure of diagnosis codes can breach research participants' privacy. *J. Am. Med. Inform. Assoc. JAMIA* **17,** 322–327 (2010).
13. Kushida, C. A. *et al.* Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies. *Med. Care* **50 Suppl,** S82–101 (2012).

14. Heatherly, R., Denny, J. C., Haines, J. L., Roden, D. M. & Malin, B. A. Size matters: How population size influences genotype-phenotype association studies in anonymized data. *J. Biomed. Inform.* (2014). doi:10.1016/j.jbi.2014.07.005

15. Jiang, X., Sarwate, A. D. & Ohno-Machado, L. Privacy technology to support data sharing for comparative effectiveness research: a systematic review. *Med. Care* **51,** S58–65 (2013).

16. Zhao, Y., Wang, X., Jiang, X., Ohno-Machado, L. & Tang, H. Choosing blindly but wisely: differentially private solicitation of DNA datasets for disease marker discovery. *J. Am. Med. Inform. Assoc. JAMIA* **22,** 100–108 (2015).

17. Newton, K. M. *et al.* Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. *J. Am. Med. Inform. Assoc.* **20,** e147–e154 (2013).

18. 42 CFR 495 - STANDARDS FOR THE ELECTRONIC HEALTH RECORD TECHNOLOGY INCENTIVE PROGRAM. at <http://www.gpo.gov/fdsys/granule/CFR-2011-title42-vol5/CFR-2011-title42-vol5-part495>

19. Turvey, C. *et al.* Blue Button use by patients to access and share health record information using the Department of Veterans Affairs' online patient portal. *J. Am. Med. Inform. Assoc. JAMIA* **21,** 657–663 (2014).

20. The My HealtheVet Personal Health Record Portal in 2013: New Features, Study Findings, and Opportunities. at <http://www.hsrd.research.va.gov/for_researchers/cyber_seminars/archives/632-notes.pdf>

21. Dolin, R. H. *et al.* HL7 Clinical Document Architecture, Release 2. *J Am Med Inf. Assoc* **13,** 30–9 (2006).

22. Common Data Model | Observational Medical Outcomes Partnership. at <http://omop.org/CDM>