Updated 5/19/2015

DRAFT

Data Management Infrastructure For a National Precision Medicine Cohort

The vision of precision medicine is of individually-tailored approaches to health maintenance, disease prevention and risk assessment, disease diagnosis, treatment and follow-up. Tasks such as medical diagnosis have always been grounded in the attempt to classify an individual as belonging to a group about which there are general principles known (e.g., 'type II diabetics') for purposes of prediction of future health states and informing actions to cure or alleviate the condition. The era of precision medicine refines the task of classification so that similarities and differences among individuals are based on dramatically larger sets of observations and personal characteristics, including molecular variation contained in an individual's genome, proteome and other biologically important molecules, combined with factors that include environmental exposures, lifestyle choices, personal preferences, and social and cultural factors.

The path to precision medicine begins with building an evidence base large enough to support pattern detection and correlations among many different types of data, potentially with varying degrees of completeness and quality, from large numbers of individuals. In cases where the condition of interest is very uncommon, or requires combinations of features (rather than single features, such as a Single Nucleotide Polymorphism) contribute to a health state of interest, the ability to find enough sufficiently similar individuals to support drawing statistically valid conclusions may require drawing from pools of millions of persons. Thus, the Precision Medicine Initiative has as a key component a large 'national cohort' of willing and engaged individuals whose data is available for analysis.

This document describes the major categories of data that are likely to be generated by a large scale PMI cohort, as well as technical and organizational approaches to managing that data and making it available for analysis. The approaches described here build upon decades of experience with NIH-funded multi-center research and also on recent innovations in information technology that make possible novel, as yet untested alternatives. Specific implementation decisions will necessarily await the final plans for how the national cohort will be constructed, however the data management infrastructure alternatives are described here in general terms, and adaptable to essentially any large scale longitudinal cohort that generates the kinds of data enumerated in Table 1:

Table 1: Anticipated Categories, Sources and Uses of Data			
Category	Source(s)	Anticipated Use	Examples
Individual demographics and contact information	Study participant, research and healthcare organizations	Participant-specific communications, analytics	Study appointment reminders, invitations to participate in substudies, risk stratification, assessment of covariates and confounds
Terms of consent and personal preferences for participation	Study participant	"Precision Participant Engagement"	Fine-grained consent for options to participate e.g., receive research results
Self-reported measures	Study participant	Many	Pain scales, quality of life measurements. environment and lifestyle
Sensor-based observations	Physiologic monitors	Functional impairment assessment	Continuous cardiac rhythm monitoring, respiratory rate, blood glucose, activity
Clinical data derived from Electronic Health Records (EHRs)	Multiple provider organizations per study participant, via institutionally-managed channels or direct from participant via personal download/upload	Correlation of clinical events with other categories of data	ICD/CPT billing codes, clinical lab values, medications, problem lists, narrative documents
Research specific observations	Study participants, research organizations	Many	Research questionnaires, whole exome/genome sequences done for research purposes
Biospecimens	Study participants, biobanks	Correlation of tissue findings with other categories of data	Blood and other body fluids, organ- and disease-specific tissue samples
Geospatial and environmental data	Public and private sources	Epidemiology, epidemic surveillance	weather, air quality, environmental pollutant levels
Other novel forms of data	Public and private sources	Predictive analytics	Social Networking e.g., Twitter feeds, OTC medication purchases

Of the categories of data listed, those derived from EHRs are associated with several key challenges but also tremendous opportunity. The rapid increase in adoption of electronic health records mean that for the first time we have a foundation of digital, coded data available for the majority of Americans. But these data are dispersed across multiple provider locations and vendor systems, with no easy, standardized way to bring them together. The holders of the data—often large delivery systems—have legitimate concerns about privacy and security but also sometimes use these challenges to mask other competitive interests. 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 these issues.

Organizational and Process Models

Organizational and process models for acquiring and managing the research data associated with a national cohort can benefit from decades of prior NIH experience with cooperative groups and research consortia, as modulated by new types of computing and communications technologies. Most notable among these is the rapid uptake of smartphones, tablets and other network-connected personal electronics and their associated 'apps' marketplace. From a research data management perspective, the majority of the US population now has a self-funded general purpose, network-connected computing device in their possession.

The 'traditional' model for an NIH-funded multicenter research consortium involves an operations and data coordination center that serves as the hub in a hub-and-spoke model of collaborating research organizations (Fig. 1). The coordination center and each of the participating organizations maintains a research data management infrastructure and dedicated personnel with expertise in acquiring, storing, and transmitting data of various types. In settings such as EHR-derived datasets, where the primary data are created in heterogeneous formats with nonstandard naming and coding conventions, additional steps of 'data normalization' are needed, and these transformations can be applied at the local institution, the central data center, or both, depending upon the consortium design and capabilities of each participating organization. It is important to understand that in this model the data source directly attached for scientific purposes is essentially never the EHR itself, but rather a 'data warehouse' and other analytic systems that contains copies of the data extracted from the operational EHR and reformatted into data structures that favor cross-tabulations and other forms of analysis. This contrasts with the organization of those same data in the EHR, which tend to be in a format optimized for 'patient-at-a-time' lookup and display. Data warehouses are also not, in general, back-compatible with clinical systems so there is seldom if ever an existing pathway to return data or analyses made in the research environment back into clinical environments directly.

Research use of EHR data inherits a policy context established by HIPAA/HITECH for the clinical use of that data, which is the principle of 'minimum necessary'. The financial penalties and public embarrassment associated with breaches of confidentiality align powerfully with healthcare organization motives to maintain control over clinical data and not share it. Thus, consortia using the hub and spoke model depicted in Fig. 1 are strongly disincentivized to send

bulk copies of the data in their data warehouses to any outside organization. Instead, 'federated query' models have become prevalent, where each participating organization receives and separately processes a study specific query sent by the coordination center (e.g., "send these 20 variables in this standard format, on all individuals who meet this set of selection criteria"). No data other than the minimum necessary to satisfy the query leaves the originating institution. This set of institutional motivations to be minimal with data releases commonly trumps individuals' consent for broad sharing of data, due to a variety of factors including the perception that liability in cases of incorrect subsequent disclosure and/or use rests with the organization, not the individual consenting for broad data release. De-identification of data to specific standards such as HIPAA "safe harbor" and "limited dataset" requirements can facilitate some types of EHR-derived data sharing, however as the richness of detail within de-identified data grows, guarantees regarding identifiability become progressively harder to establish.

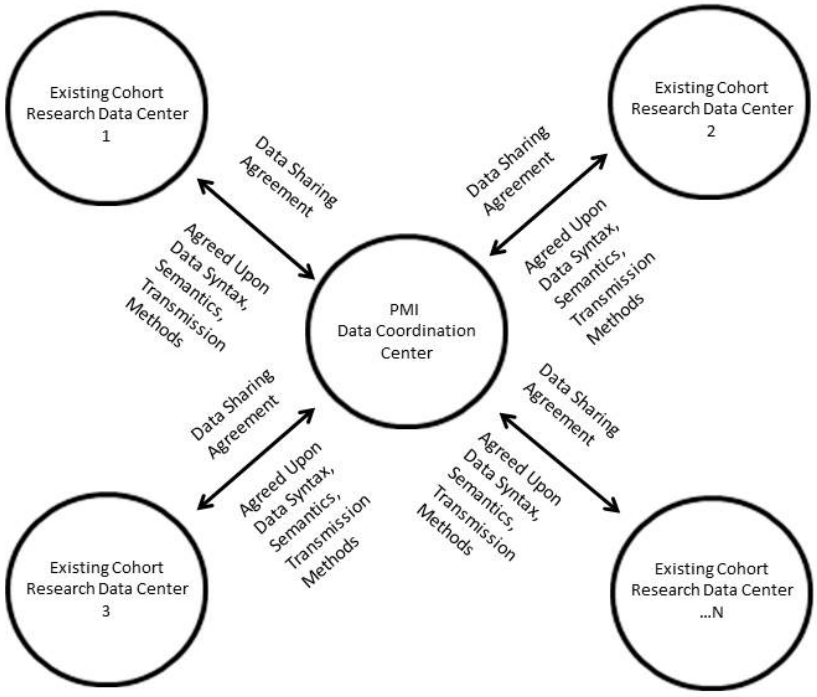

Figure 1. Hub and spoke model for EHR data pooling among existing cohorts

A PMI cohort assembled via the 'stitching together' of existing cohorts would reasonably use this well-established model of research-specific observations as well as clinical data transfer. Startup of new consortia involves both policy and technology. The policy component is the establishment of data transfer and data use agreements between each local organization and the coordinating center at a minimum (which involves N-1 agreements for a consortium with N partners), and can potentially escalate to pairwise agreements among all institutions (N(N-1)/2, or on the order of N^2) depending upon institutional risk management policies. The technology

component involves negotiations between the data coordination center and the research data management staff at each participating site to agree upon both the syntax (structure) and semantics (common naming and coding practices) of the data to be communicated. Successful models exist for both the policy and technology aspects of consortium formation for biomedical research involving EHR-derived clinical data. Though labor intensive and difficult to scale to very large numbers of organizations, a national PMI cohort assembled from the stitching together of existing cohorts would not need to break new ground in policy, data science, or information technology to achieve operational status.

Participant-centered technologies for clinical data access

A novel and as yet untested pathway for acquisition of clinical data for research is via the rights granted to each individual by HIPAA/HITECH legislation to obtain electronic copies of their EHR data. Once an individual 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. This access and download capability has been termed the "Blue Button" and the effort is now curated and managed by the Office of the National Coordinator (ONC). It is the focus of evolving technology and data standards where the Precision Medicine Initiative could be a focal point and inspire coordinated action by the federal government and EHR vendors to accelerate progress.

In addition to the HIPAA access rights, patients have additional and more timely access to their digital health information from providers and hospitals participating as a result of the EHR incentive program, dubbed "Meaningful Use". EHRs certified by ONC under current Meaningful Use Stage 2 criteria are required to be able to produce a few types of documents: a clinical care summary with coded clinical content including problems, medications and lab results (based on the consolidated 'Clinical Document Architecture' standard), an explanation of benefits, and three mechanisms to access and share the information (view only, secure download, and transmit). Enabling patients to access and share their data through the meaningful use view, download and transmit function is an intriguing possibility to build the cohort—especially since the required CCDA format includes many of the desired data types identified in the table above--but there are several challenges. First, the documents are often not sufficiently complete and standardized. Second, the transmit functions are difficult to use and confusing. Third, it is not possible to specify the granularity of the data that are needed (e.g., all my data, all my medication information, etc.), or the desired time period. And last, the patient-facing functions are not built with the attention to user experience and simplicity that Americans have come to expect from consumer products. Two additional challenges with existing methods of retrieving and sharing patient contributed data, such as via Blue Button, is they are one-time downloads of information (and thus would require manual re-uploads of new data) and they do not carry a link back to the original clinical data to enable data to be inserted easily into the EHR (e.g., from genomic testing).

Greater functionality will become available in Meaningful Use stage 3, which begins phased implementation in 2017 under current ONC and CMS rulemaking. The proposed Meaningful Use Stage 3 regulations give patients a new method to access their health data, through an

application programming interface (API). EHR vendor groups have rallied around this option and are developing pilots for API access to health records that incorporate simpler data exchange methods such as FHIR (Fast Healthcare Information Resource) and industry standard query and data representation methods such as a JSON (JavaScript Object Notation) to shorten the time to operational implementation of improved clinical data downloads from ONC-certified EHR systems. OSTP and ONC are positioned to positively impact this evolution of individually-mediated clinical data exchange on a national scale.

The direct-from-participant model of clinical data acquisition for research may tend to favor (though not require) a centralized resource as a common destination for data uploads, as the incoming data will likely arrive in a variety of formats that need quality control, reformatting and data normalization. To the extent that clinical text is included in uploaded data, natural language processing software will need to be developed and adapted to extract and synthesize additional structure from unstructured sources. Though multiple centers might be envisioned as regional destinations for aggregation and processing of individually-uploaded data (similar to Figure 1), a single organization would be likely to achieve expert level competence more quickly based on the larger volume and variety of incoming data received. 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.

Data Privacy

A national cohort that includes a highly interactive approach to communicating with and soliciting input from study participants will necessarily have to operate in two data management modes, while respecting participant preferences and terms of consent. The 'fully identified' mode of operations will be needed for messaging, study appointment reminders, phone interactions, etc. Cybersecurity of these sensitive personal data will be a high priority for the PMI cohort.

Aggregate data assembled for analysis will need to be de-identified by removal of standard classes of personal identifiers such as those specified by HIPAA Limited Data Set and Safe Harbor provisions. These are imperfect privacy standards, however, and the clinical and research-generated data are expected to be rich in features that make each individual's contribution unique. Uniqueness is not synonymous with re-identification (which requires in addition a naming source), but the proliferation of data mining methods and potential naming sources (voter lists, public registries, social media postings, etc.) means that technology alone will be insufficient to address issues of data privacy for the PMI cohort. Acceptable use policies with substantial enforceable sanctions will need to be developed or adapted from other similar research efforts.

Implementation

The strengths and weaknesses of the mechanisms by which cohort data is acquired, as described above, are summarized here:

Pathway to acquiring data	Strengths	Weaknesses
Institutionally mediated	1. Takes advantage of existing data and biobank resources and existing staff expertise managing those resources 2. Leverages an existing infrastructure for cohort re-consent where needed. 3. Provides locally-based quality control of data 4. Supports distributing the task of mapping data to preferred formats and semantics 5. Employs already established and proven methods of data normalization and secure communication. 6. Enhances local institutional prestige as a national PMI cohort participating organization. 7. Provides opportunities for translation into clinical practice, as many existing cohorts are led by large delivery systems 8. Can provide methodologies to periodically update data with new clinical encounters	1.Difficult if not impossible to scale to hundreds or thousands of participating organizations 2.May impose new burdens on already overloaded local IT staff 3.Constrained by institutional perceptions of risk of data sharing, particularly clinical data 4.Data use agreements may become more contentious and difficult as more partners added to consortium (particularly competing health systems) 5.PMI program management expense and complexity for NIH scales proportional to the number of consortium partners. 6.Limits performance of data mining techniques 7.Requires extensive inter-organizational negotiations and governance agreements before work can begin on cohort
	Strengths	Weaknesses
Direct from participants	1. Empowers participants in a direct and appealing process model 2. Exercises HIPAA/HITECH rights of access already in place. 3. Takes advantage of rapidly emerging mHealth platforms and technologies that are being purchased by	1.Requires enhancements to current ‘Blue Button’ and similar technologies, and their availability within commercial EHRs acting as data servers. 2.Requires new software development for apps downloadable by participants.

	large numbers of individuals for other purposes. 4. Potentially scales at low marginal cost to tens of millions of participants or more. 5. Reduces or eliminates need for local IRB review (though central IRB review of project design and experience still important) 6. Reduces or eliminates local IT staff workload at clinical sites. 7. Leverages participants' personal knowledge of their healthcare, for quality control of clinical data submitted 8. Potentially better as a lifetime clinical record than any single institutional EHR, if individuals were to systematically and repeatedly upload information from all their healthcare providers 9. If successful, a powerful model for other research efforts that rely upon active participant engagement.	3.Becomes a reliable comprehensive data source only after critical mass of EHRs provide the needed server side functionality. 4.Requires participant education and outreach regarding their critical ongoing, active engagement 5.Disintermediates institutions from decisions made regarding data release, which may be viewed as a threat to autonomy and income streams. 6.Adding a subcohort that uses a different set of infrastructure and procedures adds cost and complexity relative to a monolithic approach. 7.Overall, a higher risk but potentially higher payoff approach relative to institutionally-mediated access. 8.Current technologies would require participants to repeatedly download and contribute data from multiple healthcare centers, which could lead to out-of-date data 9.Could lead to a more biased enrolled population (those tech-savvy, higher socioeconomic class, younger, and potentially healthier without cognitive deficits)

The key step for both assembly of existing cohorts' data and for creating a receiving capability for individual data uploads will be creation of a detailed statement of the functional requirements to be addressed by organizations wishing to compete for the role of the PMI cohort data

coordination and operations center. The alternatives are not exclusive, and a hybrid approach may be feasible. These requirements would also become part of the evaluation criteria published as an NIH Funding Opportunity Announcement. NIH's standard peer review process is well suited to assessing the merit and feasibility of such data and operations center applications, with the goal of awarding a Cooperative Agreement as the fiscal and management vehicle for implementing the data management technologies and processes needed for the PMI cohort. Given the early developmental stage and low current usage of Blue Button technologies, pilot grants and demonstration projects may be an appropriate stimulus to assessing current capabilities and stimulating advances both in information technology and new participant roles in research.