Precision Medicine Initiative: Building a Large U.S. Research Cohort
NIH Workshop
February 11–12, 2015
Porter Neuroscience Building, Bethesda, MD

EXECUTIVE SUMMARY

Overview

On February 11-12, 2015, NIH held a workshop to explore the opportunities and identify the challenges associated with building a large research cohort as a part of the Precision Medicine Initiative (PMI). Workshop presenters and participants from a wide array of disciplines and sectors were in attendance, and included experts in privacy, patient advocacy, healthcare, epidemiology, genomics, mobile health (mHealth), computer science, and information technology. In addition to invited experts, the workshop was accessed by over 1000 viewers via videocast, and hundreds of participants joined the workshop conversation via twitter and an NIH-hosted online discussion forum.

Recurring Points Made at the Workshop

Participants as partners
Workshop attendees agreed that participants need to be partners in this research. For the PMI cohort to be successful, participants need be engaged from start to finish in all stages of research, including study design, data collection, and governance, particularly regarding decisions about return of research results or other information, including ongoing studies using the cohort, and how to sustain participant engagement over time.

Children
Although the workshop focused on adult participants, many agreed that including children in the cohort could be a valuable method to build a lifespan perspective on health and disease. However, studying children can present particular challenges related to informed consent, and often involves a long period before disease indicators appear.

Building a hybrid cohort that incorporates both existing cohorts and new participants
The merits of using current research and healthcare cohorts, enrolling participants de novo, or a hybrid approach, were discussed. Advantages of existing cohorts include established infrastructure, connection to participants, and extensive knowledge on how to run a cohort. Disadvantages include limited representation, and the challenge of establishing a "least common denominator" for cohorts to be included in the PMI.

The need to incorporate and plan for short-term and long-term goals for the cohort
The PMI cohort will be an ongoing research project with both short-term and long-term goals. The design of the PMI cohort should specify the minimum elements required for an existing cohort to participate in the national cohort, and the requisite baseline data from all research participants. These initial measures can be used to carry out short-term research goals. In the longer term, the addition of new tiers of research participants, some that may only provide select types of data, and cohorts, as well as new data types and participant information, will provide for the possibility of additional research studies and additional analyses.

Enhanced data collection

The PMI cohort has the potential to address various new research questions, and will be most successful if it is flexible and adaptive to new technologies. Novel phenotyping and other data collection technologies, such as mHealth applications to better assess disease outcomes, and behavioral and environmental exposures, are still being developed. One possible approach to build flexibility is to form a core cohort dataset, and to layer tiers of participant groups and additional measurements on top of this base.

Ensuring the cohort is useful for a wide array of research

Rather than focus on any one disease, the cohort will provide a framework for researchers in a wide variety of disciplines to test their own individual hypotheses related to precision medicine. The cohort will be valuable as a testbed for biomedical research studies, mHealth technologies, and for pilot studies for research participants to engage with research in dynamic and ongoing ways.

Evaluation of mHealth technologies

mHealth technologies offer the opportunity to collect myriad data points about a participant's health between doctors' visits, and to engage participants to deposit their own data into research studies and receive feedback. However, mHealth technologies present a number of challenges including feasibility and validation in large, diverse cohorts. The PMI cohort offers the opportunity to test the ability of mHealth technologies to both gather data from participants and relay information back to them that could improve their health outcomes. Regulatory and data security issues will need to be addressed.

Electronic Health Records and Blue Button Technology

"Blue Button" delivery of electronic health records (EHRs) directly to patients enables individuals to download their own electronic health information, and provides opportunities for them to share it with researchers. However, it is not yet fully usable for many patients, and needs to be optimized for the PMI.

<div align="center">

WORKSHOP SUMMARY

DAY 1: **Patient/Participant Centered**

</div>

Vision for the Cohort and the Precision Medicine Initiative – *Francis Collins, M.D., Ph.D.,* *Director, National Institutes of Health*

On January 30, the President announced the launch of the Precision Medicine Initiative to enhance innovation in biomedical research with the ultimate goal of moving the U.S. into an era where medical treatment can be tailored to each patient based on many factors. The vision of the Initiative is to build a broad research program to encourage creative approaches to precision medicine. This includes rigorous evaluation of new approaches, and assembly of an evidence base to inform clinical practice. Early successes of the PMI will include evaluation of the efficacy of pharmacogenomics on treatment decision making and outcomes, identification and testing of new variants that affect drug response, and tests of wearable sensors for monitoring health. In addition, a sign of early success will be a ready platform for researchers to launch observational and interventional studies.

The PMI will be comprised of two components: 1) cancer clinical trials; and 2) creation of a large research cohort to expand our knowledge of precision medicine approaches for all diseases. The cancer

component of the Initiative will test precision therapies in collaboration with private sector partners, and expand current understanding of therapeutic response. The second component provides a long-term structure to generate a knowledge base to move precision medicine to health and disease more broadly by building a national research cohort of one million or more volunteers and supporting research to develop and test technology and mHealth approaches.

The proposed PMI budget for FY16 includes $215 million total; $130M for the research cohort; $70M for cancer; $10M for the U.S. Food and Drug Administration (FDA); and $5M for the Office of the National Coordinator for Health IT (ONC). The timeline for the PMI cohort will entail planning in FY2015, including the formation of a working group of the NIH Advisory Committee to the Director (ACD), outreach, and coordination with the White House and other Federal agencies. Pending appropriations from Congress, research will begin in FY2016.

Session 1: Reports from Workshop Planning Teams – *Moderated by Richard Lifton, M.D., Ph.D., Yale School of Medicine/Howard Hughes Medical Institute*

Prior to the workshop, four workgroups were assembled to evaluate some of the key issues surrounding the implementation of a national cohort, and were tasked with developing white papers and reporting their findings to workshop attendees to engage discussion.

1. Building a consortium of cohorts: cohort identification and participant recruitment – *Workgroup chaired by Michael Lauer, M.D., National Heart, Lung, and Blood Institute, and Eric Boerwinkle, Ph.D., University of Texas Health Science Center, Human Genome Sequencing Center Baylor College of Medicine*

NIH could leverage some of the many federally funded research efforts and existing cohorts to assemble the PMI cohort. Challenges to building the cohort lay in expense, rapidly evolving technology, coordination, and incentives for participants and researchers to join. NIH will need to determine the best balance of leveraging existing cohorts and de novo recruitment for the PMI cohort.

Key Discussion Points:
- Minimal criteria of characteristics must be established for entry of existing cohorts into the PMI cohort, e.g., types of data collected, stored biospecimens, numbers of participants, ability to recontact, existence of EHR, etc.
- NIH should set clear goals for the PMI, and plans to facilitate early discoveries, while also enabling the study of complex problems in the future.
- Enrolling participants de novo could offer unique opportunities.
- A hybrid approach between leveraging existing cohorts and new enrollment is possible.

2. Participant engagement, data privacy, and novel ways of returning information to participants – *Workgroup chaired by Laura Lyman Rodriguez, Ph.D., National Human Genome Research Institute, and Pearl O'Rourke, M.D., Partners HealthCare Systems*

The workgroup provided recommendations for how the PMI cohort could engage participants and promote sustainable participation by being inclusive, flexible, and by providing updates on cohort activities and findings. The cohort should provide transparency in its activities and clear information regarding coverage and payment for any research-related procedures, or follow-up from research results. The cohort should be participant-centric, affording a sense of belonging by leveraging social media or other broadly accessible tools. Individual-level data should be provided that are relevant to a

participant's health. The workgroup suggested that the PMI both establish a participant advisory board and also incorporate participants among all advisory groups throughout PMI governance . Further challenges to promote and enhance participant protections include harmonizing state and Federal privacy requirements, , ensuring data security from data collection through data management, streamlining research oversight requirements , and new pathways to integrate appropriate FDA oversight within cohort activities.

Key Discussion Point:
- Research participant protections for the PMI cohort must be robust, thus methods to oversee cohort activities must be adapted to align with the dynamic and open-ended study envisioned. Institutional Review Boards (IRB)s are mandated to approve protocols to conduct specific research rather than malleable, evolving research relationships between participants, researchers, (and providers). Therefore, IRB review and study implementation will need to be iterative, with phasing of IRB reviews and approvals as details evolve to enable maximal flexibility of cohort plans and ensure appropriate participant protection.

3. Data collection and mobile technologies – *Workgroup chaired by Roderic Pettigrew, M.D., Ph.D., Director, National Institute of Biomedical Imaging and Bioengineering, and Kevin Patrick, M.D., M.S., University of California, San Diego*

The PMI cohort offers the opportunity to build a modern communications information technology (IT) network to seamlessly link EHR, mobile, and other databases. This network could also provide feedback to participants, and mHealth technologies present new opportunities for participant engagement. For the success of this IT network, a concerted effort to develop and share best practices is needed. The cohort will face many implementation challenges in this domain, including data standards, participant privacy, incentives for participation in the cohort, data quality, rapid proliferation of new technologies, and ownership of and access to data. NIH will need to improve ease of use for participants, access to technology for some participant populations, measurement of participant exposure to environmental factors, and interfaces between data and mobile devices.

Key Discussion Points:
- Lightweight and minimally invasive technology that accounts for user experience, design, and feedback will encourage active participation from the broadest possible group of participants.
- A requirement for smartphone usage could skew the composition of the cohort towards young, educated, employed, and healthy participants.
- Standards are needed for both data normalization and analytic techniques. Standards for the validity of data collected with mobile technologies will improve with time and as new data types are combined.

4. Opportunities and challenges related to the use of electronic health records (EHRs) data for research – *Workgroup chaired by Daniel Masys, M.D., University of Washington, Seattle, and Rex Chisholm, Ph.D., Northwestern University*

EHRs include medical records and observational data from participants, family, sensors, etc., and have been adopted by nearly 100% of hospitals and 85% of outpatient facilities. To best make use of EHRs, the workgroup proposed that the PMI cohort develop motivation for participation, finely granular consent, integration of health information and research data, industry engagement, the ability of participants to have control over their data and how it's used, and cybersecurity provisions. Because

per-participant financial incentives for participating researchers and institutions are likely to be small, information returned to them should be valuable. Possible incentives could include access to NIH-sponsored cohorts, relationships with healthcare entities, and methods to facilitate direct submission of data by individuals.

"Blue Button" technology enables individuals to download parts of their own electronic health information, and provides opportunities for them to share it with researchers. However, this technology is not yet fully usable for many patients. The workgroup believes engagement with industry, on a technology-agnostic basis, will be important to ensure EHR data availability. In addition, because cyber-attacks on healthcare data are increasing, data security is crucial, and participants should be informed of the risk of re-identification.

Interoperability is a high priority for the PMI cohort, but EHRs are not currently interoperable, so engaging with existing efforts to build standards will be helpful. Integrating and analyzing data from heterogeneous systems will require both data rich in re-identification features and Natural Language Processing (NLP) of clinical text.

Key Discussion Points:
- Inaccurate or inconsistent diagnoses in EHRs present a problem, but may be addressed by examination of multiple types of EHR data, including laboratory results, prescriptions, written notes, and other data; including participant provided information.
- Use of social media with EHRs is possible, and most EHRs have a patient portal.

The following five sessions began with two to three 10-minute presentations, followed by lengthy discussion with workshop attendees and presenters.

Session 2: Creation of a Large U.S. Research Cohort – *Moderated by Teri Manolio, M.D., Ph.D., National Human Genome Research Institute*

Report of NIH Cohort Inventory Findings – ***Dave Kaufman, Ph.D.,*** *National Human Genome Research Institute*

To gain insight into current efforts relevant to PMI, NIH sought to catalog existing large U.S. cohorts with a broad range of demographics and links to EHRs. NIH explored 69 studies, which is not an exhaustive list. This subset of cohorts collectively includes 13.1M geographically diverse participants. 32 of the studies (2.8-6.1M participants) have the ability to recontact participants, have stored biospecimens, and have consented participants for broad-data sharing. Studies that have worked with underserved and understudied populations have important experience in recruitment, and should be leveraged for the PMI. Beyond the 69 examined studies, other sources that NIH might leverage in building the PMI cohort include large-scale consortia, biobanks, the HMO research network, NCI cohort consortium, PCORnet, Patients Like Me, 23andMe, Sage, Google+, and others.

The Million Veteran Program Cohort: An example of a Large-Scale Health System-Based Cohort – ***Michael Gaziano M.D., M.P.H.,*** *U.S. Department of Veterans Affairs (VA) Boston Healthcare System*

The VA's Million Veterans Project (MVP) has already enrolled 350,000 of a projected one million participants, including underserved minority populations. For each participant, the MVP collects

biospecimens, health records, as well as individual data not available in EHRs. The MVP engages participants via individual mailings, questionnaires, and appointments. Broad consent, including consent for sharing of data with any qualified American researcher for unspecified use, is acquired at the start of participation, and participants are reminded of their participation through periodic newsletters.

The MVP operates all over the U.S., receives and processes 400-600 samples per day, and incorporates hundreds of variables and complex disease phenotypes. Participant phenotypes are determined through a three-tier approach, using an intuitive algorithm, extracts data from records, as well as a probabilistic approach to phenotyping. Challenges to this approach to data collection include limits of self-reporting and EHRs, as well as lifestyle issues, which might be addressed through mobile technology.

Building a Consortium of Cohorts – *Eric Boerwinkle, Ph.D.,* *University of Texas Health Science Center, Human Genome Sequencing Center Baylor College of Medicine*

A cohort is a sample of people sharing some characteristic. Cohort studies are longitudinal, take measurements, assess outcomes, and are not necessarily representative. The Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) consortium has gathered extensive phenotypic and genotypic data on more than 100K participants, including whole exome sequencing. Forming a truly national consortium of cohorts will provide opportunities for new research studies, and could help create a virtuous cycle of research, whereby new knowledge begets yet more knowledge about health and disease. Challenges to building a national cohort include coordination and data harmonization, data storage, analytics, and connectivity among studies and data features. It would be ideal to collect biological samples over time from the same individuals. This could be accomplished if samples taken at the doctor's office were moved to a biorepository, but would be logistically difficult.

Key Discussion Points:
- To build an investment for the future, the PMI cohort should balance the use of existing cohorts, which will not provide all of the necessary data, and new recruitment. Therefore, the PMI cohort should be proactive in planning for future research and gathering new populations. Formative work for methodological advances could start in existing cohorts, creating a platform for rudimentary strategies, which can be layered with additional elements as the PMI matures.
- Although it may not be necessary to make the cohort representative of the U.S. population, the cohort needs to be diverse. To accomplish this, the cohort can oversample some populations and make the sampling strategy clear for weighting in analyses.
- Efforts are needed to include particular groups, such as those below the poverty line or living downstream of industrial pollution. For all participants, but especially those from minority populations, it will be crucial to create trust, address ethical issues, build ongoing relationships, and allow for ways to discontinue participation at any time.
- A large sample size is important for looking at complex questions like gene/environment interactions and rare diseases.
- Not all kinds of health data are represented in EHRs, other databases may be complementary.
- Some existing cohorts that include children could be leveraged for the PMI cohort.
- Developing community among participants is important; people like participating in the MVP because it helps veterans. Mobile technology can offer a way to form communities, allow people to interact with each other, and create a sense of belonging.
- Participants could volunteer to be notified of clinical trials of potential relevance to them.

- Creating a standardized dataset to which researchers can submit queries will present a challenge. NIH is working on how to standardize data cleaning and assign metadata. Engaging people involved in data homogenization in other fields could be helpful.

Session 3: Participants as Partners – *Moderated by Pearl O'Rourke,* M.D., *Partners HealthCare Systems*

Fair Information Practices: Building Trust with Consumers– *Dixie Baker, Ph.D.,* *Martin, Blanck and Associates*
Building trust with participants is extremely important for precision medicine, as the inherent identifiability of genomic data poses a privacy risk to individuals and their families. Consumers want to control their health information. They are concerned about privacy, and appreciate both transparency and being asked for permission for their data to be shared.

Fair information practices principles (FIPPs) codify important consumer principles for information sharing practices. FIPPs were introduced in 1973, and provide 10 consistent principles relevant to information management: collection, quality, purpose, notification, uses, security, openness, access, ability to correct, and accountability. Examples of FIPP implementation include communication practices using language or media that communicate risks and benefits; enabling people to change preferences for data sharing; and avoiding "surprises" for consumers in how data about them is handled.

New Ways of Engaging Research Participants and Novel Consent Models – *Sharon Terry, Genetic Alliance*

The PMI cohort offers the opportunity to engage research participants as co-investigators, to provide relevant and valuable benefits to participants in real time, and to bring about a culture change to empower participants and build transparency and trust in the research enterprise. Consent should be an "engagement" process with participants around a range of concerns and issues surrounding privacy and data management. Trust is more important than privacy for many participants. The Platform for Engaging Everyone Responsibly (PEER) enables participants to choose which data types to allow, deny, or "ask" about sharing. Most participants in PEER to date are willing to share their data.

Participant Perspectives on Data Sharing: What is Important and Why – *Andrea Downing, Brave Bosom*

Three key themes emerged from Ms. Downing's recommendations: the need to let patients lead; the need for equal access to data; and the need to provide information to patients in a timely and effective way. Education tools could help promote data literacy, as well as enable NIH to embrace citizen science.

Key Discussion Points:
- NIH will need to carry out research and outreach activities to assess who is optimal to "speak for" participants. Given the expected diversity of the cohort, NIH should ensure that all prospective participants views are represented, not just the views of patient advocates.
- Building communities of participants and creating a sense of 'belonging' could help volunteers see participation as aiding a greater goal, and help them to engage more authentically.
- Participant perspectives need to be incorporated in the cohort from the outset, in its design and conception, priorities for funding decisions, and the learning process that informs new research.
- Consent processes should focus on participant relevant issues and understanding, and include options for participants to receive and make personal use of their individual data. Consent forms need to facilitate participant understanding of research procedures and available choices about

specific research activities. Dynamic consent models, overarching transparency, and frequent updates for participants, or participant information portals could help better engage participants from initial consent through all study interactions.

- Consent processes must enable potential participants to weigh risks and benefits of research. Broad consent could be a reasonable approach for the PMI cohort if done appropriately.
- Return of data and information to participants is important to people for different reasons. Some organizations return data to participants, and some do not due to fears about how to validate, annotate, and contextualize the returned data for participants. Including participant perspectives in the research design and development process will provide on-going insights into how best to accomplish this goal.
- There should be accountability for research conduct within the PMI cohort and for secondary data users. Consequences for misconduct or misuse of data should be clear and appropriate to particular violations of trust or rules of access. Consequences for some types of misconduct and misuse will need additional analysis to develop appropriate enforcement mechanisms.

DAY 2: Informatics/Data Centered

Session 4: mHealth Technologies – *Moderated by William Riley, Ph.D., Office of Behavioral and Social Sciences Research, Office of the Director, NIH*

How Americans Use Technology to Track and Understand Their Health – *Peter Tippett, M.D., Ph.D., Verizon*

Mobile devices and smartphones are widely used in the U.S., and there are a large number of mobile apps in the health field, many of which are not widely used. Mobile patient monitoring is growing rapidly. In the future, randomized clinical trials may not be the gold standard; a combination of observational science and big data, including inferred data, such as measurements of movement, will likely be the best approach in the future.

Choosing Wisely: What Should be Measured in a Cohort this Large? – *Kevin Patrick, M.D., M.S., University of California, San Diego*

The exposome is the sum of environmental, social, and behavioral exposures, including lifestyle factors, over someone's lifespan. Mobile devices can help monitor the exposome and related health states with a variety of sensors, and reflect the importance of place in health and in the social determinants of health. Engaging with participants through their social networks can help gather more information, improve retention, and facilitate timely communication with participants.

Testing the Devices: Using the Cohort to Assess Efficacy – *Santosh Kumar, Ph.D., University of Memphis*

Ideally, the entire cohort will carry a smartphone and wear a smartwatch, which touches the body and can collect more information than a smartphone. Continuous measurement creates temporal precision, which can lead to timely interventions through quick-response techniques such as text messages, and prevent long-term damage to participants' health. Because continuity is important, active and passive engagement are required to make the studied activities part of the participants' habits.

mHealth technologies can measure outcomes, and can move precision medicine from treatment to prevention. A good platform will account for variability between individuals and measurements, apply data from the past, and enable continuous improvement of data interpretation.

Key Discussion Points
- The market-driven culture of app creators may not take into account differences in access to technology. The use of smartphones could skew the participant sample group, particularly since there are many people without access to unlimited mobile data. Smartphones and support for data plans may need to be given to participants who cannot afford them. Cultural context for participation is also needed, and participant recruitment through places of worship and community centers could reduce bias.
- Vigorous assessment of apps and sensor data is needed to determine whether or not they can accurately assess key variables of the PMI. Apps and sensors developed for research have considerably more data on validity and efficacy than do commercial apps and sensors, and additional study is needed for many of these mHealth apps and sensors prior to deployment in a large cohort study. Examples of targets being pursued include smoking cessation, measuring stress/smoking cues, predicting smoking lapses, and developing interventions. Learning from mHealth tests will important for determining success of the PMI.
- Regulatory approval presents new challenges to mHealth. FDA released updated guidance for mHealth in February 2015. "Health" apps that don't involve diagnosis or treatment are not regulated by the FDA. Many other non-invasive apps targeting disease also are not regulated. "Medical devices," by contrast, require additional time and expense for regulation, and, if used by a HIPAA-covered entity, are subject to HIPAA requirements. As the PMI progresses toward intervention and improving outcomes, researchers will need to understand the regulatory framework, and what is required for FDA approval. ONC has regulatory purview over health IT more generally, and plays a role in setting standards.
- The PMI cohort could accept data from multiple apps for different participants and sub-studies. Open Application Program Interfaces (APIs), permissions, and policy will be critical for this goal. Apps that are useful would rise to the top, and be widely used by study participants.
- The platform needs to be secure, all data sharing should be clear to participants, and participants should be told when there is a breach.
- The PMI cohort could be a platform to improve adherence in both observational and interventional clinical trials. Because many people carry their phones with them at all times, mHealth technologies make it easier to participate in data collection, and texts to participants can be sent to encourage continued participation.
- Data generated by different apps will be heterogeneous, and will require harmonization. There is expertise to accomplish this, and the NIH Big Data to Knowledge Initiative (BD2K) could help.

Session 5: Informatics Requirements and Electronic Health Records (EHRs) – *Moderated by Daniel Masys, M.D., University of Washington*

Motivations to Participate – ***Rex Chisholm, Ph.D.,*** *Northwestern University*

Benefits of EHRs to the public include allowing people to enter their own data, and portals that provide participants with their data. "Blue button" technology enables participants to engage with and own their data, improving clinical care and the future health of participants and their families. However, limitations in current technologies could cause study fatigue and challenges to data interpretation.

Creating a cohort of cohorts for the PMI could limit the study to established ways of thinking. A middle ground for the cohort could be to link existing research data to participants' EHRs, but this would require new or additional consent from participants. The barriers for bringing this PMI cohort together are loss of control of data by existing studies, reduced engagement by investors, and lack of sustainability. The incentives to existing studies are expanded data, more statistical power, and expanded recruitment.

Clinical providers can provide extracts of EHR data for research, but these are complicated, and are hard to use. Access to IT systems could be incentivized by return of data to benefit clinical care, marketing value, expanded recruitment, and payment models to providers. Centralized organization of EHRs is efficient but unwieldy, and requires credit mechanisms to encourage participation by providers. A federated organization of EHRs for the cohort would maximize broad participation, engage local experts, and reduce concern over loss of control over data.

Technical Issues in Aggregating and Analyzing Data from Heterogeneous EHR Systems – *Josh Denny, M.D., M.S.,* *Vanderbilt University*

EHRs are becoming increasingly dense. Phenotyping algorithms for different diseases can be developed using conventional EHR data, free text in the medical record using natural language processing, billing codes, and medication information. However, the development of these algorithms takes time, and EHR data can be inaccurate, biased, fragmented between healthcare systems, and challenging to mine.

Every population can be organized into four groups for each disease: definite, possible, excluded, and controls. The Electronic Medical Records and Genomics (eMERGE) Network collects phenotypes so others can download and use validated phenotype algorithms. 70% of patients are used in more than one study. A central model for combining EHR data would require far fewer numbers of data use agreements than reciprocal agreements between every participating institution, which simplifies the process of integrating EHR data into the PMI cohort.

Enhancing "Blue Button" functionality for Research – *Douglas Fridsma, M.D., Ph.D.,* *American Medical Informatics Association*

Blue Button technology empowers patients to access their EHR data electronically. Today, more than 50% of people can get their health records via Blue Button partners. However, Blue Button should be improved to provide an option to view, download, and transmit data, to allow for triangulation with many types of data, and to facilitate full data extraction. A "sync for science" version of the Blue Button would leverage existing data exchange standards, such as Fast Healthcare Interoperability Resources (FHIR).The PMI data system should be decentralized, with continuous evolution and deployment, and participants as the common feature. Five data elements to be standardized are: meaning, content structure, transport, security, and services.

Key Discussion Points
- A federated system of EHRs and informatics tools may be required at this scale. Social and business issues also make federation attractive, vendors could be incentivized to add patients. eMERGE has given participants and providers a sense of ownership through a "coat check" model, whereby hospitals own their data, and can pull out of the consortium if they want. This model incentivizes the best structure for the consortium.
- Heterogeneity and lack of compatibility among EHRs is an obstacle for the PMI. Some experts can mine data efficiently, but EHRs currently lack a common data model, and a consistent

common data environment may not be scalable for all certified EHRs, so a common denominator must be found among systems.

- Assembling the big picture requires that data from tertiary care centers (as in eMERGE) be combined with data from primary care sites, across regional health providers, and claims. It is difficult to get valuable patient-level CMS data, especially from Medicare.
- Mobile data could enrich phenotypic data, either by a centralized or federated model. Participants stand to gain valuable information from integrating data and devices. Triangulation among multiple types of health data to develop the most accurate knowledge is important to develop a PMI knowledge base, and can be tied into participant engagement by asking participants to validate phenotyping information about themselves. This approach could help eliminate errors.
- Efficient data curation and cleaning approaches are vital for the success of the PMI. eMERGE has been successful using incrementalism: by "cleaning" (i.e., removing errors in) data for one project, the next project becomes easier. ONC is also committed to the development of NLP, because much of the important content is not represented by structured data and needs to be extracted from narrative text.
- The patient needs to be confident in the cost versus benefit of participating in research. My Research Chart uptake and user support helps participants access to their research data, while maintaining separation of research data from clinical care data. Efforts that rely on de-identified data being available to the public or qualified researchers could employ the cloud and layers, whereby a de-identified layer sits on top of the fully identified data layer. eMERGE has been exploring the capability to package programs securely and send these packages to the data in order to return results in a secure manner.

Session 6: Data Access for Researchers: Guiding Principles for Data Access and Sharing – *Moderated by Philip Bourne, Ph.D., Associate Director for Data Science, Office of the Director, NIH*

Data Access: Who Can Access the Data and How? – *Rory Collins, F.MedSci., University of Oxford*

Data in the UK Biobank are available for all academic or commercial healthcare research in the public interest, without preferential or exclusive access. Researchers pay the cost of access. The Biobank's prospective cohort was assembled by obtaining general consent from 500K men and women, along with extensive questionnaires, measurements, and samples. As interest grew, other research components were added, such as medical sensors. Participants stay informed about developments via email and the Biobank website. Foremost obstacles include to the success of the biobank are depleting samples, data inaccessibility, complicated data access processes, and insufficient specificity of disease outcomes. Perceived access is also important as the UK Biobank is rarely used outside of the United Kingdom.

What Protections are Needed – *David Ledbetter, Ph.D., Geisinger Health System*

Geisinger established a biobank in 2007 to leverage its clinical infrastructure, capture data from routine patient interactions, and to make every clinical encounter an opportunity for learning to help the next patient. The study now includes whole-exome sequencing consistent with Clinical Laboratory Improvement Amendments of 1988 (CLIA) standards and performed by Regeneron. All employees go through extensive HIPAA training, and consequences for misuse are termination of employment.

Best practices for the PMI cohort will need to take into account different standards for clinical laboratories, clinical health systems, and the research community, including international researchers

and the private sector. FIPs are valuable best practices. To facilitate participant engagement, ClinGen recently created a patient portal for test reports, which are reviewed and curated by medical genetics laboratories. It is consistent with Federal Information Security Management Act (FISMA) requirements.

Key Discussion Points
- The UK Biobank is centralized in organization, allowing for separation of research questions and approaches.
- Although commercial use of the Biobank is permitted, it is not widespread. Rather than collaborate, the Biobank helps investigators access data, and NIH should consider providing education to help the research community best use the PMI cohort.
- Identifying important genetic variants and entering them into EHRs leaves out a lot of information that clinicians might want to access, such as hypotheses about disease management. This data can be explored in a research environment accessible by the clinician. If the patient is outside the research environment, this causes CLIA problems. However, making large-scale sequencing labs CLIA-certified would be a relatively modest cost.
  Best practices are different in research, clinical, and health care settings. To join research and medical communities into a single ecosystem, these differences must be acknowledged and resolved to find a middle ground.

Comments and ideas that arose through workshop discussion, and that were submitted via twitter and the workshop's online interface, were brought forward in a final discussion session.

Emerging Ideas from the Workshop – *Moderated by Eric Green, M.D., Ph.D., Director, National Human Genome Research Institute*

Health Disparities

- In building the PMI cohort, NIH should consider focused recruitment in areas that have not been covered in other cohorts, including geographical disparity in the central U.S., minority populations, and individuals with rare diseases.
- To ensure diversity in the cohort, disparities in the use of technology and devices may be surmounted. However, disparities in access to health care are more problematic, because participant without health care cannot provide data.
- The issue of trust is paramount to underserved populations.
- To build diversity in the cohort, NIH could leverage activities of heath care providers, such as Kaiser's research into social determinants of health. The IGNITE network brings in primary care networks not affiliated with academic centers, particularly those serving underserved populations. This provides a sense of empowerment, and is encouraging for the future. Additionally, high levels of EHR adoption among critical access hospitals will reduce disparities.

Pediatrics

- Children could be valuable to the cohort because health is a continuum that starts pre-conception, and we do not understand the developmental basis of all diseases. However, recruiting women of childbearing age is not straightforward. Recruiting children presents a problem of reconsenting at ages 12 and 18. There are approaches to lessen that burden, such as systems that can automatically engage electronically when the child reaches a certain age.
- Capturing entire families would be very beneficial: patients want their genetic information shared with their families, but HIPAA does not currently allow for it.

- Younger participants would be valuable to the PMI cohort because they are comfortable with electronic devices. There is also potential for building interest in STEM careers using the cohort.

Approaches to prevent the misuse of health information and data

- Closing loopholes in the Genetic Information Non-discrimination Act (GINA) is important, but could take a long time. Policy and rule-making may be the best approaches to prevent the misuse of data. NIH, however, has limited enforcement abilities once a grant period ends, and some data users will not be NIH grantees. In addition, NIH should be careful in writing guidelines as limitations on data use could stifle innovation.
- The issue of re-identification of participants using research and clinical data is important. The PMI cohort should support a chain of trust among participants, health care providers, and researchers. The organizational structure could also affect the consequences of misuse—hacking a federal database could result in jail time.
- NIH should engage professional societies and scientific journal editors to encourage transparency in other healthcare industries, and to start a culture change.

Cohort Retention

- Leveraging investments that health systems are making in this area could aid retention of cohort participants because some people stay in the same health care system for long periods of time.
- It is critical that participants feel ownership over their data. Partnership in the design and development of the study, building trust, and feeling of personal investment are crucial. The cohort should aim to provide as much transparency as possible. Sending newsletters and having a community advisory board, among other ideas are helpful for retention and engagement.
- Active retention should be encouraged, but building a system of passive retention is essential.
- The retention model will be different for those coming from existing cohorts. In addition to retention, efforts to gather new participants are needed.

Formulating Next Steps – *Richard Lifton, M.D., Ph.D., Yale School of Medicine/Howard Hughes Medical Institute, and Kathy Hudson, Ph.D., Deputy Director for Science, Outreach, and Policy, National Institutes of Health*

To outline a plan for the PMI cohort, which will include addressing many of the issues raised during the workshop, a working group of the ACD will be established. This group will present an interim report to the ACD in September 2015 to inform NIH funding decisions for FY16. NIH will continue to reach out to experts, cohort leaders, mHealth developers, and potential participants over the next few months. Coordination of the PMI will involve inter-agency and trans-NIH governance committees.

Closing – ***Francis Collins, M.D., Ph.D.,*** *Director, National Institutes of Health*

Dr. Collins provided a brief summary of the key themes from the workshop and areas for further discussion. This is the first of many workshops NIH will convene to chart a course for the PMI, and there will be many more opportunities to engage with NIH on this important and exciting Initiative.