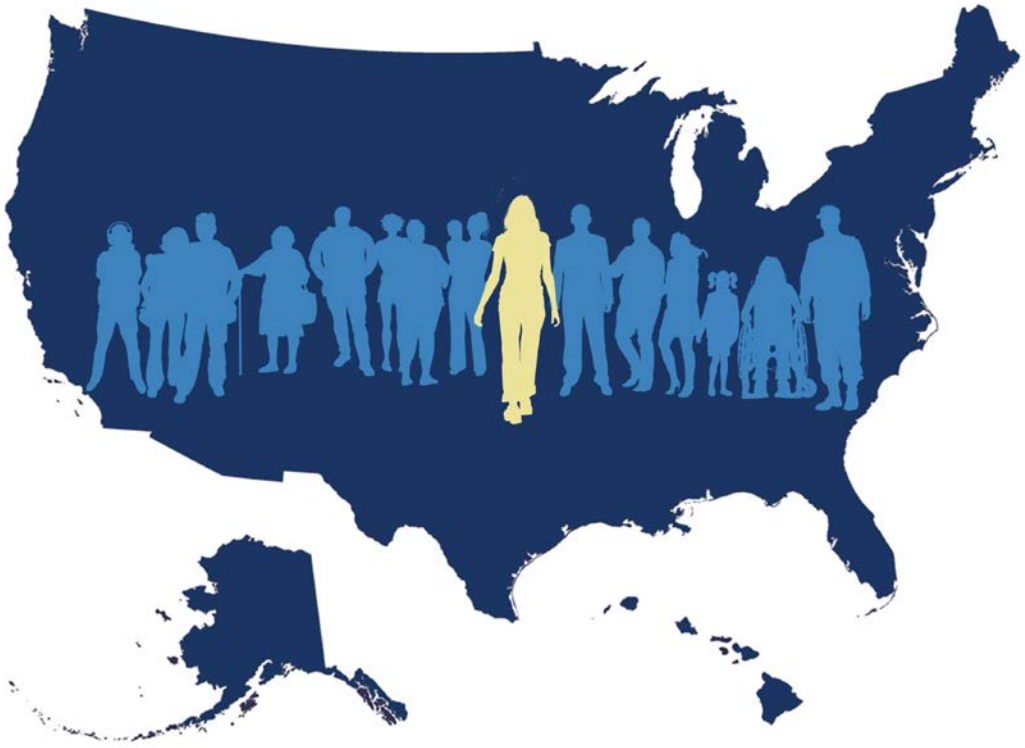


The Precision Medicine Initiative Cohort Program – Building a Research Foundation for 21st Century Medicine

Precision Medicine Initiative (PMI) Working Group Report to the Advisory Committee to the Director, NIH

September 17, 2015

Roster

Kathy Hudson, PhD (co-chair)
National Institutes of Health

Richard Lifton, MD, PhD (co-chair)
Yale University School of Medicine

Bray Patrick-Lake, MFS (co-chair)
Duke University

Esteban Gonzalez Burchard, MD, MPH
University of California, San Francisco

Tony Coles, MD
Yumanity Therapeutics LLC

Rory Collins, FMedSci
University of Oxford

Andrew Conrad, PhD
Google X

Josh Denny, MD, MS
Vanderbilt University

Susan Desmond-Hellmann, MD, MPH
Bill and Melinda Gates Foundation

Eric Dishman
Intel Corporation

Kathy Giusti, MBA
Multiple Myeloma Research Foundation

Sekar Kathiresan, MD
Harvard Medical School

Sachin Kheterpal, MD, MBA
University of Michigan Medical School

Shiriki Kumanyika, PhD, MPH
University of Pennsylvania Perelman School of Medicine

Spero M. Manson, PhD
University of Colorado, Denver

P. Pearl O'Rourke, MD
Partners HealthCare

Richard Platt, MD, MSc
Harvard Pilgrim Health Care Institute

Jay Shendure, MD, PhD
University of Washington

Sue Siegel
GE Ventures

EX OFFICIO MEMBERS

Robert Califf, MD
Food and Drug Administration

Karen DeSalvo, MD, MPH, MSc
Office of the National Coordinator

Jo Handelsman, PhD
Office of Science and Technology Policy

Timothy O'Leary, MD
Department of Veterans Affairs

Terry M. Rauch, PhD
Department of Defense

EXECUTIVE SECRETARY

Gwynne L. Jenkins, PhD, MPH
National Institutes of Health

Acknowledgements

The Precision Medicine Initiative (PMI) Working Group was privileged to be given an audacious charge from the NIH Director to lay out a blueprint for a transformational Presidential Initiative. Our work together has been enormously gratifying, bringing together a remarkable group of people of diverse expertise and experience to focus on a shared goal. Our deliberations benefitted from the remarkable contributions of hundreds of individuals from across the country who participated in the four workshops held over the last 6 months. We are deeply indebted to the panelists, speakers, and attendees at these workshops, who posed and debated critical opportunities and challenges for the design and execution of the PMI Cohort Program (PMI-CP). Considering the exceptionally short time line for planning and holding these workshops, the overwhelmingly positive response of invitees, many of whom had to change long-scheduled plans and travel great distances to participate, was extraordinary; we are deeply indebted to all for helping to shape the Working Group's thoughts on many complex issues.

These workshops, as well as the two requests for information (RFIs) and analysis of a survey by the Foundation for NIH that informed the Working Group, were developed by an outstanding group of experts at the NIH. We would like to give special recognition to the workshop planning leads: Vence Bonham, Josie Briggs, Alan Guttmacher, Regina James, Mike Lauer, Laura Lyman Rodriguez, Teri Manolio, Dan Masys, Roderic Pettigrew, Bill Riley, and Carrie Wolinetz, as well as Gary Gibbons and Eric Green who provided a liaison to NIH IC Directors, and Dave Kauffman for his analysis of FNIH survey data.

Our efforts were supported by a sensational team of NIH staff including Allison Lea and Lauren Milner, and led by Gwynne Jenkins, our Executive Secretary who provided steadfast dedication, countless hours of work, amazing intellect, exceptional judgment, steady grace and wry humor. Working Group member Josh Denny provided extensive time and commitment to drafting the report and coordinating deliberations. Our gratitude also goes to Dan Masys, who provided extensive consultation. Throughout our project, the NIH Office of Communications and Public Liaison provided support for social media and communications for our work.

We are grateful to the hospitality and collegiality of Vanderbilt University (Nashville, Tennessee) and the Intel Corporation (Santa Clara, California), which each hosted a public workshop of the Working Group.

The Working Group Co-Chairs

Kathy Hudson, Rick Lifton, and Bray Patrick-Lake

Contents

update the TOC

Executive Summary

In his State of the Union Address on January 20, 2015, President Obama announced his intention to launch a Precision Medicine Initiative (PMI) "to bring us closer to curing diseases like cancer and diabetes, and to give all of us access to the personalized information we need to keep ourselves and our families healthier." Ten days later, at a White House event with patients, advocates, scientists, and industry leaders, the President shared his vision for the 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.

Precision medicine is an approach to disease treatment and prevention that seeks to maximize effectiveness by taking into account individual variability in genes, environment, and lifestyle. Precision medicine seeks to redefine our understanding of disease onset and progression, treatment response, and health outcomes through the more precise measurement of molecular, environmental, and behavioral factors that contribute to health and disease. This understanding will lead to more accurate diagnoses, more rational disease prevention strategies, better treatment selection, and the development of novel therapies. Coincident with advancing the science of medicine is a changing culture of medical practice and medical research that engages individuals as active partners – not just as patients or research subjects. We believe the combination of a highly engaged population and rich biological, health, behavioral, and environmental data will usher in a new and more effective era of American healthcare.

In order to achieve the President's ambitious plan, the PMI Cohort Program (PMI-CP) will build a large research cohort of one million or more Americans that will provide the platform for expanding our knowledge of precision medicine approaches and that will benefit the nation for many years to come. In March of 2015, NIH Director Dr. Francis Collins formed the PMI Working Group of the Advisory Committee to the Director to develop a plan for creating and managing such a research cohort. To help carry out its charge, the Working Group engaged with stakeholders and members of the public through workshops and requests for information, focusing on issues related to the design and oversight of the cohort. Public engagement, as well as internal discussions among the Working Group, led to the vision for the design and utility of the cohort program outlined in this report. The report includes recommendations in six areas critical to the development, implementation, and oversight of the PMI-CP: cohort assembly, participant engagement, data, biobanking, policy, and governance. In addition to the recommendations, the Working Group outlined the potential utility and unique opportunities that could be addressed by the PMI cohort.

Capabilities of the PMI cohort: Thanks to advances in genomic technologies, data collection and storage, computational analysis, and mobile health applications over the last decade, the creation of a large-scale precision medicine cohort is now possible in a way that it was not before. The Working Group identified a number of high-value scientific opportunities or use cases that could be used to inform the design of the PMI cohort. These use cases include: development of quantitative estimates of risk for a range of diseases by integrating environmental exposures, genetic factors, and gene-

environment interactions; identification of determinants of individual variation in efficacy and safety of commonly used therapeutics; discovery of biomarkers that identify people with increased or decreased risk of developing common diseases; use of mobile health (mHealth) technologies to correlate activity, physiologic measures and environmental exposures with health outcomes; determination of the health impact of heterozygous loss of function mutations; development new disease classifications and relationships; empowerment of participants with data and information to improve their own health; and creation of a platform to enable trials of targeted therapy.

These scientific opportunities will be explored in stages throughout the life of the PMI cohort and should not be considered as an exhaustive list. Rather, as the richness of the data increases and technology evolves, additional capabilities of the cohort will emerge.

Specific Recommendations for the PMI-CP and PMI Cohort. The Working Group offers recommendations, major and minor, to guide the development of the PMI-CP and the PMI cohort. What follows is a summary of the major findings and recommendations. References to specific numbered recommendations, where appropriate, are included in parentheses.

Cohort Assembly. The Working Group supports the initial goal to include one million volunteers, and suggests that the cohort should continue to grow over time. The Working Group envisions the PMI cohort as a new, broadly accessible, national research resource of volunteers specifically consented as part of the PMI-CP. To be useful to the goals of PMI, the Working Group recommends that all potential participants in the PMI cohort must agree to share their health data, provide a biospecimen, and be recontacted for future research (3.1). The Working Group also recommends that the PMI cohort reflect the diversity of the U.S. (3.2).

The Working Group identifies two distinct methods to recruit participants (3.4). The first approach is designed to enable any individual living in America to volunteer for the PMI cohort (3.3). These "direct volunteers" would consent to be part of the PMI cohort, agree to be recontacted, undertake a PMI baseline health exam, and provide a biospecimen. They would share available health data by either directing their electronic health record (EHR) data to the PMI-CP and/or by undergoing an initial exam with a health care provider. The second approach would be to collaborate with healthcare provider organizations (HPOs) to recruit participants (3.5). HPOs would recruit participants, consent them for participation in the PMI cohort, conduct a PMI baseline exam, collect a biospecimen, and share EHR data with the PMI-CP. When selecting HPOs to join the PMI-CP, the Working Group recommends that NIH consider the contribution of the HPO to the overall diversity of the PMI cohort, the robustness of the EHR, and expected length of follow-up for participants. With robust implementation, the Working Group expects the PMI-CP to be able to recruit at least one million participants over about four years.

Participant Engagement. The Working Group recommends that the PMI cohort be developed using a highly interactive and proactive participation model, where cohort participants are encouraged to engage in all aspects of the cohort. In addition to providing feedback and input during planning and implementation phases of the cohort, participants should have significant representation on PMI-CP governance and oversight committees (4.1). The Working Group recognizes that building and

maintaining trust is a critical component to a successful, ongoing, collaborative relationship with participants and the public at large, and recommends the development of a set of guiding principles that apply to all PMI cohort stakeholders (4.2). To ensure that participant-related interactions remain consistent throughout the PMI-CP, the Working Group recommends that activities related to engagement and communications be organized and managed by a central entity (4.3, 4.4). However, the PMI-CP would still benefit from collaborating with a variety of organizations in support of the PMI cohort, to maximize its outreach potential. The Working Group also recommends that the PMI-CP use a standardized consent protocol to ensure consistency in the terms and conditions that all PMI cohort participants agree to. The PMI cohort consent protocol should give participants the option to join supplementary or complementary studies outside the PMI cohort (4.5, 4.6). The Working Group recommends that the PMI-CP return to each participant their own results and aggregated results from its studies to all participants (4.7). Participants should be able to set preferences to dictate how much personal information they receive, and be able to change their preferences throughout their participation in the PMI-CP. To oversee the development and implementation of policies related to the return of aggregate and individual results to participants, the Working Group recommends that a subcommittee with significant participant representation be formed as part of the PMI-CP governance structure (4.8).

Data Considerations. Successful development of the PMI-CP will require a combination of well-proven and innovative methods and technologies for data collection and management. Guided by the scientific opportunities, the PMI-CP should anticipate and collect a diverse set of data types, beginning with a core set of high-value variables to be acquired during enrollment from all PMI cohort participants. Together, these will constitute a core data set that will enable both cohort-wide analyses and identification of subcohorts eligible to participate in specialized studies (5.1). The Working Group recommends that the initial core data set acquired from all PMI cohort participants be collected and stored centrally (5.13, 5.23). The PMI-CP should seek to align its core data set with other comparable core data sets where possible. The recommended initial core data set includes data from EHRs, health insurance organizations, participant surveys, mHealth technologies, and biologic investigations, and would be expected to grow with time.

Efficient structuring and management of data is important to the success of the PMI-CP. Toward this end, the Working Group recommends use of a common data model to organize data similarly across HPOs and from direct volunteers, where possible, while recognizing that many data types useful from clinical investigation may not easily be transformed to an existing or created standard at this time (5.20, 5.23). The best approach will balance normalization of only the highest value data initially for all participants followed by on-demand data curation of other data as driven by scientific demand. In addition, the Working Group recommends that existing data standards and common data models be leveraged where possible (5.25, 5.26), while recognizing that standards do not exist for many emerging modalities, such as a number of sensor technologies. The Working Group recommends early selection of commonly used mHealth technologies to gain experience in use and integration of these new modalities.

To ensure that PMI cohort data is used responsibly and securely, the Working Group recommends a data access control approach appropriate to the level of sensitivity of the data, from open-access for summary data to role-based access for individual level data (5.16). Participants should have access to their own PMI-CP data, except when there are compelling concerns about potential harm that may arise from such access (5.20). The process for return of individual-level PMI information should be overseen by a subcommittee that includes substantial participant representation (4.8). An initial set of analysis, visualization, and dashboard tools should be offered for participants as soon as possible in development of the PMI cohort (5.16). In general, data should be accessed and analyzed in de-identified forms (5.31), and secure computing environments should be used for data access and analysis to enhance security and maintain privacy of the data (5.22, 5.28). Maintaining data security and privacy will be paramount to maintaining participants' trust and engagement in the PMI-CP, so it should engage teams of privacy experts and employ rigorous security testing models (5.32), develop participant education with regard to privacy and potential re-identification risk (5.33), and clearly articulate response plans in the case of a privacy breach (5.34).

To facilitate data access, data normalization, and participant engagement, the Working Group recommends that the PMI-CP follow a "hub-and-spoke" model that has a Coordinating Center to provide a single point of contact for coordinating data, biospecimens, participant communication and engagement, and research studies (3.8, 5.23). The Working Group encourages NIH to consider novel collaborations with not-for-profit and commercial organizations to achieve state-of-the-art analysis methods, scientific rigor, elastic storage and compute capabilities, and technological expertise (3.9, 5.21). For data storage and access, the Working Group recommends the PMI-CP pursue a hybrid data and analytics architecture that leverages both centralized data storage of core data while preserving federated access to additional data at the nodes across the network, as needed by specific studies (5.24). This hybrid model would accelerate execution of many research queries but still allow detailed data access for queries not addressable through the current data common data models.

Biobanking. Collection of biological specimens is essential to the successful development of the PMI-CP. The Working Group recognizes a number of types of biospecimens that could be collected, including blood, microbiome specimens, and nail and hair clippings, among others. The Working Group feels the highest priority for a national collection would be for blood collection (6.5). The Working Group recommends that each PMI cohort participant should provide a new blood specimen, to be collected and processed using a standard CLIA compliant procedure, in order to ensure quality control and comparability of biospecimen collection across the PMI cohort (6.3, 6.4). Finally, the Working Group notes that samples collected for the PMI cohort will be sent to a central biorepository, which will support collection, processing, storage, retrieval and biochemical analysis and/or shipment to analytic laboratories. This biobank should be in place before the start of recruitment (6.2).

Policy considerations. The success and longevity of the PMI-CP will be heavily influenced by the laws, regulations, and policies surrounding research, data security and privacy, and access and interoperability of EHRs. Gaps and conflicts in policies will need to be addressed. In addition, an internal framework of PMI-CP policies will need to be developed to address participant inclusion; Institutional Review Board (IRB) review and consent; privacy, misuse of information, and security; sharing of data and specimens

with researchers; and sharing of data and research results with participants. To achieve the goal that the PMI cohort reflect the diversity of the U.S. population, some populations, such as decisionally-impaired individuals, will require special attention to address the unique ethical, logistical, and legal implications of inclusion in the PMI cohort (7.2 and 7.3). The Working Group has a number of recommendations related to the ethical review of research and consent of all participants, including the use of a single IRB to reduce administrative burden and associated costs of the cohort, review time, and to harmonize inconsistent or conflicting policies between PMI cohort nodes (7.4). The Working Group also has recommendations related to security and privacy of individual information, including establishing safeguards against unintended release of data (7.8) and penalties for the unauthorized re-identification of participants (7.9). These recommendations are intended to ensure the proper use of the data and to set the foundation of trust between participants, researchers, and PMI-CP governance. Finally, the Working Group recommends that the PMI-CP support revisions to the Common Rule that enable broad consent for secondary use of data and biospecimens in research (7.7) which is key for optimizing utility of the PMI cohort.

Governance. The governance structure for the PMI-CP must combine effective and timely decision making with opportunities for consultation and deliberation. The PMI-CP will be large and diverse in terms of constituents, sites, participants, research use cases, and infrastructure needs. Thus, the PMI-CP will require nimble and innovative approaches. The Working Group recommends that the PMI-CP be led by a director with the institutional authority, professional expertise, and structural support to provide strong, credible, and effective leadership (8.1). The PMI-CP director should have authority to set short and long-term goals, develop solicitations and make award recommendations, and have authority over data sharing, storage, and use. The Working Group recommends the creation of a Steering Committee of critical stakeholders led by a smaller Executive Committee (8.2), and five subcommittees that would report to and inform the work on the Steering Committee and PMI Director: Return of Results and Information (4.8), Data (5.2), Resource Access Subcommittee (5.19), Biobanking (6.1), and Security (7.16). The Working Group also recommends that the creation of an Independent Advisory Board to provide external oversight for the PMI-CP (8.3). NIH should develop a mechanism to ensure that governance structures for the PMI-CP are coordinated with other federal agencies, including the Centers for Medicare & Medicaid Services, the Health Resources and Services Administration, the Food and Drug Administration, the Office of the National Coordinator for Health Information Technology, the Department of Veterans Affairs (including the Million Veteran Program), and the Department of Defense (8.4). However, final responsibility for the PMI-CP should reside with the NIH Director (8.5)

Concluding remarks. After careful consideration, the Working Group is unanimous and enthusiastic in supporting this endeavor. The Working Group is convinced that the time is right to mount this ambitious project to transform the understanding of factors contributing to individual health and disease, with conviction that success in this effort will advance the health of the United States.

Section 1 – Introduction

A. The promise of precision medicine

The surest path to advancing prevention and treatment of disease has been the detailed understanding of the factors that contribute to health and disease in individual patients. Biomedical research has made continual progress in refining the classification of disease and determining the underlying factors that contribute to it. Rigorous evaluation of the safety and efficacy of new preventive and therapeutic strategies has led to reduced morbidity and mortality. Commonly, such evidence leads to treatments that are expected to benefit the population as a whole, based on the expected response of a "typical" patient. However, despite a net benefit, it is clear that individual patients can have markedly variable responses to therapy, ranging from highly efficacious outcome, to no effect, to deleterious outcome. The roots of this variability likely include unrecognized differences in disease pathophysiology, environmental exposures, social and behavioral factors, and genetic factors.

To date, our progress in identifying optimal treatments for each individual has been modest, owing to incomplete knowledge of disease causation in individuals and the factors that dictate variable responses to therapy. Moreover, ideal approaches would prevent development of disease in the first place. This requires the ability to recognize individuals at high risk of developing specific disorders and the development of new interventions that can prevent subsequent development of overt disease. Examples of diseases of high population burden that currently do not have specific predictive biomarkers include Alzheimer's disease and type II diabetes mellitus.

We define precision medicine as an approach to disease treatment and prevention that seeks to maximize effectiveness by taking into account individual variability in genes, environment, and lifestyle. Precision medicine endeavors to redefine our understanding of disease onset and progression, treatment response, and health outcomes through the more precise measurement of potential contributors – for example, molecular measurements as captured through DNA sequencing technologies or environmental exposures or other information captured through increasingly ubiquitous mobile devices. A precise delineation of the molecular, environmental, behavioral, and other factors that contribute to health and disease will lead to more accurate diagnoses, more rational disease prevention strategies, better treatment selection, and the development of novel therapies. Coincident with advancing the science of medicine is a changing culture of medical practice and medical research that engages individuals as partners – not just as patients or research subjects. We believe the combination of a highly engaged population and rich biological, health, and environmental data will usher in a new and more effective era of American healthcare.

We have already witnessed early successes of precision medicine. These include, for example, the development of targeted treatments for cancer[1] and cystic fibrosis[2] that are effective in patients who share an underlying causal genotype. Precision medicine is also yielding a wealth of potential information to help ensure that each patient is given the right drug at the right dose the first time.[3] For example, there are clear examples of therapies where individual genetic profiling can be used to avoid

drugs likely to cause serious adverse effects[4–6] and progress is being made in understanding how to optimize therapies based on how different polymorphisms predict therapeutic response.[7–10] With individual genome sequencing, patients with previously undiagnosed genetic diseases are being successfully diagnosed.[11,12] In addition, new subtypes of disease are increasingly being defined through molecular profiling of affected tissues, an advance that is expected to lead to more focused design and testing of both therapeutic and preventative strategies – for example, treatments for specific subtypes of a disease, or behavioral interventions tailored to specific subgroups of the population.

The widespread adoption of precision medicine may have a profound impact on American competitiveness and the economy. The economic impact of investments in large-scale biomedical research has proven true in the past: the $4 billion investment in the Human Genome Project has spurred an estimated $965 billion in economic growth – a 178-fold return on investment.[13] Continued innovation in biomedical technology and research is required for the U.S. to sustain its global leadership in these fields.[14,15]

B. The time is right

The concept of precision medicine is not new.[16] However, the rising costs of drug development and healthcare in the U.S. suggest that a new model of clinical care is needed that will rely on robust and innovative health research. Drug discovery has slowed, and only a small fraction of proposed medications are successfully translated into approved and prescribed thearaputics.[17] Clinical trials of new therapeutics may often be underpowered due to unrecognized heterogeneity in disease pathogenesis among enrolled patients such that drugs that are highly beneficial for a definable subset are rejected because the majority of patients in the trial fail to respond. The discovery of genetic factors underlying disease can be used to identify drug targets[18,19] as well as to selectively give those drugs to patients that are most likely to have the greatest efficacy with the least adverse effects.[20] Understanding the genetics of disease and biomarkers will allow us to rationally select patient groups that are most likely to respond to particular agents, not only improving “numbers” (e.g., lower cholesterol) but also improving health outcomes (e.g., reduced heart attacks) and quality of life.[21,22]

Prospective cohort studies have the ability to identify biomarkers and causative factors contributing to future disease. This approach revolutionized the prevention of cardiovascular disease, the most frequent cause of death in the U.S. and worldwide.[23,24] From the prospective study of several thousand individuals, the Framingham Heart Study identified smoking, high LDL cholesterol, and high blood pressure as major independent risk factors for heart attack and stroke, ultimately leading to dramatic reductions in morbidity and mortality from these diseases by mitigation of these risk factors. This highly successful model has been challenging to replicate for less common diseases, which, nonetheless, collectively account for enormous health and economic burden on the population. For these diseases, a prospective cohort of several thousand individuals would typically yield too few incident cases to provide statistical power to detect new prospective risk factors. Moreover, establishing the necessary infrastructure to support the recruitment, data collection, and longitudinal follow-up for hundreds of such studies of individual diseases is inefficient and very costly. Instead, it would be ideal and cost-effective to study a single, very large cohort that would provide sufficient power to study ostensibly all

relatively common diseases within a single cohort.[25] The barriers to such a study have been the relatively high cost of ascertaining cohort members, collecting comprehensive clinical and experimental data, and following participants over time.

Over the last decade, however, a number of technological advances have converged to dramatically reduce the barriers to the assembly, evaluation, and analysis of cohorts of one million or more people. Firstly, the information technology (IT) revolution has provided remarkable reductions in the cost of data storage, and comparable increases in analytic capabilities. This has enabled the assembly and analysis of exceptionally large databases in biomedicine, as well as many other fields. In parallel, the raw cost of DNA sequencing has been reduced nearly 10 million-fold from the time the sequencing phase of the Human Genome Project began in 1998, with complete human genomes now being routinely sequenced and analyzed for less than $2000 in several days, and expectations for continued reductions in cost. Similar advances in mass spectrometry have drastically lowered the cost and expanded the ability to characterize the proteins and metabolites present in biological samples.

The development and wide implementation of electronic health records (EHRs) in 95% of U.S. hospitals and a number of comprehensive health care systems[26] now allows the collection and maintenance of ostensibly complete longitudinal health care records at extremely low cost. EHRs provide a rich source of clinical data that have been used by researchers to examine biological and environmental contributions to a wide array of conditions and health outcomes. The accessibility and interoperability of EHRs – e.g., the ability to exchange and combine health data within and across healthcare systems, as well as the ability of individual patients to access their own data, is a vital present-day challenge. However, if this can be overcome, the networking of large cohorts of individuals with EHR data has the potential to dramatically accelerate biomedical research, while also providing passive follow-up on individual health for longitudinal studies.

Personal mobile technologies have been adopted at an exponential rate, with more than seven billion cellular phone subscriptions worldwide as of 2014.[27] In the U.S., 91% of adults have a mobile phone, and 64% have a smartphone.[28,29] In addition, many medical technologies traditionally found only in hospitals and clinics have become mobile, home-based, and/or consumer-operated (e.g., blood pressure devices, home defibrillators, heart rate monitors), enabling a wide range of telehealth and remote monitoring scenarios. There is a growing interest in using phones, wearables, in-home devices, and related mobile technologies as a novel way to collect health information (sometimes called “mHealth”), to improve patient care as well as to advance research. Data from sensors and software applications can enrich self-reported data on lifestyle and environment, giving researchers a clear view into these factors that have previously been difficult to capture with accuracy.

Equally importantly, in addition to technological advances, patients have become more engaged in healthcare and health research, more connected and organized through social media, and more impatient as they wait for better treatments and cures for themselves and the people they love. These trends denote a cultural shift that is critical to the success of precision medicine.

Given these rapid and ongoing transformations in medicine, technology, and society, it is the Working Group's assessment that the time is right for the U.S. to undertake an ambitious research agenda focused on development and implementation of precision medicine to improve the health of the nation.

C. Presidential announcement of the Precision Medicine Initiative

In his State of the Union Address on January 20, 2015, President Obama announced his intention to launch the Precision Medicine Initiative (PMI) "to bring us closer to curing diseases like cancer and diabetes, and to give all of us access to the personalized information we need to keep ourselves and our families healthier."[30] Ten days later, at a White House event with patients, advocates, scientists, and industry leaders, the President shared his vision for the PMI 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.

The President's PMI is comprised of many components, incorporating efforts from across the government, including the Food and Drug Administration's (FDA's) effort to develop a nimble framework for the regulation of genomic technologies; the Office of the National Coordinator for Health Information Technology's (ONC's) work to identify data standards critical for precision medicine; the Department of Veterans Affairs (VA) commitment to the Million Veteran Program (MVP); and efforts from the Department of Defense (DoD) to bring active duty women and men into PMI research.

Table 1.1: Proposed PMI Budget Allocations for FY 2016
Department of Health and Human Services

Investment	Agency	Purpose
$130 million	National Institutes of Health	To develop a voluntary national research cohort to propel our understanding of health and disease and set the foundation for a new way of doing research.
$70 million	NIH National Cancer Institute	To scale up efforts to identify genomic drivers in cancer and develop more effective approaches to cancer treatment.
$10 million	Food and Drug Administration	To acquire additional expertise and advance the development of high quality, curated databases to support the regulatory structure needed to advance innovation in precision medicine.
$5 million	Office of the National Coordinator	To support the development of interoperability standards and requirements that address privacy and enable secure exchange of data across systems.

NIH's investment in PMI will focus on: 1) PMI-Oncology, an effort to advance precision oncology and 2) the PMI Cohort Program (PMI-CP), the creation of large, voluntary national research cohort. PMI-Oncology will test precision therapies for cancer, including targeted agents and immunotherapies, while also developing a national cancer knowledge system.[32] PMI-CP will lay the foundations for precision medicine approaches more broadly, by building a national research cohort of one million or more volunteers who are engaged as partners in a longitudinal, long-term effort to identify the molecular, environmental and behavioral factors that contribute to diverse diseases; to facilitate the development

and testing of novel therapies and prevention approaches; and to pioneer mHealth strategies for improving the efficacy of health care.

The President's proposed PMI budget for fiscal year (FY) 2016 is $215 million,[33] and is allocated to different agencies as noted in Table 1.1 (above).

D. Charge and activities of the ACD PMI Working Group

Charge

On March 30, 2015, Dr. Francis Collins, NIH Director, established the Advisory Committee to the NIH Director (ACD) PMI Working Group[34] to devise a blueprint for the creation of the PMI-CP. The charge to the PMI Working Group was to develop a vision for how to harness recent advances in technology, scientific understanding, and participant engagement to develop a platform for precision medicine research. To ensure that its contributions can be incorporated into the funding plan for FY 2016, the Working Group was tasked with developing a plan for creating and managing a large research cohort with data and specimens that can be accessed by all researchers for studies to understand the variables that contribute to health and disease and ultimately translate that knowledge into treatments tailored to individuals. The charge for the Working Group asked its report to:

- Articulate the vision for the PMI cohort, including what can be learned from a study at this scale and what success might look like five and ten years out.
- Detail the issues that will need to be examined in developing the study design.
- Consider how to incorporate rapidly evolving technology into the cohort design.
- Develop criteria for the inclusion of mobile health technologies into the cohort, both for baseline and ongoing data collection.
- Outline a path forward for getting the EHRs associated with the cohort interoperable or for standardizing data that can be extracted in response to a query.
- Consider what laboratory components can be incorporated, including various –omics and imaging approaches.
- Assess ways to recruit visionary experts from multiple disciplines to join the PMI cohort, including potential support of pilot projects to test new ideas and new technologies.
- Identify the pros and cons of different models for enrolling participants into the cohort, including consideration of methods for ensuring that underserved populations are included.
- Review existing models of participant engagement and recommend a path for establishing the best strategy for ensuring participants are involved in the planning, building, and management of the cohort.
- Identify any gaps in existing policies on data privacy, security, and misuse that must be filled to support the PMI cohort, and propose a route to develop appropriate solutions.
- Recommend approaches to work with the International community, both for learning as we build the cohort and for possible collaborations.
- Describe methods for evaluating the success and utility of the cohort for all parties – participants, researchers, healthcare providers, funders, etc.

Activities

In carrying out its charge, the Working Group engaged many individuals with a wide variety of expertise and experience through four public workshops on issues critical to the design and vision for the cohort and two Requests for Information (RFIs). The Working Group also met a number of times in person and by conference call to discuss and deliberate specific recommendations regarding the development and implementation of the PMI-CP.

Workshops

On April 28-29, 2015, the Working Group held its first workshop, *Unique Scientific Opportunities for the National Research Cohort,* at the NIH in Bethesda, Maryland.[35] The purpose of this workshop was to consider visionary scientific questions that could be addressed by the national research cohort proposed under the President's PMI. The workshop featured speakers and panelists with a variety of expertise, including genomics, metabolomics, environmental and behavioral health, and informatics. The workshop resulted in a series of use cases describing the distinctive science that the cohort could enable in the near term and longer term.

On May 28-29, 2015, the Working Group held the *Digital Health Data in a Million-Person Precision Medicine Initiative Cohort* Workshop at Vanderbilt University in Nashville, Tennessee.[36] The purpose of this workshop was to discuss scientific and methodologic considerations for cohort design and collection of detailed health information in the development of a national research cohort of one million or more Americans. The workshop featured panelists and speakers from many disciplines and sectors, including epidemiologic and health care delivery network cohorts, data integration, health advocacy, and information technology, as well as EHR systems. Two special guests, U.S. Senator Lamar Alexander and U.S. Representative Marsha Blackburn, participated and shared visions for precision medicine and improved health. Workshop discussions informed Working Group considerations on the inclusion of specific participant groups in the cohort and the most appropriate model for PMI cohort data aggregation and sharing.

On July 1-2, 2015, the Working Group held the *Participant Engagement and Health Equity* Workshop at the NIH in Bethesda, Maryland.[37] The purpose of this workshop was to discuss participant engagement and health equity as they relate to the proposed PMI-CP. The workshop focused on the design of an inclusive cohort, building and sustaining public trust, direct-from-participant data provision, and effective and active participant engagement characteristics of a national research cohort. The workshop featured panelists and speakers with expertise in participant engagement, recruitment and retention, research with underserved and underrepresented populations, health equity and health disparities, large cohort and long-term longitudinal studies, and health advocacy. The workshop also included White House representatives, including Brian Deese, Senior Advisor to President Obama, who reiterated the President's passion for precision medicine. Workshop discussions directly informed Working Group considerations on specific strategies and best practices for recruiting and engaging participants and communities in the PMI-CP.

On July 27-28, 2015, the Working Group held its fourth and final workshop, *Mobile and Personal Technologies in Precision Medicine,* at the Intel Corporation in Santa Clara, California.[38] This workshop focused on the scientific, methodological, and practical considerations to inform the incorporation of mobile and personal technologies in the PMI-CP. The workshop featured panelists and speakers from academic and industry organization that have expertise in utilizing mobile health technologies in a wide range of research activities, including participant engagement, data collection and return of information. Workshop discussions directly informed Working Group considerations on incorporation of sensor and mobile technology data, their data standards, and the types of studies they enable.

Requests for Information

NIH published *Request for Information: NIH Precision Medicine Cohort* (NOT-OD-15-096) on April 20, 2015, seeking public input on the characteristics, purpose, and overall aspects in the development and implementation of a national research cohort, as well as information about existing or new research entities that might be combined to form the cohort.

There were a total of 152 respondents to the RFI. Comments were received from a variety of stakeholder groups who provided a number of suggestions, including making the cohort diverse and inclusive, collecting a wide range of biological and environmental samples from cohort participants, and ensuring that participants provide informed consent with the ability to be recontacted. The RFI also identified a number of research entities that stated their ability to and interest in joining the cohort, including healthcare systems and networks, universities, research consortia, and biorepositories registries.[39]

NIH published *Request for Information: NIH Precision Medicine Cohort - Strategies to Address Community Engagement and Health Disparities* (NOT-OD-15-107)on May 29, 2015, seeking public input on effective community engagement strategies to advance the ability of the PMI-CP to address health disparities, factors and incentives to participate, and strategies to address participation barriers for participants from underrepresented populations.

There were a total of 69 respondents to the RFI. Comments were received from a variety of stakeholders, including researchers, patient advocates, and the general public, who provided a number of suggestions on how to effectively engage underserved communities, including reaching out to trusted community members/organizations for support, ensuring trust in participating communities by involving them in discussions about research questions for the cohort, and ensuring that logistical details are addressed (e.g., covering transportation costs, providing childcare during research visits) to encourage participation.[40]

E. Conclusions

Informed by stakeholder input, consultations with experts in specific topic areas, and its closed session deliberations, the Working Group developed a collective vision for the development, implementation, and oversight of a national PMI-CP. The Working Group first considered the scientific opportunities associated with such a national PMI cohort. These scientific opportunities (Section 2) formed the

foundation for its further deliberations and recommendations with respect to cohort assembly (Section 3), participant engagement (Section 4), data (Section 5), biobanking (Section 6), policies (Section 7), and governance (Section 8). The specific Working Group recommendations are detailed within these sections and summarized in the Executive Summary.

Section 2 – Utility and Unique Opportunities of the PMI Cohort

The path to advancing human health has been built upon a detailed understanding of the specific factors that contribute to both health and disease. Longitudinal cohort studies have provided many critical insights into the pathogenesis of human diseases and the existence of genetic and environmental risk factors. However, to date, owing to the cost of enrolling participants and collecting relevant clinical data, the vast majority of longitudinal cohort studies of general populations have typically been of modest size, limiting analytic power, or have targeted specific classes of disease, biasing aspects of what can be learned.

Over the past few years, advances including the proliferation of EHRs, inexpensive technologies for DNA sequencing and metabolite profiling, and the ubiquity of mobile, wearable, and home-based devices that can be used for capturing health information and environmental exposures, have converged to create an unprecedented opportunity to assemble a cohort vastly larger and with far more comprehensive data than was heretofore imaginable. The PMI-CP, by enrolling and studying one million or more participants in the U.S., will comprise an accessible resource for researchers and participants to work in partnership to accelerate our understanding of health and disease. Thousands of researchers will be able to explore the underpinnings of disease, while participants will have the opportunity to use their health data in innovative ways for their own betterment.

It is anticipated that insights stemming from the PMI-CP will broadly inform opportunities for disease prevention and treatment. Below, we describe examples of specific questions that can be addressed by the PMI cohort. These use cases are not intended to be exhaustive, but reflect a range of important questions that the Working Group anticipates the PMI cohort can address. Critically, the clinical outcomes and richness of the collected data will grow with time, providing continuing opportunities for advancing knowledge. Moreover, as science and technologies advance, the PMI cohort will provide an ongoing venue for testing new hypotheses, as well as for determining whether results from smaller cohorts generalize to the broader U.S. population. Further, the PMI cohort will provide a great case study as well as a "leading wedge" for exploring how EHR interoperability, mHealth technologies, a nimble regulatory framework for genomic technologies, and facilitating individuals' access to their own health data can be transformative forces for biomedical research.

A. Utility of the PMI cohort

The goal of the President's PMI is to enable a new era of medicine through research, technology, and policies that empower patients, researchers, and providers to work together toward development of individualized care. This includes the discovery of environmental, genetic, biochemical and other factors predictive of disease risk, response to therapy, and disease outcomes. These discoveries will favorably affect the health of the American population by identifying new opportunities and optimized strategies for both the prevention and treatment of disease, with the ultimate goal of improving clinical outcomes.

To achieve these goals, the PMI-CP will enroll one million or more volunteers that are inclusive of the diversity of the U.S. population, and will follow their health and clinical outcomes over time. Participants

will be asked to provide permission for recontact for follow-up of clinical or research findings. Baseline and longitudinal clinical data will be collected from EHRs and will be supplemented by biological, environmental, and behavioral data in order to build a more comprehensive set of personal health data. These data will be stored in a secure computing environment under the most rigorous standards to protect individual privacy. Because of its scale as well as the breadth and depth of information represented, the PMI cohort dataset will comprise an unprecedented resource for the research community to apply innovative strategies to identify new determinants of health and disease. The PMI cohort, comprising a large cadre of individuals with comprehensive and detailed clinical information, may also provide unique opportunities for preventive and therapeutic trials based on selection of groups of individuals with distinct factors contributing to disease.

The PMI cohort will provide the information needed to address a wide range of scientific questions. Examples of some of the scientific opportunities that we anticipate can be addressed include the following:

1. Developing quantitative estimates of risk for a range of diseases by integrating environmental exposures, genetic factors, and gene-environment interactions. Environmental and genetic risk factors have been found for many rare and common diseases and traits. However, rigorous conclusions about disease etiology and population impact require population-based cohort studies, in which samples of a population are followed prospectively for the development of specified endpoints. Phenotypic, genotypic and environmental information collected in a standardized manner are crucial to such efforts. The PMI cohort will provide a broadly useful resource for rigorously validating and quantifying the contributions of genetic and environmental risk factors, as well as their interactions with one another, in a large, diverse population. This will certainly include risk factors that have been proposed from smaller studies, but the comprehensiveness of the PMI cohort dataset will also allow for data scientists to identify new and unexpected associations. As it grows in breadth and depth, the PMI cohort will allow for these estimates on uncommon as well as common diseases.

2. Identifying the determinants of safety and efficacy for commonly used therapeutics. U.S. retail pharmacies fill more than four billion prescriptions each year;[41] 49% of American adults take at least one medication and 22% take three or more.[42] While these therapeutics may demonstrate net benefit in the populations studied in the rigorous clinical trials that led to their FDA registration, it is recognized that there is wide variation in response for many commonly used drugs, including some individuals who have no response, contributing to unnecessary expense and worse health outcomes. Moreover, other individuals have adverse reactions to specific medications, sometimes with severe clinical outcomes. These result in 4.5 million ambulatory visits annually and thousands of hospitalizations.[43]

Currently, more than 150 FDA-approved drugs include genomic information in their labeling to guide their prescription and use based on observed associations between genotypes and treatment outcomes.[44] By analyzing participant genotypes in relation to prescription patterns and other health care data, the PMI cohort could be used to validate these associations across a larger and more diverse population than the originating studies, while also expanding our knowledge of pharmacogenetic interactions. Additionally, while genetic effects have been identified that predict adverse reactions or

lack of efficacy to some drugs, the predictors of individual response for most commonly used therapeutics are poorly understood. Having a very large data set that includes self-reported and health data-derivable individual responses to therapy with commonly used medications, along with extensive genetic and metabolic measures, will provide the opportunity to comprehensively identify predictors of individual response to therapy. By allowing clinicians to tailor medication and dosage to each individual's profile, these discoveries have the potential to increase the efficacy of prescribed therapeutics, to reduce the incidence of adverse effects, to improve health outcome and to reduce overall healthcare costs.

3. Discovering biomarkers that identify individuals with an increased risk of developing common diseases. We presently lack the ability to identify individuals with a high risk for future development of a wide range of common diseases, thwarting prevention efforts. Recent technological advances provide new opportunities for inexpensive genome and cell-free DNA sequencing for measurement of a wide range of metabolites and signaling molecules, and for measurement of immune system activity. Dense genotyping or genome sequencing can reveal the inherited contribution to traits while other biomarkers can integrate the effects of both inherited and environmental influences. Data analytic methods applied to medical imaging more accurately catalogue risk and development of certain diseases. Furthermore, links between biomarkers and genetic variation can be used to establish causal relationships between specific metabolites and disease pathogenesis. By establishing a large cohort with broad baseline measurements of genomic and metabolic factors along with longitudinal follow-up to monitor for emergence of disease, new biomarkers of future health outcomes can be identified, providing new insights into disease pathogenesis as well as opportunities for disease prevention and/or early therapeutic intervention.

4. Using home and mobile health (mHealth) technologies to correlate body measurements and environmental exposures with health outcomes. Dramatic advances in mHealth technologies currently afford continuous monitoring of activity, behavior, and aspects of cardiovascular function, including heart rate and rhythm. Additional innovative applications with clinical implications, including ambulatory blood pressure and measures of mental health, are being devised at a rapid pace. Similarly, global positioning system (GPS) monitoring can be integrated with environmental pollutant and activity data to provide improved measures of environmental exposures. Because these data may be collected continuously and passively at very low cost, they afford new opportunities for understanding the relationships between these measures and health outcomes. Better measurement of nutrition and physical activity over time may improve our understanding of their relationship to obesity, type II diabetes mellitus, and a wide range of health outcomes. Changes in aspects of activity, sleep behavior, and/or social interactions may provide new insight into development of neurodegenerative and neuropsychiatric diseases as well as response to therapeutics. For example, typing speed has been shown to be predictive of development and treatment efficacy in Parkinson's disease.[45] Similarly, identification of previously healthy participants with sensor-detected paroxysmal atrial fibrillation may identify those at increased risk of embolic stroke.[46]

5. **Determining the clinical impact of loss-of-function mutations.** Recessive loss-of-function mutations in many genes cause specific genetic diseases, often with early and severe phenotypes. We commonly

maintain that the heterozygous state for such mutations is without clinical consequence; however, moderate effects of such mutations have not been well investigated for the vast majority of genes. Most recessive diseases compatible with survival to birth have population incidence greater than 1 in 250,000, implying that more than 1 in 250 participants must have heterozygous loss-of-function mutation in each specific gene associated with a recessive disease. Consequently, a cohort of one million or more participants is expected to include thousands with heterozygous loss-of-function mutations in each such gene. This will allow robust exploration of the phenotypic consequence of the heterozygous state. For example, recessive loss-of-function mutations in nearly 30 genes have been shown to cause microcephaly and severe neurodevelopmental problems due to impaired genesis of cerebrocortical neurons.[47] A study of one million or more participants would be very well powered to determine whether the heterozygous state for these mutations predisposes to early onset Alzheimer's disease, impaired cognition, or other neurodevelopmental/psychiatric abnormalities. Similarly, the PMI cohort will include many healthy individuals, affording the opportunity to identify loss-of-function mutations in the heterozygous, or even the homozygous state that impart desirable effects on health. A recent example of this principle is the discovery of loss-of-function mutations in the gene *PCSK9*; individuals with loss of function mutations have very low LDL cholesterol levels and greatly reduced risk of myocardial infarction.[48] Such protective mutations can identify new therapeutic targets for disease prevention or management. The ability to look broadly across the deep phenotypes measured in the PMI cohort will provide an excellent opportunity to identify both anticipated and unanticipated impacts of loss-of-function mutations. Lastly, the ability to recontact and perform further studies of individuals whose genotype predicts a specific Mendelian disease but who remain unaffected may afford an opportunity to identify environmental, behavioral or genetic factors that mitigate disease severity.

6. **Developing new disease classifications and relationships.** The fundamental basis of rational therapeutics is accurate diagnosis. Current classifications of disease typically group symptoms, signs, and laboratory results into a discrete diagnostic category. Underlying these structures is a disease nomenclature anchored in centuries of observation prior to the current era of molecular characterization. A large and complex set of data points from one million or more participants, including comprehensive clinical records, a broad range of laboratory and molecular investigations, and clinical diagnoses and health outcomes, provides the opportunity to discover unexpected connections within the data as well as new subtypes of disease. Dense molecular and/or clinical information has recently been used to identify new clinical subtypes of cancers,[49,50] autism[51] and heart disease,[52] as well as unanticipated links between genes mutated in autism and congenital heart disease.[53,54] In other disciplines, analytic techniques such as machine learning have enabled the development of robust prediction tools from complex data sets without *a priori* assumptions. The PMI cohort offers the possibility of extending these advances to the field of human health. Psychiatric illnesses may be particularly well positioned for the development of a new disease taxonomy, and the PMI cohort could accelerate that opportunity.

7. **Empowering participants with data to improve their own health.** PMI cohort participants will have access to their own data, including data from structured EHRs, laboratory tests, molecular investigations, and results from real time wearable sensors and from a range of applications on their

phones, tablets, and laptops. For example, participants with asthma could wear sensors to detect environmental pollutants and use mHealth apps to record their ease of breathing, leading to both new knowledge and the potential to take action to modify their exposure. This information may promote healthier behavior, reduce unnecessary testing, and improve medication compliance. The PMI cohort can provide a platform for formal investigations of whether and how data are used by participants and ways to promote information use by individuals.

8. Enrolling PMI cohort participants in clinical trials of targeted therapies. With a very large, well-characterized and motivated cohort of participants with a wide range of disorders and underlying biomarkers, PMI cohort participants may be very well-suited to academic or industry-led clinical trials targeting participants with disease subsets characterized by particular biomarkers. Such targeted clinical trials can potentially be well powered with small sample sizes due to the expectation of higher response rates in such biomarker-driven subsets. Examples of this principle include selective responses of melanoma and lung cancer patients with activating mutations in *BRAF* and *EGFR* to agents targeting these mutant proteins,[55,56] respectively, and selective response of patients with cystic fibrosis to therapeutics targeting specific disease-causing mutations in *CFTR*.[57,58] Execution of such trials is often challenging, owing to the difficulty of enrolling enough participants with desired biomarkers; however, the size of the PMI cohort, coupled with the ability to recontact participants to invite their participation in further clinical studies, provides a significant opportunity to promote the timely and cost effective execution of such studies.

These expected scientific opportunities will be realized over different timeframes as biological, clinical, and environmental data are generated and collected. We anticipate both near-term and long-term benefits arising from the PMI cohort (see Table 2.1 below). Moreover, discoveries of biomarker-defined disease subsets in the cohort may lead to targeted clinical trials of cohort members selected or stratified for such biomarkers, allowing for more rapid development of new therapeutics.

Table 2.1: Timeline when expected PMI cohort capabilities will be realized. The estimated timeline for focused research for each type of investigation is indicated by the number of "+" characters in each cell.					
		Time in years			
		0-2	3-5	5-10	>10
Cohort Capabilities	1. Discovery of disease risk factors	+	+++	+++	++++
	2. Pharmacogenomics	+	+++	+++	+++
	3. Discovery of disease biomarkers	+	++	+++	+++
	4. mHealth connections with disease outcomes		+	++	++++
	5. Impact of loss-of-function mutations		+	+++	+++
	6. New classifications of diseases		+	+++	++++
	7. Empowering participants	+++	+++	+++	+++
	8. Clinical trials of targeted therapies		+	+++	+++

B. Unique and critical attributes of the PMI cohort

The challenges of comprehensive medical record collection, the expense of genomic and other molecular and imaging studies, difficulty in enrolling participants, and the absence of robust mHealth capabilities have, to date, resulted in cohort studies that are typically orders of magnitude smaller in size and lacking in the rich detail envisioned for the PMI cohort. As a result, it is expected that the PMI cohort will be a transformative platform for the understanding of health and disease.

While a number of large cohorts are being developed for related purposes in other countries[59–61] and even as the PMI-CP will seek opportunities for international collaboration, the PMI cohort has several important attributes that make it of fundamental importance specifically to the health of the U.S.

The proposed size of the PMI cohort – one million or more participants – makes it potentially the largest national or international cohort, with substantial statistical power to allow robust conclusions. A cohort of this size will capture a wide range of diseases and will be sufficient to detect genetic, behavioral and environmental effects and interactions that are difficult or impossible to identify in smaller cohorts (more detail on cohort size can be found in Section 3). Moreover, the comprehensive collection of electronic health data along with diverse genomic, biomarker, imaging, mHealth, environmental and self-report data will provide an extremely rich source of data from which to identify new factors contributing to health and disease (details of data collection can be found in Section 5).

The population of the U.S. is remarkably diverse. The PMI cohort is designed to be statistically powered to find factors contributing to differences in health and disease among major demographic groups in the U.S., including participants of diverse age, sex, as well as diverse racial/ethnic, socioeconomic, and

geographic groups. This will be critical to ensure that the benefits of PMI cohort research will be applicable to the broad population of the U.S.

The collection of large sets of data of orthogonal types in members of the cohort, along with longitudinal follow-up, is designed to ensure that PMI cohort will not only be able to support the identification of factors contributing to prevalent diseases, but also biomarkers before diseases are manifested, affording important opportunities for disease prevention. In addition, by consenting participants for recontact, participants can be approached for enrollment in follow-up studies to establish further links between biomarkers, health and disease, as well as clinical trials of therapeutics tailored to participant's specific disease and biomarker profile.

Data in many cohort studies are not broadly or readily accessible by the research community, limiting the ability of investigators with important research questions to analyze cohort data. The PMI-CP is designed to allow access to cohort data to the broad research community, including citizen scientists, with the fewest impediments possible while maintaining appropriate participant privacy. It is expected that the results of such studies will be published and results used to further enrich the PMI cohort data set, ensuring that participants and the public have the opportunity to receive health benefit from discoveries in the cohort, and increasing the utility of the PMI cohort data set.

Finally, the PMI-CP will emphasize engaged participants and open, responsible data sharing with researchers and study participants with strong privacy and security protections. NIH will work closely with other participating federal agencies and the White House to ensure that the PMI-CP adheres to the privacy and trust principles recently developed by an interagency task force to apply broadly to PMI and which are currently undergoing revisions following a public comment period (closed on August 8, 2015).[62] Participants and their representatives will play an integral role in the PMI-CP's governance by working with other stakeholder groups to oversee cohort design, all aspects of data collection, use, management, security, dissemination and access. Transparency regarding data access and use will be emphasized, with return of information to participants, including aggregate data and return of participant's personal data as desired.

Section 3 – Assembling the PMI Cohort of One Million or More Volunteers

A. Rationale for one million or more volunteers

Recommendation 3.1: The PMI cohort should assemble one million or more individuals who agree to share their longitudinal health data and biospecimens for research and to be recontactable.

In order to address the wide range of capabilities and use cases set forth in Section 2, the Working Group recommends the PMI-CP seek to achieve a cohort size of one million or more individuals, with an infrastructure that could scale to allow for many more participants as described in greater detail in Sections 5 and 6. A central research goal of the PMI-CP is to evaluate the impact of biology and environment on phenotypes. Thus, a key requirement of participation in the PMI cohort is that individuals share biospecimens. Indeed, as described below, the Working Group recommends that this requirement be met at the time of enrollment, given that prospective collection is essential to identify risk factors for development of future clinical outcomes. The Working Group anticipates follow-up for at least 10 years to allow accrual of new incident disease as well as exposure to environmental factors and therapies.

Critically, each participant will share a range of health data. A substantial source of health data for the PMI cohort should derive from EHRs but will also include self-report data and the ability to collect data from a number of other sources, such as sensors and mobile devices. EHR data provide detailed current and past medical data, as well as a means for passive, cost-effective follow-up after enrollment.

Finally, the ability to recontact individuals is also a requirement to achieve several of the goals expressed in Section 2, including the potential to deeply explore phenotypes of subsets of individuals based on genotypes or other characteristics, to return information to participants if they desire, and to explore the possibility of targeted clinical trials. Thus, participants must share identifying information and contact methods. Nonetheless, willingness to be recontacted does not imply agreement or obligation to participate in any future study.

Statistical Power Considerations for the PMI-CP

The PMI cohort should be designed to support diverse use cases, as outlined in Section 2. In its deliberations, the Working Group considered different sampling strategies that have been adopted in other cohorts to guide design of the PMI cohort. For example, the UK Biobank enrolled individuals aged 40-69 years to achieve a cohort size of at least 500,000 individuals.[59] Other cohorts, such as MVP and the Geisinger MyCode initiative, recruited from their general populations not restricted based on age.

To provide guidance on different recruitment strategies, we estimated prevalent and incident disease frequencies in a million Americans from existing EHR data from a large EHR-associated biobank that accrues from among any consenting patients in the health system.[A] In this setting, 261, 136, and 59 prevalent diseases exceeded 5,000, 10,000, and 20,000 cases, respectively (Table 3.1). Similarly, over a period of 5-10 years, we will expect to exceed 20,000 incident cases for many common diseases with significant mortality and morbidity, including both adult and pediatric diseases (Table 3.2). These estimates are sensitive to age at, and means of, ascertainment.

The Working Group considered the statistical power to detect the association of an exposure (genotype or environmental) with health outcomes in the PMI cohort using a nested unmatched case-control design. Such power calculations have been well considered and described in detail by the UK Biobank.[66] A main effect (genetic or environmental) or a gene-environment interaction term, along with binary exposure variables (genetic and/or environmental), were considered; power calculations were based on simulations. Unconditional logistic regression was used to test association of binary exposure variables (genetic and/or environmental) with case status. These calculations assumed that interaction terms reflect departures from additivity on the log-odds scale (i.e., departures from a multiplicative model) and that each nested case-control study contains four unmatched controls for each case.

Table 3.3 details statistical power for the binary main effect (genetic or environmental) analysis, reporting the minimum detectable odds ratio (MDOR) that can be detected with 80% power given N cases and M unmatched controls. MDORs are stratified by the number of cases available for study in the nested case-control design, prevalence of the exposure, and three alpha thresholds. For example, if one tests an exposure present in 10% of the PMI cohort (e.g., an allele carried by 10% of participants or a behavior present in 10%) for association with a disease seen in 25,000 cases and applies a stringent alpha of 10^{-7} (as might be applied for a genome-wide genetic discovery study), the cohort would provide 80% power to detect an odds ratio of at least 1.16.

Table 3.4 details statistical power for a gene-environment interaction term. The interaction odds ratio reflects the magnitude of departure from the odds ratio based solely on a simple multiplicative model using the main effects. So, for example, if the odds ratio associated with the binary genetic determinant in individuals that are unexposed to the "at risk" level of the environmental exposure is 1.6, while the equivalent odds ratio in those that are exposed to that environmental determinant is 2.0, the

---

[A] To identify diseases in the population, we pulled all International Classification of Diseases, 9th edition codes from individuals who consented to be part of the Vanderbilt BioVU DNA biobank (which includes all life stages),[63,64] and mapped these into a disease classification system with over 1,600 manually-curated hierarchical disease groups using previously-validated methods.[65] Incident diseases were estimated as new diseases occurring after 2003 (designated as "time zero" to provide for 10 years of follow-up) in individuals clinical data for at least 1 year prior to this. Prevalent data were extrapolated to 1 million individuals. These data were then compared to data from other sites in eMERGE, whose available data contained individuals selected for particular diseases. These eMERGE estimates, given the selection of older and sicker individuals, provided higher disease rates than the BioVU estimates.

interaction odds ratio would be 2.0÷1.6=1.25. For example, if one tests the interaction between an allele carried by 10% and an environmental exposure present in 10% of PMI for association with a disease seen in 25,000 cases and applies a stringent alpha of 10^{-7} (as might be applied for a genome-wide genetic discovery study), the PMI cohort would provide 80% power to detect an interaction odds ratio of at least 1.52.

The above estimates suggest that the PMI cohort will have 80% statistical power to detect scenarios where a binary exposure seen in 10% of the sample increases odds by 35% or 15% for a health outcome observed in 5,000 or 20,000 cases, respectively (at a genome-wide level of significance). The PMI cohort will be similarly powered to detect gene x environment interactions that have odds ratios of 2.3 or 1.5 for a health outcome observed in 5,000 or 20,000 cases, respectively (at a genome-wide level of significance). All else being equal, statistical power will typically be higher if the exposure variable is continuous.

From these considerations, the Working Group concludes that with one million or more participants, the PMI cohort will be well powered to detect relationships between exposures and hundreds of health outcomes.

A. Vision for the PMI cohort sample

Recommendation 3.2: The PMI cohort should broadly reflect the diversity of the U.S.

The U.S. is a diverse society. Americans comprise diverse social, racial/ethnic, and ancestral populations living in a variety of geographies, social environments, and economic circumstances. The U.S. population includes people with extreme wealth and people living in abject poverty and with varying access to social, educational, and other resources. As a nation that historically includes indigenous people, people who originally migrated or were brought to this country from different parts of the world, and more recent and new migrants from numerous countries, the U.S. benefits from the richness of ancestral diversity. The PMI cohort is intended to be a national resource that will, over time, benefit the entire U.S. population. It will leverage our diversity to provide new insights to factors determining health and disease, as well as prevention and therapeutic strategies. Critically, the PMI cohort will be designed to ensure that people historically underrepresented in biomedical research are included in sufficient numbers to allow robust inferences in these groups. The Working Group gave considerable thought to the steps that must be taken to ensure the cohort becomes such a national resource.

Table 3.1: Estimated number of prevalent diseases in a population of one million. Each number represents the number of unique diseases expected at each threshold.

Participants	**Case count threshold:**			
	>2,500	**>5,000**	**>10,000†**	**>20,000‡**
Diseases among participants, any age	420	261	136	59
Diseases among participants, aged 40-69	413	267	151	68

†Threshold for study of genetics on particular phenotypes
‡Threshold for phenotype×gene×environment interaction (e.g., pharmacogenomics) studies.

Table 3.2: Estimated disease incidences and prevalences in one million people.			
Disease	**Expected prevalent cases**	**Incident cases**	
		5 years	**10 years**
Myocardial infarction	39,273	14,981	27,112
Thrombosis	26,746	11,559	21,169
Type 2 Diabetes	135,658	40,411	123,196
COPD	48,728	15,396	33,584
Stroke	16,016	8,969	15,598
Melanoma	6,109	2,727	3,873
Breast cancer (female)	20,470	12,068	21,382
Lung cancer	11,432	2,866	4,828
Prostate cancer	13,861	6,241	13,848
Colorectal cancer	9,407	3,745	6,844
Rheumatoid arthritis	45,835	11,466	23,875
Dementia	13,373	7,028	9,656
Parkinson's disease	4,311	2,127	4,032
Abdominal aortic aneurysms	8,729	2,451	5,518
Atrial fibrillation	78,272	35,047	56,292
Congestive heart failure	73,723	21,315	40,322
Cystic fibrosis	3,799	231	849
Asthma	62,149	17,292	44,036
Type 1 diabetes	40,047	8,600	26,315
Lupus	14,659	3,283	6,738
Eating disorders	3,775	1,665	2,971
Epilepsy	33,426	4,161	11,248
Attention deficit hyperactivity disorder	13,039	7,213	13,582

Table 3.3: **Minimum detectable odds ratios associated with 80% statistical power for main effects by exposure prevalence and alpha threshold. Based on nested case-control study design with 4 controls analyzed for every case.**[66]

Exposure prevalence	Critical P-value	Minimum detectable OR for main effect (4 controls per case)			
		2,500 cases	5,000 cases	10,000 cases	20,000 cases
0.5	0.01	1.16	1.11	1.08	1.06
0.5	10^{-4}	1.23	1.16	1.11	1.08
0.5	10^{-7}	1.32	1.22	1.15	1.10
0.25	0.01	1.19	1.13	1.09	1.06
0.25	10^{-4}	1.28	1.19	1.13	1.09
0.25	10^{-7}	1.37	1.25	1.17	1.12
0.1	0.01	**1.28**	**1.19**	**1.13**	**1.09**
0.1	10^{-4}	1.39	**1.26**	**1.18**	**1.12**
0.1	10^{-7}	1.54	1.36	**1.24**	**1.16**
0.05	0.01	1.39	1.26	1.18	1.12
0.05	10^{-4}	1.59	1.39	1.26	1.80
0.05	10^{-7}	1.80	1.51	1.34	1.23
0.01	0.01	1.99	1.63	1.41	1.28
0.01	10^{-4}	2.50	1.91	1.58	1.38
0.01	10^{-7}	3.16	2.26	1.78	1.51

Table 3.4: **Minimum detectable odds ratios associated with 80% statistical power for gene-environment interaction effects by exposure prevalence and alpha threshold**.[66]

Genotype prevalence	Environmental prevalence	Critical P-value	Minimum detectable OR for interaction effect (4 controls per case)			
			2,500 cases	5,000 cases	10,000 cases	20,000 cases
0.5	0.5	0.01	1.37	1.25	1.17	1.12
0.5	0.5	10^{-4}	1.54	1.36	1.24	1.16
0.5	0.5	10^{-7}	1.80	1.51	1.34	1.23
0.25	0.25	0.01	1.46	1.31	1.21	1.14
0.25	0.25	10^{-4}	1.69	1.45	1.30	1.20
0.25	0.25	10^{-7}	1.96	1.61	1.40	1.27
0.1	0.1	0.01	2.07	**1.67**	**1.44**	**1.29**
0.1	0.1	10^{-4}	2.62	**1.98**	**1.62**	**1.41**
0.1	0.1	10^{-7}	3.28	2.31	**1.81**	**1.52**
0.05	0.05	0.01	3.42	2.39	1.85	1.54
0.05	0.05	10^{-4}	5.02	3.13	2.24	1.77
0.05	0.05	10^{-7}	7.24	4.05	2.69	2.01
0.05	0.5	0.01	1.88	1.56	1.37	1.25
0.05	0.5	10^{-4}	2.34	1.82	1.53	1.35
0.05	0.5	10^{-7}	2.89	2.12	1.70	1.46
0.5	0.05	0.01	1.88	1.56	1.37	1.25
0.5	0.05	10^{-4}	2.34	1.82	1.53	1.35
0.5	0.05	10^{-7}	2.89	2.12	1.70	1.46

The Working Group considered the impact of population biases in prior research, especially when planning for the PMI-CP's capacity to serve the broad American population with meaningful findings. Ninety-six percent of individuals included in recent genome-wide association studies are of European descent.[67] Racial and ethnic diversity do not just reflect genetic ancestry; they are social constructs rooted in cultural identity and shaped by historic and current events, which influence an individual's behavior, place of residence, and life opportunities. Because genetic variation correlates with self-identified race,[68] a racially/ethnically diverse cohort will provide unprecedented opportunities to examine the complex relationship of ancestral influences, environmental exposures, and social factors. In turn, understanding the interaction between the social and environmental milieu with an individual's biologic profile and genetic ancestry can extend our understanding of disease pathology and extend the application of best practices in therapeutics to more diverse groups.

Health disparities, defined as significant differences in health between populations that are more or less socially advantaged or disadvantaged, persist across the U.S. population.[69] Examples include differential disease prevalence, unequal access to treatment, and variable response to therapy. For example, up to 75% of Pacific Islanders are unable to convert the antiplatelet drug clopidogrel into its active form and are at higher risk for adverse outcomes following angioplasty.[70–72] Moreover, healthcare providers and scientists are informed by research extrapolated from a largely homogeneous population, usually white, male, urban, and of higher socio-economic status.[73–75] In the U.S. alone, African-Americans represent 12% of the nation's population but only 5% of clinical trial participants; Hispanics make up 16% of the population but only 1% of clinical trial participants. It is consequently unclear whether the biology of complex traits and the response to therapy, largely defined in the white male population, is directly applicable to others. This has the potential to maintain or aggravate health disparities.

Health disparities are prevalent in many chronic diseases, such as asthma and diabetes. Chronic diseases are major sources of morbidity and mortality among disadvantaged populations across the U.S. These health disparities have multiple determinants, including socioeconomic status, reduced access to health care and quality health care providers, health literacy, cultural beliefs, individual choices, social and family situations, legal and structural constraints, sex and gender inequality, and racial/ethnic discrimination. Many diseases of the respiratory system, for example, are linked in their causation and exacerbation to environmental exposures that disproportionately affect disadvantaged populations: in the U.S., racial/ethnic minority and socioeconomically disadvantaged children are disproportionately exposed to higher levels of air pollution[76–78] and are more likely to have serious asthma complications compared with white children.[79,80] These disparities have economic consequences. For example, eliminating health disparities would have reduced total medical costs during 2003-2006 by more than $1.2 trillion.[81] While health disparities may be attributable, in part, to socioeconomic status, discrimination based on race/ethnicity remains a major determinant of economic opportunities for minorities.[82] Thus, in order to be a resource that will truly benefit the entire U.S., the Working Group recommends that the PMI-CP leverage America's rich diversity, thereby increasing scientific rigor that accounts for individual variation while providing opportunities to advance research that may reduce disparities and move towards health equity.

As a prospective cohort, the PMI-CP should not focus on any particular set of diseases or health status for potential recruitment, but should include the broad range of health and disease experienced by Americans. Inclusion of healthy individuals promotes efforts to identify new risk factors predictive of future incident disease. Inclusion of those with disease allows systematic study of pharmacogenetics and disease outcomes across a wide range of diseases and drugs. Including individuals diverse in age, sex, race/ethnicity, socioeconomics, and geographic location provides new opportunities to understand the factors that contribute to health and disease as well as response to therapies.

The Working Group acknowledges that its proposed accrual methodology would lead to a number of very rare diseases that would not be represented in adequate numbers in the PMI cohort to allow for robust study. According to the Rare Diseases Act of 2002, rare diseases are defined as those affecting <200,000 individuals in the U.S.[83] There are currently more than 6,000 known rare diseases.[84] Given the size and diversity of needs in studying rare diseases (which have very diverse phenotypes, clinical assessments, treatments, and outcomes), specific disease repositories and recruitment efforts independent of the PMI-CP are likely to be the ideal method to study rare diseases.

In addition, the Working Group recommends that the PMI-CP include individuals from all life stages. The Working Group gave special consideration to including children, decisionally impaired adults, and participants who may become incarcerated after enrollment. There are scientific, ethical and policies issues surrounding these populations that warrant further discussion. Therefore, the Working Group recommends that NIH consider the safeguards necessary to ensure the appropriate enrollment, retention, and protection of these groups into the PMI cohort.

Although the process needed to ensure sufficient inclusion of diverse participants could increase the cost and complexity of the PMI-CP, the long-term scientific and health benefits can far outweigh short-term expenses. Modest and mindful changes toward recruitment efforts may involve using culturally competent study designs, providing access to health care resources, community involvement and participation of community leaders, and inclusion of multilingual and multicultural investigators and recruiters.

B. Strategy for assembling the PMI cohort

Recommendation 3.3: All individuals living in the U.S. should be able to volunteer for the PMI cohort.

Recommendation 3.4: The PMI-CP should adopt two distinct recruitment approaches that leverage the strengths of healthcare provider organizations with existing relationships with individuals, as well as the enthusiasm of individuals who wish to volunteer directly.

Recommendation 3.5: Healthcare provider organizations participating in the PMI cohort must be able to consent volunteers to the PMI cohort, to share participants' electronic health record data, and to collect new biospecimens. Healthcare provider organizations that capture more comprehensive health care data that have a record of longitudinal follow-up of participants, that serve the diversity goals of the PMI cohort, and that can contribute significantly to the size of the PMI cohort should be preferred.

The PMI-CP presents a unique challenge to develop a research cohort that is reflective of the breadth of the U.S. population in all its diversity and builds statistical power over time, while concomitantly having

impact in the nearer term. Of key importance among the Working Group considerations were: the ability to collect comprehensive health data; the ability to collect, store, and transfer biospecimens; the effectiveness of mechanisms to consent and recontact participants; the ability to be open to all volunteers; the embrace of diversity; and the speed, ease, and cost of recruitment. The Working Group discussed the desired data elements for the PMI cohort, which are described in detail in Section 5.

Among these are important abilities to collect self-report data (at baseline and periodically as driven by scientific opportunities), to link with mobile-technology and wearable sensor data, and to collect environmental exposures via location data, derived from a variety of sources. Each of these items requires an engaged population that can be recontacted.

The Working Group believes that one of the best modalities to collect prevalent and ongoing health data is available EHR and health insurance data. EHRs include a robust longitudinal record of diagnoses, medications, procedures, medication response, analyte measurements, demographics, some exposures, and some social information. The rapid adoption of EHR systems, especially since adoption of the Health Information Technology for Economic and Clinical Health (HITECH) Act, has led to U.S. hospital implementation rates of 95%.[26] While much work remains to be done to optimize EHRs for clinical care and research use, existing publications support the efficiency and efficacy of EHR data from healthcare provider organization (HPO)-associated individuals for use in genomic, pharmacogenomics, and clinical research.[85–88] Moreover, sites with EHR-associated biospecimens have combined into networks and effectively shared phenotype-defining algorithms to study common and rare disease and drug response.[89–93] Studies comparing these EHR-derived studies to non-EHR derived genetic results have often demonstrated similar results.[65,94,95] Since a participant may have multiple health care providers, combining EHR data with insurance claims data increases the comprehensiveness of disease, procedure, and outpatient prescription medications records. Claims data have also been a powerful tool for clinical research.[96,97] Given the dramatic growth in EHR use, their demonstrated value for research in a variety of research models, and the promise of increasing the portability and interoperability of EHRs, the Working Group considered the availability of EHRs for participants to be essential to the success of the PMI-CP as a cost-efficient method to provide for collection of health record data and health outcome follow-up.

In addition, and as discussed previously, providing biospecimens is an essential requirement for participants. Thus, the ability to collect and transfer biospecimens is a requirement in the key considerations for the assembly of resources and approaches to the PMI cohort.

In its discussions of the PMI-CP's role in consent, recruitment, and contact, the Working Group considered an arc of possibilities. At one end of the spectrum, the PMI-CP could merely be an aggregator of extant cohorts, and the original cohort could be the arbiter of communications and uses of the participant data. In this model, the participant's consent and contact would primarily be through the existing cohort or entity participating in the PMI-CP. Such a model would limit the ability of the PMI cohort to serve as a national resource accessible by many researchers, and would hinder standardization of participant communication and engagement, sample collection, and data collection. At the other end of the spectrum is a model in which existing research cohorts or organizations serve primarily as a

means to identify potential participants for the resource, and the PMI-CP would directly recruit participants and contact would be primarily through the PMI-CP. The Working Group feels that the PMI cohort design should fall somewhere in between these two ends of the spectrum: it should collaborate with entities with existing relationships with individuals to identify and consent volunteers for the PMI cohort, while requiring that those individuals be recontactable by the PMI-CP for future research. Individuals would be newly consented specifically for PMI-CP, but the relationship could still preserve the existing entity's relationship with the participant for communication (Section 4), healthcare data gathering (Section 5), and biospecimen collection (Section 6).

The Working Group considered the challenges and opportunities presented by directly recruiting and enrolling individuals into the PMI-CP, thereby ensuring that the PMI cohort is open to all volunteers. A consistent theme throughout the four PMI workshops (see Section 1) was the enthusiasm of individuals to engage with the PMI-CP and the commitment on the part of NIH and the White House for this resource to be available to everyone. Thus, another cornerstone for the PMI cohort design is the requirement that any individual living in the U.S. be able to volunteer for the PMI cohort. This could be one way to enhance diversity, another cornerstone of cohort design. While it will not be possible to address every question of importance in every population, the PMI-CP, under direction of the Steering Committee, should seek to oversample diverse populations to increase the statistical power to make robust inferences within each group.

Taken together, the Working Group emphasizes the critical importance of recruiting underrepresented populations, which places a high priority on working productively with health care providers that can engage significant numbers of such individuals in a timely and cost-effective fashion. Balancing the participant characteristics sought in the design of the PMI cohort with the cost, scalability, and speed of recruitment, as well as the desirability of accepting all volunteers for the PMI cohort, will be critical to the success of the PMI-CP

Approaches considered by the Working Group for assembling the PMI cohort

One approach to assembling a PMI cohort of one million or more volunteers is to leverage existing cohorts or entities with key relationships with potential participants and resources to facilitate recruitment. The Working Group considered the many types of organizations that could be leveraged for assembling the PMI cohort, including healthcare provider organizations (HPOs), non-provider research cohorts, patient and health advocacy groups, and community groups.

HPOs are characterized by patients who receive care over time from an institution, thus, producing a longitudinal record of care increasingly available in an electronic format with on-going, documented follow-up. Examples of HPOs include academic medical centers, Federally Qualified Health Centers (FQHCs), vertically integrated private health care organizations (e.g., Kaiser Permanente), and vertically integrated governmental organizations (e.g., VA).

Non-provider research cohorts include any collection of research participants that have engaged in clinical studies for a given disease, exposure, or outcomes but are not a characterized by central coordination by a common healthcare system with primary access to EHR data (e.g., Framingham Heart

Study, Multi-ethnic Study of Atherosclerosis, etc.). They can have high-quality research data but those data are usually focused on particular diseases or outcomes. Moreover, data are typically actively collected, leading to a high cost per participant and more difficulty scaling to a broad range of outcomes and comprehensive record of health.

Patient and health advocacy groups include organizations that provide support, education, and advocacy for patients and families of a broad range of healthcare conditions. Their involvement in research typically involves raising public awareness of and funding for a particular disease or condition. Many groups are also taking an increasingly engaged role in research activities, including: establishing disease-specific biobanks and patient registries; working with researchers or research institutions to design; develop, and disseminate research; directly funding research; and, in some cases, operating their own research networks.

Community groups frequently include religious organizations and other affinity groups. These organizations typically do not have healthcare information on individuals but have historically played important roles in engagement and recruitment with certain clinical research studies.

The Working Group recognizes the long and important role non-provider research cohorts, patient and health advocacy groups, and community groups have played in clinical and biomedical research. These organizations often have long-standing and trusting relationships with communities and individual participants. However, they are unlikely to have primary, continuous access to a wide variety of health data and may lack the ability to readily collect and transfer biospecimens. While some research cohorts do receive EHR data from local providers, systems that do not have primary access to EHR data are considered less desirable by the Working Group, since secondary EHR data access may be disrupted by a number of external influences. The Working Group feels that non-provider research cohorts, patient and health advocacy groups, and community organizations will play an important role in promoting the recruitment of individuals as PMI cohort participants as either direct volunteers (those with no affiliation to an PMI HPO, described below) or through their involvement in HPOs (both of which are discussed in more detail below). Leveraging the energy and member relationships of these non-HPOs to encourage their members to become involved in the PMI cohort would form an important part of a recruitment strategy, especially with more diverse communities. The Working Group feels that the PMI-CP encourages continued involvement from patient advocacy groups after recruitment for participant engagement (see Section 4). However, these organizations would not have ongoing relationships with participants in the context of their involvement in the PMI-CP.

In addition, the Working Group discussed future generations of the PMI cohort and envisions the development of application programming interfaces and data standards that would allow integration of diverse research data collected on limited sets of participants (such as members of patient and health advocacy groups and non-provider research cohorts) to coordinate with the PMI-CP (see Section 5A and Table 5.1).

Of all the types of organizations that could be leveraged, the Working Group determined that recruiting participants from HPOs is likely to represent the fastest and most cost efficient approach to enrolling

large numbers of participants with existing dense healthcare data, and with an expectation of easily acquiring follow-up health data. The Working Group concludes that HPOs with primary access to and the ability to share ongoing, comprehensive health data have exceedingly strong potential to be highly effective partners with the PMI-CP, functioning as sites or "nodes" within the PMI cohort for recruitment, communication, biospecimen collection, and healthcare data collection (through their clinical care relationship). The Working Group expects that the majority of the participants will be recruited from HPOs. Diversity and inclusion of historically understudied populations should be considerations in selecting HPO cohorts for inclusion in the PMI cohort, as described above. The HPOs may continue to be the primary contact with the volunteers they recruit from within their patient population. Recruitment of the individuals, collection of biospecimens, and regularization and transmittal of participant EHR data represent significant, primary efforts to build a valuable PMI cohort. These organizations will need funding to support these efforts.

One of the principles of the President's announcement of the PMI is the ability for all Americans to volunteer for the PMI cohort. The PMI cohort is intended to be a broad, public resource to advance the health of all Americans. Limiting participation to HPOs would prevent individuals living in many parts of the country from enrolling in the cohort, since the number of organizations contributing to the PMI cohort is unlikely to cover the U.S. uniformly. Indeed, even in geographic areas in which a PMI-participating HPO is present, many individuals may not seek care from that or any other organization.[98] Thus, to ensure broad availability of the PMI cohort to Americans, to maximize recruitment for the cohort, and to capitalize on the extraordinary energy around individuals to volunteer for this effort, the Working Group recommends also allowing any individual to volunteer. For the purposes of this report, these individuals are referred to as direct volunteers. Together, direct volunteers and HPOs will form the PMI-CP's two distinct approaches to recruitment. The recommended eligibility requirements for both are provided below.

By having volunteers join that do not have existing affiliations with HPOs, the Working Group acknowledges that such direct volunteers will be in control of their data flow and responsible for sharing data with the PMI-CP. For these individuals, the PMI-CP is the primary "membership group."

Recommendations for eligible Direct Volunteers

The Working Group discussed qualifications for individuals who will participate through direct recruitment. First, the individual should be living in the U.S. or a U.S. Territory and should expect to remain in the U.S. or U.S. Territory for as long as they are enrolled in the PMI cohort. Accordingly, individuals must be able to visit a U.S. health care provider. Scientifically, individuals would not necessarily have to be a U.S. citizen or permanent resident, and as such enforcing strict verification of a participant's citizenship or residence was not considered a priority for the PMI-CP.

Second, the Working Group recommends that direct volunteers provide a certain amount of core data to meet the overall goals and use cases for the PMI cohort (see also Sections 2 and 5). For instance, the Working Group expects a baseline of phenotype assessment and data quality for direct volunteers. The

Working Group recommends that direct volunteers meet the following requirements to be enrolled in the PMI cohort:

- They must be recontactable and provide identifying information (e.g., email, phone number, address, and a close contact) to allow recontact by the PMI-CP.
- They must provide a biospecimen (see Section 6).
- If they have EHR data, they must agree to share it with the PMI-CP (see Section 5A for methods of linking EHR data). EHR data is not required, however, to be a PMI direct volunteer.
- They must see a healthcare provider for an initial screening exam, whose details are specified in Section 5. If they do not have a provider, the Working Group recommends that the PMI-CP establish a process to allow the participant to receive a standardized exam through a provider that could be reimbursed by their health insurance (if they have it), or be evaluated by a provider affiliated with the PMI-CP. A PMI provider network could be formed by contracting with fast-access convenience provider networks (e.g., MinuteClinic, The Little Clinic, CareSpot). The protocol here would be to collect biospecimens and perform an initial history and physical exam. However, a challenge to using convenience providers is that most currently do not have the capacity to draw peripheral blood samples as would be required for the collection of biospecimens. An alternative that was considered was the use of blood banks to collect biospecimens and healthcare data; however, they typically do not have providers on site to conduct a PMI baseline health exam.

The enrollment of direct volunteers should take place through a focused recruitment site(s)/network that are separate from existing HPOs participating in the PMI cohort given the unique challenges of recruiting and enrolling these individuals. Direct volunteers could be recruited through a number of technologies, such as internet, social media, and mobile technologies. See Section 4 for Participant Engagement recommendations.

Recommendations for eligible HPOs

The Working Group anticipates that a wide range of HPOs may be responsive to the PMI cohort requirements. Such HPOs may include healthcare delivery systems, academic medical centers, and FQHCs, among others. The Working Group recommends the following requirements for HPOs enrolling PMI cohort participants:

- All participants must be consented for the PMI cohort and be recontactable, sharing identifying information as for direct volunteers (see above).
- The HPO must agree to share each participant's core data (see Section 5B) and collect and send biospecimens (see Section 6) at or near the time of enrollment to the PMI cohort for use by PMI investigators, and to allow the PMI-CP to manage the primary process of approving access to and use of these biospecimens.
    - Organizations may also maintain their own parallel biospecimen collections (which would not be subject to PMI use), but these do not replace or substitute for the PMI cohort collections.

- The HPO must have an EHR system that meets at least Meaningful Use Stage 2 standards, unless the HPO is an FQHC or similar entity (such as Rural Health Clinics), in which case the HPO must have adopted, implemented or upgraded to a certified EHR. In either case, the HPO should be able to share the core data elements defined in Section 5B from their EHR to the PMI cohort.
- The HPO must perform a PMI baseline health exam, ensuring collection of certain baseline measures and history data (see Section 5B).

Each of these requirements must be met for each participant in order for them to be considered enrolled in the PMI cohort. Because of these requirements, the Working Group recommends that applicants should provide evidence of capabilities in several key areas:

- Anticipated number of participants that can be recruited and the rate of accrual. The Working Group notes that data harmonization issues will be easier when fewer entities each recruit a large number of participants (e.g., ten HPOs each recruiting 100,000 participants or more). Therefore, larger cohorts should be preferred, though HPOs with access to specific populations needed to meet diversity goals may be required.
- Population demographics.
- Evidence of EHR robustness, comprehensiveness, and its length of use. Specific questions include but are not limited to whether they have availability of inpatient and outpatient data, electronic prescribing and medication fill data, laboratory, and electronic clinical documentation.
    - Availability of all of these EHR elements are not required (e.g., an outpatient-only environment would be acceptable), and other factors, such as the diversity of the sample included, length of anticipated follow-up, or anticipated recruitment speed, may balance specific weaknesses in these areas.
    - Familiarity with reusing EHR data for research and informatics expertise would be beneficial, as data will have to be queried and standardized for transmittal to the PMI cohort (see Section 5).
- Ability to follow participants longitudinally for health outcomes using an EHR, with minimal loss to follow-up. HPOs should be evaluated by their ability to follow patients over time and should detail expected participant dropout rates due to leaving the HPO. However, the Working Group recognizes that recruiting understudied groups may present challenges owing to higher losses to follow-up in some cases, and potentially less robust EHR data among providers in some settings, such as FQHCs.
- Existing procedures, if any, in place to collect biospecimens on individuals.

The Working Group discussed the importance of availability of passive follow-up data for individuals coming from HPOs. It is true that in any healthcare system, individuals may leave the HPO for a variety of reasons, including moving, changes of insurance and/or employment, etc. In addition, some systems will provide only inpatient or outpatient care, or may represent one of many HPOs in the region. Such data fragmentation can play a role in limiting the ability to accurately determine phenotypes and clinical outcomes. The Working Group recognizes the importance of determining follow-up, and recommends

that for patients coming from more fragmented healthcare delivery systems or potential for inconsistent follow-up, attempts should be made to identify those subsets most likely to receive longitudinal care at a given institution. However, the Working Group recommends that follow-up considerations should be relaxed for diverse populations (such as in FQHCs).

The Working Group recognizes that integration of the EHR data across contributing HPOs is nontrivial; these issues are discussed in more detail in Section 5.

Facilitating research for HPO-associated participants

Recommendation 3.6: The PMI -CP should share PMI cohort-generated research data with participating healthcare provider organizations that are providing ongoing data and biospecimens to the PMI cohort, according to participant preferences.

A key factor to the success of building the PMI cohort will be effective partnership with HPOs. Thus, the Working Group favors a partnership model in which biologic and phenotype data generated by subsequent research in the PMI-CP could be shared back to the HPO. These data could enrich local HPO research cohorts by integrating data across different healthcare systems, through new data generated from biological testing (e.g., genotyping), surveys, mobile technologies, and linkage with national data sets (e.g., vital records) that may not be found in the local HPO's data warehouses. Another benefit is that it could provide an opportunity to shorten the time between start of an intervention (prevention or treatment) and documentation of its effectiveness by either individual participant feedback on their treatments through PMI cohort interfaces or through specific PMI cohort research studies.

PMI cohort participants should be in control over whether to share such data back to their HPOs. The Working Group agrees the default position should be that data generated from the PMI cohort should be sent back to HPOs unless the participants opt not to share such data. Participants should have easy control and access to their data to restrict data sharing options.

The balance of use cases and data generation methods, as well as the overall goals of the PMI cohort as a research enterprise, and the lack of existing robust data transfer protocols for diverse data types into clinical systems, leads the Working Group to favor that PMI cohort-generated data should be treated as research (and not clinical) data. Thus, HPOs would not be expected to integrate such data into their clinical record nor surveil for clinically actionable information generated as a product of PMI cohort research.

Indeed, participants may be less likely to share certain types of information about themselves if they believe these data might be shared back to their physicians, and there is no current EHR workflow and procedures that could anticipate the diversity and actionability of data that could be generated by the PMI-CP. However, participants would be able to share their data themselves with their HPOs, if they so desire, outside of the context of the PMI cohort. It will be important that it be made clear to participants that their data remain in the PMI cohort unless they transmit these results (e.g., specific genetic test results) back to their physician.

A related topic is that participants should be able to use the PMI cohort as a vehicle to aggregate and view their clinical data as well as access their individual research data that are developed in accordance with the standards specified in the Clinical Laboratory Improvement Amendments (CLIA). This topic, as well as related topics of breadth of and procedures to access participant data, is discussed more in Sections 4 and 5.

What happens if an HPO leaves the PMI cohort?

It is possible for a variety of reasons that HPOs could cease recruiting individuals into the PMI cohort and stop providing data in a proactive way to the Coordinating Center (discussed below). If a HPO leaves the PMI-CP, the participant, who has consented to be part of the PMI cohort, will remain a part of the PMI cohort. That data, which have already been shared by the HPO centrally, can remain within the Coordinating Center. Future data sharing, contact, engagement, biospecimen collection can all be handled through the Coordinating Center as if the participant was recruited as a direct volunteer. New data from the participant would accrue via direct methods, including by the participant exercising their rights under the Health Insurance Portability and Accountability Act (HIPAA) of 1996 to access and share their health data (see Section 5).

Timeline for recruitment

Recommendation 3.7: The PMI-CP should seek to recruit individuals rapidly once necessary infrastructures (both physical and computational) are in place to an initial goal of one million or more participants, and continue to accrue participants throughout the lifespan of the PMI-CP.

Given the Working Group's recommendation that recruitment should occur as rapidly as possible, it will be important that the site(s) that will be accepting data and individuals via direct volunteer recruitment have the necessary software, data storage, and biobank specimen protocols (see Sections 5 and 6) in place before recruitment, and have the capacity to handle a potentially large number of individuals contributing data quickly. Recruitment of individuals from HPOs may occur at a more predictable rate, driven in part by the number of sites involved. Sites should be equipped to engage in novel and broad consent mechanisms using electronic means to facilitate recruitment (see Section 4).

Review of several HPOs with active recruitment programs reveal that they commonly recruit ~20,000 individuals per year; however, two larger HPO-based cohorts (Kaiser Permanente and MVP) have recruited at a pace of 85,000 individuals per year (see Table 3.5). A common theme is an efficient consent process and biospecimen collection during clinical visits, though some sites also used asynchronous means of contacting participants outside of clinic visits using mail or phone calls, and then collecting biospecimens when they are in the clinic. Collections of biospecimens would likely require participant visits, which can be a rate-limiting step. If 10 sites each enrolled an average of 25,000 participants per year, the PMI cohort would reach one million participants in four years. The Working Group, however, advocates for as large a cohort as possible. Thus, recruitment should continue throughout the lifespan of the PMI-CP, beyond one million participants. Indeed, many of the infrastructure components could scale with larger populations, yielding greater efficiency as the size of the PMI cohort grows.

Table 3.5: Select existing biobanks with healthcare provider data. Participants in all biobanks listed below are recontactable.				
Biobank	**HPO system size**	**Current Biobank Size**	**Recruitment method**	**Time to achieve size (during active enrollment)**
Million Veteran Project	6 million	400,000	Mailed veterans info about MVP and enrolled at visit	4 years in 54 sites
Kaiser Permanente	10.1 million	245,000 (goal of 500,000)	Mailed consent and mailed saliva sample (N = 189,500); electronic or in-person consent and blood samples (N= 50,000)	3.5 years using direct mail to 2 million
Partners Healthcare Biobank	6 million	>30,000	In-person at outpatient visits and inpatient floors; Electronic consent via emails using patient portal	5 years since launch: 2 year pilot study; 3 years via in person recruitment; eConsent for past 1 year; current rate is 1100/month
Geisinger MyCode	1.3 million with an EHR encounter in last 10 years	>86,000	In-person during routine outpatient visit; Electronic consenting pending	10 years; however, current rate is 1000/ week
Marshfield Clinic Personalized Medicine Research Program	>2 million	20,000	In-person, recruited via phone and mailers	16 months at 4 sites of Marshfield clinic
Mayo Clinic	2 million	>60,000	In-person consent at clinic	7 years, current rate about 8000-9000/year
Children's Hospital of Philadelphia	2.5 million	110,000	In-person consent at clinics	9 years
Cincinnati Children's Hospital Medical Center	670 thousand	>56,000	Hospital-wide consent by registrars at registration	4 years

C. The organization of PMI cohort entities and communication between them

Recommendation 3.8: The PMI-CP should have a Coordinating Center that serves as a hub for accessing data and biospecimens and participant communication and engagement.

Recommendation 3.9: Novel collaborations of traditional and nontraditional NIH-funded organizations should be pursued to achieve state-of-the-art analysis methods, scientific rigor, elastic storage and compute capabilities, and technological expertise.

Recommendation 3.10: The PMI-CP should create a specific organizational infrastructure to recruit and assess direct volunteers.

In evaluating an organizational structure for the PMI-CP, the Working Group brought to bear knowledge and experience gained from a number of existing large networks. Examples include PCORnet, FDA Sentinel, the Electronic Medical Records and Genomics (eMERGE) Network, Multicenter Perioperative Outcomes Group (MPOG), and Health Care Systems Research Network. Typically, these networks follow a hub-and-spoke organization with a central coordinating center along with a number of data-producing sites with enrolled participants. In most NIH networks, these operate in a federated model such that each site contains the majority of the data and shares as necessary. However, to facilitate PMI-CP operation and the speed to which it can respond to requests, the Working Group recommends a hybrid model that centralizes core data elements across all sites and all participants in a Coordinating Center. These recommendations are discussed in Section 5D.

An important aspect of the PMI-CP is that there will need to be easy flow of data to the Coordinating Center and between any sites that may need to share data functionally on a given participant. By forming agreements between the Coordinating Center and the data containing nodes instead of between all data-generating or consuming nodes directly, the PMI-CP can facilitate data sharing, standardization, and communication. It is important that a single data sharing policy be formed that allows free data exchange according to participant preferences across sites.

A Coordinating Center will be needed to host and store PMI direct volunteer data, while also providing interfaces and technical support to facilitate transfer of their EHR data and entry of other data. Publicity for the PMI direct recruitment and coordination activities could be facilitated via the Coordinating Center or specific grants (i.e., PMI direct volunteer nodes) directed at participant recruitment.

Finally, the PMI cohort will generate very large amounts of data in many novel formats (discussed more in Section 5A). The data should be widely accessible and support the cutting-edge compute and scalable storage technologies that may not be found within traditional non-profit awardee organizations. Novel collaborations with for-profit and non-traditional technology companies skilled in securing large-scale cloud technologies would benefit PMI-CP capabilities and use. However, the Working Group also recognizes that expertise in biological knowledge, clinical informatics, analysis methods, and scientific investigation typically resides in academic centers. Thus, collaborations should be encouraged.

Patient engagement and communication functions are discussed in Section 4 and 5; data storage, methods for collecting, and processing in Section 5; and biospecimen collection and provenance in Section 6.

Section 4 – Engagement

Participant engagement and empowerment are core values for the PMI-CP. Whereas the majority of clinical research has been transactional in nature, with unidirectional data sharing from the individual to the study, the PMI-CP seeks true partnership between participants and researchers. The Working Group feels that the PMI-CP should strive to be an exemplar in this consent for the research in which they participate, and to receive results and information from the research conducted in the PMI cohort. From their deliberations, the Working Group developed a number of recommendations relating to participant partnerships, building and retaining trust, communication, consent, and return of results. The recommendations are focused on developing partnerships that unite the expertise of researchers and clinicians with the goals, energy, and perspective of the participants to create a thriving PMI cohort.

A. Participants as partners

Recommendation 4.1: Research participants and their advocates should be central partners in the governance, design, conduct, oversight, dissemination, and evaluation activities of the PMI-CP.

The Foundation for the National Institutes of Health (FNIH) recently completed a public opinion survey about participation in a large precision medicine cohort study. Of the 2,601 individuals sampled in the survey (oversampling for underrepresented populations), 79% of respondents reported that a precision medicine cohort study should be done, and over half of the respondents would either "definitely or probably participate in a study if asked." Opinions on the study and willingness to participate increased with education level and decreased with age; however, no significant differences were observed between racial/ethnic groups for these questions. In addition, 71% of respondents agreed that participants and researcher should be equal partners in such a precision medicine cohort study, and over half of the survey respondents reported that cohort participants should be involved in multiple aspects of the cohort study development, including selecting the kind of research questions that should be answered in the cohort, and deciding how study results should be disseminated.[99]

The FNIH survey results are consistent with other studies in terms of the public's interest in participation in research that advances biomedical discovery and the high value placed on receiving information back from the studies in which they participate.[100,101] Consistent with public opinion and the values of the PMI-CP, participant input and feedback should be sought and considered throughout the design, implementation, and oversight of the PMI cohort. Participants and their advocates should be significantly represented throughout the PMI-CP governance structure. By making participant perspectives essential to all facets of the PMI-CP, the Working Group seeks to ensure that the PMI-CP is a truly collaborative effort between scientists and participants.

B. Building and retaining trust

Recommendation 4.2: The PMI-CP must prioritize building and maintaining trust with participants and communities by operationalizing the best approaches for participant engagement and scientific integrity.

Researchers and the communities they serve must have strong, collaborative relationships. The expectation for such relationships as a foundation to biomedical research has markedly increased, as has the recognition of the value of these relationships.[102,103] The PMI-CP should identify and prioritize effective engagement strategies and operational principles for engendering public trust, maximizing the potential benefits of a large national research cohort, and minimizing the risks inherent in large-scale data collection, analysis, and sharing. The Privacy and Trust Principles developed by the White House as part of the PMI should provide a "north star," guiding the PMI cohort's implementation and providing the expectations and requirements of stakeholders engaged in the PMI-CP (see Section 7).

Among the opportunities to effectively build and maintain trust, it will be critical that educational efforts and general safeguards are undertaken to limit the likelihood of investigators unintentionally breaching trust. These goals form the basis for recommendations regarding data access that are presented in Section 5. Trust also requires clarity and realistic expectations related to anticipated scientific outcomes and the timeframe in which they might be expected. The PMI-CP should avoid over promising and under delivering in the eye of participants.

C. Engaging participants in the PMI-CP

Recommendation 4.3: Engagement and communications with participants should be managed through a single entity that organizes communications with participants across recruitment sites and the Coordinating Center.

A goal of the PMI cohort is to empower individuals to understand potential opportunities to manage their health offered through genomic sequencing, aggregation of longitudinal health information, and sharing of data with researchers, under a cooperative model of partnership and trust. A consistent challenge faced by science and medicine, however, is effectively communicating with the public about how research advances benefit health and the important role individuals play in furthering progress by participating in research.[104–106] But the paradigm is shifting: clinician, researcher, patient, and caregiver communities are increasingly engaged in active discussions about ways to build shared responsibility for health knowledge, empowering individuals, families, and communities to make the best decisions about their health and well-being.[107,108] The Working Group expects the PMI-CP to exemplify engagement at its best.

To accomplish this, the Working Group believes that a single entity should lead the development, dissemination, and implementation of communication and engagement activities across the PMI-CP. The entity should effectively engage a variety of skilled experts to mobilize grassroots and community engagement, develop social marketing and branding, and promote volunteerism. A coordinated approach will ensure consistent messaging, returning information, and approaches to engagement across the wide variety of sites enrolling participants into a single PMI cohort. This approach, however, should also provide the latitude to leverage the existing successes in relationships and engagement enjoyed by HPOs with their participants in the PMI cohort. The route of access or recruitment into the PMI cohort should not change the participant's expectations of partnership. To ensure this expectation is met, evaluation and continuous quality improvement should be intrinsic to the PMI-CP engagement and communication activities. Indeed, the Working Group anticipates that the PMI cohort will become a

critical test bed for advancing research on participant engagement and the biomedical communication with the lay public.

The Working Group urges the PMI-CP to promote creativity in how it approaches engagement and communication. For example, the PMI-CP should identify ways to leverage grassroots movements and cohort "champions" (including health advocacy, community engagement, and community organizing experts). Involving such stakeholders in developing an outreach plan would shift the paradigm of research engagement away from a disease focus to a focus on all members of the public. Support for educational activities related to personal genetics and health data management will also be critical, particularly for models or programs designed to reach diverse and underserved populations. This commitment may entail funding or partnering with entities that are not part of the traditional research ecosystem; room should be made for such entities as well as for new and traditional funding opportunities. In addition to conducting research on engagement among the entities funded by the PMI-CP, NIH should consider, for example, providing funding for research on community education and engagement within the scope of the PMI-CP.

Early in the process of implementation, the PMI-CP should seek to raise potential participants' awareness of the benefits of generating, aggregating, and sharing data: many may not realize the challenges of data held behind the walls of hospitals, offices, health centers, and industry, and the critical role they can play by sharing that information for research. The PMI-CP can help communicate the importance of large, high quality datasets in understanding disease, and can potentially inspire individuals to want to become involved. The PMI-CP will benefit from working with a number of organizations to generate awareness of the PMI cohort and its objectives. Academic medical centers, community health systems, private and government organizations, health advocacy groups, and community organizations, such as churches or other faith organizations or racial/ethnic identity groups, should be encouraged to play critical roles in disseminating information about the PMI cohort and the importance of joining this effort. It will also be important to keep participants informed of the goals, structure, and governance of the PMI-CP. Progress toward each goal must be clearly and consistently communicated in an effort to demonstrate the value of the data provided by participants.

D. Communicating with participants

Recommendation 4.4: The PMI-CP's approach to communication with potential participants, enrolled participants, and the public should utilize multiple technologies, including internet, telephone, and mobile-based communications. Information should be provided at the point of initial engagement and periodically thereafter in culturally appropriate manners and through languages reflective of a diverse cohort.

Information should be provided to potential participants, enrolled participants, and the public in culturally appropriate, responsible, and predictable manners. A dynamic information sharing process should be utilized to enable participants to actively engage in an informed and voluntary manner, and to modify their own preferences in a simple and straightforward way as data sharing, use requirements, and technology evolve.

In order to attract the broadest range of potential participants, the Working Group recommends that the PMI-CP use a range of media and existing organizations to raise awareness about the study. In order to reach the broadest audience, it is important that a carefully constructed marketing approach be created that includes the use of social media, direct mail, and other outreach approaches. Ambassadors and advocates for the PMI-CP should extend beyond the recruiting sites to include unaffiliated academic medical centers, community health systems, health advocacy groups, community organizers, or other organizations such as churches or social groups.

Since the PMI-CP hopes to engage participants for many years, a dynamic information sharing system should be utilized that enables participants to actively engage in an informed, voluntary, and ongoing manner. One aspect of this should be a centralized, bidirectional participant portal available to all participants regardless of what means are used to enroll participants. The portal should be designed simply, using appropriate languages, and should facilitate and motivate each person to actively participate in the PMI cohort. The centralized portal should be consumer friendly and clearly state why participant data are needed and how our understanding of health and disease will benefit from the use of their data. The portal should use online tools to assist the participants when they have questions or concerns, and should also enable participants to provide requested data when it is convenient for them. PMI-CP governance should monitor portal metrics, and use these and other data to update the portal design based on utilization and effectiveness. Again, partner organizations, especially those already working in underserved areas, can play a critical role, and the PMI-CP should work closely with such groups.

The PMI-CP's centralized portal must actively deliver results of studies and updates of the status of the project to participants in order to keep them informed and engaged, while allowing participants to opt out of such communication if they wish. Participants should be reminded of the value they provide through PMI-CP-generated emails and texts. They will feel part of the PMI cohort family if they are receiving correspondences recognizing goals met, anniversaries to joining, or birthdays. They will appreciate simple acknowledgments for taking the time to build on their dataset and complete surveys. Of equal importance are communications for the PMI cohort community as a whole. As new ideas and discoveries are generated that support better health and wellness, it is important that this information be communicated back to the participants who helped make the observations possible.

Finally, the PMI-CP must recognize diversity in its communications with participants. The PMI-CP should have a help desk within the Coordinating Center that can provide support via phone, text and mobile devices. The help desk must be staffed with appropriately trained individuals who can address specific language needs and obstacles participants may be facing.

The PMI-CP should recognize diversity in the amount of information a participant may want. Some participants may want frequent updates and others less frequent. In general, and according to individual preferences, information should be communicated to participants clearly and regularly indicating:

- How, when, and what information and specimens will be collected and stored;
- Details of specific studies and options about receiving research results;

- The types of studies for which the individual's data and specimens may be used now and in the future;
- The goals, potential benefits, and risks of participation; and
- The privacy, security measures, and governance systems that are in place to protect the participant's data and specimens.

Both active and passive modes of delivering such information to participants should be pursued, based on user preferences.

E. Consenting participants

Recommendation 4.5: A standardized and centralized electronic consent protocol should be used across all PMI cohort participants to ensure consistency, minimize organizational burden, and maximize participant recruitment.

Recommendation 4.6: Participants should be able to set preferences for invitations to participate in supplementary or complementary studies that are outside the general PMI-CP protocol.

The Working Group believes that electronic consent across digital platforms will be a core component of PMI cohort recruitment. There should be a general PMI-CP consent for all participants that include the following elements:

- Collection of their identifying information, with the understanding that the PMI-CP will keep it secure but it could be shared for recontact for specific, approved PMI cohort studies;
- Permission to use their identifiers to link across disparate data sources using identifiers (see Section 5);
- Permission to recontact participants;
- Permission to share specimens and data for future research uses; and
- Clarification that data already contributed to research studies or included in existing aggregate data sets cannot be withdrawn if they withdraw from the PMI cohort.

Participants should be provided choices as to whether they would like to be contacted to participate in supplementary or complementary studies to the PMI cohort protocol.

F. Returning results and information

Recommendation 4.7: The PMI-CP should ensure the responsible return of personal results and information to individual participants and sharing of aggregate findings from its investigations with participants so all volunteers may have opportunity to benefit from the science.

Recommendation 4.8: A Return of Results and Information Subcommittee that includes substantial representation from the participant community should be established to oversee the development and implementation of policies related to the return of aggregate and individual results to participants.

Results and information should be made available to participants in a responsible, respectful manner. In the FNIH survey discussed above, 90% of the respondents noted that learning about their health information would be a prime motivation for participation in a precision medicine cohort.[99] However, as discussed in Section 5C, returning results and information responsibly may be complicated. Results may

be not be actionable or may be of unknown significance. Moreover, participant preferences may be heterogeneous when it comes to return of results, ranging from full return of all information to return of minimal information. Therefore, dynamic and responsive preference settings should be made available. For example, a participant may not want to know individual results upon entry into the cohort but a change in health status may motivate the participant to seek additional health information. PMI-CP should allow participants to change settings to receive individual results throughout the duration of the study and access to past findings if they so choose.

Aggregate Results: The Working Group recommends that aggregate results for all studies be made available to all participants at the conclusion of individual investigations, and that individual data from which these results were obtained be archived, together with details of methodology (including required algorithms and/or software), for re-analysis by the research community. In general, results intended for presentation to participants and the public should be presented following a process of scientific peer review and a publication intended for a scientific audience. However, a lay summary of this research should be made available essentially contemporaneously with scientific publication and after a reasonable amount of time if no publication results from the research.

Individual Results: Most individuals want and deserve to know the results of their participation in the PMI cohort. Basic health information such as serum cholesterol and blood pressure should be provided. Genomic data can present more of a challenge, however. The return of unannotated genetic data to participants could prove to be frustrating, and the return of data sufficiently annotated to facilitate use in a medical care environment is governed by a medical practice and legal framework that specify sample collection, handling, and analysis requirements. These requirements, specified under CLIA, tend to significantly increase costs but, importantly, increase the certainty that the results which are presented to a participant represent their results (i.e., they have not been “mixed up” with someone else’s) and that they are analytically valid. In addition, clinical laboratory results may require the use of FDA-approved or cleared medical devices, also increasing cost, and, in the case of clinically actionable genetic test results, may require interpretation by a medical practitioner. Interpretation of genetic test results is not simply a matter of knowing genetic sequence; clinical and family history information is typically used to facilitate clinical interpretation. Because of these complexities, as well as the nature of decisions that may result from having genetic test information, genetic counseling services are typically made available by the clinical laboratories or health care organizations providing genetic testing services. For these reasons, it is important that the PMI-CP properly budget for specimen collection, genetic testing and return of results.

An alternative strategy is clinical retesting of participants whose research results suggest an actionable abnormality. Such a strategy may be acceptable under CLIA, but would require the development of an infrastructure to assure that this type of follow-up is available for all participants in the PMI cohort. The challenges of returning genetic information and policy advances needed to address CLIA are discussed in Section 7 (see Recommendation 7.21).

Return of Results and Information Subcommittee: This subcommittee should explore the following issues related to return of individual results and clinical standards:

- Preference setting and assessment of whether or not an individual wants to see their results;
- Preference setting for return of results to proxy and/or posthumous sharing of participant data;
- Advice to participants about potential implications of return results and legal protections, such as Genetic Information Nondiscrimination Act (GINA) and issues related to life or disability insurance;
- Platform for returning results;
- Approach for determining what types of results and interpretation should be returned to participants;
- Assurance that recontact for future studies does not inadvertently reveal information about one's health that they prefer not to have;
- Whether there is an obligation to assure the availability of disease surveillance and/or treatment;
- PMI-CP-wide policies for return of results that are highly predictive adverse outcomes (e.g., BRCA 1 or BRCA 2 mutations), in ways that respect participant preferences while ensuring that participants receive and understand the implications of results;[109] and
- Assurance that the proper security framework is in place to protect the information that is returned.

Section 5 – Data

As detailed in the Participant Engagement section, the PMI-CP proposes a highly interactive and proactive participation model. Participants will be the primary source of many research observations, co-designers of studies, mediators of access to their healthcare data, contributors to overall data quality control, donators of data from mobile and wearable devices, and recipients of their own as well as aggregate data and analysis results. Enabling this vision will require a combination of well-proven and innovative methods and technologies for data collection and management. Since the data, compute, networking, and storage technologies will change rapidly throughout the study, the PMI-CP will need ongoing technical and innovation expertise to revisit and revise these systems as it progresses.

A. Acquiring research data

Types of data to be acquired

Recommendation 5.1: Guided by the scientific use cases and consideration of value to participants, the PMI-CP should anticipate and collect a diverse set of data types, beginning with a more limited set of high-value variables to be acquired primarily at entry from all PMI cohort participants, but also including a limited set of longitudinal variables. These data will constitute a core PMI dataset that will enable both cohort-wide analyses and identification of subcohorts eligible to participate in specialized studies.

Recommendation 5.2: A Data Subcommittee should be formed to consider and evaluate core data elements.

Accurate data is indispensable for research and the volume and variety of data that can be feasibly acquired will both enable and circumscribe the scientific possibilities for the PMI cohort. The issues related to data, in turn, derive at least initially from the scientific use cases outlined in Section 2 of this report. The Working Group used the presentations and discussions at each of the PMI workshops (outlined in Section 1), as well as relevant literature and the professional experience of its members, to formulate a vision and specific recommendations related to the data that will become the foundation of the PMI cohort research resource. For most (but not all) of the categories of data that the PMI-CP will need to achieve its goals, there is substantial prior experience among Working Group members and generally acknowledged best practices. For novel classes of data, such as those derived from physiologic sensors communicating wirelessly via smartphones, technical methods and data standards are either nonexistent, early in development, or largely proprietary.

Table 5.1 summarizes the categories and likely sources of research data needed to support the scientific opportunities detailed in Section 2. Over the course of its deliberations, the Working Group developed the concept of a "core" data set that would be obtained from all PMI cohort participants and stored in a Coordinating Center using a Common Data Model (CDM). Specific recommendations for the initial core data set are found in Section 5B. These core data would be supplemented as needed by study-specific data queries to participating HPOs and individuals, in a process described in Section 5D. We have also labeled, in Table 5.1, the Working Group's initial assessment of whether each category is likely to be part of the core dataset (labeled C), or part of a subgroup dataset (labeled S) needed only for specific studies

on specific subgroups of the overall PMI cohort. It is anticipated, in general, that the amount and types of data collected on broad sets of participants will grow over time. The anticipated uses listed are meant to be representative only, and not an exhaustive catalog of all possible uses. The Working Group anticipates that creation of a large, recontactable cohort with diverse available information will lead to new avenues of investigation and data types heretofore unimagined.

Table 5.1: Categories, Sources, and Uses of Data

Category	Examples	Source(s)	Example Uses	Core/ Subgroup
Individual demographics and contact information	Date and place of birth, sex and gender, detailed and multiple races/ethnicities (e.g., Asian of Indian descent, Asian of Chinese descent), name, mailing address, phone number, cell phone number, email address, marital status, educational status, occupation/income	Study participant, healthcare provider organizations	Participant-specific communications, analytics, risk stratification, assessment of covariates and confounds, study appointment reminders, invitations to participate in sub-studies	C
Terms of consent and personal preferences for participation in the project	Fine-grained consent for options to participate e.g., receive research results	Study participant	"Precision Participant Engagement"	C
Self-reported measures	Pain scales, disease-specific symptoms, functional capabilities, quality of life and well-being, gender identity, structured family health history	Study participant	Many	C/S
Behavioral and lifestyle measures	Diet, physical activity, alternative therapies, smoking, alcohol, assessment of known risk factors (e.g., guns, Illicit drug use)	Study participant (retrospective and prospective) and healthcare provider organizations	Correlation with clinical events, drug response, and health outcomes	C/S
Sensor-based observations through phones, wearables, home-based devices	Location, activity monitors, cardiac rate and rhythm monitoring, respiratory rate	Smartphone sensors, commercial and research-grade physiologic monitors	Functional ability and impairment assessment	C/S
Structured clinical data derived from Electronic Health Records (EHRs)	ICD/CPT billing codes, clinical lab values, medications, problem lists	Multiple provider organizations per study participant, via institutionally managed channels or direct from	Correlation of clinical events with other categories of data	C

		participant via personal download/upload		
Unstructured and specialized types of clinical data derived from EHRs	Narrative documents, images, EKG and EEG waveform data	Multiple providers, via federated queries rather than inclusion in core dataset	Correlation of clinical events with other categories of data	S
PMI baseline health exam	Vital signs, medication assessment, past medical history	Study participant interacting with healthcare provider organization	Provides baseline measures on all participants	C
Healthcare claims data	Periods of coverage, charges and associated billing codes as received by public and private payers, outpatient pharmacy dispensing (product, dose, amount)	CMS and other federal sources, private insurers, pharmacy benefits management organizations	Assessments requiring complete longitudinal record of exposures/outcomes during specific periods, e.g., within X years of a diagnosis or medication exposure; health services research, exposure and outcomes assessment	C
Research specific observations	Research questionnaires, ecological momentary assessments, performance measures (six minute walk test), disease specific monitors (e.g. glucometers, spirometers)	Study participants, research organizations	Many	S
Biospecimen-derived laboratory data	Genomics, proteomics, metabolites, cell-free DNA, single cell studies, infectious exposures, standard clinical chemistries, histopathology	Study participants, provider organizations, outsourced laboratories	Correlation of tissue findings and high throughput biomolecular data with other categories of data	C
Geospatial and environmental data	Weather, air quality, environmental pollutant levels, food deserts, walkability, population density, climate change	Public and private sources not directly part of PMI	Epidemiology, epidemic surveillance	C/S
Other data	Social networking e.g., Twitter feeds, social contacts from cell phone text and voice, OTC medication purchases	Public and private sources not directly part of PMI	Predictive analytics	S

EHR and imaging data, mHealth and sensor data, and biologic data are discussed in more detail in following sections. The Working Group also noted that there are special considerations for several specific categories of data as listed, and offers its observations on these special cases:

- The Working Group heard considerable interest in considering the impact of the "exposome" on health. There exists a vast amount of available environmental and exposure data that could be linked with geospatial data collected from mobile devices and home/work addresses to enable study of pollution and weather data, as well as socio-environmental phenomena such as food deserts and walkability. In addition, the impact of personal exposures could be measured via mobile technologies and at-home sensors as well. Mobile technologies and sensor data are discussed more below. The ability to comprehensively assess exposure to infectious agents similarly has the potential to provide unanticipated links to specific disease outcomes.
- Collecting and understanding family relationships among related individuals adds significant statistical power in genetic analyses. In cases where more than one family member becomes a PMI cohort participant, the PMI-CP should acquire and maintain explicit linkage between participant records, assuming that each participant agrees to be associated with their relative. The primary relationships desired will be first and second degree relationships; more distant relationships may be hard to ascertain and link prior to genetic studies. The Working Group understands that widely available dense genomic data will enable discovery of biological familial relationships not manually determined, some of which may disagree with relationship data the participant provides. It is generally the position that biologically defined relationships discordant with reported family history should not be returned to participants.
- Gender identity is often suboptimally represented in research databases as well as in EHRs. Constituencies such as the transgender community are particularly affected by lack of clarity and common approaches to coding. Guidance for knowledge representation and database design has recently been published in the Institute of Medicine's "Capturing Social and Behavioral Domains in Electronic Health Records"[110] and the Working Group offers this as a potentially useful reference for data infrastructure design. One example the Working Group reviewed that has captured gender identities well in a structured format included the Genetic Alliance's Platform for Engaging Everyone Responsibly (PEER).[111]

Motivations to contribute data to the PMI cohort

Recommendation 5.3: The PMI-CP should develop and publicize sustainable value propositions for individuals and healthcare provider organizations to participate in the PMI cohort as data providers.

The success of building PMI cohort data resources will depend critically upon the alignment of incentives for individuals and organizations to contribute to the greater community good of a national research resource. The Working Group heard presentations in several of the PMI workshops (see Section 1) that pointed to the value of quid-pro-quo models that leverage the value of return of research results, for both recruitment and long term retention of individual and organizational participants.

Of the categories of data listed in Table 5.1, those derived from EHRs and insurers/payers are arguably associated with the greatest number of challenges. Although data arising as a byproduct of healthcare delivery have the potential to support both discovery science and healthcare improvement, they are encumbered by a set of hurdles related to information technology, policy, organizational motivations, and concerns over data privacy and security that may grow in proportion to the number of organizations involved in their creation, storage, transmission, and analysis. The effective use of routinely collected

electronic 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 upon the ability to address these issues.

EHR data provided by participating HPOs

Recommendation 5.4: The PMI-CP should develop standardized and, to the extent possible, automated mechanisms to acquire clinical data efficiently from participating healthcare provider organizations.

Recommendation 5.5: The PMI-CP should develop and implement a rigorous data curation program for the core datasets to create analysis-ready datasets for a broad range of uses.

Recommendation 5.6: The PMI-CP should periodically implement and revise phenotype algorithms for a number of core diseases and outcomes of interest. As data are cleaned or algorithms are evaluated and implemented, these should be shared back to the Coordinating Center for reuse by others.

Recommendation 5.7: The PMI-CP should develop central resources to support data curation and implementation of algorithms to identify common phenotypes of interest. This process is a bidirectional process involving the Coordinating Center and PMI nodes.

Recommendation 5.8: The PMI-CP should support development and adoption of automated text analytics that can be used both centrally and locally to extract for research purposes the information contained in narrative clinical documents.

Recommendation 5.9: The Coordinating Center should interface with the CMS and other insurers to retrieve medical claims data for integration with participant EHR data.

Routinely collected electronic health data, including both EHR and insurance claims data, are a valuable source of information on real world health events, and include both structured and unstructured components. Healthcare data accrued over time are an important method for ascertaining incident diseases and health outcomes experienced from prevalent disease. The major unstructured component is narrative text as created by providers in documents such as history and physical exams, operative notes, discharge summaries and outpatient clinic visit notes. EHR data have been very useful for a broad range of clinical and genomic research into diseases[86,112] and drug response.[87] The experience of NIH-supported research groups, such as the eMERGE consortium, has demonstrated that identifying specific clinical phenotypes from EHR data require use algorithms incorporating demographic data, diagnostic and procedure codes, lab values, medications, and natural language processing (NLP) of text documents.[113] Administrative and claims data are an essential complement to EHR data, enabling assessment (for insured patients) of essentially all medically attended care without regard to the HPO. Given that integration of medical claims data would be valuable for all participants, it would be most efficiently undertaken through the Coordinating Center. Another potential but yet untested complement to EHR data are data from mobile and wireless sensor technologies which may further refine clinical phenotype algorithms and offer the potential for both earlier diagnosis and recognition of undiagnosed diseased (discussed more below).

A fundamental truth about current EHR systems is that the primary data are created in heterogeneous formats, frequently with institution-specific nonstandard naming and coding conventions. In order to use such clinical data for research, additional steps of data normalization are needed that address both

the syntax (structure) and semantics (common naming and coding practices) of the data to be communicated. The process of data curation – translating and reformatting heterogeneous clinical data to a high quality research resource – essentially always requires knowledge of local conventions and coding practices. The Working Group extensively discussed the pros and cons of doing these quality control, data normalization, and harmonization tasks at the time of initial acquisition of the data (a process named "early binding"), versus a "late binding" approach that applies these curation steps only at the time data is actually needed for a research study. Typically, early-bound curation of data to a common standard would occur at the HPOs before data are shared. There is consensus that demographics data, insurance data (including dates and coverage type), diagnosis and procedure codes, care setting (inpatient, ambulatory, etc.), medication data, vital signs (height/length, blood pressure), and selected laboratory test results are high value structured data that should be curated as acquired to the PMI cohort core data set (see Section 5B).

Clinical laboratory results encompass potentially thousands of different test methods and results, only a few of which any single patient would be expected to have received, so there are mixed opinions as to which laboratory tests might be subjected to early curation, versus those that could be quality controlled only as needed for specific research studies. The general sense of the Working Group is that common clinical laboratory results that contribute to establishing common diseases, such as coronary heart disease, diabetes, hypertension, and liver and kidney disorders, should be included in the core dataset with early binding. However, exhaustive quality control and inclusion of repeated measures of such common tests (e.g., blood glucose measurements in participants with diabetes) would not be useful in the initial core dataset. The Working Group also endorses a quid-pro-quo model of clinical data use: researchers who add value to clinical data received from PMI cohort by quality control and curation steps necessary for their own research, and develop algorithms to identify specific groups of individuals with clinically relevant conditions, should be required to return to the Coordinating Center a copy of their curated data and their algorithms (with an embargo period on secondary publications resulting from these works, also discussed in Section 5D). In this fashion, research users of the resource become co-developers of the resource and invested in its continuous improvement. Public libraries of NIH-supported EHR phenotype selection logic exist, such as PheKB;[114] the PMI-CP should use these resources and extend them. To facilitate research by others, the PMI-CP should consider implementing such algorithms centrally for easy accessibility to users (e.g., via a web-based query tool). Making a core set of algorithms for common diseases, such as diabetes or cardiovascular disease, and health outcomes may accelerate research. For any such available algorithm, however, the Working Group discussed the need to make the algorithm logic and performance characteristics easily accessible to researchers.

Other major classes of EHR data are imaging and waveform data, such as recorded from radiological, hospital-based sensor technologies (e.g., telemetry) and other testing procedures. Radiology images typically follow the Digital Imaging and Communications in Medicine (DICOM) standard and are aggregated in a picture archiving and communication system within a healthcare system. DICOM facilitates portability of a particular image from one location to another, though large-scale transfer of radiology data is untested. Other waveform or imaging data may purport to be in proprietary formats

(e.g., telemetry or echocardiography data). Physician-generated reports made from these images are often stored as narrative text documents within EHR systems and are abundantly available.

Many clinically important observations are present only in narrative text in EHR data. When the initial research use cases of precision medicine are considered, such as identification of detailed disease phenotypes or adverse drug events, NLP methods will be needed to extract content. Many text analysis methods benefit from access to as large a corpus of text as possible. For this reason, centrally implemented NLP methods may be preferred, and there would be economies of scale to having a Coordinating Center with advanced NLP expertise apply those methods to large numbers of clinical documents submitted by collaborating institutions.

However, there are technical and policy challenges to aggregating such text in a centrally managed database. The technical challenges derive from the Working Group's experience that local adaptation of specific NLP methods is required by individual sites due to variabilities in narrative generation, hospital practices, differential uses of *ad hoc* acronyms and abbreviations, and geographic differences in disease patterns. The policy challenges derive from the potential of narrative text to reveal the identity of a research participant inadvertently, and lack of computational methods that can give assurances of anonymity when applied to unstructured data sources. The PMI-CP will depend upon participant informed consent to release all of their data from EHR systems as the antidote to this problem as well as ongoing collaboration from HPOs to address practice questions, data format and definitions, and engage in data cleaning procedures. Organizational willingness to deliver the totality of an individual's medical care record including all clinical documents is untested on a national scale.

Participant-centered technologies for clinical data access

Recommendation 5.10: The PMI-CP should support development and evaluation of tools that enable individuals to acquire, transmit, and continuously update their EHR data to the PMI cohort from multiple provider organizations.

Recommendation 5.11: The PMI-CP should support development of tools for individual participants to review, annotate, and contextualize the clinical data provided by them. Annotations and revisions of clinical data will be for the purposes of PMI cohort and research use, not clinical care.

Individuals often receive care from more than one HPO, and the PMI-CP will invite participation from persons unaffiliated with participating HPOs, in addition to participants in HPOs. 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 health record, share it with their provider, or provide it to researchers or other third parties. "Blue Button" is a term used by ONC[115] and others for such patient online access to healthcare data with download ability and, in some cases, transmittal to a third party application or service of the patient's choice. Over 600 public and private sector organizations have committed to make health information more easily available electronically to individuals and to encourage its use, providing many avenues for patients to access their health data from various sources.[116] Individual access and donation of their health data as a

core method of health data accrual in PMI-CP could make it a catalyst to accelerate progress in technology and data standards among both EHR vendors and federal programs.

EHRs and other types of health IT certified by ONC under its current program, which is aligned with Stage 2 of the CMS EHR Incentive Program, are required to have only a few types of documents: a clinical care summary format (based on the consolidated Health Language 7 [HL7] "Clinical Document Architecture" 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, detailed, or comprehensive for research-quality cohort identification and case selection.

More finely detailed data specifications are in place for Stage 3 of the CMS EHR Incentive Program, which begins phased implementation in 2017, and ONC's 2015 Edition Health IT Certification Criteria. Both proposed rules[117,118] included new requirements for EHR systems to implement application programming interface (API) functionality to better facilitate individual access to and interoperability of health information. The possibility exists for EHR vendor groups, such as the Argonaut Project,[119] to adopt simpler data structures and exchange methods, such as FHIR (Fast Healthcare Information Resource), and industry standard query methods, such as JSON (JavaScript Object Notation), to shorten the time to operational implementation of improved granularity and completeness of individuals' clinical data downloads from ONC-certified EHR systems. 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 EHR. The White House Office of Science and Technology Policy (OSTP) and ONC are positioned to positively encourage the evolution of individually mediated clinical data exchange on a national scale.

The Working Group envisions an adaptation of this download-and-forward capability as a "Sync for Science" (S4S) application and protocol that enables participants to acquire and review their EHR data as maintained by the HPO from which they receive services. Since it will be important to detect and forward clinical data as new medical events occur, a full implementation of the S4S concept will require coordinated action by federal agencies to design and incentivize adoption of EHR technologies that enable individuals to transmit and store their preference for automated data updates to be sent by their providers to the Coordinating Center.

This direct volunteer model of clinical data acquisition for research favors 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, NLP software will need to be developed and adapted to extract and synthesize additional structure from unstructured sources.

One attractive quality control aspect of this process model is that participants could view the data of interest and verify that it indeed belongs to them either prior to or after receipt by a Coordinating

Center. The availability of clinical data to individuals can also serve as a starting point for participants to review and annotate their clinical data as submitted to the PMI cohort. This contextual information can expand the depth of their clinical data and its research utility; for instance, indicating age of onset of a given disease, details of treatment, and any family history of the disease. However, given the complexities and potential legal issues to implementing these data into clinical care, the Working Group recommends that these annotations be part of the PMI cohort research record and not attempted to be pushed back "upstream" into clinical care systems.

Because all of the needed technologies for individual participant-mediated clinical data access are in various stages of development or uptake and not sufficiently mature for immediate implementation, the PMI-CP will benefit by participating in prototype development and demonstration projects to advance the state of the art. These activities should be undertaken so long as they do not delay the PMI-CP's central research activities. The long-term benefits are potentially colossal: a research data infrastructure whose reach extends beyond the PMI-CP and that aspires to serve many millions of research participants proactively engaged in advancing the understanding of human health and disease.

The strengths and weaknesses of the institutionally mediated and direct volunteer mechanisms by which cohort clinical data is acquired, as described above, are summarized in Table 5.2 (below).

Table 5.2: Strengths and weaknesses of individually and organizationally provided healthcare data.

Pathway to acquiring clinical data	Strengths	Weaknesses
HPO mediated	1. Takes advantage of existing data resources and existing staff expertise managing those resources 2. Leverages an existing infrastructure for PMI cohort consent 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	1. Difficult if not impossible to scale to hundreds of participating organizations 2. May impose new burdens on already overloaded local clinical, administrative, and IT staff 3. Although there may be institutional perceptions of risk of data sharing, particularly clinical data, these should be addressed by the participants' consent agreements with the PMI cohort 4. Data use agreements may become more contentious and difficult as more partners added to consortium (particularly competing health systems). These should also be addressed by the participants' consent agreements 5. PMI-CP management expense and complexity for NIH scales proportional to number of consortium partners
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	1. Requires enhancements to current "Blue Button", "View , Download, and Transmit", patient portals, APIs, and similar technologies, and their availability within commercial EHRs acting as data servers

	mHealth platforms and technologies that are being purchased by large numbers of individuals for other purposes 4. Potentially scales at low marginal cost to tens of millions of participants or more 5. Simplifies IRB review as participant directly contributes data (single IRB review of project design and experience is still required) 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 personal health record than any single institutional EHR, due to fragmentation of care delivery from multiple providers 9. If successful, a powerful model for other research efforts that rely upon active participant engagement	2. There is currently no effective approach to curating EHR data from so many sources, both because of the number of sources (potentially tens of thousands), and because there is no originating clinical or technical partner to work with to address data anomalies 3. Requires robust API implementation by EHR vendors and new software development for apps downloadable by participants 4. Becomes a reliable comprehensive data source only after critical mass of EHR data provided 5. May require higher central PMI support and outreach to maintain active data uploads 6. Disintermediates institutions from decisions made regarding data release, which may be viewed as a threat to autonomy and income streams 7. Adding a subcohort that uses a different set of infrastructure and procedures adds cost and complexity relative to a monolithic approach 8. Overall, a higher risk but potentially higher payoff approach relative to institutionally mediated access 9. The cost per participant for acquiring and curating data will be much greater

Because these mechanisms are in many ways complementary, *and because the PMI cohort design is intended to embrace both HPO-based cohorts and direct volunteers, the Working Group believes that the most robust strategy will be to pursue both data aggregation pathways concurrently.*

Genomics and other specialized laboratory data

Recommendation 5.12: Biomolecular data should be acquired and maintained permanently, with attention to including metadata that describes the methods and technologies used to produce it.

Recommendation 5.13: The PMI-CP should generate a core set of biologic data at scale for the PMI cohort, such as specific analytes and broad genomic data, as soon as it is feasible to do so.

High throughput laboratory methods for genomics, proteomics, and other classes of biologically important molecules provide opportunities for both discovery science and predictive analytics. Moreover, the costs of omics assessments are decreasing and will continue to do so over the course of the PMI-CP, making cohort-wide studies increasingly affordable. New laboratory methods are in development continuously, and this phenomenon will continue for the duration of the project, so that "best of breed" analysis technologies available at the start of the project will almost certainly be eclipsed by subsequent methods.

The Working Group notes that, in contrast to standard clinical laboratory methods, essentially all high throughput laboratory methods generate imperfect datasets. For this reason, maintaining metadata about the methods used to generate each set of individual results will be important. As technology marches on, existing PMI cohort data should not be discarded, regardless of whether the technology used to produce it becomes obsolete. Newer instrumentation and methods should be linked to current PMI cohort instrumentation and methods such that outputs are harmonized on the same metric or scale (e.g., mmHg even though current blood pressure instruments use no mercury) to maintain trend analyses as instrumentation precision improves.

The Working Group discussed whether PMI-CP should perform a standard set of laboratory assessments, such as a high-density SNP array or genome-scale sequencing, for all participants. Because of the cost, imperfect results and expectation of technology obsolescence noted above, the Working Group believes that very large investments such as whole genome sequencing for large numbers of individuals need to be keyed to specific scientific use cases that are compellingly cost-effective at the current time. For this reason, the Working Group recommends that the PMI-CP's scientific leadership establish a mechanism to obtain ongoing expert advice on when the costs and capabilities of laboratory methods have achieved a "tipping point" where scientific value justifies whole-cohort sequencing, proteomic, or other omic assessments. It is likely that an early investment in high-density genotyping arrays surveying common and rare variants across the genome will provide some utility at affordable cost (<$50 per sample). Given that unique opportunity to pursue pharmacogenomics study for many disease therapies early in the lifespan of the PMI-CP, these arrays should also prioritize inclusion of known pharmacogenomics variants. Even so, certain benefits arise from processing samples in batches, such as process consistency and the ability to randomize samples by site to lessen the possibility of batch effects leading to spurious results in later analyses. In addition, biospecimens collected at enrollment (see Section 6) can also be subsequently analyzed in subsets of individuals who have developed diseases for discovery of prospective biomarkers.

In cases where there is a reasonable expectation that research results will have clear and immediate utility for healthcare decisions made by participating individuals and organizations, the Working Group recommends that wherever feasible, laboratory data created by the PMI cohort should be acquired and maintained using CLIA-compliant methods and standards. To the extent that the PMI-CP desires to return research data and/or develops interpretive services that assist participants to understand their health status using their data, CLIA-compliant methods will be required. However, the Working Group believes that the data generated by the PMI cohort should be treated as research data by the providing HPOs, due to the considerations discussed in Section 3. Participants can, however, share their PMI cohort-generated data with their providers if they so choose.

Participant-collected survey data

Many of the existing biobanks have leveraged participant-collected structured interview data extensively. The Working Group has intimate knowledge of both the MVP and UK Biobank enrollment protocols. The UK Biobank questionnaire used a touch-screen system that took an average of about 50 minutes for participants' responses. UK Biobank investigators found that it was cost-efficient compared

to a structured interview and that users were more comfortable at times answering sensitive questions in an electronic system than verbally to an interviewer. Diet, traditionally challenging to assess, was queried as a "one-day diet diary" developed for use as a web-based questionnaire. Participants completed it first before they left the assessment center and then multiple times once every three to four months during the subsequent year. The advantage of such a system is that it enables building up a diet diary for each participant that is automatically encoded (by contrast to standard paper diaries) as they are completed. MVP has used similar participant-collected data using the food frequency questionnaire methodology.

Self-reported measures could profit from nesting strategies to acquire more detailed information about health conditions that are either known at time of study entry or develop while in the PMI cohort. Such strategies have been employed by the UK Biobank in its tablet-based enrollment form. For example, when EHR data indicates a new diagnosis of asthma, a tailored query could be sent to that participant to ask more specific relevant history, such as "lifelong" home-address list (for environmental exposures) and "lifelong" employment and occupational exposure list, deeper family history, etc. relating to asthma. These tailored queries could also leverage mobile technologies to administer these self-report questionnaires, not just statically, but dynamically and experientially via random or event-based administration. For example, instead of or in addition to the one-day diet diary, participants could be prompted randomly throughout a week at times and locations in which dietary intake is likely to have recently occurred, and query users about the foods and their amounts consumed at that meal.

Mobile health technology and sensor data

Applications of mobile technology represent a novel modality of data collection that requires more testing, but for which the PMI-CP can make significant contributions to science and health. While mHealth and other sensor technologies are rapidly proliferating in the consumer domain, their formal application and study in health research has occurred only in the last few years. The application of mHealth to facilitate the detail and frequency of data gathering is staggering. Indeed, demonstration projects have suggested mHealth technologies could help ascertain exposures, mood, and activity levels. The clinical relevance of temporally dense data from mobile technologies, however, is largely unknown. Identifying the patterns and context of frequently changing clinical variables, such as heart rate, physical activity, mood, and substance use, may better characterize both clinical endpoints and influences on these clinical endpoints. A cautionary tale in medicine, however, is the observation that not all aberrancies not previously measured should be treated, as doing so may cause unintended adverse consequences, such as in the Cardiac Arrhythmia Suppression Trial (CAST) trial.[120] By collecting dense clinical data articulated with sensor and mHealth data, the PMI-CP will have the ability to link, at a very large scale, mHealth measures to clinical endpoints to measure the utility of mHealth data.

A variety of mHealth apps and devices exist, each with different performance characteristics, upgrade paths, and "structured" output. In many cases, developers support downloading these data. The heterogeneity of devices, protocols, and outputs challenges aggregation and interpretation. The Working Group discussed consolidating toward specific devices used for specific applications, such has been the practice of current research groups heard from at the *Mobile and Personal Technologies in*

Precision Medicine workshop (see Section 1). An alternative approach could be the creation of a "PMI-CP-approved device" program. These devices would have base set of capabilities, capture and report metadata, confidence levels in measurements, and have sufficient security capabilities. Device manufacturers might need to create special device drivers that allow their devices to meet PMI-CP requirements. To ensure data reliability and integrity, the PMI-CP should consider requiring ambulatory/at-home measurement devices to be "validated" – defined as either FDA approved or with comparable evidence of accuracy and precision for those devices and applications that do not require FDA approval. Device characteristics would preferably be documented via publication in a peer-reviewed journal or made publicly available through other means. Given the promising but nascent research on mobile and wireless health monitoring, the PMI cohort has the potential to serve as a test bed for validation of new and emerging technologies.

Given that penetration rates in the U.S. for cell phone use approaches 100%, including extensive use of cell phones in underrepresented and underserved populations, data readily collected from these feature phones should be the minimal data set considered by the PMI-CP for mobile phone collection. Such data include experience sampling (random or event-based self-report items) via text message, location data (via triangulation vs. global positioning system data) and social contact range and frequency (number of calls/texts, unique others called/texted).

Nearly two-thirds of the U.S. population now uses smartphones[29] and, with assistance from mobile phone device makers and network providers, smartphones could potentially be made available to all PMI cohort participants. Via smartphones, the PMI-CP could obtain experience sampling, location, and social contact rates more precisely and readily. In addition, smartphones would provide an array of additional resident sensors that could be used to assess movement and to prompt active data collection by participants of variables such as heart rate, pictures of foods served at a meal, and/or brief motor or performance tests. Smartphones can serve as the conduit for collecting, aggregating, and securely sending data from any peripheral wireless sensors (e.g., wrist accelerometers, glucometers, wireless weight scales) that the PMI-CP might utilize. Most importantly, the smartphone, in addition to mail, email, web sites, etc., can serve as an important connection between the PMI-CP and participants, providing feedback to participants in a timely manner and maintaining ongoing engagement. There is also the opportunity for authorized "software sensors" that collect application data that participants may volunteer to share (e.g., card game playing, mouse clicks, amount of social interaction via phone, etc.).

Beyond the smartphone, the PMI-CP should consider the use and integration of select wireless sensors, either worn or employed in the home, that have been validated and that would provide useful health data for specific subgroups of the PMI cohort. These sensors may include research grade or commercial grade wrist-worn accelerometers, wireless weight scales, movement sensors in the home, continuous heart rate and pulse oxygen monitors, respiration monitors, glucometers, spirometers, and other FDA approved wireless medical monitoring devices used in the home.

Record linkage technologies

Recommendation 5.14: The PMI-CP should employ probabilistic record linkage strategies, including privacy-preserving methods for sensitive information.

To reduce ambiguity, the PMI-CP should assign a research unique identifier (RUID) to each participant. This RUID could be used in data transfer between data providing organizations and in subsequent research using the PMI cohort. However, given the number of data providers anticipated for the PMI cohort (including different HPOs that may share patients, insurers, social network data, and patients themselves), the PMI-CP will also need formal processes of aggregating data from different sources that have not *a priori* linked their data with the PMI cohort RUID. As most data sources will not use a common identifier, the PMI-CP will need to employ record linkage strategies. Experience with health information exchanges and other data networks has shown the need for record linkage strategies that take into account the possibility of differences in identifying fields. For example, last names can change with marital status, first and middle names can be interchanged or abbreviated, and typographical errors can occur within just about any field. The most robust patient matching algorithms incorporate a number of demographic fields (e.g., name, date of birth, gender) and identifying numbers (such as Social Security number) in probabilistic algorithms to identify likely matches.

Given that inclusion of certain identifiers represents sensitive information that individuals may not want to share, the Working Group recommends consideration of privacy preserving linkages for those information that are not strictly needed for the PMI cohort. For example, social security numbers will not be needed, but represent a viable piece of information for record linkage. Thus, Social Security numbers could be shared securely by data providers using a hashing algorithm before transmitting to the Coordinating Center to allow for record linkage on this attribute without storing the actual identifier within the Coordinating Center. In addition, sensitive identifiers such as these can be permuted to several different possible values that are each hashed before transmittal to allow for matching across possible typographical errors.

B. Recommendations for an initial PMI core data set

Recommendation 5.15: The PMI-CP should seek to align its core data set with existing large-scale biobanks where possible. To enable asking of diverse research questions and facilitate infrastructure development, the recommended initial core data set includes data from EHRs, health insurance organizations, participant surveys, PMI baseline health exams, mHealth technologies, and biologic investigations. Data collection should proceed with a phased approach that can grow over time with new types of data in response to scientific opportunities.

Having a central core data set for querying will be essential to enable use of the PMI cohort for research and to allow it to be broadly accessible by a diverse set of researchers. Here we are guided by the significant efforts undertaken already by such cohorts as the MVP and UK Biobank. The Working Group's desire to leverage the prior work of these groups is two-fold. First, careful vetting and research into designing these programs have led to a robust set of measures that have been tested in the field. Second, and perhaps more importantly, aligning with these cohorts would be an important step toward a truly global ability to understand the determinants of human health and disease. For these reasons,

the Working Group advocates that the PMI-CP should try to create an initial dataset that is as comparable as possible (at the individual data variable level) to the baseline survey data collected by the MVP and UK Biobank. Some of the observations (particularly self-report measures) will likely need to be rephrased or otherwise modified to accommodate geographic, racial/ethnic, and cultural differences.

In considering harmonization with existing cohorts, the PMI-CP should also be careful not to allow the PMI cohort to become locked-in to dated forms of data collection. New and potentially more precise and rich forms of data collection (instrumentation and methods) for a specific variable can be co-calibrated and linked to the metrics or scales of the existing datasets. This is particularly important for variables currently assessed via retrospective self-report, some of which now or in the near future can be measured more precisely and objectively via mobile and wireless technologies.

In developing the recommendation for data elements to include in a PMI cohort core data set from healthcare-generated data, the Working Group studied the CDM from PCORnet[97] and the existing algorithms used in the eMERGE network.[93] To date, eMERGE has studied >40 phenotype algorithms, all of which are shared on PheKB. To evaluate the general usefulness of EHR components in phenotyping, the Working Group evaluated 92 total algorithms present on PheKB, which include algorithms from eMERGE, PCORnet, and several other networks and research groups. The most common components used were International Classification of Disease (ICD) codes (used for 70% of the phenotype algorithms), medications (62% of algorithms), text querying/NPL components (46% of algorithms), Current Procedural Terminology (CPT) codes (41% of algorithms), and laboratory tests (38% of algorithms). ICD and CPT codes are part of the PCORnet data model and eMERGE shared data. While ICD and CPT codes are easily shared, medications and laboratory data can require significant normalization efforts to make them interoperable. However, medication data are readily available in highly standardized form from insurers for most ambulatory dispensing and for therapeutics captured by procedure billing codes. Medications can be mapped, through either use of NLP or electronic prescribing tools, to standards such as RxNorm to share between sites. Additionally, networks such as eMERGE, MPOG, and PCORnet have successfully normalized and shared select labs of interest.

Another guiding principle for development of a core data set is to include elements of data from each of the major data categories from which we seek data from the outset: including self-report data, healthcare-related data, mobile/sensor technologies, and collection of biospecimens for batched biologic data (e.g., omic assessment).

Specifically, the Working Group recommends that the following elements be part of the initial PMI core data set:

1. **A set of standard self-report measures via direct patient assessment (all participants)**. The Working Group recommends reviewing and adapting questions from the MVP Baseline Survey and UK Biobank protocol. MVP collected much of its data from the healthcare setting with the addition of a baseline survey, which overlaps significantly with data elements from the UK Biobank. Mobile and web technologies to collect such information are encouraged so that structured data are

derived at the initial collection and so that less stable variables (e.g., substance use) can be more frequently and prospectively assessed.

Self-reports can be used to collect exposure and health data. The Working Group considered diet and substance use as some of the key exposures to ascertain through self-report. The UK Biobank assesses diet through a one-day diet survey that is randomly sent at different times to participants that asks each to recall their diet for the previous day. The Working Group also considered it an important opportunity to use self-report disease and symptom questionnaires to complement use of EHR data. For example, the repeated remote completion of cognitive assessment or mood questionnaires may detect elements before they are clinically apparent to healthcare providers. Self-report health data may be especially important for direct volunteers, which will likely not have as comprehensive EHR data until new data sharing protocols are developed.

2. **A brief, standardized baseline health exam at enrollment for all PMI cohort participants**. To ensure that all participants in the PMI cohort have a core set of baseline measurements, the Working Group recommends that the PMI-CP should specify a set of vital measurements and history and physical exam data that should be collected at enrollment into the PMI cohort. These elements should be ascertained from both HPO-based participants and direct volunteers, the latter collected via their visits with local providers or contracted convenience providers (see Section 3).

3. **Structured clinical data.** The methods to collect these data from HPOs and direct volunteers are described in Section 5A. Expected clinical data for each participant, especially initially, differs based on whether they are from a HPO or a direct volunteer, as detailed below.

   Data expected from HPOs (all participants from these organizations):

   a. All ICD codes with dates.
   b. All CPT codes with dates.
   c. Select, high-value clinical laboratory results in a structured form. Given that many historically are not standardized, current networks such as eMERGE, PCORnet, Sentinel, and MPOG have structured and standardized specific sets of labs. The Working Group recommends a similar approach for the PMI cohort. In addition, each laboratory value must be sent with units of measure and reference ranges. The list of laboratory results shared in the CDM can grow over time as scientific opportunities arise. Retrieving and curating lab data for direct volunteers at the current time may not be possible until more robust transfer protocols, such as S4S, are developed and adopted. Even then, curating such data will be more challenging with HPO-derived data, but this task will become easier over time as EHR data and protocols standardize more.
   d. All available lifetime medication data including start date, and if available, stop date, if inpatient or outpatient, and dose, route, frequency and strength. For inpatient medications, the list should focus on those medications actually administered (where known). For outpatient medications, this list should include all medications prescribed. If pharmacy fill data is available for outpatient medications (likely from health insurance claims data), such

information is desirable. If the individual does not have any medication data, the HPO should provide to the PMI cohort a status of "no current medications" at entrance to the cohort.

e. Vital measurements, including all weights, heights, heart rate, blood pressure, and pain score values. For direct volunteers, the initial assessment should contain a set of measurements and vital signs, such as pulse, blood pressure, weight, height, and waist and hip measurements. Although individuals may be recruited through a number of physical locations, care should be taken to attempt to standardize these assessments (e.g., height without shoes).
f. A record of all encounters (e.g., dates of clinic visits, inpatient visits, ER visits). These records provide important context for vital sign measurements, labs, etc.
g. For health plan data, enrollment and disenrollment dates, and whether the coverage included medical benefits, pharmacy benefits or both.

HPOs would also be required to send the core data elements specified above for the individual to be considered enrolled in the PMI cohort. For certain data classes, data shared would be already available, prevalent data, so that the occurrence of any particular laboratory test or the past history of patient encounters, for instance, may be highly variable in the population. Specific history and exam elements and baseline measures will be collected for all individuals via the PMI cohort baseline health exam. Augmentation of certain tests on a systematic basis could be performed from stored biospecimens in a central laboratory, which would provide both better data quality and cost efficiency. Data would need to be updated regularly, likely on a set schedule at least several times annually.

Clinical Data from Direct Volunteers:
It is expected that most direct volunteers will likely have some contact with healthcare systems and thus have available EHR data that they could transmit to the PMI cohort through data transfer protocols as outlined in Section 5A. Current Blue Button functionality at many locations (e.g., from CMS) provides for ICD, CPT, some problem list data, and/or medication data. These data would be aggregated at the Coordinating Center. The Working Group recognizes that, at the current time, the data obtained from EHR data export functionalities will be very heterogeneous (thus difficult to normalize) and highly fragmented, without verifiability of its completeness or accuracy. However, the Working Group expects that the data quality will improve over the time course of the PMI-CP, with the potential for the PMI cohort to catalyze development of these technologies.

4. **Biospecimen-derived data (all participants)**. Specific biospecimens to collect are detailed in Section 6, and all participants will provide biospecimens. As feasible, the Working Group recommends generation of dense omic data, as described above.

5. **mHealth data (many participants)**. The Working Group feels that sharing of mHealth data by participants should not be a requirement to be enrolled in the PMI cohort. However, despite limited

standardization of mHealth technologies, particularly commercial sensors, and a number of competing technologies, early acquisition of such data will enable exploration of use cases and facilitate building the infrastructure to handle the scale and collection of such data. The Working Group considered several early possible opportunities for early integration of mHealth data. Two platforms emerged as early potential targets. The first was acquisition of the sensor data that can be passively obtained from smartphones, including location, movement, and social connections. Existing research application platforms for smartphones also provide the basis for integrating various participant data collection and performance measures relevant to PMI use cases for select subgroups (e.g., motor and cognitive tests). Configurable ecological momentary assessment systems can be incorporated for researchers to design the prospective administration of selected self-report measures at random intervals or event-based (e.g., if participant is in the location of a hospital for more than 6 hours, query if being treated, having a procedure, or hospitalized). The PMI cohort smartphone base can also serve as the conduit for wearable sensors, both research and consumer grade. A second early target for mHealth data is the data available from current generation commercial activity monitors. These commercial platforms could be validated against research grade activity monitors distributed to a subset of the PMI cohort, and data from these monitors could be aggregated to estimate daily steps, calories expended, hours of sleep and sleep quality. The combination of wrist-worn activity sensors, coupled with location, also provide a rich data set for characterizing discrete health-related activities such as smoking, eating, walking, and time in sedentary states. Importantly, these data combined with the health data derived for the PMI cohort would allow for testing the ability of mobile and wireless data to predict disease and health outcomes.

In its deliberations, the Working Group identified a key design principle for building big data analytic systems: design data systems with an expectation that future growth that may be dramatic. In order to do this, the data systems' functional requirements should incorporate at least a small number of extreme use cases of large data sources. A useful method to doing so is to implement one or two key use cases in the initial phase of the PMI-CP, while designing infrastructure anticipating data volumes that are two or more orders of magnitude larger. In the context of the PMI-CP, candidate categories of data for which this design philosophy is particularly relevant include full genomes and proteomes, streaming data such as real time physiologic sensors and video, as well as the unstructured and semi-structured data found in large volumes of clinical documents.

The PMI-CP will inevitably find that some data of contemporary interest becomes superseded by new methods and new scientific questions. For this reason, the choice of types of data and their sources will be a dynamic process, and prospective collection of data found to be of little value may cease. The Working Group recommends that, in such cases, and for all data acquired by the PMI cohort, there be a commitment to permanent archival preservation of data.

C. Data access, use and analysis

Principles for data access

Recommendation 5.16: The PMI-CP should create a variety of data access and analysis services that would encourage broad use of aggregated data and credentialed, role-based access to individual level data. Varying levels of access control should be applied that are appropriate for data of different levels of sensitivity. An initial set of analysis, visualization, and dashboard tools should be offered as soon as possible in development of the PMI cohort.

Recommendation 5.17: New data created by researchers using PMI cohort biospecimen and/or data resources, should, as a condition of use, be shared back to the Coordinating Center for reuse by other investigators. To facilitate this, the PMI-CP should create mechanisms for sharing and reuse of analytics and analytical results with other researchers.

Recommendation 5.18: The PMI-CP should be resourced to provide for technical support for both participants and researchers to use PMI cohort data resources.

Recommendation 5.19: The PMI-CP should create a Resource Access Subcommittee to evaluate and approve access by researchers to data and biospecimens.

The Working Group envisions PMI cohort data as a public resource that should foster a broad range of research. To accomplish this, there will need to be a substantial variety of data access and analysis services and support to help researchers of varying levels of sophistication, including "citizen scientists" and study participants, achieve their research goals. The PMI cohort data resource will be multifaceted and have internal complexity that predicts that most or all levels of users will need assistance in designing and implementing studies that use PMI cohort data, and the Coordinating Center will need to have a research support services component to provide this assistance. A Resource Access Subcommittee, which would function under the PMI Steering Committee (see Section 8), should be formed to evaluate and approve applications for data and biospecimen (see Section 6) access by researchers to ensure resources are used in an appropriate and ethical manner (Table 5.3).

In the spirit of transparency and collaboration, individuals and organizations that provide data to the PMI cohort should, as a general policy, have unrestricted rights of access to their own submitted data. Individual participants will have varying levels of health and science literacy, and will need assistance with interpretation of their data.

The Working Group also envisions community-facing tools that would allow all individuals some basic level of access to PMI cohort core data resources. Web-based query tools, such as the NIH-supported Informatics for Integrating Biology & the Bedside (i2b2) data workbench,[121,122] provide models for easy-to-use "drag and drop" graphical query builders that enable individual end-users to rapidly query research repositories. An exemplar includes the UK Biobank's Data Showcase, which allows anyone to freely investigate what types of data and biospecimens are available on anyone in the cohort.[123]

The PMI cohort's data will have varying levels of sensitivity, ranging from aggregate numerical data to personal demographics and individual contact information such as name, phone numbers, and email addresses.

The Working Group recommends a strategy of access control that is keyed to the types of data:

- **Public information**: Public information about the PMI-CP, online newsletters, links to publications (the "Internet storefront" of the PMI-CP). These types of information should be available without requirement for login.
- **Public querying with login**: Queries generating aggregate data in the form of record counts meeting the search criteria. This is a mode of query that is commonly used to determine the initial feasibility of the PMI cohort resource being useful to answer a particular research question. For this type of query, which should be enabled for a broad community of users, the Working Group recommends using self-created online accounts and simple verification of identity approaches that are used commonly for e-commerce systems, e.g., a valid email address to which an authentication request is sent. This will facilitate metrics of individual users and reduce the risk of anonymous "bot attacks" against the aggregate data. Policy protections against re-identification of individuals are also recommended (see Section 7C).

The Working Group notes in this context that even aggregate data are susceptible to re-identification attacks when the data returned references small subsets of individuals, particularly if the attacker already possesses certain information about the participants, and genomic data are a component of either the query or the result set.[124] Strategies, such as setting a minimum threshold of records in the results, below which no results are returned, can reduce but not eliminate this risk.

- **Individual-level, de-identified data**: Cross-tabulations and other tests for associations in the data will require individual level, de-identified data. For this type of access, the PMI-CP should require submission of a brief study description, and use a process of account creation that includes human verification of identity. Users would agree to a data user agreement, part of which would specify that the researcher would agree not to try to intentionally re-identify individuals, and would not use any unintentionally identified data for other purposes. Existing NIH resources, such as the database of Genotypes and Phenotypes (dbGaP), provide a reference model for this type of access control that can inform the specific design. (The Working Group notes that dbGaP approach has both enthusiastic proponents and vocal detractors, but it has substantial operational experience that should be leveraged by PMI-CP.)
- **Individual level data with identifiers**: This type of access will be needed for recruiting to and conducting some types of prospective studies, particularly interventional studies. This type of data access and use will constitute human subjects research and may require IRB review, in addition to the type of identity verification and project description submission described for de-identified data. Users would agree to a data use agreement, specifying that they would not store or use identity information beyond purposes described in the project.

Queries of individual level data, whether identified or de-identified, should be conducted within the secure computing environment provided by the Coordinating Center, as outlined in Section 5D; downloading of individual level data outside the secure computing environment should be prevented to minimize the risk of data privacy loss. As noted elsewhere, to build and enhance the data resource over time, the Working Group recommends that as researchers use PMI cohort data and biospecimens, any

new data generated on PMI cohort individuals be shared back to the Coordinating Center as a quid-pro-quo condition of using PMI cohort resources. The Working Group recommends an embargo period on secondary publications, similar to employed in dbGaP, arising from data contributed to the PMI cohort to allow primary researchers to have the opportunity to publish their work while also encouraging rapid sharing of their data back to the PMI cohort.

Table 5.3: Types of Access. RAS=Resource Access Subcommittee

	Open Access (no login required)	Anyone with a login (no application necessary)	After approval of brief project review	With RAS and IRB review	With RAS and IRB review and additional participant consent
Newsletters, ongoing PMI studies and general updates	X				
Aggregate counts of individuals	X				
Graphical query to assess study feasibility using counts		X			
Query interface exact counts of rare events			X		
Access to de-identified individual-level data				X	
Access to identified data				X	
Recontact individuals				X	X
Clinical trials					X

Participant access to data

Recommendation 5.20: Participants should have access to their PMI cohort data, except where there are compelling concerns about harm that may result from such access. Participants should be in control over whether their data are shared back to their healthcare provider organization for research use.

An important component of the PMI-CP will be to facilitate participant access to their PMI-generated data. This process begins with a commitment to make readable summaries of research activities pursued in the PMI cohort and aggregate study data found using cohort data. Individuals could be

notified of how their data were used, as a listing of particular studies in which their data contributed. Individuals should also have full access to the downloaded healthcare data and CLIA-generated laboratory data (including omic data).

As discussed in Section 4, the Working Group recommends formation of a Return of Results and Information Subcommittee with community stakeholders to monitor and advise on return of results issues. In accordance with general clinical policy, non-CLIA test results (e.g., especially new research methodologies) usually should not be returned, but should be monitored by this Subcommittee as well. In general, the PMI-CP should attempt to use CLIA testing procedures and facilities when possible.

Participant data can be made available via participant-oriented websites or mobile applications. To encourage sustained engagement, participants should also be able to configure the frequency of feedback received regarding the data they provide, including the ability via smartphones to receive immediate, real-time feedback where appropriate. For security purposes, applications that allow participants to access and download data should be encrypted with two-factor authentication to gain access to raw data (e.g., both phone passcode to get to the PMI mobile application, and login for the application). All data should be retained at the server with no data provided to the participant that remains resident on the access device (phone, computer, etc.).

Computational capacity for data storage and analysis

Recommendation 5.21: The PMI-CP should leverage existing public and private expertise to design and implement large scale, elastic data storage and analysis capabilities to serve PMI cohort's needs.

Recommendation 5.22: The PMI-CP should develop protected environments for data access and analysis that incorporate best practices for data with varying levels of sensitivity.

The PMI cohort is being launched at a time of explosive growth in the number, size, and complexity of potentially relevant data resources. The "big data" of human biology, such as full genomes and high resolution digital images, may be combined with other novel forms of equally large or larger data, such as weather patterns, environmental monitoring, and streaming physiologic sensor data from study participants.

Some of the PMI cohort's scientific use cases will create computational challenges that require highly elastic data storage and computing resources (i.e., able to scale orders of magnitude on demand for specific research projects). Cloud computing and similar approaches can provide an analytical environment where "tools go to the big data" rather than data being downloaded to the tools, and NIH has made recent investments in prototype research computing environments (e.g., National Cancer Institutional Genomics Cloud Pilot projects[125]) that can help inform PMI cohort design decisions for creating an appropriate and effective data storage and analysis infrastructure. As discussed in Section 3, the PMI-CP should take advantage of state-of-the-art cloud computing environments that could involve novel public-private and academic-commercial partnerships.

Because PMI cohort research will involve sensitive personal information and require secure computing environments, it will benefit from contemporary technologies such as "data enclaves" that provide both data access and computational tools for analysis within a secure environment. The CMS Virtual Research

Data Center[126] provides a contemporary technical and process model that has high relevance for the PMI cohort. In general, the Working Group feels that the PMI-CP should require researchers to use data within centrally maintained secure computing environments for most if not all use cases, though should consider alternate data management strategies if compelling scientific use cases arise. As new security and enclave technologies become commonplace over time, the PMI-CP should review and revisit its architectural approaches to leverage mass market solutions that make secure sharing of data easier to use and affordable.

D. Technology infrastructure and operations

Data management architecture

Recommendation 5.23: The PMI-CP should have a Coordinating Center that will centrally store a curated, analysis-ready, core data set on all participants using common data models, so that many queries and analyses can be executed at the Coordinating Center and not require local site/network involvement.

Recommendation 5.24: The PMI-CP should pursue a hybrid data and analytics architecture that leverages both centralized data storage of a growing amount of core data and federated access to additional data at the nodes across the network, as needed by specific studies. To ease application of federated queries and analyses across sites, the sites should work toward use of a common data model, recognizing that not all types of queries and analyses may be performed using a common data model.

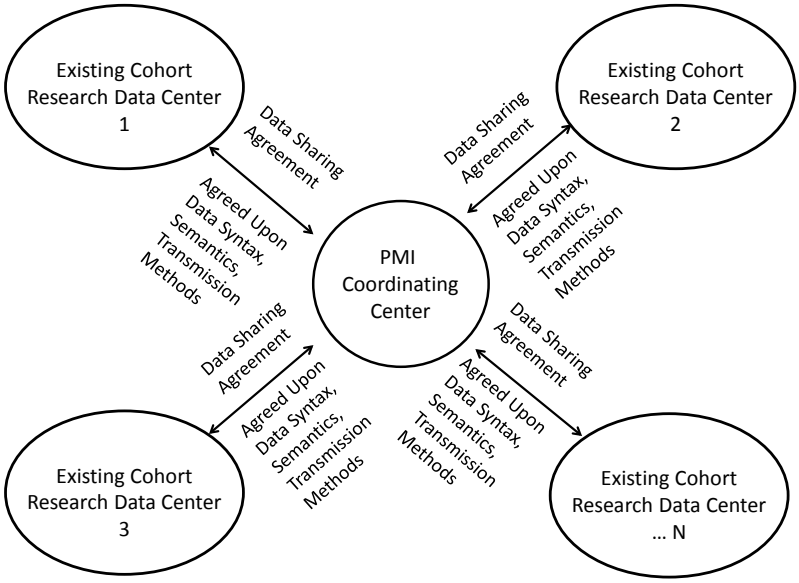
The "traditional" model for an NIH-funded multicenter research consortium involves a Coordinating Center that serves as the hub in a hub-and-spoke model of collaborating research organizations (Fig. 5.1, below). Nodes could include HPOs and site(s) that collect and aggregate direct volunteers. The Coordinating Center and each of the participating HPOs maintains a research data management infrastructure and dedicated personnel with expertise in acquiring, storing, and transmitting data of various types.

Given the novelty and complexity of mobile and sensor data inputs, a technology core or center should be a critical "spoke" for the Coordinating Center. All data, including mobile and sensor data, should be sent to, stored by, and managed by the Coordinating Center, but the technology center can develop the mobile and sensor data collection applications, test and document algorithms for any non-raw data being provided to the Coordinating Center, and negotiate with various outside devices and data source entities.

It is important to note that the Coordinating Center need not be a single entity, yet there does need to be a single entity that ultimately is a coordination point for all data and its access. For instance, the Working Group notes that the harmonization of omic data, EHR data, mobile data, and collection of direct volunteer data could all be performed within separate sites (but with each individual type of data being housed, curated, and accessed via a single entity). Indeed, the facilities and expertise for storing and handling biological samples, curating clinical and biological data, and managing sensor data may all be different. Importantly, data across all participants (both HPOs associated and direct volunteers) should be combined. Ultimately, these data should be aggregated to a single, unified interface for

querying. The Working Group has termed this these collective functions (or cores) as the Coordinating Center. It is possible it could either be separate awards by molecular data, clinical data, mobile and sensor data, and patient engagement, or the Coordinating Center could be a single award that contains multiple cores, potentially spanning multiple sites.

Fig. 5.1: Organizational relationships



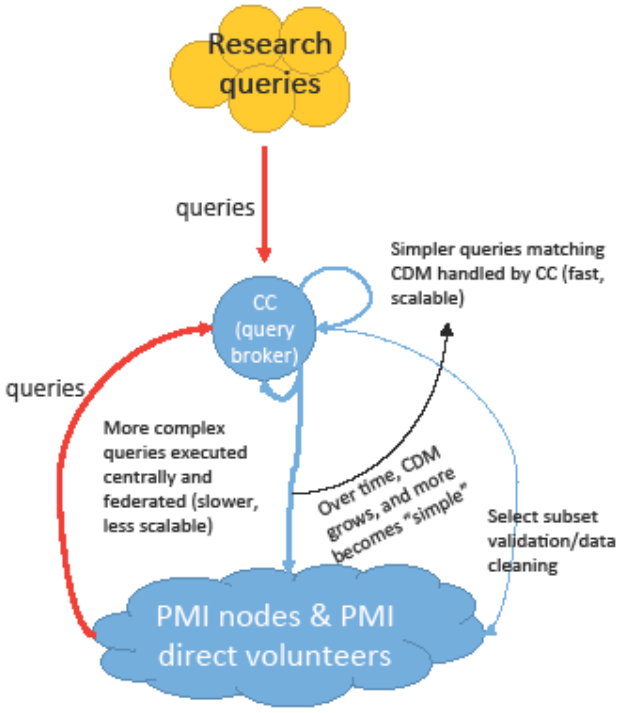
It is important to note that the Coordinating Center need not be a single entity, yet there does need to be a single entity that ultimately is a coordination point for all data and its access. For instance, the Working Group notes that the harmonization of omic data, EHR data, mobile data, and collection of direct volunteer data could all be performed within separate sites (but with each individual type of data being housed, curated, and accessed via a single entity). Indeed, the facilities and expertise for storing and handling biological samples, curating clinical and biological data, and managing sensor data may all be different. Importantly, data across all participants (both HPOs associated and direct volunteers) should be combined. Ultimately, these data should be aggregated to a single, unified interface for querying. The Working Group has termed this these collective functions (or cores) as the Coordinating Center. It is possible it could either be separate awards by molecular data, clinical data, mobile and sensor data, and patient engagement, or the Coordinating Center could be a single award that contains multiple cores, potentially spanning multiple sites.

Centralized resources for data acquisition and management, with a Coordinating Center serving as the host for a core dataset and home of user services that support both participants and researchers' access to the data, represents the most efficient architecture. Data queries that are answerable via core data at the Coordinating Center could essentially be scaled to very high throughput.

In cases where specific studies require additional data not already contained in the core datasets maintained by the Coordinating Center (but available at the sites), a distributed or "federated" model for data query and analysis is necessary and appropriate. In this model, one node (i.e., a participating

organization) of the network – most commonly but not exclusively the Coordinating Center – is the source of a data query that is broadcast to all participating nodes in the network, and each of those nodes separately executes the query and returns results in a specified format to the requesting node. The requesting node then merges the query results and presents them to the user and/or saves the concatenated results from the network for further processing. In the process of executing a more detailed federated query, many may require further data curation and the application of custom algorithms. Curated data could be shared back to the Coordinating Center to facilitate subsequent queries. As such, these queries will be more analysis-intensive and fewer of these queries will be accomplishable per unit time. The Working Group believes that this hybrid architecture of centrally hosted queries against a well-curated core dataset, supplemented by the ability to invoke federated query approaches for questions that need data not contained in the core, will serve the goals of efficiency and scientific flexibility over the life of the PMI-CP (Figure 5.2).

Fig. 5.2: Proposed data flow between the Coordinating Center (CC) and PMI cohort network nodes and volunteers



Use of data standards

Recommendation 5.25: The PMI-CP should leverage prior investments in data standards, and adopt existing national and international data standards wherever feasible. Where new data standards are needed, PMI-CP should collaborate with national and international standards organizations to extend and enhance existing resources.

Recommendation 5.26: The PMI-CP should use Common Data Models (CDMs) for structured data that are contained in the core PMI dataset, wherever feasible. Building upon and extending existing CDMs is preferable to creating new ones.

As a design philosophy in the creation of PMI cohort data resources, the Working Group notes the self-evident truth that use of data standards promotes quality science, consistency in data use and re-use, and meta-analysis. In this regard, there are several levels at which data standards are relevant. Common Data Elements (CDEs) are standards that, in general, define individual variables. CDMs provide additional structure to organize CDEs, data communications and database design. There has been substantial federal investment by the Patient Centered Outcomes Research Institute (PCORI), NIH, FDA, and ONC to build libraries of standardized data elements.

Use of CDEs and CDMs incurs additional cost at the point of acquisition of data into research systems, and it is almost always easier to create a new variable name and definition than to take the time to search and evaluate existing published definitions. An additional complexity for the PMI-CP is that, given the extensive variety of useful data sources, no single CDM is likely to suffice for all PMI cohort research. These difficulties notwithstanding, the Working Group believes that the PMI-CP should use existing standards wherever feasible, and not create new data standards if existing ones will suffice to support its scientific agenda. Several exemplars of CDMs exist, including those being created by PCORnet, the National Patient-Centered Clinical Research Network, and the Observational Health Data Sciences and Informatics collaborative.[127] Learning from these and other successful models, such as the use of semantic web standards in projects like PubChem[128] will make it easier for users to find, share, and combine information from PMI cohort data resources and other, related research information not maintained by the PMI-CP. To provide for the greatest number of use cases for PMI data, the PMI-CP should keep abreast of ongoing standards development and catalyze their growth.

Data security

Recommendation 5.27: The PMI-CP should designate a formal process and personnel with security expertise to create and maintain appropriate physical, technical and policy safeguards that will ensure state-of-the-art security for all PMI-CP data and systems, and comply with applicable regulations.

Recommendation 5.28: The PMI-CP should provide levels of data protections that are appropriate for de-identified data and sensitive personally identifiable data, using a risk-based approach.

Recommendation 5.29: In general, the PMI-CP should discourage data from being copied outside the PMI secure computing environment, while allowing outside data to be imported into the PMI cohort computing environment.

Recommendation 5.30: The PMI-CP should establish procedures and backups to ensure continued access to the data.

Data security will be essential to establishing and maintaining the trust of study participants and collaborating organizations. The goals of the administrative and technical components of data and communications security are to ensure data integrity and to limit data access to authorized users for authorized purposes. It is a foundational truth in the realm of data security that perfect access and perfect security cannot co-exist: all data security is a compromise between usability and effective safeguards.

The PMI-CP will depend critically upon the use of the Internet and cell phone networks, personal workstations and smartphones, and involve transmission, storage and analysis of sensitive personal

data. Thus it will inherit the security vulnerabilities of all of those technologies simultaneously. A hard fact of modern society is that cyber threats are real, pervasive, and will continue to evolve during the duration of any longitudinal study. In addition, the PMI cohort data and communications will be an appealing target for hackers and for well-resourced criminals up to and including state-sponsored cyberattacks, with motivations that may be personal, financial and/or political.

Detailed standards and implementation practices for protecting the privacy and security of shared genomic and clinical data have been developed and published by the Global Alliance for Genomic Health (GA4GH), which is an international partnership of more than 300 organizations, including NIH. The Working Group recommends the GA4GH reference and design documents[129] as a well-crafted set of policies and technologies to guide formulation and implementation of data security within the PMI-CP. PMI-CP systems should consider relevant provisions of the Federal Information Security Management Act[130] as one guide, which outlines levels of protection based on sensitivity of data and consequences of breaches of security. The PMI-CP should ensure regular testing, monitoring, and public reports at some interval after a major attack or quarterly/semi-quarterly basis. Such reports should indicate results of penetration and internal security ("redhat/whitehat") testing, attacks, etc. In keeping with current best security practices, the PMI-CP should also enable a "bug bounty" program to encourage the best minds to find deficiencies in the PMI cohort and constructive ways to implement solutions. Developing and implementing security safeguards for the PMI cohort will be a complex, ongoing operational issue.

As noted above, providing a secure environment for centrally managed computational resources may benefit from computing "enclaves" such as those developed by CMS and the DoD. In this regard, the PMI-CP should leverage the technologies and lessons learned by those federal organizations.

Another important component of data security is assured ongoing access to and maintenance of the integrity of the participant data and scientific resources that will be created as a result of PMI-CP investment of public funds. The Coordinating Center will need to have a data and applications backup infrastructure that is not dependent upon a single commercial or academic institution's local facilities. One such possibility would be the National Library of Medicine, which has been a steward of biomedical data resources for more than 175 years, as a natural home for a suitably encrypted and protected backup copy of all PMI cohort data and applications resources, updated at regular intervals appropriate to the pace of expansion of the PMI-CP's digital resources.

Data privacy

Recommendation 5.31: The PMI-CP should create and use de-identified data for research whenever feasible to do so.

Recommendation 5.32: The PMI-CP should engage data privacy experts to create an effective combination of technology and policy to minimize risks of re-identification of de-identified data.

Recommendation 5.33: The PMI-CP should develop educational materials for participants that explain the principles of data privacy, its limitations, and their role in helping to maintain it.

Recommendation 5.34: The PMI-CP should have a clearly articulated plan in case of a privacy breach, which includes notification to participants.

The goal of data privacy is assurance to participants that their personal identity will not be subject to breach of confidentiality, nor able to be inferred against their wishes from features of the data they provide to the PMI cohort. Data privacy is built on the foundation of effective data security (whose goal as noted above is to limit data access to authorized users for authorized purposes, and ensure data integrity). Data privacy extends these concepts into the realm of considering the risks associated with authorized users having access to potentially sensitive information.

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-CP, as outlined above.

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, ancestry web sites, etc.) means that technology alone will be insufficient to address issues of data privacy for the PMI cohort. Expert testimony presented at the *Digital Health in a Million-Person Precision Medicine Initiative Cohort* workshop (see Section 1) brought forth the view that de-identification should not be thought of as a guarantee of anonymity, but rather simply "another disincentive to attempting re-identification of individuals." Acceptable use policies with substantial enforceable sanctions will need to be developed or adapted from other similar research efforts to complement the technical approaches to de-identification of data.

Study participants can inadvertently contribute to increased risk of personal identification via online publication of personal information via social media. For this reason, educational materials will be needed for study participants that explain these types of risks, and how they can reduce the likelihood that their identity could be discovered by data mining of publically accessible data resources.

Section 6 – Biobanking

As discussed in Sections 2 and 3, the collection of biological specimens is essential to the mission of the PMI-CP. The Working Group recognizes a number of types of biospecimens that could be collected, including DNA, RNA, plasma, microbiome specimens, and nail and hair clippings, among others, but feels the highest priority should be to obtain blood specimens from each participant. Although some HPOs recruiting participants into the PMI cohort may have their own biobanks, the Working Group agrees that each participant joining the PMI cohort must provide a new specimen. Specimens should be collected and processed using a standard CLIA compliant procedure to ensure quality control and comparability of biospecimen collection across the PMI cohort. It will be necessary to establish a central biorepository, the PMI biobank, to support collection, processing, storage, retrieval and biochemical analysis and/or shipment to analytic laboratories of all biospecimens. Ideally, the PMI biobank would be in place before the start of recruitment.

A. Biobank Subcommittee

Recommendation 6.1: The PMI-CP should establish a Biobank Subcommittee that will oversee all aspects of PMI cohort biobanking. This Subcommittee reports to the overall PMI-CP Steering Committee.

The Biobank Subcommittee will be responsible for implementing the PMI cohort protocol that defines how and what biospecimens should be collected and the specific tasks of collection, processing, storage, analysis, and sending out of those specimens. The initial focus of this Subcommittee will be on biospecimen collection, processing, and storage. It will need to establish quality standards and assurance processes. Later, the Biobank Subcommittee will oversee the use and preservation of the banked specimens. The Biobank Subcommittee will need to establish the procedures by which investigators can gain access to and use stored specimens, ensuring limited resources, such as stored plasma, are used parsimoniously. Sample use will need to be monitored throughout the life of the PMI-CP, and access procedures potentially revised periodically to respond to changing resources, participant engagement, and scientific opportunities.

B. Central biobank

Recommendation 6.2: Before participant recruitment, the PMI-CP should establish a full service, central biobank that manages all aspects of collection, processing, storage, retrieval, sample tracking, and biochemical analysis. The capacity of this facility should be for at least several million specimens and utilize state-of-the-art robotic processing and storage systems to facilitate high-volume acquisition and retrieval.

A central biorepository should be established very early in the formation of the PMI cohort. The tasks that would likely be in the scope of the PMI biobank would be all aspects of specimen collection, processing, storage, retrieval and analysis. This includes communication with the recruitment centers; establishing a specimen tracking database that interfaces with the recruitment database; setting up incoming specimen processing; purchasing all processing and storage supplies; working with vendors to purchase the necessary robotics equipment; sending specimens out for analysis at other approved

vendors and, if desired, establishing in-house high-throughput analysis platforms. For samples that are sent out, the biobank staff need to be able to communicate with all analytic labs and may need to do early blinded testing of analytic procedures. All processes will need to be tracked using a secure informatics system that interfaces with and meets the standards of the data system in place for the PMI-CP.

The biobank will manage both the collection process, taking into account all locations for recruitment, and the shipping of all specimens. This will include instructing all sites on how to collect and ship samples and managing shipping contracts. The management of collection is a critical task, as the collection and shipping practices determine the quality of the specimens as they arrive in the lab. It is also crucial to establish the types of procedures needed to accurately assign specimen codes to each specimen linking them to the right enrollee using CLIA standards. The biobank needs to have a system in place that can evaluate the quality of specimens being received and a system for providing feedback to staff in the field to resolve any protocol incongruences or shipping irregularities.

It is essential that the laboratory design a system that can efficiently and accurately handle the anticipated daily volume of specimens. Thus, the biobank will need to establish high-throughput procedures for the efficient handling and processing of each type of specimen coming into the biobank. This will include acquisition and maintenance of robotic systems to separate, label, and store biospecimen components (e.g., DNA or serum), as directed by the Biobank Subcommittee. Various options for automated DNA extraction will need to be considered as well.

The biobank will be responsible for establishing an automated storage and retrieval system that can store the anticipated numbers and types of specimen aliquots that will result from the processing described above. The large biorepositories currently in existence have generally adopted robotic storage and retrieval systems that move coded specimen tubes into and out of freezers. Decisions regarding the type of robotic storage will involve exploring existing facilities and creating a design for storage and retrieval system that can accommodate the needed capacity and the anticipated frequency of retrieval.

The biobank should oversee sending out specimens to analytics labs and return of data to and from analytic laboratories using blinded specimen numbers. A critical task of the biobank is to get specimens of all types to the appropriate place for analysis and ensure data transferred to the Coordinating Center in a secure manner. Specialized vendors for analysis of specimens including sequencing and genotyping may need to be engaged. The biobank should be responsible for understanding the requirements (concentrations, volumes, etc.) and providing those to the particular vendor using unique codes for each specimen shipped. Then, working with the Coordinating Center, the biobank should maintain coded cross-walk databases in a secure fashion. After analysis is done by outside vendors, the biospecimens should be returned to the biobank or destroyed, according to a process determined by the Biobank Subcommittee. It may make sense to consider setting up some high-throughput analysis capabilities in the biobank.

The biobank should develop and maintain a computerized system for tracking of all specimens at all steps along the process and maintain metadata and identifiers about specimens (location, volume, etc.),

working in close collaboration with the Coordinating Center. All activities, including collection, processing, storage, retrieval and send-out for analysis, need to be tracked using a database that can interface with the main PMI cohort database at the Coordinating Center.

The biobank must maintain security, backup, and alarm systems to insure that the specimens are secure and maintained at appropriate conditions. Security systems must be in place that only permit appropriate personnel have access to the biobank. In order to ensure that specimens are maintained in optimal conditions, backup electrical systems and alarms should be in place that will notify responsible personnel when freezers are out of prespecified temperature range. Staff should be available to respond to an alert at all times. In addition, the system should be safeguarded against fire, floods, and water intrusion (e.g., routing away sprinkler system water or a broken water line).

C. Specimen collection and storage

Recommendation 6.3: Only new specimens should be collected for the PMI cohort and banked.

Since there are many challenges using previously collected and stored samples, it is preferable to collect new biospecimens from each participant at the time of enrollment into the PMI cohort. This will ensure consistent quality and availability of specimens across the PMI cohort.

Recommendation 6.4: Specimens should be collected, shipped, processed, and stored using CLIA procedures.

Collection and handling of biospecimens using CLIA-compliant procedures will insure that specimens will be attributed to the correct participant with a high degree of certainty. This will involve strict procedures at the time of collection and at every step along the path of processing, storage, retrieval and analysis.

Recommendation 6.5: The PMI-CP should collect blood and other samples that anticipate current and future uses for baseline samples and samples collected at subsequent intervals. These include serum and plasma to support analysis of routine and advanced analytes (e.g. metabolites, cell-free DNA), leukocyte nucleic acids, blood cells stored to permit future sorting and analysis, samples for potential exposure studies (e.g. hair, nails), and samples for potential microbiome studies.

Initial collection should include the specimens of the expected greatest utility over the course of the PMI cohort and for which there is demonstrated stability. These include plasma, serum, buffy coat, red blood cells, and genomic DNA. Consideration needs to be given to the many potential uses of stored biospecimens, potentially requiring specialized procedures at the time of collection, and the fact that plasma and serum cannot be replenished. Many additional types of specimens can be collected including urine, stool, nails and hair. For example, the UK Biobank has collected blood and urine, while MVP collected blood only. Kaiser Permanente started with saliva and then moved to blood collection. The Working Group felt peripheral blood collection should be required for all participants. The approximate number and size of aliquots to be stored for each participant will need to be determined prior to finalizing agreements with the sites that will be collecting samples.

Recommendation 6.6: The PMI-CP should establish a backup facility that can provide storage of a portion of the biospecimens.

A backup facility for biospecimen storage is essential to prevent a catastrophic event from causing irreparable loss of or damage to the collected biospecimens. This backup facility should be in a different geographic location than the primary production facility. The Working Group discussed different approaches to a backup facility, ranging from a fully redundant backup facility (with robotic freezers, etc.) to purely archival storage that would allow disaster recovery but would likely result in some downtime in access of the samples. The Working Group feels that the latter option is likely sufficient for most cases as expected loss of the primary PMI facility (e.g., natural disaster, terrorist act, etc.) should be unlikely. It is not necessary that equal volumes of material be distributed between the archival and production biobank facilities.

Section 7 – Policy Considerations

The success of the PMI-CP will depend, in part, on the legal, regulatory, and policy landscape surrounding research, EHR access and interoperability, regulation of genomic – and other omic – technologies, and data security and privacy. For the PMI-CP to achieve the vision described here, policy gaps and conflicting policies will need to be identified and corrected. In addition, and more directly, the PMI-CP will need to develop an internal policy framework that will apply to individuals participating in the PMI cohort as research participants, data collectors, or data users. Just as the PMI-CP will advance a new model of participatory research, addressing any existing policy issues and gaps will benefit research and the delivery of care beyond the PMI cohort.

Many of the policy challenges and needs confronting the PMI-CP were discussed at the PMI workshops (see Section 1) and Working Group meetings. Some of the policy recommendations can be implemented by the PMI-CP itself, while others may require action by NIH, the Administration broadly, Congress, or the private sector.

Box 7.1: The legal, regulatory, and policy landscape affecting PMI-CP

As a publicly funded research initiative involving human research participants, the PMI-CP is subject to a number of laws, regulations, and policies, generally including:

- Healthcare Insurance Portability and Accountability Act of 1996 (HIPAA) Privacy Rule
- Health Information Technology for Economic and Clinical Health (HITECH) Act
- The Privacy Act of 1974
- Clinical Laboratory Improvement Amendments (CLIA)
- HHS regulations for the meaningful use of electronic health record technology
- FDA framework for regulation of genomic technologies
- FDA regulations for medical devices (including *in vitro* diagnostics and mobile health technologies)
- Federal Policy for Protection of Human Research Subjects (the Common Rule)
- NIH Grants Policy Statement (http://grants.nih.gov/grants/policy/nihgps/nihgps.pdf)
- NIH data sharing policies:
    - NIH Data Sharing Policy
    - NIH Genomic Data Sharing Policy
    - (Draft) NIH Policy on Dissemination of NIH-Funded Clinical Trial Information
- State laws as applicable

The Working Group benefited from the Proposed Privacy and Trust Principles developed for PMI and put forth by the White House.[131] The principles, which are overarching to cover all components of PMI, describe a set of core values that will help guide the creation of a large national precision medicine

research cohort. The principles describe the vision of PMI, one that prioritizes public trust, data privacy and security, and responsible data sharing to maximize the possible benefits of the PMI cohort. The inter-agency group that developed the principles aimed to maximize the benefits of precision medicine research at the proposed scope and scale while minimizing the inherent risks in data collection, storage, analysis, and sharing. The draft principles cover governance, transparency, reciprocity, respect for participant preferences, data sharing, data quality and integrity, and security. These principles were posted for public comment and will be finalized over the coming months.

Gaps in the existing legal and regulatory framework are described below with recommendations on what will be needed to achieve the objectives of the PMI cohort. Recommendations are also offered on policies applicable only to the PMI-CP that should be developed by NIH.

A. State laws

Recommendation 7.1: NIH should analyze relevant state laws to ensure that the PMI-CP is in compliance and barriers identified.

Because the PMI-CP will be a nationwide effort, there are diverse state laws that will become important for the conduct of the PMI-CP. NIH should analyze the most relevant of these laws to ensure that the policies and governance established as the PMI cohort is created comply with local laws whenever possible.

B. Inclusion

Recommendation 7.2: NIH should carefully examine issues related to inclusion of three special populations: children, decisionally impaired individuals, and PMI cohort participants who become incarcerated after enrollment. NIH should develop specific approaches to address the needs of these individuals so that they may be included and retained in the cohort.

Recommendation 7.3: The PMI-CP should anticipate the need for special provisions to allow for continued engagement and follow up with participants who have undergone life events or other changes that alter their participation status or capacity.

The PMI cohort is intended to reflect the diversity of the U.S. and thus, NIH should ensure that all who want to participate in the cohort have the opportunity to do so (Section 3). Without broad inclusion, the PMI-CP will not fulfill its most important goal, to increase the health of all Americans through new understanding of the development and progression of disease, optimal treatment, and prevention strategies. The goal of inclusion must be matched by the goal that participation be informed, voluntary, and not subject to coercion. Thus, the PMI-CP will need to develop special considerations and safeguards for those who may not be able to provide fully informed, voluntary consent.

This issue is addressed in part by existing federal rules for the protection of human subjects, which include special requirements for the inclusion of children, prisoners, and individuals who are decisionally impaired.[132] Including these populations, among others, in the PMI cohort is important because they represent a substantial proportion of Americans. Excluding them would limit the scientific validity and utility of the cohort, deprive PMI cohort participants of opportunities to benefit from research, and worse, could increase health disparities for these groups.

While the Working Group supports the inclusion of children and decisionally impaired adults, it was unable to consider policy issues fully surrounding inclusion of these groups. To address these questions, NIH should consider how best to incorporate the necessary safeguards into the PMI cohort in order to ensure the appropriate enrollment of, retention of, and protection for children, decisionally impaired adults, and participants who become incarcerated.

C. Institutional review board and consent

Recommendation 7.4: The PMI-CP should have a single Institutional Review Board (IRB) (to the extent permitted by law) constituted to ensure prompt and thoughtful consideration of the evolving protocols in the PMI cohort and the central importance of participants as research partners. NIH should consider whether such an IRB would best be located at the NIH or at the Coordinating Center.

Recommendation 7.5 NIH should work with FDA to seek to remove the statutory requirement for local IRB review of multi-site device trials in order to facilitate such research using the PMI cohort.

Recommendation 7.6: The PMI-CP IRB should include a substantial number of members of the public and representatives of the participant community.

Recommendation 7.7: The PMI-CP should support revisions to the Common Rule that enable use of broad consent for secondary use of data and specimens to allow knowing, willing participation and facilitate research.

The Working Group identified a number of issues related to the ethical review of research and the informed consent of participants that should be addressed. Studies have shown that the use of multiple IRBs in large, multi-site studies can impose significant burden and delay while introducing variability in consent processes and documents.[133,134] Central IRBs may lower administrative costs, reduce review time, and reduce variability between research sites.[135] Given the scope and scale of the PMI cohort, and the evolving nature of the research that will be conducted with the cohort over time, there could be considerable benefit in adopting a single IRB. Furthermore, many features of the PMI cohort will be novel and a single IRB will be better equipped to ensure consistency with the overall vision and principles for the PMI cohort. A single IRB is consistent with recent regulatory and policy proposals such as the 2015 Notice of Proposed Rule Making (NPRM) for the Common Rule[136] and the 2014 NIH draft policy for the use of single IRBs.[137] NIH should also consider the pros and cons for maintaining the single IRB at NIH, using a model analogous to the National Cancer Institute's Central IRB Initiative,[138] or at an institution involved in the PMI-CP, such as the Coordinating Center. Regardless of where the IRB resides, it needs to be independent from, but able to work closely with the overall governance of the PMI-CP. The single IRB should also serve as the HIPAA Privacy Board in order to retain the benefits of PMI-CP IRB centralization.

The Working Group recognizes that there are laws, regulations, or special considerations that require the use of local IRBs. Notably, the 1976 Medical Device Amendments Act of the Federal Food, Drug, and Cosmetics Act requires review by a local institutional review committee of any proposed clinical testing of a device. This provision will increase the burden for conducting device trials through the PMI cohort.[139] The PMI cohort represents a unique resource for conducting trials of new medical devices that may help to speed their development. Furthermore, new medical devices may be able to contribute important information to the PMI cohort, through baseline information on all participants or through

trials done on relevant subsets of participants. A legislative change will be needed to amend the Medical Device Amendments Act to allow the use of single IRB review for medical device trials in the PMI cohort. Such a change would also improve the ability to conduct device trials nationally. The House Energy and Commerce Committee included such a proposal in their 21st Century Cures Bill, which passed the House on July 10, 2015.[140]

Developing a more participatory model of research requires that individuals be given greater choice over how their information is used and other preferences, including modes of communication with the PMI-CP, whether they can be recontacted for participation in new studies, and what types of information they receive about themselves. In order to ensure that review by the PMI-CP IRB reflects the diverse background and preferences of PMI cohort participants, the IRB will need to include a substantial number of representatives from the public, including the unique perspective of patients and their advocates. These views will be essential in ensuring that the conduct of research in the PMI cohort meets not only the ethical standards and regulatory requirements of the Common Rule, but is also responsive to the needs of patients and populations.

Respect for individual autonomy and rejection of paternalism is a paramount concern of the PMI-CP and is a motivation underlying the participatory model. Implementation that fails to uphold and actively promote these values will be inconsistent with the vision of PMI cohort. Obtaining consent for research uses of biospecimens and data is respectful of research participants, and it is increasingly clear that people expect to be asked for their permission to use and share their specimens for research, even if the specimen cannot be readily identified.[141,142] Furthermore, because full cooperation by participants will be necessary to obtain the level of information that the PMI cohort will collect, and because it is a goal of the PMI-CP to provide research results and other information to participants, the informed and willing engagement of participants is the only practical approach for the PMI cohort. [142,143]

Currently, the Common Rule incentivizes the use of specimens and information by allowing their use without the knowledge or consent of the individual once they have been stripped of identifiers. As a result, participants are removed from the equation, are unaware of how their specimens are being used, and no longer have the opportunity to withdraw from research. In addition, in today's data-rich environment and with the technologies and computing that make re-identification increasingly possible, there are informational harms that can result from data generated through "de-identified biospecimens" unbeknownst to the individual. While that approach is allowed under today's regulations, this would not be consistent with the PMI-CP's goal of advancing a more participatory model of research. In addition, this is inconsistent with the NIH Genomic Data Sharing Policy, which expects that genomic research involving de-identified specimens comes with consent for future research use and broad sharing.[144] The recently proposed modifications to the Common Rule include a requirement for consent for the use of biospecimens, whether identifiable or not, and would allow that consent to be obtained at the time of collection by using a general, open-ended consent for future unspecified research.[145] NIH should support the proposed revisions to the Common Rule, and if a final rule containing this requirement is not in place in time, the PMI-CP should consider how to adopt a policy of requiring consent for all research uses, and allow the use of a broad consent when appropriate

(e.g., for aggregate data that will be made publically available, for de-identified data used for secondary research, etc.).

D. Privacy, misuse of information, and security

Recommendation 7.8: To safeguard against unintended release of the information, NIH should seek to establish an exemption under the Freedom of Information Act (FOIA) for release of genomic and other data held by the federal government.

Recommendation 7.9: Unauthorized re-identification or recontacting of participants should be expressly prohibited in agreements for the use of specimens and data, and NIH should pursue legislation penalizing such actions.

Recommendation 7.10: The PMI-CP should only enter into agreements with sensor technology developers that have appropriate security and privacy measures in place to safeguard device users' data.

Recommendation 7.11: The PMI-CP should only partner with sensor technology developers that agree not to sell or use information generated from the PMI cohort unless these uses are agreed to by participants.

Recommendation 7.12: To protect individual PMI cohort data from disclosure in civil, criminal, administrative, legislative, or other proceedings NIH should require all users of identifiable data to secure a certificate of confidentiality from NIH (as authorized under section 301(d) of the Public Health Service Act). Further, NIH should seek legislation to strengthen Certificates of Confidentiality to ensure that disclosure by researchers is not optional, other than with consent or for certain public health exceptions. This will be critical to ensure that data on PMI cohort participants is not used for any purpose other than research.

Recommendation 7.13: NIH should encourage issuance of an Executive Order to ensure that research information from the PMI cohort is not used by any executive agency to deny federal benefits.

Recommendation 7.14: Participants should be informed that some uses of their genetic information are prohibited by the Genetic Information Nondiscrimination Act, and informed of the limitations of the Act's protections.

Recommendation 7.15: Participants should be notified promptly by the PMI-CP following discovery of a breach of privacy. Notification should include, to the extent possible, the types of information involved in the breach, steps individuals should take to protect themselves from potential harm, if any, and steps being taken to investigate the breach and mitigate losses.

Recommendation 7.16: The PMI-CP should establish a Security Subcommittee of the Steering Committee composed of leading experts on cyber security and the management of large amounts of data to ensure that the PMI-CP is incorporating cutting edge security measures and actively monitoring the strength of the data systems.

Maintaining robust privacy protections and ensuring appropriate use of the data that participants provide to the PMI cohort will be essential to engendering trust; the PMI-CP cannot be successful without a foundation of trust between participants, providers, researchers, private sector partners, and the PMI cohort governance. The PMI-CP is asking individuals to provide substantial information about themselves and their health. In return, we must ensure that the PMI-CP has all the tools needed to protect those data from breach and unintended uses. Moreover, use of data generated through sensor

technology, like other data generated from the PMI cohort, should only be used for purposes for which the participants provided consent.

Through public consultation, Working Group meetings, and with the benefit of the Proposed Privacy and Trust Principles put forth by the White House, the Working Group highlighted several notable gaps in current privacy protections and recommends changes needed to provide a robust privacy framework under the information that will be collected and maintained by the PMI-CP.

The Freedom of Information Act (FOIA)[146] allows the public to request any government record, and provides certain narrow exemptions including for information that, if released, would constitute a clearly unwarranted invasion of personal privacy. Although NIH has denied FOIA requests for genomic data, including for de-identified genomic data, using current exemptions, the decision could be challenged in court and may not stand up. The lack of an appropriate exemption for de-identified genomic information jeopardizes the ability to control access to information held by the PMI cohort, undermining both respect for participant's wishes and the confidentiality of their information. Legislation that would prohibit releasing genomic information in response to a FOIA request would strengthen the ability of the government to deny such requests.

Data obtained by investigators from the PMI cohort will, in most cases be de-identified, which offers a degree of protection to research participants. However, there is a growing field of literature demonstrating myriad ways that individuals can be re-identified using "de-identified" information of various types, particularly through combining large amounts of information from multiple sources, and often involving genomic data.[147] In order to obtain such data, investigators should be required to agree not to re-identify or to attempt to recontact individuals. Unfortunately, the mechanisms available to the PMI-CP for enforcement of these agreements could be limited, especially if the data users are not funded by NIH. In order to provide the strongest enforcement mechanisms, legislation establishing penalties for violation of these agreements will be needed.

The PMI-CP may partner with developers of sensor technology and other devices to utilize or test these devices in the PMI cohort. The PMI-CP has a responsibility to ensure that data collected from those devices are protected from breach and misuse. There are currently very few regulations governing commercially available sensor devices and how the data they collect is used and protected. As their use continues to rise, the security and privacy of these data could become a concern for the public and policy makers. In order to encourage adoption of appropriate security and privacy measures, the PMI-CP should only enter into agreements with sensor technology developers that have appropriate security and privacy measures in place to safeguard device users' data.

Certificates of Confidentiality, which are issued by NIH upon request to institutions for individual research projects, allow them to refuse to disclose names or other identifying characteristics and data about research participants, including genetic information, in response to legal, civil, or administrative demands.[148] However, NIH-funded investigators are not required to obtain a Certificate of Confidentiality and even when they do, having one does not prevent investigators from voluntarily releasing any of the above-mentioned information or data. Both of these gaps weaken the overall

privacy protection afforded by Certificates of Confidentiality, and should be closed. NIH should modify its policy to require that any research using identifiable information be required to obtain a Certificate of Confidentiality, and should seek legislation amending the authority for certificates that would prohibit investigators holding a certificate from voluntarily releasing information other than with consent or for certain public health exceptions. This model would be similar to the Privacy Certificates issued by the National Institute of Justice.[149] Participants of the PMI cohort should feel confident that their information will not be used against them in a civil or criminal legal proceeding. Similarly, participants will need to be confident that their participation will not jeopardize their eligibility to receive benefits from government agencies to which they are entitled.

No security measures that permit the use of information can ever completely safeguard against the possibility of release of information or inappropriate use of information. There have been a number of very large information breaches recently, which have put the spotlight on cybersecurity and the trade-off between broad access and usability of information and strong data security. Participants need to be aware that there is a risk that their information may be disclosed or used inappropriately. Educating prospective participants about the extent to which existing laws protect them from misuse of their information, and the extent to which this legislation does not protect against misuse will allow them to make a more informed choice about participation. To prepare for any privacy or security breaches of information held by the PMI-CP, there will need to be a clear protocol for promptly notifying participants of the breach and steps individuals can take to lower the risk of harm resulting from that breach.

E. Sharing of data and specimens for research

Recommendation 7.17: Consistent with efforts outlined in Section 5, NIH should support ongoing Administration and private sector efforts to promote patient access to their health records, the exchange of health information across providers, and broad system interoperability of electronic health records.

Recommendation 7.18: The PMI-CP should only enter into agreements with sensor technology developers that have policies in place that allow individuals a right of access to their sensor device data and allows them to contribute it to the PMI. NIH should use the PMI cohort to facilitate the voluntary adoption of such policies among sensor technology developers.

Recommendation 7.19: The PMI-CP should only enter into agreements with sensor technology developers that will provide access for participants to, and interpretations of, sensor device data, including environmental data.

Recommendation 7.20: NIH should seek the explicit statutory authority to allow the NIH Director to require data sharing in the PMI-CP.

The PMI cohort will be the first of its kind in scale and scope in terms of the types and breadth of data collected. Providing researchers from all sectors with access to specimens and data from the PMI cohort will provide the greatest opportunity for learning about disease and health, treatment and prevention. This will also provide a unique opportunity for data science, advancing our collective understanding of managing and using data optimally for science.

The foundational source of information for the PMI cohort will be information from and access to participants' EHRs. Currently, there are diverse standards for EHRs that may prohibit direct combination of datasets and require curation. We have seen widespread adoption of EHRs in recent years, largely due to the efforts of ONC and incentive programs established by CMS. However, despite such progress, HPOs are still lacking broad data and system interoperability. Local examples of successful interoperability can provide important case studies but will need to expand greatly for the PMI cohort to operate in a seamless data exchange environment. The PMI-CP should take advantage of all progress in this area and should be used as a catalyst for moving the field to the extent possible. NIH should be at the table for policy decisions that affect the accessibility and portability of health data and encourage the use of the PMI cohort as a test bed for successful case studies.

Sensor devices are generating a wide variety of health information and health-related information (e.g., environmental data) that could be useful both for participants and for the PMI cohort to collect and provide to investigators for research. A more thorough consideration of the types of data that might be collected and the types of devices that might be used in the PMI cohort is made by the Working Group in Section 5. There is considerable variation in the extent to which these developers' policies permit individuals to access their own data, and to control the use or sharing of their data. PMI-CP should only enter into agreements with or accept data from developers that have policies in place that provide a right of data access to PMI cohort participants and allows them to contribute their sensor device data to the cohort. Furthermore, NIH should facilitate the voluntary adoption of such policies among sensor technology developers. In addition, participants in the cohort should receive access to data generated from them by sensor devices through accessible apps or interfaces that provide an interpretation of the data. In particular, the Working Group feels that participants should receive access to and interpretations of data generated from environmental sensors that could have an impact on their health.

Sharing research data is fundamental to the advancement of science and knowledge and is consistent with longstanding values and federal policies, including NIH policies, such as the 2003 Data Sharing Policy[150] and 2014 Genomic Data Sharing Policy.[151] The PMI-CP should adopt a position consistent with NIH policies and require data sharing for investigators who publish results of research conducted with specimens or data obtained from the PMI cohort (and with the limitation that data obtained from the cohort should not be shared further). Currently, NIH can require data sharing through terms and conditions established as part of funding, but this only becomes a requirement when an awardee accepts funding. Further, it is difficult to enforce once the funds have been spent, effectively removing the enforcement hook. Statutory authority that would allow the NIH Director to require data sharing related to all aspects of the PMI-CP, as necessary, would strengthen NIH's authority broadly and be beneficial for the PMI cohort and the advancement of all biomedical research funded by NIH.

F. Sharing of data and research results with participants

Recommendation 7.21: NIH should work with CMS, OCR, FDA, and others, and in coordination with the research and participant community, to enable PMI cohort participants to access research data about themselves. Specifically, clarity and consistency are needed about PMI-CP goals, the rights afforded by

HIPAA for access to research results, and restrictions under the Clinical Laboratory Improvement Amendments.

Building a true partnership with participants requires respecting participant preferences about the return of information and avoiding large disparities in data and information access between researchers and participants. Sharing information, including both data generated from biospecimens and new interpretations of data resulting from research, is critical to the success of the PMI-CP, and we know that this is what potential participants value most. In a recent survey of a representative U.S. population supported by the FNIH, 90% of respondents indicated that "learning information about my health" would be very or somewhat important in deciding whether or not to participate in a cohort study like the PMI cohort.[99] On July 22, 2015, Secretary's Advisory Committee on Human Research Protections (SACHRP) held a public discussion on the "Return of Research Results and Emergent CLIA and HIPAA Issues." This conversation focused on the February 6, 2014, rule entitled "CLIA Program and the HIPAA Privacy Rule: Patients Access to Test Reports," published jointly by CMS and the Office for Civil Rights (OCR). There have been recent attempts to clarify that HIPAA's right of access to data extends to laboratory test results, and SACHRP suggested that the U.S. Department of Health and Human Services (HHS) may have intended to allow an individual to request laboratory test results for a purpose other than treatment, which creates an apparent contradiction with CMS's interpretation that test results from non-CLIA-certified laboratories should not be shared with research participants.

NIH should support conversations within HHS to ensure that the core tenets of progress in precision medicine and the values of giving individuals access to their data are represented appropriately and responsibly in HHS regulations, specifically for the Privacy Act, HIPAA and CLIA, and that these policies do not conflict with each other. More information on Working Group recommendations for returning individual PMI cohort participant results can be found in Section 4.

G. Areas for study/evaluation of participant preferences and for identification of policy gaps

Recommendation 7.22: The PMI-CP should conduct or support research to understand participants' preferences. These evaluations will help to elucidate any additional policy gaps and responses needed to protect participants and encourage strong collaborative participation.

- *Determine mode and frequency of communication about the PMI-CP that participants desire.*
- *If families are recruited, consider how to maintain autonomy of individual family members, for example, if remote monitors are used in the home.*
- *Consider how to protect the privacy and autonomy of "collateral participants" in addition to family members that might be created through the unintentional collection of identifiable information of non-participants.*
- *Consider how to provide meaningful contacts for participants, such as whether the PMI-CP should provide central as well as local contacts. Participants must feel that they have a comfortable contact process for questions and concerns.*
- *Consider how to ensure an honest and fair recruitment process.*

The nature of the relationship between participants and researchers has changed dramatically in recent decades. This evolution is one of the reasons cited for the proposed revisions to the Common Rule and has been noted in many recent policies, reports, and in the academic literature. It can also be seen in the establishment of novel research programs that put a premium on participant and patient engagement, such as the PCORnet Patient-Powered Research Networks and PatientsLikeMe, and in private sector efforts, including 23andMe and Sage Bionetworks. We continue to learn more about how different types of partnerships work and where a new model is needed. The PMI cohort will provide a unique platform from which to learn about participant preferences for various aspects of research, and to identify policy gaps or hurdles that affect the vision of this highly evolving research environment. NIH should set the priorities for and support data-driven research studies that can drive policy change, at a local and national level.

Section 8 – Governance

The governance structure for the PMI-CP must combine effective and timely decision making with opportunities for consultation with stakeholders. This will require nimble and innovative approaches to governance. The Working Group considered the governance structures of a variety of other large projects and cohorts, such as the UK Biobank[152] and other international cohorts, the National Children's Study, and PCORnet. The Working Group recognizes that NIH will need to explore the relationship between the governance structure proposed here and the structure typically found in NIH's cooperative agreements and other funding mechanisms. Given the pace of scientific developments in the area of genome sciences and the complex challenges that establishment of the PMI cohort will require, success will be highly dependent upon innovation, entrepreneurship, and creative problem-solving. The governance must be nimble to ensure that the project can take advantage of new technologies and novel research techniques, adjust to new vulnerabilities in data security, and respond changes in participant preferences. This may require a new model for defining how awardees to relate to each other and to the NIH oversight structure described in this section.

A. PMI Cohort Program director

Recommendation 8.1: The PMI-CP should be led by a director with the institutional authority, professional expertise, and structural support to provide strong, credible, and effective leadership.

The PMI-CP will require a leader who is granted the institutional authority and structural support and who has the appropriate scientific expertise needed to effectively make decisions, and to create and sustain momentum. For this reason, the Working Group recommends that NIH appoint a director to lead the PMI-CP. The Working Group recommends that the director have demonstrated interdisciplinary expertise, strong management skills, exceptional talents working with diverse stakeholders, and experience in the promotion of effective collaboration with multiple stakeholders (especially participants and their representatives). In addition, the director should also have a demonstrated track record of scientific accomplishment and vision, with success in tackling large-scale problems with new and innovative approaches.

The director will play a central role in building the capacity, credibility, and prestige of this critical national effort, and hold the responsibility for advancing the scientific vision and ensuring the highest standards for science conducted in the PMI cohort. The director's roles will include (1) leadership of the Steering Committee for the PMI cohort and the Executive Committee for the Steering Committee; (2) directing the NIH PMI-CP office; (3) leading planning and decision making about new funding opportunities, collaborations, and funding plans; and (4) coordinating with NIH Institute and Center (IC) leadership and offices, external stakeholders, and other federal agencies. The director will be responsible for making all final decisions on the PMI cohort, taking into account the advice and inputs of the committees described here, and with the concurrence of the NIH Director:

- The director must have authority to develop, in consultation with the Steering Committee (including NIH staff), both short-term and long-term plans regarding the timing of the creation

of the PMI cohort's required capabilities, as well as the scope of individual components, in order to balance the PMI's aims within the available personnel and financial resources.

- The director must be able to work in close coordination with the Steering Committee, IC directors, and NIH Director at all stages of planning, program implementation, funding, and evaluation to ensure that the PMI cohort fulfills its role to generate critical knowledge to advance American health.
- The director must have substantial input into the requirements included in solicitations, and should have final signoff on awards to be recommended to Council.
- The director must have authority to determine whether awardees have met applicable performance requirements and to authorize continuation of their awards.
- The director must have authority regarding setting implementation standards (for instance including the content of a CDM), developing and maintaining the system infrastructure, including both the centralized and distributed components.

Based on these roles and responsibilities, the Working Group considered possible relationships of the director to the NIH, most fundamentally whether the director would be hired as federal staff or not. Given the high-profile nature of the PMI-CP and the national interest in its success, and the critical importance of the director's ability to determine the vision, pace, funding, evaluation, and other key decisions for the PMI-CP, the Working Group recommends that NIH ensure that the mechanism for hiring the director allow him or her to fully meet all the functions and responsibilities detailed here. The Working Group also recommends that the director be responsible to the NIH Director. An office with dedicated staff to support the PMI-CP director and to oversee all aspects of programmatic implementation should be established. This might include the creation of specific roles reporting to the PMI-CP director, such a chief engagement officer, chief technology officer, chief marketing officer, chief medical officer, etc. NIH should consider where such an office should be situated. If located in one of the Institutes and Centers (ICs), the Working Group recommends an IC with substantial programmatic expertise in the oversight of large clinical research projects providing support for program implementation (including interfacing with IC grants management and peer review), development of public private partnerships, policy development, program evaluation, and communications and outreach.

B. PMI-CP Steering Committee

Recommendation 8.2: The PMI-CP director must provide leadership for a Steering Committee representing critical stakeholders, and consider the advice and recommendations of a small Executive Committee selected from Steering Committee members.

The Working Group recommends the creation of a Steering Committee to provide coordination of the activities of the PMI-CP. The Steering Committee should include awardees, research participants and their representatives, academic and private researchers who will use the PMI cohort platform, and NIH programmatic staff. The Steering Committee will be chaired by the PMI-CP director and will report to an Executive Committee, which will be a small group of the Steering Committee members and also chaired by the PMI-CP director. The Executive Committee will ensure seamless development and

implementation of the PMI cohort across awardees, identify and explore solutions to challenges and obstacles of this goal, and make recommendations to the PMI-CP director based on proposed research in the cohort.

The Steering Committee will form a number of subcommittees that report to the Steering Committee. Several subcommittees have been identified in this report: Data Subcommittee, Resource Access Subcommittee, Return of Results and Information Subcommittee, Biobanking Subcommittee, and Security Subcommittee. Each subcommittee will be responsible for a given operational domain and must work in concert with other committees under the direction of the Steering Committee to achieve the operation objectives. The Working Group considers cohort participants to be exceptionally important to the Steering Committee and Executive Committee, and recommends strong participant representation on both committees and any subcommittees or working groups they establish.

C. Coordination and oversight of the PMI cohort

Recommendation 8.3: An Independent Advisory Board should be established to provide external oversight for the PMI-CP.

An Independent Advisory Board known for rigor and integrity will be essential for refining and reinvigorating over time the PMI-CP's vision, scientific and clinical goals, and operations. Such an advisory group should be composed of experts in areas of relevance to the PMI-CP and should be charged with performing the functions undertaken by councils for the ICs, including recommendations for funding plans and secondary review, as well as on-going advice and evaluation. Such a group should report to the PMI-CP director and the NIH Director. The Working Group recommends that NIH consider the multicouncil working group structure established to provide oversight of the NIH Brain Research through Advancing Innovative Neurotechnologies (BRAIN) Initiative.

Recommendation 8.4: Cross-agency coordination is essential and should continue to be a component of the governance structure.

Cross-agency coordination is a central feature of the PMI and has been critical to planning the PMI cohort. Providing a mechanism for continued cross-agency planning, coordination, and implementation will be critical for the success of the initiative. Key agencies for such interactions include the NIH (NIH Director and PMI-CP director), the Health Resources and Services Administration, FDA, ONC, VA including the MVP, and DoD. Such cross-agency coordination would be advisory on policy development and provide active consultation to identify opportunities to leverage resources, and programs, but should not have a direct governance role in the PMI cohort implementation.

Recommendation 8.5: Final oversight authority for the cohort should reside with the NIH Director.

Final authority for policy determination, priority setting, and oversight of the implementation should reside with the NIH Director, as advised by the PMI-CP director and the governance bodies described above.

Concluding Remarks

Advances in health require understanding the factors contributing to wellness and disease in individuals, coupled with the ability to use this knowledge to develop new effective means of disease prevention and therapy, along with the ability to deliver the fruits of these advances to the people and populations that will benefit. The Working Group has considered a broad range of issues related to the utility, feasibility, and execution of a cohort study, the Precision Medicine Initiative Cohort Project (PMI-CP), of one million or more engaged participants that is inclusive of American demographic diversity. The Working Group concluded that dramatic advances in technology over the last decade have now made cost-effective and feasible the recruitment of these participants, the collection in electronic form of their comprehensive health records, the collection of diverse types of experimental data relevant to understanding current health and predictive of future health outcome in individuals, and the ability to perform innovative analyses of these very large orthogonal data sets to identify fundamental new mechanisms that contribute to individual health outcomes.

With a plan to follow health outcomes of participants over many years, the Working Group anticipates the PMI-CP will be powered to identify biomarkers that are predictive of future development of a large number of diseases, affording new opportunity for disease prevention and therapy, as well as to provide new understanding of the factors that predict variation in response to current therapies for prevalent disease. Moreover, a design that allows participants to be recontacted for further study based on individual findings provides an invaluable opportunity to understand biological mechanisms that link biomarkers to traits in individuals.

Critical to the success of this effort will be effective engagement and empowerment of PMI cohort participants to be full partners in the design and execution of the PMI-CP. By encouraging ostensibly any individual in the population to volunteer to participate, we believe the PMI-CP has the potential to galvanize a national effort focused on advancing individual health through collective efforts at a national scale.

Similarly, the wealth of data that will reside in the PMI cohort will provide exceptional opportunity innovative analyses, and will require mechanisms to provide ready access to data to the diverse investigator community while maintaining the highest standards for data security and maintenance of privacy of participants.

A project of this scale and scope has myriad details to consider and address that will require exceptional organization and leadership. The PMI-CP will undoubtedly need to draw upon diverse talents in academia, industry, health care organizations, government, and the participant communities to further advance the planning, design and execution of this project. The PMI-CP will also require a long-term budget commitment in order to succeed as a research foundation upon which to advance precision medicine. After careful consideration, the Working Group is unanimous and enthusiastic in supporting this endeavor. We are convinced that the time is right to mount this ambitious project to transform the

understanding of factors contributing to individual health and disease, with conviction that success in this effort will advance the health of the U.S.

References

1. Roychowdhury S, Chinnaiyan AM. Translating genomics for precision cancer medicine. Annu Rev Genomics Hum Genet 2014;15:395–415.

2. Eckford PDW, Li C, Ramjeesingh M, Bear CE. Cystic fibrosis transmembrane conductance regulator (CFTR) potentiator VX-770 (ivacaftor) opens the defective channel gate of mutant CFTR in a phosphorylation-dependent but ATP-independent manner. J Biol Chem 2012;287(44):36639–49.

3. Food and Drug Administration. Genomics - Table of Pharmacogenomic Biomarkers in Drug Labeling [Internet]. [cited 2015 Aug 21];Available from: http://www.fda.gov/drugs/scienceresearch/researchareas/pharmacogenetics/ucm083378.htm

4. Mallal S, Phillips E, Carosi G, et al. HLA-B*5701 screening for hypersensitivity to abacavir. N Engl J Med 2008;358(6):568–79.

5. McCormack M, Alfirevic A, Bourgeois S, et al. HLA-A*3101 and carbamazepine-induced hypersensitivity reactions in Europeans. N Engl J Med 2011;364(12):1134–43.

6. Relling MV, Gardner EE, Sandborn WJ, et al. Clinical Pharmacogenetics Implementation Consortium guidelines for thiopurine methyltransferase genotype and thiopurine dosing. Clin Pharmacol Ther 2011;89(3):387–91.

7. Shuldiner AR, O'Connell JR, Bliden KP, et al. Association of cytochrome P450 2C19 genotype with the antiplatelet effect and clinical efficacy of clopidogrel therapy. JAMA 2009;302(8):849–57.

8. The International Warfarin Pharmacogenetics Consortium. Estimation of the Warfarin Dose with Clinical and Pharmacogenetic Data. N Engl J Med 2009;360(8):753–64.

9. Kimmel SE, French B, Kasner SE, et al. A Pharmacogenetic versus a Clinical Algorithm for Warfarin Dosing. N Engl J Med 2013;369(24):2283–93.

10. Pulley JM, Denny JC, Peterson JF, et al. Operational implementation of prospective genotyping for personalized medicine: The design of the Vanderbilt PREDICT project. Clin Pharmacol Ther 2012;92(1):87–95.

11. Bainbridge MN, Wiszniewski W, Murdock DR, et al. Whole-genome sequencing for optimized patient management. Sci Transl Med 2011;3(87):87re3.

12. Worthey EA, Mayer AN, Syverson GD, et al. Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. Genet Med Off J Am Coll Med Genet 2011;13(3):255–62.

13. Tripp S, Grueber M. The Economic Impact of the Human Genome Project [Internet]. Battelle Technology Partnership Practice for United for Medical Research; 2011. Available from: http://battelle.org/docs/default-document-library/economic_impact_of_the_human_genome_project.pdf

14. Dzau VJ, Fineberg HV. Restore the US lead in biomedical research. JAMA 2015;313(2):143–4.

15. Moses H, Matheson DHM, Cairns-Smith S, George BP, Palisch C, Dorsey ER. The anatomy of medical research: US and international comparisons. JAMA 2015;313(2):174–89.

16. National Research Council (US) Committee on A Framework for Developing a New Taxonomy of Disease. Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease [Internet]. Washington (DC): National Academies Press (US); 2011 [cited 2015 Aug 21]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK91503/

17. Plenge RM, Scolnick EM, Altshuler D. Validating therapeutic targets through human genetics. Nat Rev Drug Discov 2013;12(8):581–94.

18. Okada Y, Wu D, Trynka G, et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. Nature 2014;506(7488):376–81.

19. Sanseau P, Agarwal P, Barnes MR, et al. Use of genome-wide association studies for drug repositioning. Nat Biotechnol 2012;30(4):317–20.

20. Nelson MR, Tipney H, Painter JL, et al. The support of human genetic evidence for approved drug indications. Nat Genet 2015;47(8):856–60.

21. Myocardial Infarction Genetics Consortium Investigators, Stitziel NO, Won H-H, et al. Inactivating mutations in NPC1L1 and protection from coronary heart disease. N Engl J Med 2014;371(22):2072–82.

22. Holmes MV, Asselbergs FW, Palmer TM, et al. Mendelian randomization of blood lipids for coronary heart disease. Eur Heart J 2015;36(9):539–50.

23. Heart Disease Facts & Statistics | cdc.gov [Internet]. [cited 2015 Sep 8];Available from: http://www.cdc.gov/heartdisease/facts.htm

24. WHO | Cardiovascular diseases (CVDs) [Internet]. [cited 2015 Sep 8];Available from: http://www.who.int/mediacentre/factsheets/fs317/en/

25. Collins FS. The case for a US prospective cohort study of genes and environment. Nature 2004;429(6990):475–7.

26. Hospitals Participating in the CMS EHR Incentive Programs [Internet]. [cited 2015 Aug 12];Available from: http://dashboard.healthit.gov/quickstats/pages/FIG-Hospitals-EHR-Incentive-Programs.php

27. ICT Facts and Figures – The world in 2015 [Internet]. ITU. [cited 2015 Aug 21];Available from: http://www.itu.int/en/ITU-D/Statistics/Pages/facts/default.aspx

28. Duggan M. Cell Phone Activities 2013 [Internet]. Pew Res. Cent. Internet Sci. Tech. [cited 2015 Aug 21];Available from: http://www.pewinternet.org/2013/09/19/cell-phone-activities-2013/

29. Smith A. U.S. Smartphone Use in 2015 [Internet]. Pew Res. Cent. Internet Sci. Tech. [cited 2015 Sep 2];Available from: http://www.pewinternet.org/2015/04/01/us-smartphone-use-in-2015/

30. Remarks by the President in State of the Union Address | January 20, 2015 [Internet]. whitehouse.gov. [cited 2015 Sep 8];Available from: https://www.whitehouse.gov/the-press-office/2015/01/20/remarks-president-state-union-address-january-20-2015

31. Obama calls for major new personalized medicine initiative [Internet]. Reuters. 2015 [cited 2015 Aug 21];Available from: http://www.reuters.com/article/2015/01/21/us-usa-obama-genomics-idUSKBN0KU06L20150121

32. NCI and the Precision Medicine Initiative [Internet]. Natl. Cancer Inst. [cited 2015 Aug 21];Available from: http://www.cancer.gov/research/key-initiatives/precision-medicine

33. The President's Budget for Fiscal Year 2016 [Internet]. White House. [cited 2015 Aug 21];Available from: https://www.whitehouse.gov/node/18050

34. PMI Working Group - Precision Medicine Initiative - National Institutes of Health (NIH) [Internet]. [cited 2015 Aug 21];Available from: http://www.nih.gov/precisionmedicine/working-group.htm

35. ACD Precision Medicine Initiative Working Group Public Workshop: Unique Scientific Opportunities for the Precision Medicine Initiative National Research Cohort [Internet]. [cited 2015 Aug 21];Available from: http://www.nih.gov/precisionmedicine/workshop-20150428.htm

36. ACD Precision Medicine Initiative Working Group Public Workshop: Digital Health Data in a Million-Person Precision Medicine Initiative Cohort [Internet]. [cited 2015 Aug 21];Available from: http://www.nih.gov/precisionmedicine/workshop-20150528.htm

37. ACD Precision Medicine Initiative Working Group Public Workshop: Participant Engagement and Health Equity Workshop [Internet]. [cited 2015 Aug 21];Available from: http://www.nih.gov/precisionmedicine/workshop-20150701.htm

38. Mobile and Personal Technologies in Precision Medicine Workshop - Precision Medicine Initiative Cohort [Internet]. [cited 2015 Aug 21];Available from: http://www.nih.gov/precisionmedicine/workshop-20150727.htm

39. NIH. Summary of Responses from the Request for Information on Building the Precision Medicine Initiative National Research Participant Group [Internet]. 2015. Available from: http://www.nih.gov/precisionmedicine/2015-05-15RFISummary.pdf

40. NIH. Request for Information: NIH Precision Medicine Cohort - Strategies to Address Community Engagement and Health Disparities. 2015.

41. United States Tops 4 Billion Annual Prescriptions: Is Our Health Improving? | HealthyConsumer.com [Internet]. [cited 2015 Aug 21];Available from: http://www.healthyconsumer.com/911/united-states-tops-4-billion-annual-prescriptions-is-our-health-improving/

42. FastStats [Internet]. [cited 2015 Aug 21];Available from: http://www.cdc.gov/nchs/fastats/drug-use-therapeutic.htm

43. Sarkar U, López A, Maselli JH, Gonzales R. Adverse drug events in U.S. adult ambulatory medical care. Health Serv Res 2011;46(5):1517–33.

44. Pharmacogenomic Biomarkers in Drug Labels. [Internet]. FDA Pharmacogenomic Biomark. Drug Labeling. [cited 2014 Jan 22];Available from: http://www.fda.gov/Drugs/ScienceResearch/ResearchAreas/Pharmacogenetics/ucm083378.htm

45. Giancardo L, Sánchez-Ferro A, Butterworth I, Mendoza CS, Hooker JM. Psychomotor impairment detection via finger interactions with a computer keyboard during natural typing. Sci Rep 2015;5:9678.

46. Go AS, Mozaffarian D, Roger VL, et al. Executive summary: heart disease and stroke statistics--2014 update: a report from the American Heart Association. Circulation 2014;129(3):399–410.

47. Alcantara D, O'Driscoll M. Congenital microcephaly. Am J Med Genet C Semin Med Genet 2014;166C(2):124–39.

48. Cohen JC, Boerwinkle E, Mosley TH, Hobbs HH. Sequence Variations in PCSK9, Low LDL, and Protection against Coronary Heart Disease. N Engl J Med 2006;354(12):1264–72.

49. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. Nature 2012;490(7418):61–70.

50. Chapman PB, Hauschild A, Robert C, et al. Improved survival with vemurafenib in melanoma with BRAF V600E mutation. N Engl J Med 2011;364(26):2507–16.

51. Doshi-Velez F, Ge Y, Kohane I. Comorbidity Clusters in Autism Spectrum Disorders: An Electronic Health Record Time-Series Analysis. Pediatrics 2013;peds.2013–0819.

52. Ahmad T, Pencina MJ, Schulte PJ, et al. Clinical Implications of Chronic Heart Failure Phenotypes Defined by Cluster Analysis. J Am Coll Cardiol 2014;64(17):1765–74.

53. Vanlerberghe C, Petit F, Malan V, et al. 15q11.2 microdeletion (BP1-BP2) and developmental delay, behaviour issues, epilepsy and congenital heart disease: a series of 52 patients. Eur J Med Genet 2015;58(3):140–7.

54. Zaidi S, Choi M, Wakimoto H, et al. De novo mutations in histone-modifying genes in congenital heart disease. Nature 2013;498(7453):220–3.

55. Ellis PM, Coakley N, Feld R, Kuruvilla S, Ung YC. Use of the epidermal growth factor receptor inhibitors gefitinib, erlotinib, afatinib, dacomitinib, and icotinib in the treatment of non-small-cell lung cancer: a systematic review. Curr Oncol Tor Ont 2015;22(3):e183–215.

56. DePeralta DK, Boland GM. Melanoma: Advances in Targeted Therapy and Molecular Markers. Ann Surg Oncol 2015;22(11):3451–8.

57. O'Reilly R, Elphick HE. Development, clinical utility, and place of ivacaftor in the treatment of cystic fibrosis. Drug Des Devel Ther 2013;7:929–37.

58. Solomon GM, Marshall SG, Ramsey BW, Rowe SM. Breakthrough therapies: Cystic fibrosis (CF) potentiators and correctors. Pediatr Pulmonol 2015;50 Suppl 40:S3–13.

59. Sudlow C, Gallacher J, Allen N, et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLoS Med 2015;12(3):e1001779.

60. Chen Z, Chen J, Collins R, et al. China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up. Int J Epidemiol 2011;40(6):1652–66.

61. Leitsalu L, Alavere H, Tammesoo M-L, Leego E, Metspalu A. Linking a population biobank with national health registries-the estonian experience. J Pers Med 2015;5(2):96–106.

62. The White House. Precision Medicine Initiative: Proposed Privacy and Trust Principles [Internet]. Available from: https://www.whitehouse.gov/sites/default/files/docs/pmi_privacy_and_trust_principles_july_2015.pdf

63. McGregor TL, Van Driest SL, Brothers KB, Bowton EA, Muglia LJ, Roden DM. Inclusion of pediatric samples in an opt-out biorepository linking DNA to de-identified medical records: pediatric BioVU. Clin Pharmacol Ther 2013;93(2):204–11.

64. Roden DM, Pulley JM, Basford MA, et al. Development of a large-scale de-identified DNA biobank to enable personalized medicine. Clin Pharmacol Ther 2008;84(3):362–9.

65. Denny JC, Bastarache L, Ritchie MD, et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. Nat Biotechnol 2013;31(12):1102–11.

66. UK Biobank. UK Biobank: Protocol for a large-scale prospective epidemiological resource [Internet]. Available from: http://www.ukbiobank.ac.uk/wp-content/uploads/2011/11/UK-Biobank-Protocol.pdf

67. Bustamante CD, Burchard EG, De La Vega FM. Genomics for the world. Nature 2011;475(7355):163–5.

68. Burchard EG, Ziv E, Coyle N, et al. The importance of race and ethnic background in biomedical research and clinical practice. N Engl J Med 2003;348(12):1170–5.

69. Braveman PA, Kumanyika S, Fielding J, et al. Health Disparities and Health Equity: The Issue Is Justice. Am J Public Health 2011;101(S1):S149–55.

70. Shuldiner AR, O'Connell JR, Bliden KP, et al. Association of cytochrome P450 2C19 genotype with the antiplatelet effect and clinical efficacy of clopidogrel therapy. JAMA 2009;302(8):849–57.

71. Wu AH, White MJ, Oh S, Burchard E. The Hawaii clopidogrel lawsuit: the possible effect on clinical laboratory testing. Pers Med 2015;12(3):179–81.

72. Mega JL, Simon T, Collet J-P, et al. Reduced-function CYP2C19 genotype and risk of adverse clinical outcomes among patients treated with clopidogrel predominantly for PCI: a meta-analysis. JAMA J Am Med Assoc 2010;304(16):1821–30.

73. Chiao JY, Cheon BK. The weirdest brains in the world. Behav Brain Sci 2010;33(2-3):88–90.

74. Henrich J, Heine SJ, Norenzayan A. The weirdest people in the world? Behav Brain Sci 2010;33(2-3):61–83; discussion 83–135.

75. Williams DR, Wyatt R. Racial bias in health care and health: Challenges and opportunities. JAMA 2015;314(6):555–6.

76. Miranda ML, Edwards SE, Keating MH, Paul CJ. Making the environmental justice grade: the relative burden of air pollution exposure in the United States. Int J Environ Res Public Health 2011;8(6):1755–71.

77. Mott L. The disproportionate impact of environmental health threats on children of color. Environ Health Perspect 1995;103(Suppl 6):33–5.

78. Metzger R, Delgado JL, Herrell R. Environmental health and Hispanic children. Environ Health Perspect 1995;103 Suppl 6:25–32.

79. Leong AB, Ramsey CD, Celedón JC. The challenge of asthma in minority populations. Clin Rev Allergy Immunol 2012;43(1-2):156–83.

80. Akinbami LJ, Moorman JE, Bailey C, et al. Trends in asthma prevalence, health care use, and mortality in the United States, 2001-2010. NCHS Data Brief 2012;(94):1–8.

81. LaVeist TA, Gaskin DJ, Richard P. THE ECONOMIC BURDEN OF HEALTH INEQUALITIES IN THE UNITED STATES [Internet]. 2009 [cited 2015 Aug 16];Available from: http://health-equity.pitt.edu/3797/

82. Auerbach JA, Krimgold BK, editors. Income, Socioeconomic Status, and Health: Exploring the Relationships. Washington, D.C: Academy for Health Services Research and Health Policy; 2001.

83. Public Law 107 - 280 - Rare Diseases Act of 2002 [Internet]. [cited 2015 Sep 8];Available from: http://www.gpo.gov/fdsys/pkg/PLAW-107publ280/content-detail.html

84. Rare Disease Information - NORD (National Organization for Rare Disorders) [Internet]. NORD Natl. Organ. Rare Disord. [cited 2015 Sep 8];Available from: http://rarediseases.org/for-patients-and-families/information-resources/rare-disease-information/

85. Kho AN, Pacheco JA, Peissig PL, et al. Electronic Medical Records for Genetic Research: Results of the eMERGE Consortium. Sci Transl Med 2011;3(79):79re1.

86. Crawford DC, Crosslin DR, Tromp G, et al. eMERGEing progress in genomics-the first seven years. Front Genet 2014;5:184.

87. Bowton E, Field JR, Wang S, et al. Biobanks and electronic medical records: enabling cost-effective research. Sci Transl Med 2014;6(234):234cm3.

88. Kohane IS. Using electronic health records to drive discovery in disease genomics. Nat Rev Genet 2011;12(6):417–28.

89. Richesson RL, Hammond WE, Nahm M, et al. Electronic health records based phenotyping in next-generation clinical trials: a perspective from the NIH Health Care Systems Collaboratory. J Am Med Inform Assoc 2013;20(e2):e226–31.

90. Collins FS, Hudson KL, Briggs JP, Lauer MS. PCORnet: turning a dream into reality. J Am Med Inform Assoc JAMIA 2014;21(4):576–7.

91. Kheterpal S. Clinical Research Using an Information System: The Multicenter Perioperative Outcomes Group. Anesthesiol Clin 2011;29(3):377–88.

92. Health care systems research network [Internet]. [cited 2015 Aug 12];Available from: http://www.hcsrn.org/en/

93. Gottesman O, Kuivaniemi H, Tromp G, et al. The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. Genet Med Off J Am Coll Med Genet 2013;15(10):761–71.

94. Ritchie MD, Denny JC, Crawford DC, et al. Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. Am J Hum Genet 2010;86(4):560–72.

95. Kho AN, Hayes MG, Rasmussen-Torvik L, 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 2012;19(2):212–8.

96. Kiyota Y, Schneeweiss S, Glynn RJ, Cannuscio CC, Avorn J, Solomon DH. Accuracy of Medicare claims-based diagnosis of acute myocardial infarction: estimating positive predictive value on the basis of review of hospital records. Am Heart J 2004;148(1):99–104.

97. PCORnet Common Data Model (CDM) - PCORnet [Internet]. [cited 2015 Aug 20];Available from: http://www.pcornet.org/pcornet-common-data-model/

98. Green LA, Fryer GE, Yawn BP, Lanier D, Dovey SM. The Ecology of Medical Care Revisited. N Engl J Med 2001;344(26):2021–5.

99. Kaufman D. Public Attitudes about the PMI Cohort Study: Results of a National Survey. Participant Engagement and Health Equity Workshop. [Internet]. Available from: http://videocast.nih.gov/summary.asp?Live=16498&bhcp=1

100. Kaufman D, Murphy J, Scott J, Hudson K. Subjects matter: a survey of public opinions about a large genetic cohort study. Genet Med 2008;10(11):831–9.

101. Murphy J, Scott J, Kaufman D, Geller G, LeRoy L, Hudson K. Public Expectations for Return of Results from Large-cohort Genetic Research. Am J Bioeth AJOB 2008;8(11):36–43.

102. Wallerstein N, Duran B. Community-Based Participatory Research Contributions to Intervention Research: The Intersection of Science and Practice to Improve Health Equity. Am J Public Health 2010;100(Suppl 1):S40–6.

103. Jagosh J, Macaulay AC, Pluye P, et al. Uncovering the Benefits of Participatory Research: Implications of a Realist Review for Health Research and Practice. Milbank Q 2012;90(2):311–46.

104. Nass SJ, Levit LA, Gostin LO, Rule I of M (US) C on HR and the P of HITHP. The Value, Importance, and Oversight of Health Research. 2009 [cited 2015 Sep 8];Available from: http://www.ncbi.nlm.nih.gov/books/NBK9571/

105. Responsible Research: A Systems Approach to Protecting Research Participants [Internet]. [cited 2015 Sep 8];Available from: http://www.nap.edu/openbook.php?record_id=10508&page=2

106. Tanner A, Kim S-H, Friedman DB, Foster C, Bergeron CD. Barriers to medical research participation as perceived by clinical trial investigators: communicating with rural and african american communities. J Health Commun 2015;20(1):88–96.

107. Holman H, Lorig K. Patients as partners in managing chronic disease. BMJ 2000;320(7234):526–7.

108. Crossing the Quality Chasm: A New Health System for the 21st Century [Internet]. [cited 2015 Sep 8];Available from: http://www.nap.edu/openbook.php?record_id=10027&page=R3

109. Green RC, Berg JS, Grody WW, et al. ACMG Recommendations for Reporting of Incidental Findings in Clinical Exome and Genome Sequencing. Genet Med Off J Am Coll Med Genet 2013;15(7):565–74.

110. Capturing Social and Behavioral Domains and Measures in Electronic Health Records: Phase 2 - Institute of Medicine [Internet]. [cited 2015 Aug 20];Available from: http://iom.nationalacademies.org/Reports/2014/EHRdomains2.aspx

111. Platform for Engaging Everyone Responsibly | GeneticAlliance.org [Internet]. [cited 2015 Aug 22];Available from: http://www.geneticalliance.org/programs/biotrust/peer

112. Wei W-Q, Denny JC. Extracting research-quality phenotypes from electronic health records to support precision medicine. Genome Med [Internet] 2015 [cited 2015 Jun 23];7(1). Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4416392/

113. Newton KM, Peissig PL, Kho AN, et al. Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. J Am Med Inform Assoc 2013;20(e1):e147–54.

114. What is the Phenotype KnowledgeBase? | PheKB [Internet]. [cited 2015 Aug 20];Available from: https://phekb.org/

115. Your Health Data | Patients & Families | HealthIT.gov [Internet]. [cited 2015 Aug 20];Available from: http://healthit.gov/patients-families/your-health-data

116. Pledge Members | | HealthIT.gov [Internet]. [cited 2015 Sep 10];Available from: http://www.healthit.gov/patients-families/pledge-members

117. Federal Register | 2015 Edition Health Information Technology (Health IT) Certification Criteria, 2015 Edition Base Electronic Health Record (EHR) Definition, and ONC Health IT Certification Program Modifications [Internet]. [cited 2015 Sep 10];Available from: https://www.federalregister.gov/articles/2015/03/30/2015-06612/2015-edition-health-information-technology-health-it-certification-criteria-2015-edition-base

118. Federal Register | Medicare and Medicaid Programs; Electronic Health Record Incentive Program-Stage 3 [Internet]. [cited 2015 Sep 10];Available from: https://www.federalregister.gov/articles/2015/03/30/2015-06685/medicare-and-medicaid-programs-electronic-health-record-incentive-program-stage-3

119. Argonauts - FHIR v0.4.0 [Internet]. [cited 2015 Aug 20];Available from: http://hl7.org/implement/standards/fhir/2015Jan/argonauts.html

120. Preliminary report: effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction. The Cardiac Arrhythmia Suppression Trial (CAST) Investigators. N Engl J Med 1989;321(6):406–12.

121. Murphy SN, Weber G, Mendis M, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc JAMIA 2010;17(2):124–30.

122. i2b2: Informatics for Integrating Biology & the Bedside [Internet]. [cited 2012 Dec 29];Available from: https://www.i2b2.org/work/i2b2_installations.html

123. http://www.ukbiobank.ac.uk/data-showcase/ [Internet]. [cited 2015 Sep 8];Available from: http://www.ukbiobank.ac.uk/data-showcase/

124. Erlich Y, Narayanan A. Routes for breaching and protecting genetic privacy. Nat Rev Genet 2014;15(6):409–21.

125. NCI Cancer Genomics Cloud Pilots — CBIIT: Welcome to the NCI Center for Biomedical Informatics and Information Technology [Internet]. [cited 2015 Aug 20];Available from: https://cbiit.nci.nih.gov/ncip/nci-cancer-genomics-cloud-pilots

126. CMS Virtual Research Data Center (VRDC) [Internet]. [cited 2015 Aug 20];Available from: http://www.resdac.org/cms-data/request/cms-virtual-research-data-center

127. OHDSI/CommonDataModel · GitHub [Internet]. [cited 2015 Aug 21];Available from: https://github.com/OHDSI/CommonDataModel

128. PubChemRDF Release Notes [Internet]. [cited 2015 Aug 21];Available from: https://pubchem.ncbi.nlm.nih.gov/rdf/

129. Work Products & Demonstration Projects | Global Alliance for Genomics and Health [Internet]. [cited 2015 Aug 20];Available from: https://genomicsandhealth.org/work-products-demonstration-projects

130. Federal Information Security Management Act (FISMA) | Homeland Security [Internet]. [cited 2015 Aug 20];Available from: http://www.dhs.gov/federal-information-security-management-act-fisma

131. White House. Precision Medicine Initiative: Proposed Privacy and Trust Principles [Internet]. Available from: https://www.whitehouse.gov/sites/default/files/docs/pmi_privacy_and_trust_principles_july_2015.pdf

132. OHRP. Federal Policy for the Protection of Human Subjects ('Common Rule') [Internet]. [cited 2015 Sep 2];Available from: http://www.hhs.gov/ohrp/humansubjects/commonrule/

133. Green LA, Lowery JC, Kowalski CP, Wyszewianski L. Impact of Institutional Review Board Practice Variation on Observational Health Services Research. Health Serv Res 2006;41(1):214–30.

134. Dziak K, Anderson R, Sevick MA, Weisman CS, Levine DW, Scholle SH. Variations among Institutional Review Board Reviews in a Multisite Health Services Research Study. Health Serv Res 2005;40(1):279–90.

135. Wagner TH, Murray C, Goldberg J, Adler JM, Abrams J. Costs and Benefits of the National Cancer Institute Central Institutional Review Board. J Clin Oncol 2010;28(4):662–6.

136. Federal Register | Federal Policy for the Protection of Human Subjects [Internet]. [cited 2015 Sep 8];Available from: https://www.federalregister.gov/articles/2015/09/08/2015-21756/federal-policy-for-the-protection-of-human-subjects

137. NOT-OD-15-026: Request for Comments on the Draft NIH Policy on the Use of a Single Institutional Review Board for Multi-Site Research [Internet]. [cited 2015 Sep 2];Available from: http://grants.nih.gov/grants/guide/notice-files/NOT-OD-15-026.html

138. Welcome to the NCI Central Institutional Review Board (CIRB) Initiative [Internet]. [cited 2015 Sep 3];Available from: https://ncicirb.org/cirb/default.action

139. Medical Device Amendments of 1976 [Internet]. Available from: http://www.gpo.gov/fdsys/pkg/STATUTE-90/pdf/STATUTE-90-Pg539.pdf

140. 21st Century Cures Act [Internet]. Available from: http://docs.house.gov/billsthisweek/20150706/CPRT-114-HPRT-RU00-HR6.pdf

141. Kaufman DJ, Murphy-Bollinger J, Scott J, Hudson KL. Public Opinion about the Importance of Privacy in Biobank Research. Am J Hum Genet 2009;85(5):643–54.

142. Trinidad SB, Fullerton SM, Ludman EJ, Jarvik GP, Larson EB, Burke W. Research Practice and Participant Preferences: The Growing Gulf. Science 2011;331(6015):287–8.

143. Vermeulen E, Schmidt MK, Aaronson NK, et al. A trial of consent procedures for future research with clinically derived biological samples. Br J Cancer 2009;101(9):1505–12.

144. National Institutes of Health. GENOMIC DATA SHARING (GDS) Policy [Internet]. [cited 2015 Sep 3];Available from: https://gds.nih.gov/03policy2.html

145. OHRP. Regulatory Changes in ANPRM [Internet]. [cited 2015 Sep 2];Available from: http://www.hhs.gov/ohrp/humansubjects/anprmchangetable.html

146. Freedom of Information Act, 5 U.S.C. § 552 [Internet]. Available from: http://www.gpo.gov/fdsys/pkg/STATUTE-80/pdf/STATUTE-80-Pg250.pdf

147. Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y. Identifying personal genomes by surname inference. Science 2013;339(6117):321–4.

148. Certificates of Confidentiality (CoC) Kiosk | grants.nih.gov [Internet]. [cited 2015 Sep 3];Available from: http://grants.nih.gov/grants/policy/coc/index.htm

149. Confidentiality and Privacy Protections [Internet]. Natl. Inst. Justice. [cited 2015 Sep 3];Available from: http://www.nij.gov/funding/humansubjects/pages/confidentiality.aspx

150. National Institutes of Health. NIH Guide: Final NIH statement on sharing research data [Internet]. [cited 2015 Sep 3];Available from: http://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html

151. NOT-OD-14-124: NIH Genomic Data Sharing Policy [Internet]. [cited 2015 Sep 3];Available from: http://grants.nih.gov/grants/guide/notice-files/NOT-OD-14-124.html

152. UK Biobank Ethics and Governance Framework [Internet]. Available from: https://www.ukbiobank.ac.uk/wp-content/uploads/2011/05/EGF20082.pdf