Summary of Responses from the Request for Information on Building the Precision Medicine Initiative National Research Participant Group

A total of 152 people and groups responded to the RFI for the Precision Medicine Initiative. The number of responses per question is shown below:

The NIH seeks comments on any or all of, but not limited to, the following topics:

A. General topics on the development and implementation of this large U.S. cohort.

1) The optimal study design and sample size for a large U.S. precision medicine cohort.

98 responses

2) Data to be collected at baseline and follow-up, including mode of collection and frequency and length of follow-up.

101 responses

3) Potential research questions that could be uniquely or more efficiently and effectively pursued in a large U.S. precision medicine cohort.

89 responses

4) Any other suggestions for NIH to consider in the development and implementation of such a research cohort.

110 responses

B. Suggestions for existing or potentially new research entities (a health care system, research network, cohort study or consortium, or other entities such as longitudinal studies using digital-based platforms) **that might be combined into a large U.S. cohort. Providing the following information would be useful when suggesting research entities.**

1) The capability of the existing or potentially new research entity to efficiently identify and follow 10,000 or more participants who are likely to consent to providing their medical and other health-related data, biospecimens, and genomic data for broad research use, including in sub-group analysis that could help assess various treatment effects and outcomes. It would also be useful to provide the rationale that potential participants are likely to consent, as well as experience with and ability to participate in central IRB and a master contract agreement to streamline enrollment of the precision medicine cohort.

99 responses

2) The capability for the research entity to provide individual-level participant data, particularly those from electronic health data (including both electronic health record and payer data), that can be integrated into a standard format to create a combined large longitudinal precision medicine cohort.

88 responses

3) The capability for the research entity to track and retain the participants for several years of follow up. The race/ethnic composition, sex, and age distribution of participants from the research entity likely to consent, by standard U.S. Census categories, would also be helpful. The NIH especially seeks information about studies of populations underrepresented in research and those with phenotypes or disorders of high public health and human impact. Additional information that would be of use includes: for health care systems, the current patient turnover rate and efforts that can be made to capture longitudinal data from clinical visits outside of the system and to continue follow participants who leave the system entirely; and for ongoing cohort studies, the retention rate to date.

80 responses

Question A.1: The optimal study design and sample size for a large U.S. precision medicine cohort.

There were 98 comments on Question A1. The responses fell into five themes: 1. Size of the study; 2. Demographic information; 3. Types of medical conditions; 4. Genotypes and Phenotypes; and 5. Study design considerations.

There were 45 comments regarding the projected sample size (1 million) of the study; 27 thought the sample size was appropriate, while 10 thought it was too small, and that a much larger population should be studied to ensure adequate representation of rare disorders and small populations. Seven responders advocated for using smaller samples that included family studies, specific rare conditions (e.g. sickle cell anemia) and case-control studies. One respondent asserted that the poorly-designed PMI will be largely descriptive and yield little important information on the etiology and pathophysiology of complex conditions.

Eighty-four comments addressed demographic information. Fifty-six comments touched upon ensuring that the cohort reflects the diversity of the U.S. population. Specific demographic comments centered on including: expectant mothers (as it is hypothesized that several chronic adult-onset diseases have fetal origins), veterans, infants, children and adolescents, as well as people of all ethnicities, social economic status, geographical locations and sexual and gender minority status. Many of these respondents also encouraged oversampling of minority groups and 16 comments advocated for collecting sociodemographic data on study participants.

Thirty-seven respondents commented on types of diseases/medical conditions to include in the data, including 19 (who wanted to ensure that data on rare diseases were included to facilitate research studies and 10 who advocated for asking about chronic conditions (e.g. cardiovascular disease, diabetes, etc.) and/or ensuring that the data broadly capture disease status. Other topics included: mental illness, dental information and participants who are critically ill.

Two hundred and eighteen comments addressed genotypes and phenotypes. Almost half of the responders (40) thought biospecimens should be collected on participants. Suggested molecular studies included: genetic evaluation (37), DNA variant analysis (34), biomarker identification (23), metabolomics information (14), transcriptome reporting (8), microbiome evaluation (3) and DNA methylation assessment (1). Almost half of the responders also recommended phenotypic studies, with many respondents arguing for deep phenotyping of subjects. These respondents often were the same ones who advocated for including case-controlled studies, small cohorts and people with rare conditions/diseases. Phenotyping comments recommended collecting information on the ‘environment’ (23) and argued for including behavioral data (13) and assessing (epi)genotype/phenotype correlations (13).

There were 214 comments that addressed the specifics of study design. Thirty-four respondents thought that the design should incorporate precisely-defined cohorts and/or family studies, to gain more precise phenotypic, genetic and environmental information on specific diseases. Thirteen comments were directed at sampling. Different suggestions included: probability sampling, stratified sampling, statistically-representative sampling, data combing and the use of randomized control trials. The use of longitudinal studies (30), electronic medical records (24), standardized data collection methods (15), already developed infrastructures (9; CTSAs, PCORI, SEER, blood banks, etc.) and mHealth-based strategies (10) were also recommended. The use of both retrospective (or existing) studies (17) and prospective studies (19) were both recommended. Several people thought a combination approach that used both would be beneficial. Other comments (9) centered on the ethics of such a large study, including

consultation of the general public and establishment of policies on data access, as well as pilot studies (6), the ability to return results (3), adequate consenting (7) and the ability to recontact participants (9).

Question A.2: Data to be collected at baseline and follow-up, including mode of collection and frequency and length of follow-up.

A total of 152 people and groups responded to the PMI RFI, with 101 comments on Question A2. The responses fell into seven themes: 1. Data collection methods; 2. Length of study; 3. Follow up interval; 4. Demographic data; 5. Lifestyle factors; 6. Clinical measures and 7. Genetic and molecular studies.

There were 102 comments regarding the data collection methods. Several respondents (22) suggested using online questionnaires and self-reported health information and outcomes data, while 12 thought this information should be done in more formal settings. One respondent warned about the inaccuracies of self-reported medical information, especially in regards to lifestyle information (e.g. substance abuse, diet, physical activity, etc.). A total of 36 respondents recommended using electronic health records to integrate data, 28 mentioned using computer-based, mobile, and wearable devices to track information and aid in data collection. Four respondents brought up ethical considerations, including HIPAA laws and ensuring consent.

Twenty-eight comments addressed length of study. The vast majority (23) recommended following the cohort for more than five years. Thirty-five respondents addressed follow up intervals. Twenty-three recommended follow up collections either annually or every one to two years.

Forty-three respondents wanted to ensure that demographic data (e.g. age, birth information, ethnicity, geographic information, social economic status, education, environmental conditions, etc.) were collected at baseline and updated at subsequent timepoints. Lifestyle factors (diet, physical activity, social support, sleep patterns, as well as alcohol, illicit drug and tobacco use) were important considerations for 43 respondents.

Forty-one respondents addressed the types of clinical measures that should be reported. These included: personal health history, family health history (up to three generations), standard blood and urine-based clinical labs, blood pressure, heart rate, height, weight, BMI, cardiovascular and pulmonary function, cognition and mental health. Other measures included, disease state, chronic illness assessments, new medical conditions, imaging studies, infection exposure, medication use and fitness tests. Respondents overwhelmingly suggested that these data be collected at each visit.

There were 169 comments regarding genetic and molecular studies. Over half (53) respondents thought that biospecimens should be collected on all individuals. Thirty-three respondents recommended analysis of the microbiome (e.g. oral, nasal, rectal, skin and genitourinary samples). Thirty-one respondents argued for including genetic information (including whole genome analysis by SNP studies, whole exome or whole genome sequencing). Other common genetic and molecular analyses included metabolome studies (16), transcriptome analysis (12), epigenetic measures (9), proteome studies (9), Gene x Environment correlations (6) and pharmacogenomics information (2).

Question A.3: Potential research questions that could be uniquely or more efficiently and effectively pursued in a large U.S. precision medicine cohort.

There were 89 responses to question A.3. The majority of comments suggested research questions about genetics or epigenetics (61), and a portion of those research questions focused on gene by environment interactions (17). Other common suggestions included:

- the study of drug and treatment response differences or drug development (46)
- the study of risk or protective factors for various diseases (47)
- environmental factors (36)
- modifiable lifestyle factors/behaviors (27)
- "omics" including metabolomics, proteomics (22)
- health disparities (21)
- study of pregnancy and/or children (19)
- healthcare delivery or healthcare systems (18)
- biomarkers (13)
- comorbidities (9)
- the microbiome (8)

There was support for the study of both rare diseases (15) and prevalent diseases (10). Some responses identified specific diseases, in descending order of frequency: cancer, obesity, diabetes, brain/neurological diseases, heart disease/cardiovascular disease, dental/periodontal disease, mental illness, infectious disease, bone disease, hearing loss, respiratory disease, sickle cell anemia, lung disease, arthritis, allergy, autoimmune disorders, and chronic pain.

A small number of responses touched on dissemination and implementation considerations, but there were too few of these comments to be coded. Some responses also touched on certain study design considerations, such as the study population, sampling strategies, analysis strategies, predictive tools, and study management strategies. Most of these comments are already captured in the summaries for other questions.

Question A.4: Any other suggestions for NIH to consider in the development and implementation of such a research cohort.

There were 110 comments for question A.4. The most common topics addressed in the responses included issues related to the study population (23), engaging participants/patients (21), engaging healthcare professionals (12) and other partners (9), management or communication of the study (19), and study design/study planning (31). Some comments overlapped with other questions on data/biospecimen management (31), study questions (14), and leveraging existing studies or systems (18). All comments about study questions (14) are captured in question A.3, and all comments for leveraging existing studies or systems are captured in the summary for question B.1. A summary for each topic is included below.

Study population (23)

- Include children, pregnant women, families, women, racial/ethnic minority populations, socio-economic groups, sexual and gender minorities
- Include populations in remote and rural areas
- Distinguish cohorts designed to address issues of prevention versus clinical epidemiology of diagnosis, treatment, and outcome
- Target under-served populations or populations with unique exposures
- Consider differences in populations that may need more follow-up and others that may need less follow-up
- Enroll active blood donors

Engaging participants/patients (21)

- Enable patients to be full partners, use web/mobile technologies, community forum sessions, and devices
- Obtain meaningful consent; give participants choice regarding what data, by whom, and for what purposes, and assurance of protection and transparency
- Assess the population's attitudes about privacy of health-related data, models of consent, oversight and infrastructure of PMI, and return of research findings
- Reach and educate diverse populations (including participants, healthcare providers, public) about the benefits of clinical trial participation; send a "What To Tell Your Doctor" pamphlet to every participant
- All cohort participants, regardless of how you assess their intelligence or educational level, should be permitted to see all of the data (even raw data) that have been captured about them

Engaging healthcare professionals (12) and other partners (9)

- Establish relationships with community clinics, advocacy groups, organizations, healthcare systems, and study participants built on trust and respect
- Include the following experts in the study: genetic counselors, pediatricians and child health experts, experts in hearing and cognition
- The project should include an active education component; clinicians and others in the healthcare industry should be kept apprised of this study so that if their patients bring them questions or requests to submit samples, these people/facilities will be able to assist

- Training of health researchers and practitioners, including hospital administrators, must incorporate PMI-relevant content: genomics, big data analytics, interprofessional education (IPE) in population health and methods, and approaches for community engagement
- Consider consulting with: the HHS Office of Disease Prevention and Health Promotion (HP2020) and other HHS agencies, including the National Pain Strategy; the Institute of Medicine (soon to be National Academy of Medicine)
- Work out ways for commercial partners to help bear some of the expense

Management or communication of the study (19)

- Employ a centralized management model
- Develop a quality management system (QMS) to enable NIH to establish policies, processes, and procedures; identify quality indicators; and provide means to identify, track, and report non-conforming data.
- Develop publically accessible educational resources for both patients and providers (e.g. video tutorials on precision medicine, family history, and genome sequencing)
- Develop carefully-planned, coordinated recruitment and community awareness campaigns supported by high level government officials, community leaders, and celebrities (First Lady, members of congress, movie stars, athletes) and are integrated into a variety of media (internet, radio, television, print, 24/7 information centers).
- Don't cut the budget or it won't be done right and if it's not done right, it's a waste of money.
- Consider a contract funding mechanism, with clear deliverables and timelines, rather than grant funding to support the cohort. Grants-based level of effort funding seems likely to have a higher potential for project delays and missed deadlines.
- The team developing the plan of action should consist of a broad representation of disciplines and senior and junior investigators within each discipline, including statisticians, data scientists, and marketing and communication specialists.

Study design/study planning (31)

- Research priorities are likely to change over time - use an approach that sets clear short-term quantifiable objectives but also allows those objectives to evolve over time
- Incorporate the ability to scale up efficiently, with minimal marginal costs per measurement, per participant, per enrolling site/investigator, and per study
- Use an electronic consent system; solicit an “open” unprotected status, as the combination of electronic health record, birth year and month (necessary for many studies) and some location information will make a record so unique that complete confidentiality can’t be guaranteed.
- Use robust technology, including mobile and tablet modes of data acquisition, wearable sensors and smart phones
- Collect enough sample from each participant to allow for multiple center analysis
- Incorporate innovative use of social media to allow individuals to volunteer, participate, share their information and possibly allow linkage to their electronic health records

- Consider the impact of actionable results from genetic testing or other aspects of the precision medicine initiative on medical services. Prior to the onset of the initiative, the projected uptake of services should be determined and an appropriate plan for managing the increase in service utilization should be created.
- Allocate some of the population to a “reference group” chosen without regard to health status in a manner guaranteeing adequate representation of typically understudied segments of the U.S. population. This group can provide baseline information about the frequency of genetic markers in the population.

Data/biospecimen management (31)

- Have a central hub for participant access and central storage of data from other databases (existing or new)
- Establish a participant web-based portal accessible via multiple devices (computers, tablets, smartphones)
- Use of existing data standards for harmonization of data from existing or new databases
- Data must be open-access: consider publicly and freely releasing a de-identified dataset; open-source approach for developing apps, tools, and techniques
- Data infrastructure should leverage HIPAA compliant cloud computing

Question B.1: The capability of the existing or potentially new research entity to efficiently identify and follow 10,000 or more participants who are likely to consent to providing their medical and other health-related data, biospecimens, and genomic data for broad research use, including in sub-group analysis that could help assess various treatment effects and outcomes. It would also be useful to provide the rationale that potential participants are likely to consent, as well as experience with and ability to participate in central IRB and a master contract agreement to streamline enrollment of the precision medicine cohort.

Of the 152 responses received to this RFI, 99 addressed this question. Approximately 80 responses named a research entity. Thirty of those research entities have the capability to identify and follow 10,000 or more participants who are likely to consent. Some organizations have participants enrolled in registries, but did not comment on ability to recruit those participants for further studies. Eighteen respondents provided various comments.

Of all the responses to the question:

- 8 provided rationale that potential participants are likely to consent
- 2 mentioned sub-group analysis
- 21 have experience with and ability to participate in IRB
- 8 have Master contract to streamline enrollment

The following entities were mentioned more than once:

- PCORnet
- HMORN
- Mayo Clinic--through the Center for Individualized Medicine
- eMERGE
- Nurses' Health Study III
- Duke University's MURDOCK study
- Kaiser Permanente

Nine state cancer registries from Kentucky, New Jersey, Texas, New Hampshire, Iowa, Oregon, Delaware, Massachusetts, and Idaho responded to this question with the same exact language as to their capabilities to provide data.

Question B.2: The capability for the research entity to provide individual-level participant data, particularly those from electronic health data (including both electronic health record and payer data), that can be integrated into a standard format to create a combined large longitudinal precision medicine cohort.

Of the 152 responses received to this RFI, 88 addressed this question. Approximately 60 responses named a research entity that is or will be capable of providing electronic health data in a standard format that can be used to create the precision medicine cohort. About 20 responses, on the other hand, had suggestions for other considerations when determining which research entities will participate. Finally, 11 responses were not useful.
Of the responses that named a potential research entity:

- 59 can provide individual-level participant data
- 49 referenced having electronic health records, a clinical data warehouse, etc.
- 53 indicated their data can be integrated into a standard format
- 16 mentioned having payer data
- 9 have a biobank
- 15 stated the importance of patient engagement (e.g., collecting data directly from participants, sharing personal health information with participants, etc.)

Research entities that deemed themselves capable of providing the proposed data were mostly large health plans, health care systems, research networks, or large research studies. The National Patient-Centered Clinical Research Network (PCORnet) and the HMO Research Network (HMORN) were mentioned several times as being ideal candidates to involve.
Of note, nine state cancer registries from Kentucky, New Jersey, Texas, New Hampshire, Iowa, Oregon, Delaware, Massachusetts, and Idaho responded to this question with the same exact language as to their capabilities to provide data.

The importance of including family histories was stated several times; the Utah Population Database sounded unique in this regard.

About 20 responses made suggestions for additional considerations when selecting research entities. The suggestions included:

- Payer/claims data may not have sufficient detail to correctly classify the disease status of individuals and may lead to biased or spurious results
- Should incorporate technology in the form of wearable health tracking devices and mobile platforms
- Potential research entities need to provide individual-level data in real-time and on-demand
- Potential research entities must also have the capability to communicate with and provide care for individual participants
- Since the key is to have EHRs integrated into a standard format, perhaps the PMI can move the nation forward in accomplishing a standard format EHR that has utility for research, clinical care, and billing
- Given the extensive collection of personal data that is proposed, a federal health data protection act should be considered (similar to the Genetic Information Nondiscrimination Act of 2008)
- Dental records have not been integrated into EHRs which may hinder research into dental and medical co-dependent diagnoses

- The PMI should set up a single data coordination center for management of EMR data and payer data associated with the PMI cohort; this would be the most efficacious way to collect quality data across multiple data sources
- Leverage data from blood centers which maintain complete health histories of blood donors
- It is vital that participants can access their health data and be included in the feedback loop as the study goes forward

Question B.3: The capability for the research entity to track and retain the participants for several years of follow up. The race/ethnic composition, sex, and age distribution of participants from the research entity likely to consent, by standard U.S. Census categories, would also be helpful. The NIH especially seeks information about studies of populations underrepresented in research and those with phenotypes or disorders of high public health and human impact. Additional information that would be of use includes: for health care systems, the current patient turnover rate and efforts that can be made to capture longitudinal data from clinical visits outside of the system and to continue follow participants who leave the system entirely; and for ongoing cohort studies, the retention rate to date.

There were 80 total responses for question B.3.

- cohorts with extensive follow up time and a good record of tracking and retaining participants were mentioned a good deal **(64)**
- the need for large cohorts with thousands of participants was mentioned a few times
- ability to retain and track participants including the use of medical records, tracking insurance claims data, in person interviews, increasing the quality of patient-provider relationships and communication, as well as using mobile clinics, cell phones, and other electronic devices
- large integrated healthcare systems like Kaiser were mentioned for their ability to track, and retain large amounts of patients
- for larger state-based cohorts, the ability of the cohort to mirror the ethnic/racial, socioeconomic, age, and other demographic characteristics of the state and counties was emphasized
- expressed a need to include cohorts with large minority participation including African Americans and Latinos and other underrepresented minorities **(53)**
- less mention of making sure women were a large proportion of the cohort **(28)**
- its was moderately mention that children should be included to study early childhood exposures and older adults so that chronic diseases such as cancer could be examined, so a large proportion of the population could be covered **(35)**
- only a few participants mentioned the cohort should be large so that rare diseases could be studied (power)
- significant mention of including studies from low SES areas so that all underserved communities would be included and studied **(21)**
- a few mentions about rural vs urban study settings and the need to include more rural populations **(10)**