ASPE
ASSISTANT SECRETARY FOR PLANNING AND EVALUATION
OFFICE OF HEALTH POLICY

OS PCOR TF
OFFICE OF THE SECRETARY
PATIENT-CENTERED OUTCOMES RESEARCH TRUST FUND

Challenges and Improvements for PCOR Data Infrastructure: Results from a Stakeholder Prioritization Activity

PRESENTED TO:
Office of the Assistant Secretary for Planning and Evaluation (ASPE)
U.S. Department of Health & Human Services

PRESENTED BY:
NORC at the University of Chicago
AcademyHealth

November 10, 2020

Acknowledgments

This report was prepared under contract under Contract No. HHSP233201600020I, Task Order No. HHSP23337001T between the U.S. Department of Health and Human Services' (HHS) Office of the Assistant Secretary for Planning and Evaluation (ASPE) and NORC at the University of Chicago (NORC).

We extend our sincere appreciation to all the participants who volunteered their time to participate the virtual Listening Session and online prioritization activities that inform this report. The views expressed are the participants' and do not necessarily represent the views of ASPE, HHS, or the U.S. Government.

Table of Contents

Executive Summary

The U.S. Department of Health and Human Services (HHS) Assistant Secretary for Planning and Evaluation (ASPE) receives a portion of the Patient-Centered Outcomes Research Trust Fund (PCORTF) to build data infrastructure that enhances the conduct of patient-centered outcomes research (PCOR). To achieve this, the Office of the Secretary Patient-Centered Outcomes Research Trust Fund (OS-PCORTF) portfolio supports approximately 50 cross-agency projects—of which 27 are currently active—that simultaneously respond to major federal legislation, the HHS Secretary's priorities, individual agency data strategies, and evolving patient-centered research needs.[i] ASPE has organized these projects around five functionalities that form the basis for robust PCOR data infrastructure and are central to ASPE's strategic framework for building data capacity for PCOR:[ii]

- Using Clinical Data for Research: Optimizing data for research by improving access, enhancing quality, and promoting interoperability of clinical data across multiple sources.
- Standardizing Data Collection: Better defining and standardizing key data terms and concepts (i.e., common data elements) to more effectively and efficiently share, link, and aggregate across data sources.
- Linking Data: Linking clinical data with other data types (e.g., claims data, program data, and participant-provided information) in order to track patients across the continuum of care and/or capture a range of health-related outcomes.
- Collecting Participant-Provided Information (PPI): Developing and using new standards and technologies to collect PPI so that participants can participate more fully in clinical research.
- Using Federal Databases for Research: Enhancing federal and state-level data systems to enable greater access, use, linkage, and analysis of publicly-funded data for research.

In December 2019, Congress reauthorized the PCORTF for 10 additional years, which will allow for additional OS-PCORTF-funded projects over the next decade. Subsequently, ASPE sought to gather perspectives on challenges and improvements for PCOR data infrastructure from a diverse group of stakeholders—with a wide range of occupational backgrounds including policy, health care delivery, research and informatics—through an online prioritization activity. Participants first generated challenges and improvements for the five functionalities. Participants then voted on challenges and improvements within each functionality, generating a ranked list of the participant-generated ideas.

The online prioritization activity generated a total of 87 data infrastructure challenges and 76 data infrastructure improvements. Across the five functionalities, participants returned to five

[i] Dullabh P, Dhopeshwarkar R, Heaney-Huls K, et al. Building the Data Capacity for Patient-Centered Outcomes Research: The 2019 Annual Report. Prepared by NORC at the University of Chicago under Contract No. HHSP233201600020I. Washington, DC: Office of the Assistant Secretary for Planning and Evaluation. https://aspe.hhs.gov/system/files/pdf/259016/2019OS-PCORTFPortfolioReport.pdf

[ii] Dhopeshwarkar R, Dullabh P, Dungan R, et al. Building Data Capacity for Patient-Centered Outcomes Research Portfolio Highlights (2016 – 2019) Impact, Opportunities, and Case Studies. Washington, DC: Office of the Assistant Secretary for Planning and Evaluation. https://aspe.hhs.gov/system/files/pdf/259016/OS-PCORTFImpactReport508.pdf

common themes that focused on the need to: 1) enhance consistency in data standardization; 2) improve access to social determinations of health (SDOH) data that are not routinely collected during care delivery; 3) improve ability to access, integrate and use PPI, particularly those data generated from medical devices and wearables; 4) increase access to federal data sets, with an emphasis on access to de-identified data sets; and 5) expand collaboration across organizations at the local, state, and federal level.

Enhance consistency in data standardization. Participants raised multiple challenges and improvements related to the issue of consistency or transparency in data standardization. Participants focused on the need for consistent processes for collecting, cleaning, and presenting data. They also highlighted the importance of promoting adoption of and adherence to standards across the health system after they are developed.

Improve access to SDOH data that are not routinely collected during care delivery. Participants sought resources to support the standardized collection of SDOH data, and expressed desire for expanded access to federal data sets to support research inquiries related to SDOH, including zip code level data on neighborhood characteristics.

Improve ability to access, integrate, and use PPI, particularly those data generated from medical devices and wearables. Stakeholders noted the importance of accessing PPI, including patient-reported outcomes (PROs) and patient generated health data, from medical devices to support their research inquiries. Participants focused on the need to develop and disseminate standards to support PPI data collection and analysis (including the collection of PROs) and the aggregation and integration of PPI into electronic health records. Participants also sought mechanisms to promote collection and use of PROs among patients and clinicians.

Increase access to federal data sets, with an emphasis on access to de-identified data sets. Across the functionalities, access to data sources was a prominent theme. As previously noted, participants focused on access to SDOH data resources, including federal data sets with SDOH data, across multiple functionalities. Participants also broadly highlighted the need for increased access to federal data resources. This topic was particularly prominent for the topic of using federal databases for research, where participants focused on the need for easily accessible, de-identified federal data sets. Participants also underscored the need for access to surveillance data, an emerging topic given the current coronavirus disease 2019 (COVID-19) global pandemic.

Expand collaboration across organizations at the local, state and federal level. Participants highlighted the need for collaboration to leverage and enhance existing data sources and infrastructure. Collaboration was discussed at both the meso-level (e.g., collaboration to enable cross-sector data sharing) and the macro-level (e.g., regulatory frameworks, enhanced federal data assets, and development of and incentives for standards adoption). Based on the challenges and improvements they submitted, participants foresee the need for widespread cooperation to make data available and useful for research, while maintaining the privacy and security of patient health information.

Introduction

Background

As a principal advisor to the Secretary of the U.S. Department of Health and Human Services (HHS), the Assistant Secretary for Planning and Evaluation (ASPE) conducts strategic planning, economic analysis, legislation development, and policy research and evaluation. ASPE receives approximately 4 percent of the Patient-Centered Outcomes Research Trust Fund (PCORTF) to build data infrastructure that enhances the conduct of patient-centered outcomes research (PCOR). To achieve this, the Office of the Secretary Patient-Centered Outcomes Research Trust Fund (OS-PCORTF) portfolio supports approximately 50 cross-agency projects—of which 27 are currently active—that simultaneously respond to major federal legislation, the HHS Secretary's priorities, individual agency data strategies, and evolving patient-centered research needs.[1] ASPE has organized these projects around a guiding framework of five functionalities that form the basis for robust PCOR data infrastructures (Table 1). Project activities include the development and use of clinical registries and health outcomes research networks—in order to develop and maintain a comprehensive, interoperable data network to collect, link, and analyze data on outcomes and effectiveness from multiple sources including electronic health records (EHRs).

> In alignment with broader national health policy and strategy objectives, OS-PCORTF projects tend to focus on developing resources that can improve the quality of PCOR on priority topics: COVID-19, opioid use, value-based care, mortality data, real world evidence, and the interoperability of electronic health records.

Table 1. OS-PCORTF Data Infrastructure Functionalities

Functionality 1: Using Clinical Data for Research – Optimizing data for research by improving access, enhancing quality, and promoting interoperability of clinical data across multiple sources.
Functionality 2: Standardizing Data Collection – Better defining and standardizing key data terms and concepts (i.e., common data elements, or CDEs) to more effectively and efficiently share, link, and aggregate across data sources.
Functionality 3: Linking Data – Linking clinical data (e.g., EHR data, clinical registries) with other data types (e.g., claims data, program data, participant-provided information) in order to track patients across the continuum of care and/or capture a range of health-related outcomes.
Functionality 4: Collecting Participant-Provided Information (PPI) – Developing and using new standards and technologies to collect PPI so that participants can participate more fully in clinical research.
Functionality 5: Using Federal Databases for Research – Enhancing federal and state-level data systems to enable greater access, use, linkage, and analysis of publicly funded data for research.

[1] Dullabh P, Dhopeshwarkar R, Heaney-Huls K, et al. Building the Data Capacity for Patient-Centered Outcomes Research: The 2019 Annual Report. Prepared by NORC at the University of Chicago under Contract No. HHSP233201600020I. Washington, DC: Office of the Assistant Secretary for Planning and Evaluation. https://aspe.hhs.gov/system/files/pdf/259016/2019OS-PCORTFPortfolioReport.pdf

Purpose

In December 2019, the Congress reauthorized the PCORTF for 10 additional years, which will allow HHS to add OS-PCORTF projects to the portfolio over the next decade. Subsequently, ASPE sought to gather perspectives on PCOR data infrastructure needs and priorities from a diverse group of stakeholders through an online prioritization activity. The goal of the activity was to provide feedback on PCOR data infrastructure gaps and priorities as stakeholders consider areas in need of progress during the next decade.

Approach

The prioritization activity occurred in three parts: 1) a virtual Listening Session, 2) an idea generation activity using the Codigital platform, and 3) an online prioritization activity also using Codigital. Below, we briefly describe how we selected participants and the process for the prioritization activity.

Participants

The virtual Listening Session convened stakeholders with expertise related to numerous aspects of PCOR (e.g., methods and data, public health interventions). These individuals also have experience in a wide range of occupational backgrounds including policy, research and informatics, and work in a variety of sectors including—but not limited to—government, academic, and health system settings. Participants, listed by name in **Appendix A**, belonged to the following stakeholder groups:

- Patient-Centered Outcomes Research Institute (PCORI) stakeholders (Patient Engagement Advisory Panel; Methodology Committee; PCORnet leadership)
- Members of AcademyHealth interest groups (e.g., Health Information Technology; Learning Health System; Public Health Systems Research; Quality & Value)
- Health system representatives (Department of Veterans Affairs Health Services Research & Development Service; Kaiser Permanente; Geisinger; Sanford Health)
- Technical advising bodies (e.g., Electronic Data Methods Forum; Methods & Data Council; State University Partnership Learning Network; Medicaid Medical Directors Network)
- Industry representatives (e.g., Cerner; Epic)
- Public health communities (e.g., American Medical Informatics Association ® 2020 Scientific Program Committee Leadership members, National Interoperability Collaborative; Data Across Sectors for Health)
- Members of the OS-PCORTF Technical Expert Panel (TEP)

Activity Process

In the first stage of the activity, we convened participants for a virtual Listening Session which presented an overview of the activity, the background on OS-PCORTF, and a draft set of questions that would guide the idea generation and prioritization activities. The draft questions are provided below:

- What are your most pressing data infrastructure challenges for conducting PCOR?
- What improvements over the next 3-5 years would address your most pressing data infrastructure challenge?
- What outcomes should the field (broad lens) strive for in the next 10 years to advance data infrastructure for PCOR?
- What are the 2-3 most impactful intermediate steps?
- How do we define success? What indicators best demonstrate progress in building data infrastructure capacity for PCOR?

The Listening Session used a moderated discussion guide to elicit feedback that informed refinement of these questions. The ultimate goal was to identify guiding questions that would help ASPE better understand immediate priorities for data infrastructure facing PCOR stakeholders. During the session, participants noted overlap across certain questions and suggested revisions that would introduce more specificity, such as: 1) identifying the acting or implicated parties for each question; 2) defining key terms and concepts such as *data infrastructure, outcomes,* and *the field*; and 3) anchoring the discussion in how we define and conceptualize PCOR data infrastructure.

In light of this feedback, we reframed the questions in terms of the five OS-PCORTF functionalities. A total of ten questions (**Appendix B**) were used in the subsequent phases of the activity on the Codigital platform. For each of the five functionalities, participants were invited to respond to the following questions:

- What are the most pressing data infrastructure challenges for [the functionality] in the next 10 years?
- What improvements are needed (e.g., policies, governance, standards, services, technology) for [the functionality] over the next 10 years?

Using the Codigital platform, participants were asked to provide answers to all 10 questions during the idea generation phase. We provided sample responses (3-5) for each functionality to start the conversation, and participants could edit and refine these responses to add nuance or clarity. They could also refine ideas provided by other participants. Then, during the prioritization phase, participants completed a guided, pairwise voting activity to identify the top challenges and improvements within each functionality. To do so, participants were presented with pairs of responses that had been generated during Phase 1, and asked to select their preferred response. Participants repeated this process until they completed voting on all of the ideas or closed out of the window.

The Codigital voting process generated a ranked list of responses for each functionality based on the collective voting patterns. While responses could be edited by multiple participants within the activity, the design of this virtual platform does not allow any further reviews and edits for clarification or direct follow-up with participants at the conclusion of the ranking process to clarify responses. While this is a limitation of our approach, this ensures that results reflect the direct feedback of participants without post-hoc alterations.

Thus, the results reflect the information directly entered by participants in the Codigital platform. In some cases, there was duplication of ideas across challenges and improvements—a byproduct of having multiple participants provide responses through a virtual activity. We did not go through the process of de-duplicating responses prior to the voting phase of the activity, but rather allowed participants to rank all inputs in the form they were received.

Results

The Listening Session was attended by nearly 40 participants, over half of whom provided verbal or written feedback during the session. This level of engagement carried on to the Codigital activities. Each question engaged an average of 27 contributors. Overall, participants were highly engaged and reacted positively to the use of the Codigital platform.

Across functionalities, the Codigital activity generated 87 data infrastructure challenges and 76 data infrastructure improvements. While we did provide sample responses for each functionality, participants were able to edit and refine the sample responses and vote on all of the responses. Thus, the results of the activity fully represent the input of the participants.

In the sections below, we present and contextualize the top five challenges and the top five improvements participants reported for each functionality to demonstrate their relevance in the PCOR data infrastructure landscape. We also discuss additional key themes that emerged within each functionality. Given that we did not confer with participants following the idea generation and prioritization activities, in this report we do not interpret the challenges or improvements, attempt to generate implications from the responses, or add concepts not raised directly by participants. A list of all challenges and improvements organized by functionality is provided in **Appendix B.**

Functionality 1: Use of Clinical Data for Research

Use of Clinical Data for Research refers to the need for researchers to be able to access and analyze clinical data that are routinely collected during clinical care. These data are rich in information that is critical to research and the generation of scientific knowledge but they are often siloed within different, incompatible systems that make the data challenging for researchers to access and use. The development and use of standards, policies, services, and analytic tools can allow researchers to more easily access, aggregate, and analyze data from multiple sources in innovative ways.

This functionality also refers to the infrastructure needed to ensure that data are high quality and fit-for-use in research. This includes standards, policies, and services to encourage data completeness, comprehensiveness, and representativeness to the population being studied, as well as tools to assess these attributes.

Participants identified 21 challenges and 20 improvements related to the use of clinical data for research, many of which focused on the need for high quality, standardized interoperable data sets.

Challenges

Using the Codigital platform, participants were asked to identify and rank challenges related to *Use of Clinical Data for Research.* From this ranked list, we present the top five challenges.

Top Five Challenges

1. Disconnect between the data necessary for many research questions (e.g., Social Determinants of Health (SDOH), Patient-Reported Outcomes (PROs) and the data routinely recorded in EHRs, particularly the data recorded in structured form.
2. Lack of interoperability across EHR software platforms, settings, and sectors (e.g., between health care, social service and public health sectors).
3. Issues with data quality, such as data inaccuracy or incompleteness, with EHR data.
4. Lack of alignment across standards efforts, with Office of the Secretary (OS), Office of the National Coordinator (ONC), Federal Drug Administration (FDA), etc. all pushing efforts for slightly different purposes. Need to focus in order to increase adoption.
5. Lack of standardized data definitions/elements and audits for interrater reliability in individual registries, research studies, and harmonization across registries/other data sources will impact the reliability/usability of data from these sources.

Challenge 1. Disconnect between the data necessary for many research questions (e.g., SDOH, PROs) and the data routinely recorded in EHRs, particularly the data recorded in structured form. Participants noted the importance of social determinants of health data (SDOH) and patient-reported outcomes (PROs), which are not routinely collected during clinical encounters and integrated into EHRs. In addition, to the extent that these data may captured in EHRs, it is often not collected in a structured way, thereby limiting its ability to be used for secondary purposes such as quality improvement and research.

Challenge 2. Lack of interoperability across EHR software platforms, settings, and sectors (e.g., between health care, social service and public health sectors). Participants expressed that the variability in the systems and standards that the myriad of stakeholders use to capture health-related data continues to create barriers to data sharing. Variability exists across health systems, specialties and sub-specialties, research networks, etc., as well as across sectors that are increasingly trying to share information to coordinate patient care and address social needs. This variability makes it extremely difficult for researchers to aggregate and link data from multiple sources.

Challenge 3. Issues with data quality, such as data inaccuracy or incompleteness, with EHR data. Participants frequently focused on issues related to data quality and completeness. When aggregating clinical data for analysis, researchers often find that—because of differing priorities, workflow, clinical burden, and lack of incentives—the data may not be complete, accurate, or readily analyzable. Problems can range from partial demographic information, to variable coding, to missing clinical information on the outcomes of interest.

Challenge 4. Lack of alignment across standards efforts, with the Office of the Secretary (OS), Office of the National Coordinator of Health Information Technology (ONC), the Food and Drug Administration (FDA), etc. all pushing efforts for slightly different purposes. Participants expressed a desire for more coordinated efforts across the federal

agencies that are working towards developing consensus on existing standards, identifying gaps, and disseminating data standards. Each agency may have a slightly different set of priorities for standardization efforts or may be working on different but related data standards; and this could result in confusion and a lag in the adoption of standards.

Challenge 5. Lack of standardized data definitions/elements and audits for interrater reliability in individual registries, research studies, and harmonization across registries/other data sources will impact the reliability/usability of data from these sources. Because researchers rely on increasingly diverse data sets to conduct their analysis, standardization efforts need to further extend to ensure consistent definition and use of standards once they are adopted. Once a standard is defined, audits or other implementation management activities may be needed to ensure that studies and registries are using standards as written. For example, registries or studies can use different definitions of standards (e.g., medications prescribed vs. medication dispensed for "drug utilization") or they can modify standardized measures, making it difficult for researchers to utilize data sets. A lack of data harmonization, or alignment across data models, or lack of common data elements negatively impacts the reliability of the data, and the ability for researchers to access and use data across multiple sources.

Improvements

A total of 32 distinct contributors were engaged in adding, refining, and prioritizing 20 submitted ideas focused on improvements that could be made to enhance the use of clinical data for research. The top five improvements are discussed below.

Top Five Improvements

1. Create policy or financial incentives for EHR vendors to adopt consistent standards in how they structure key patient demographic and utilization data to facilitate comparisons across sites and systems.
2. Further development and dissemination of standards, services, and policies to assure data quality and metadata descriptions for research and practice.
3. Rigorously assess investments in large electronic data warehouses (e.g., PCORnet, Sentinel, NIH CRN) to consider synergies, gaps, and avoiding wasteful redundancies for further building out of infrastructure.
4. Develop and implement core outcome sets to build connections across research, clinical practice, and quality improvement and improve the relevance of routinely recorded clinical data for research.
5. Better incentives for adoption of standards, services, etc. within health systems and health plans.

Improvement 1: Create policy or financial incentives for EHR vendors to adopt consistent standards in how they structure key patient demographic and utilization data to facilitate comparisons across sites and systems. Participants advocated for the creation of effective policy and financial incentives designed to promote adherence to standards. Programs like Promoting Interoperability incentivize providers to use EHRs. This improvement suggests an approach that incentivizes EHR vendors to adopt consistent clinical data collection standards;

standardizing patient demographic and health care utilization data would facilitate researcher efforts to draw needed comparisons across sites and systems.

Improvement 2: Further development and dissemination of standards, services, and policies to assure data quality and metadata descriptions for research and practice. As described in the challenges section, participants overwhelmingly recognized the need for the expansion of standards across settings, clinical conditions and outcomes, standards that improve interoperability and usability for research-centric outcomes, and standards that improve the comparability of data from different sources.

Improvement 3: Rigorously assess investments in large electronic data warehouses (e.g., PCORnet, Sentinel, NIH CRN) to consider synergies, gaps, and avoiding wasteful redundancies for further building out of infrastructure. Several distributed research networks (e.g., PCORnet, Sentinel, and the National Institutes of Health [NIH] Clinical Research Network [CRN]) currently exist to support the use of clinical data for research. However, these entities do not always work in streamlined or synergistic ways. Conducting a targeted evaluation of these distributed research networks could flag redundancies and enable creation of efficiencies across them.

Improvement 4: Develop and implement core outcome sets to build connections across research, clinical practice, and quality improvement and improve the relevance of routinely recorded clinical data for research. For example, agreed upon outcomes could include data on SDOH and other non-clinical outcomes that are increasingly important in PCOR. Development and dissemination of standards, services, and policies would help assure data quality while core outcome sets could help improve the relevancy of clinical data for research.

Improvement 5: Better incentives for adoption of standards, services, etc. within health systems and health plans. Participants advocated for incentivizing the adoption of standards as part of multiple suggested improvements. This improvement focuses specifically on health systems and health plans. Incentives that target the health system and health plans, where clinical data often originate, could improve adoption of data standards that make data interoperable and useful for research.

Summary of Themes in Top Challenges and Improvements

Challenges and improvements related to the *Use of Clinical Data for Research* focused primarily on a lack of completeness in the data and, more specifically, a lack of coordination across efforts to address lack of standardized data definitions. In some cases, additional standards must be developed, but the suggested improvements related to the additional steps of disseminating and incentivizing the adoption of and use of these standards across key stakeholder groups.

Additional Key Themes

Beyond the top five challenges and improvements detailed above, other themes emerged within this functionality that warrant discussion.

Difficulty Using Data from Multiple Sources. Stakeholders agreed that challenges related to data quality, standardization, and interoperability render data difficult to link across sources. Linkages are necessary both to ensure that research is conducted on a large, representative sample, and to overcome the challenges reported above (poor quality data, missing information). Data housed outside of clinical settings, such as SDOH or PROs, can be particularly difficult to link, harmonize, and analyze due to lack of standardization.

Standardization of demographic data to support record linkage across disparate data sources. Participants consistently focused on the need for standards across challenges and improvements. Even data fields commonly used for linking data sets lack standardization, which can make it difficult to accurately determine whether records match. For example, multiple formats for recording name, birthdate, or zip code can mean that when comparing data from a community health clinic and an emergency room, it is not clear whether the records belong to a single patient (e.g., John Doe, John A. Doe) or multiple patients with the same name.

Need for timely, publicly accessible, nationally representative data sets for research use. Participants reported multiple challenges related to the lack of large data sets to support research. Clinical outcomes research, PCOR, and quality improvement activities benefit from cross-institutional data sets from complementary data sources (e.g., EHR and claims data). Research also requires nationally representative data and data that allow for sub-group analysis (e.g., based on location, age, race, clinical condition, etc.). Participants also mentioned the need for improvements to existing data warehouses and data sets (e.g., Sentinel, PCORnet, Centers for Disease Control and Prevention [CDC] Vaccine Safety Datalink, NIH's Clinical Research Network, Mental Health Research Network [MHRN])—including more timely data, and removing the financial barriers that limit data access.

Functionality 2: Standardizing Data Collection

The goal of *Standardizing Data Collection* is to propagate more shareable, useful data across the health system. Health data collected as part of a clinical encounter have the potential to support clinical quality improvement and large-scale clinical research studies. However, differences in clinical data definitions and elements across health information technology (IT) systems and the resulting variability can challenge the meaningful interpretation of study results and use of the results to improve patient outcomes. In order to support comparability and analysis across data sources, researchers need standard definitions of the data.

Challenges

The challenges-related question for *Standardizing Data Collection* in the Codigital prioritization activity engaged 34 distinct contributors in adding, refining, and prioritizing 19 submitted ideas. Below, we discuss each of the top five challenges in more detail.

Top Five Challenges

1. Creating data standards that EHR vendors adhere to and are required across EHR platforms and establishing infrastructure to test and ensure EHR vendors adhere to standards.
2. Lack of investment in development/deployment of technical infrastructure to support social services, which generate SDOH data just as clinical services generate EHR and clinical claims data.
3. Limited access to data not commonly collected in the course of clinical care in a study (e.g., social determinants of health data).
4. Lack of standardized methods for cleaning clinical data. Standardizing clinical data capture is an important but much larger task. Agreeing on standardized methods for cleaning and mapping clinical/SDOH data (some exist already) would be helpful.
5. Varying levels of local implementation of standardized terminology for labs (LOINC), medications (RxNorm, CUI), diagnoses (ICD-10 and SNOMED CT).

Challenge 1: Creating data standards that EHR vendors adhere to and that are required across EHR platforms, and establishing infrastructure to test and ensure EHR vendors adhere to standards. Similar to the challenges identified in the use of clinical data for research (Functionality 1), stakeholders found the lack of data standardization across EHR vendors problematic. Even if EHR data standards exist, the lack of adherence to a common set of data standards among EHR vendors creates barriers to successful data mapping, or linking across systems.

Challenge 2: Lack of investment in development/deployment of technical infrastructure to support social services, which generate SDOH data just as clinical services generate EHR and clinical claims data. This encompasses not just the capture of SDOH data, but the infrastructure to share information with the community-based organizations who provide social services, and mechanisms to follow-up on referrals to ensure patients' needs were met.

Challenge 3: Limited access to data not commonly collected in the course of clinical care in a study (e.g., SDOH data). Data on patients' SDOH, such as housing information or socio-economic status, are critical for researchers to understand the full spectrum of a patients' health. However, stakeholders highlighted the lack of technical infrastructure to collect and generate usable SDOH data. There is a need to integrate data into EHRs in a standard way to make them accessible for care and research.

Challenge 4: Lack of standardized methods for cleaning clinical data. Acknowledging that standards take time to develop, build consensus around, and disseminate broadly, in the meantime researchers' must clean data sets by removing incorrect, duplicate, ill-formatted, or corrupt data. Participants identified a lack of available tools and standardized methods for cleaning clinical, SDOH, and other types of data so that they can be reliably used for research. Broadly disseminating tools and best practices for assessing and improving data quality would support researchers in their quest to obtain high quality and useable data.

Challenge 5: Varying levels of local implementation of standardized terminology for labs (LOINC), medications (RxNorm, CUI), diagnoses (ICD-10 and SNOMED-CT). Stakeholders identified the varying levels of implementation of standardized terminology for labs, medications,

and diagnoses as a challenge. Despite existing guidance on using standardized terminology, the use of local terminology or codes persists. This inconsistent usage of terms necessitates post-hoc mapping of local codes to standardized terminology, which reduces the fidelity of the data and thereby their utility for research.

Improvements

A total of 30 distinct contributors were engaged in adding, refining, and prioritizing 15 submitted ideas focused on improvements that could be made to data standardization. The top five improvements are discussed below.

Top Five Improvements

1. Federal policies and financial incentives to promote the adoption and use of standards across sectors.
2. Standardize measurement of critical clinical covariates, such as obesity, smoking, substance abuse, medications.
3. Integration of different data sources, including patient-provided data, and real-time data reconciliation to address: 1) missing data and 2) conflicting data.
4. Development and dissemination of standardized patient outcome measures that can be used in research and captured and used across clinical care settings (e.g., primary and specialty care).
5. Establish best practices for the standardized collection of patient generated health data (PGHD), including patient-reported outcomes (PROs), SDOH, and other data.

Improvement 1: Federal policies and financial incentives to promote the adoption and use of standards across sectors. Echoing suggestions for Functionality 1, participants emphasized the need for standards and the role of incentives in accelerating their widespread adoption and use.

Improvement 2: Standardize measurement of critical clinical covariates, such as obesity, smoking, substance abuse, medications. Standardizing measurement of critical clinical covariates, such as obesity, smoking, substance abuse, medications—to reduce ambiguity in how these concepts are being measured and captured—would help researchers to better understand nuances in the relationships between conditions.

Improvement 3: Integration of different data sources, including patient-provided data, and real-time data reconciliation to address: 1) missing data and 2) conflicting data. Using data from complementary data sources including PPI is a common strategy among researchers to fill gaps or address inconsistencies in single source data sets. Researchers would benefit from standardized tools for integrating different information to produce higher quality data sets.

Improvement 4: Development and dissemination of standardized patient outcome measures that can be used in research and captured and used across clinical care settings (e.g., primary and specialty care). While patient outcomes are often defined as the result of care rendered, having a more standardized measure for assessing changes in important indicators of health would allow for more specific research questions to be examined. A standards-based approach for collecting and reporting patient outcomes information to

clinicians within their workflow, across primary and specialty care settings, would improve the data available for research.

Improvement 5: Establish best practices for the standardized collection of patient-generated health data (PGHD), including PROs, SDOH, and other data. Echoing the other improvements, informants suggested establishing best practices for the standardized collection of PGHD will improve linkages to existing clinical data and help ensure the data is complete and fit for use in conducting PCOR.

Summary of Themes in Top Challenges and Improvements

The top five challenges and improvements for *Standardizing Data Collection* highlighted opportunities to improve access and use of non-clinical data. The standardization and integration of data related to both SDOH and patient-provided information was a common theme across challenges and improvements. Similar to suggestions related to the use of clinical data for research, stakeholders highlighted the need to incentivize adherence to standards among EHR vendors.

Additional Key Themes

Beyond the top five challenges and improvements detailed above, participants consistently raised concerns as to how the widespread need for standards—to govern a range of settings, conditions, outcomes, and data types—ultimately affects researchers' ability to contribute to the evidence base.

Widespread Need for Standards to Help Build the Evidence Base. When discussing standardization, participants raised multiple needs and suggestions for making sure there is rich, robust data to build the research evidence base. They mentioned a need for standardized measures of:

- Covariates that are useful to control for confounding and improve methods for causal inference in observational studies using EHRs, such as obesity, smoking, substance abuse, and medications (one of the top five improvements)
- Validated clinical phenotypes for diseases
- Primary, specialty, and sub-specialty care
- Specific diseases and conditions (e.g., traumatic brain injury, autism, etc.)

Functionality 3: Linking Clinical and Other Data for Research

Linking Clinical and Other Data for Research refers to the process of combining disparate data sources such as claims, survey data sets, SDOH data, and physician practice data sets. Combining these data sources requires unique identifiers to facilitate the matching of patients and clinicians across data sets. Linking clinical and other data sources enables PCOR on multi-pronged interventions across health care settings.

Challenges

The challenge-related question engaged 25 distinct contributors in adding, refining, and prioritizing 13 challenges for *Linking Clinical Data and Other Data for Research*. Below, we discuss the details of the top five challenges for this functionality.

Top Five Challenges

1. Challenges matching patients across data sets that come from different sectors (e.g., social service and health care) due to privacy constraints and immature master data management practices.
2. Linkage to data sets that support understanding of environment exposures and the relationship of home and work locations to health.
3. Access to data not commonly collected in the course of clinical care in a study (e.g., vital records, social determinants of health).
4. Decentralized systems to connect disparate data sets for deployment of remote automated differentiated analytics that deliver only combined derived data and keep data at its local source.
5. Lack of practical guidance for the harmonization of different common data models.

Challenge 1: Challenges matching patients across data sets that come from different sectors (e.g., social service and health care) due to privacy constraints and immature master data management practices. Potential barriers to linking data sets due to regulatory policy is a recurring theme across this functionality. Participants specifically highlighted the difficulty of navigating privacy constraints when attempting to link data and match patients across disparate data sources. Given the volume and diversity of patient data, linking patients across different sectors can be a complex process that requires technological resources and well-developed data management practices to ensure the quality and accuracy of datasets.

Challenge 2: Linkage to data sets that support understanding of environment exposures and the relationship of home and work locations to health. In this challenge, participants highlighted the need for environmental data at multiple geographic levels, which are often not available. Access to this type of environmental data would allow researchers to ask important questions about how environmental exposures (e.g., air pollution), workplace factors (e.g., noise at a construction site) or housing conditions (e.g., pests or asbestos) contribute to health outcomes. In addition to navigating the need for multiple geographic levels, combining data sources to examine the role of these exposures would be challenging due to different standards and measures across sectors, technological barriers, and concerns about data quality.

Challenge 3: Access to data not commonly collected in the course of clinical care in a study (e.g., vital records, SDOH). Participants noted that access to data is broadly a challenge for data infrastructure improvement. While data sets produced by research studies can provide a rich data source for pre-specified outcomes, these sources may need to be linked to other types of data to facilitate more robust analyses. For SDOH data, researchers must navigate data privacy and security policies, and consider the financial and technological resources needed to obtain data sets. In addition, data owners may be reluctant to share data resources without clear incentives. In contrast, vital records data are more readily available, though barriers do exist. Some states may charge fees for access, and researchers do not necessarily

have the funds or tools to link their data to these records, so representation of these data in the clinical record would address data access issues.

Challenge 4: Decentralized systems to connect disparate data sets for deployment of remote automated differentiated analytics that deliver only combined derived data and keep data at its local source. This challenge reflects a critical issue of data ownership in the conduct of PCOR. Data collection and use efforts must account for people's individual levels of comfort with openly sharing their data for research. There is a growing interest and need to support distributed research networks, such as PCORnet and Sentinel, where data holders maintain control of their data, allowing access to discrete portions of the data set for data queries and analysis. However, distributed data analysis models may present challenges to certain research efforts, particularly those that utilize artificial intelligence for analysis; these often require access to large amounts of data to train algorithms and then conduct predictive analysis.

Challenge 5: Lack of practical guidance for the harmonization of different common data models (CDMs). A CDM contains multiple data schemas that enable the compilation and translation of data from various sources, in ways that ensure consistency and content and structure. Networks such as PCORnet and Sentinel use CDMs to harmonize data within their systems; however, participants indicate that there is a lack of practical guidance for how to harmonize CDMs across systems. Successful CDM harmonization across systems allows researchers to aggregate data across research networks in order to conduct research on a more expansive network of patients, and to be confident that the data returned is in a format fit for purpose. This supports evidence generation based on data that is more accurately representative of the overall population. Over time, more widespread adoption of Fast Healthcare Interoperability Resources (FHIR®) standards will offer an additional solution for interoperable data sharing as they require data exchange using standardized vocabulary, helping to reduce the reliance on CDMs and mappings.

Improvements

A total of 24 distinct contributors generated 13 ideas focused on improvements to data linkage. Below, we discuss the details of the top five improvements for this functionality.

Top Five Improvements

1. Standardization of core record linkage variables (names, addresses, phone numbers, other pseudo-identifiers).
2. Clarity of roles, responsibilities, and regulations related to cross sector and cross-jurisdiction (e.g., states) data sharing.
3. Work with states to break down barriers to obtaining identifiable Medicaid data for PCOR, with appropriate protections, to enable linkage with EHRs and survey data.
4. Ability to capture longitudinal SDOH and share these data across settings.
5. Updated, public data sets that provide useful zip code information on SDOH-related neighborhood characteristics (food desert, food swamp, walkability, transportation, public parks and green spaces, safety) for use by clinicians and researchers.

Improvement 1: Standardization of core record linkage variables (e.g., names, addresses, phone numbers, other pseudo-identifiers). Data sets often have variables that facilitate the linking of data resources. Record linkage variables include names, addresses, phone numbers, and other identifiers from disparate data sources. However, these variables are not standard across sectors. Efforts from ONC like the United States Core Data for Interoperability (USCDI) and Interoperability Standards Advisory (ISA) are addressing the need for data standardization. Standardized variables would facilitate linkages between clinical and other types of data from disparate sources. Standardization of core record linkage variables would improve capacity for robust data linkage and minimize incorrect or missing data linkages.

Improvement 2: Clarity of roles, responsibilities, and regulations related to cross sector and cross-jurisdiction (e.g., states) data sharing. Cross-sector data sharing often requires researchers to navigate a complex matrix of regulations with multiple actors and data owners. Navigating cross-jurisdiction data sharing between the federal, state, and local level adds another level of complexity. In addition to federal data sharing regulation and policies, each state may have additional policies that guide the process of data governance and data-sharing. Researchers may need support to better understand these roles, responsibilities, and regulations. Easily accessible resources and guidance documents may reduce the burden of cross-sector and cross-jurisdiction data for researchers by providing clear information about the steps and processes that are needed.

Improvement 3: Work with states to break down barriers to obtaining identifiable Medicaid data for PCOR, with appropriate protections, to enable linkage with EHRs and survey data. Medicaid data provides information on health care utilization and expenditures, which can be linked with EHR and survey data to enhance PCOR. However, Medicaid programs are unique to each state, and thereby subject to that specific states' data governance and reporting policies to protect patient privacy. In conducting single-state or multi-state studies, variations in policies can complicate researchers' efforts to obtain identifiable Medicaid data. Other researchers who are unfamiliar and/or uncomfortable with variability in state laws may shy away from multi-state studies because of perceived barriers. Working with states to demystify regulations and other perceived barriers, while developing solutions to data sharing barriers related to privacy, will improve researchers' access to data while also ensuring that data are appropriately protected.

Improvement 4: Ability to capture longitudinal SDOH and share these data across settings. A consistent theme across this functionality is that supporting cross-sector data sharing of SDOH is a critical area for data infrastructure improvement. The collection of longitudinal SDOH data—repeatedly measuring variables relevant to SDOH over periods of time—would allow for a better understanding of the effects of SDOH on health and health outcomes, as well as understanding how outcomes change over time.

Improvement 5: Updated, public data sets that provide useful zip-level information on SDOH-related neighborhood characteristics (e.g., food desert, food swamp, walkability, transportation, public parks and green spaces, safety) for use by clinicians and researchers. Social risk factors play an important role in contributing to observed patient health outcomes. Updating public data sets with specific information on SDOH-related neighborhood

characteristics at the zip code-level can provide a more complete picture of residents' social and environmental risk factors. By indicating whether a community is a food desert, or whether residents face inaccessibility of public transportation, researchers' work may be better informed and more effective in helping to interpret observed patient or population level outcomes. While some data sets already provide this type and level of information, participants emphasized across several challenges and improvements, that more access to SDOH data is needed.

Summary of Themes in Top Challenges and Improvements

The top five challenges and improvements for *Linking Clinical Data and Other Data for Research* highlighted issues related to SDOH data, access to standardized data variables that can support data linkages, and policy barriers related to privacy and data sharing. Data linking challenges and improvements echoed the recurring theme of accessing and using SDOH data more effectively. Stakeholders noted both the lack of access to and accuracy of data that are not commonly collected in the course of clinical care. Priorities in data linking highlighted issues around data privacy and governance as well and pointed to opportunities for better coordination between state agencies to share data.

Additional Key Themes

Two additional themes emerged across this functionality: quality of clinical data and linking clinician data.

Quality of Clinical Data. Several challenges and improvements touched on issues related to the quality of clinical data. Two challenges specifically highlighted the quality and format of EHR and claims data as an important data infrastructure challenge. Stakeholders noted that EHR and claims data are focused on meeting regulatory and billing requirements. Consequently, the data are often not in a format that facilitates PCOR. In addition, these data may be missing key variables that are necessary to answer research questions. One improvement focused on the quality of clinical data outside of traditional health care settings (e.g., minute clinics, pharmacies), calling for policies that would establish minimum standards to ensure that the data collected in these settings can be used for research.

Linking Clinician Data. While patient matching was considered the most important challenge for this functionality, stakeholders also raised the issue of linking clinician data to support observational comparative effectiveness research. One challenge focused on the difficulties linking clinician data, noting the value of linked clinician data to control for clinician prescribing preferences. Participants provided improving unique clinician identifiers as a potential area of improvement.

Functionality 4: Collection of Participant Provided Information

Collection of Participant Provided Information refers to the collection of PGHD from wearable devices, PROs, and patient-reported health data (e.g., reported hospitalizations). PPI provides a robust set of data that directly reflects the patient-experience. These data inform clinical care practice, the interpretation of results from clinical trials, and drug and device surveillance.

Challenges

For *Collecting Participant Provided Information,* 25 distinct contributors added, refined, and prioritized 16 challenges. Below, we provide additional detail on the top five challenges.

Top Five Challenges

1. Lack of standards for collecting patient provided information from medical devices, wearables etc.
2. Lack of guidance and standards for balancing privacy and security concerns for patient data in PCOR with enhanced access to such data.
3. Difficulties integrating PROs and patient-generated health data into EHRs.
4. Inconsistent standards make comparisons of similar data collected from different devices (e.g., glucometers, blood pressure (BP) meters) and mobile applications or wearables difficult.
5. Difficulty collecting PROs especially in populations that have less access to or familiarity with the technology needed to report their outcomes. Lack of a clear ROI for patients to report and attrition in longer term PRO studies impact usability.

Challenge 1: Lack of standards for collecting patient provided information from medical devices, wearables etc. Technological advancements allow for the collection of data directly from patients through medical devices like glucometers and wearable technology like smartwatches and fitness trackers. Multiple challenges for this functionality emphasized the need for increased standardization of PPI data collection. For example, if a patient's resting heart rate is tracked regularly using a smartwatch, data from that source may not be captured in a standardized form that aligns with other clinical records. Inconsistent standards hinder data integration and increase the difficulty of analyzing similar data collected from disparate devices.

Challenge 2: Lack of guidance and standards for balancing privacy and security concerns for patient data in PCOR with enhanced access to such data. The digital health landscape is rapidly evolving, with new technologies and capabilities to capture PPI. Consequently the regulatory landscape is also evolving to ensure patient privacy and data security during the collection and use of PPI. Researchers conducting PCOR may require additional guidance or resources regarding how to maximize the use of PPI while maintaining data privacy and security.

Challenge 3: Difficulties integrating PROs and patient-generated health data into EHRs. The ability to collect data from patients in the form of PROs and PGHD from medical devices and wearables is a significant advancement in PCOR data infrastructure. These data are most impactful when they are integrated into EHRs. Integrating PROs and PGHD facilitates the use of these data during clinical visits and also provides robust clinical data derived from the patient experience within the EHR for PCOR.

Challenge 4: Inconsistent standards make comparisons of similar data collected from different devices (e.g., glucometers, BP meters) and mobile applications or wearables difficult. Mirroring the top challenge for this functionality, this challenge again highlights how a lack of standards and/or inconsistency in standards hinders the comparison of similar data that are collected from different devices. While PPI can be a rich data source, inconsistent standards may reduce the overall usability of PPI for PCOR.

Challenge 5: Difficulty collecting PROs especially in populations that have less access to or familiarity with the technology needed to report their outcomes. Lack of a clear ROI for patients to report and attrition in longer term PRO studies impact usability. Broadly, patients need clear incentives and buy-in to report PROs. This is especially true for longer-term studies where patients may lose interest or feel burdened by providing PROs over an extended period of time. Several challenges within this functionality specifically spoke to barriers to PROs collection from vulnerable or underserved populations. Among these patients, difficulty accessing technology needed to provide PROs (e.g., smartphones) and limited incentives for participation are notable barriers to involvement in research. This creates a cycle of exclusion, leaving certain groups out of important research activities and limiting the generalizability of research findings.

Improvements

For *Collecting Participant Provided Information,* 23 distinct contributors submitted 14 improvements. Below, we discuss the top five improvements.

Top Five Improvements

1. Generate core list of key information (data and metadata) needed from wearables, devices, and mobile apps for researchers to be able to understand which data points are comparable and use them to study patient engagement and impact on outcomes.
2. Disseminate technology standards to support consistent recording and extraction of PRO data in the EHR.
3. Develop and disseminate standards to support the aggregation of and integration of medical device data with other data sets including electronic health record (EHR) data.
4. Need to develop strategies to encourage adoption and use of PROs by clinicians and patients (even if the infrastructure is there, we still need clinicians and patients to use the PROs to have sufficient data for research).
5. How to incorporate this information into clinical practice. Are the data useful, do clinicians know how to interpret, etc.

Improvement 1: Generate core list of key information (data and metadata) needed from wearables, devices, and mobile apps for researchers to be able to understand which data points are comparable and use them to study patient engagement and impact on outcomes. Researchers can most effectively examine the effect of patient activation or adherence on observed outcomes by comparing data from disparate sources. Generating a list of core data elements from these digital health sources will ensure that the PPI captured across these sources is useful and usable for researchers.

Improvement 2: Disseminate technology standards to support consistent recording and extraction of PRO data in the EHR. In addition to developing standards that support the consistent collection and integration of PROs into EHRs, there is a need to disseminate these standards to ensure consistency in the field of PCOR. Dissemination activities may speed up standards adoption.

Improvements 3: Develop and disseminate standards to support the aggregation of and integration of medical device data with other data sets including EHR data. In addition to identifying key data and metadata, participants indicated that there is a need for improved development and dissemination of standards to support aggregating of and integrating information from digital devices (e.g., wearable technology), medical devices, and other data sources like EHRs. Identifying core data elements, and developing standards for their capture and communication would improve the completeness and consistency of data used for research.

Improvement 4: Need to develop strategies to encourage adoption and use of PROs by clinicians and patients (even if the infrastructure is there, we still need clinicians and patients to use the PROs to have sufficient data for research). The collection of PROs is dependent on the participation of both patient and clinicians. Patients and clinicians may be reluctant to use PROs due to perceived burdens. Clinicians may have concerns about time constraints and difficulty fitting PROs into the clinical workflow. For patients, recording and reporting on outcomes may also be perceived as too time consuming. Developing strategies to encourage adoption and use of PROs by both clinicians and patients could ensure sufficient availability of PPI for research.

Improvement 5: How to incorporate this information into clinical practice. Are the data useful, do clinicians know how to interpret, etc. This improvement also touches on the topic of adoption and use, specifically focusing on the need for guidance or other resources to help clinicians use PPI in clinical practice. As noted above, clinicians can struggle with incorporating PROs or other forms of PPI into the clinical workflow. In addition, given the various ways PPI is gathered and the various measures used for PROs, clinicians may be uncertain of how to interpret the data and determine their clinical relevance.

Summary of Themes in Top Challenges and Improvements

The top five challenges and improvements for *Collecting Participant Provided Information* related to data collection and standards, data integration and governance, and adoption of PROs. Four of the top five challenges related to difficulties around standardizing and integration of PPI. Given that use of PPI from medical devices in research is relatively novel compared to other sources of data, it is understandable that the need for data standards is such a prominent theme. Improvements pointed to opportunities to establish common data elements for emerging digital health sources, and technical and non-technical considerations for improving uptake of PROs among both care providers and patients. Multiple challenges and improvements raised the issue of access to technology as a barrier to participation for vulnerable or underserved populations.

Additional Key Themes

Two additional themes were frequently discussed for this functionality: stakeholder engagement for the adoption and use of PROs and patient data sovereignty.

Stakeholder engagement for the adoption and use of PROs. The top challenges and improvements emphasized the importance both patient and clinician buy-in for the adoption of PROs. Stakeholders noted the need to consistently engage patients and make PROs and

PGHD clinically meaningful to patients, and the a need to engage clinicians about how to effectively use PPI. Participants also indicated interest in broader stakeholder engagement to encourage the adoption and use of PROs. Two improvements recommended convening stakeholders to determine specific PROs to be collected on routine basis (e.g., through adoption of the 12-Item Short Form Health Survey [SF-12] or a single item health status measure). One of the improvements specifically suggested engaging PCORI or National Committee for Quality Assurance (NCQA) to convene stakeholders.

Patient Data Sovereignty. One of top five challenges highlighted the need to balance privacy and security concerns to help the research community navigate the use of PPI, while other challenges and improvements focused on patient sovereignty or ownership of data. One challenge specifically stated that data should be fully patient controlled. One improvement focused on the need for technology and cryptography to allow patients' self-sovereignty over their data.

Functionality 5: Use of Enhanced Publicly-Funded Data Systems for Research

Across multiple departments and agencies, HHS collects data that can inform clinical practice and health policy. *Use of Enhanced Publicly-Funded Data Systems for Research* refers to the enhancement of publicly-funded data assets for research by improving the retrieval, linkage, aggregation, and use of this data.

Challenges

For Use *of Enhanced Publicly-Funded Data Systems for Research,* 24 distinct contributors identified and prioritized 16 challenges. Below, we discuss the top five challenges.

Top Five Challenges

1. Difficulties promoting and navigating cross-sector collaboration to bring data sets together at the local level (e.g., county or state).
2. A shared vision and coordinated leadership to promote existing data assets and resources to leverage enhanced data systems.
3. Lack of publicly available data assets to facilitate SDOH analysis.
4. Limited access to data from regional health information exchanges that aggregate data from multiple payers.
5. Need of an infrastructure to allow for federated learning across all publicly funded data where access is allowed without ever moving the data. "Use don't move" with deployed analytics and learning models will rapidly expand research.

Challenge 1: Difficulties promoting and navigating cross-sector collaboration to bring data sets together at the local level (e.g., county or state). Limitations on effective, efficient sharing of data and present a significant challenge for PCOR. Several of the prioritized challenges pointed to a lack of collaboration to leverage publicly-funded data systems and facilitate cross-sector learning. This challenge specifically references the need for increased cross-sector collaboration, at the state and local level.

Challenge 2: A shared vision and coordinated leadership to promote existing data assets and resources to leverage enhanced data systems. Leveraging existing data resources is difficult without a centralized or coordinated leadership body to organize disparate efforts and promote the use of existing data assets and resources. Coordinated leadership—meaning collaboration across organizations and efforts to create a shared vision and approach to promote data resources—could improve dissemination of existing data assets and reduce the risk of duplicative efforts. This could also increase opportunities to leverage publicly-funded data systems for research.

Challenge 3: Lack of publicly available data assets to facilitate SDOH analysis. Several challenges for this functionality focused on the issue of data access. This challenge specifically focuses on the lack of available assets that incorporate SDOH data. Use of SDOH data for research often requires cross-sector data sharing.

Challenge 4: Limited access to data from regional health information exchanges that aggregate data from multiple payers. While this challenge refers to regional health information exchanges (HIEs), the intent is most likely focused on health insurance exchanges that can support PCOR by providing aggregated data from multiple payers. Health information exchanges do not typically bring together payer data, but can provide aggregated data from providers and laboratories. Access to data from health insurance exchanges is governed by both federal and state policies for data privacy and security.

Challenge 5: Need of an infrastructure to allow for federated learning across all publicly funded data where access is allowed without ever moving the data. "Use, don't move" with deployed analytics and learning models will rapidly expand research. Analyzing data collected and stored in various locations traditionally requires that all the data are sent back to a central server for processing. This movement of the data not only limits real-time learning, but also complicates efforts to maintain privacy and security of the data. Federated learning is a machine learning technique that involves a central server that uses training algorithms to develop models across data sets, without requiring full aggregation and centralization of those data assets. The research models' learnings, instead of the individual data assets, are then fed back into the central server. This could expand research capacity to address novel questions, without requiring complicated data transfers.

Improvements

A total of 22 distinct contributors identified and prioritized 14 improvements for the use of federal databases for research. Below, we discuss the top five improvements.

Top Five Improvements

1. Increase ability to readily use, retrieve, link, and aggregate publicly-funded data (federal/state) for research.
2. Develop cross-agency and public "sandboxes" to allow for innovation and exploration of cross-agency data products.
3. Improve access to federal data assets related to public health surveillance to facilitate rapid analysis of clinical and other data for public health response.
4. Improve awareness of and access to federal data assets through creation and maintenance of a single catalog of federal data assets.
5. Increase availability and discoverability of de-identified open data sets that can be readily accessed by diverse users.

Improvement 1: Increase ability to readily use, retrieve, link, and aggregate publicly-funded data (federal/state) for research. Increased access and easier use of publicly-funded data sets would expedite the research process by increasing the availability of clean, quality data that can be easily linked. For example, researchers are currently working to identify and mitigate barriers to accessing the National Death Index (NDI) data. Increasing access and use of these data could expand the research community's ability to examine research questions related to mortality.

Improvement 2: Develop cross-agency and public "sandboxes" to allow for innovation and exploration of cross-agency data products. Creating an environment for researchers to explore cross-agency data, or to collaborate on cross-agency projects, would support learning and information sharing. For example, opioid epidemic data are collected by multiple federal agencies; offering a designated, shared space for compiling cross-agency data exploration would give researchers the opportunity to share resources and combine efforts.

Improve 3: Improve access to federal data assets related to public health surveillance to facilitate rapid analysis of clinical and other data for public health response. The current global pandemic has demonstrated the need for access to surveillance data to inform responses to public health events.

Improvement 4: Improve awareness of and access to federal data assets through creation and maintenance of a single catalog of federal data assets. Researchers may need to navigate multiple websites, data use agreements and request processes to determine what federal data assets are available and how to access them. Creating a single catalog of the process for acquiring federal data assets may facilitate use of available data assets. Promotion and dissemination of such a catalog might also encourage federal data utilization by diverse users.

Improvement 5: Increase availability and discoverability of de-identified open data sets that can be readily accessed by diverse users. Access to publicly-funded data sets often require an approval process, data use agreement, and payment. Several challenges and improvements specifically alluded to the cost of accessing data. Creating and promoting de-identified, open data sets—available without complex approval processes or access fees—

would reduce barriers to access and the overall burden of obtaining data, especially for researchers with resource constraints.

Summary of Themes in Top Challenges and Improvements

The top five challenges and improvements for *Use of Enhanced Publicly-Funded Data Systems for Research* emphasized the importance of collaboration across levels of government, within the federal government, and across sectors. Consistent with the other functionalities, the top five challenges highlighted the issue of SDOH data as well as the technical barriers encountered when engaging in cross-sector collaboration. Suggested improvements focused on increased access to publicly-funded data, opportunities to encourage cross-agency innovation, and ideas for how to make data systems more user-friendly.

Additional Key Themes

One additional key theme emerged for this functionality: the timeliness of data.

Timeliness of Data. Multiple challenges and improvements discussed the timeliness of data assets. Several improvements focused on the need for federal policies and practices to reduce the latency for these data assets. Stakeholders specifically called out the need to address the two-year lag in the NDI and the four-year lag in Medicare encounter data, as well as lags in Medicaid data (e.g., Transformed Medicaid Statistical Information System) and the Area Resource File.

Discussion

Looking to the future, stakeholders generated numerous suggestions for challenges and improvements for PCOR data infrastructure. Using Codigital as a mechanism for soliciting this feedback resulted in high levels of engagement, with 40 attendees for the virtual Listening Session and an average of 27 contributors per question. The results do not represent an exhaustive list of challenges and improvements for PCOR data infrastructure, but rather the viewpoints of the activity's participants. Within and across functionalities, the themes from this exercise are in strong alignment with ASPE's Strategic Framework and offer insight on areas for expansion and development.[2]

Below, we discuss themes that apply across the functionalities, implications for ASPE, and areas for further exploration.

Cross-Functionality Themes

Across the five functionalities, participants returned to five common themes that expressed the need to: 1) enhance consistency in data standardization; 2) improve access to SDOH data that are not routinely collected during care delivery; 3) improve ability to access, integrate and use PPI, particularly those data generated from medical devices and wearables; 4) increase access

[2] Dhopeshwarkar R, Dullabh P, Dungan R, et al. Building Data Capacity for Patient-Centered Outcomes Research Portfolio Highlights (2016 – 2019) Impact, Opportunities, and Case Studies. Washington, DC: Office of the Assistant Secretary for Planning and Evaluation. https://aspe.hhs.gov/system/files/pdf/259016/OS-PCORTFImpactReport508.pdf

to federal data sets, with an emphasis on access to de-identified data sets; and 5) expand collaboration across organizations at the local, state, and federal level.

Enhance consistency in data standardization. In addition to establishing standards, multiple challenges and improvements that participants raised related to the issue of consistency or transparency in data standardization. Participants focused on the need for consistent processes for collecting (Functionality 1 and 4), cleaning (Functionality 2), and presenting data (Functionality 1 and 4). They also highlighted the importance of promoting adoption of and adherence to standards across the health system after they are developed.

Improve access to SDOH data that are not routinely collected during care delivery. Access to SDOH data was a consistent theme across the challenges and improvements for use of clinical data for research (Functionality 1), standardizing data collection (Functionality 2), linking clinical data with other data for research (Functionality 3), and use of enhanced publicly-funded data systems for research (Functionality 5). Specifically, participants sought resources to support the standardized collection of SDOH data, and expressed desire for expanded access to federal data sets to support research inquiries related to SDOH, including zip code level data on neighborhood characteristics.

Improve ability to access, integrate, and use PPI, particularly those data generated from medical devices and wearables. Similar to the theme around SDOH data, stakeholders noted the importance of accessing PPI, including PROs and PGHD, from medical devices to support their research inquiries. This theme was noted among the top five ideas of three functionalities (Functionality 1, 2 and 4). Participants focused on the need to develop and disseminate standards to support PPI data collection and analysis (including the collection of PROs) and the aggregation and integration of PPI into EHRs. Participants also sought mechanisms to promote collection and use of PROs among patients and clinicians, suggesting that PROs must be clinically meaningful to patients and relevant to workflow.

Increase access to federal data sets, with an emphasis on access to de-identified data sets. Across the functionalities, access to data sources was a prominent theme. As previously discussed, participants focused on access to SDOH data resources across multiple functionalities (1, 2, 3, and 5). Participants also highlighted the need for increased access to federal data resources. This topic was particularly prominent for Functionality 5, where participants focused on the need for easily accessible, de-identified federal data sets, as well as the need for access to surveillance data, an emerging topic given the current global pandemic.

Expand collaboration across organizations at the local, state and federal level. Under every functionality, participants highlighted the need for collaboration to leverage and enhance existing data sources and infrastructure. Collaboration was discussed at both the meso-level (e.g., collaboration to enable cross-sector data sharing) and the macro-level (e.g., regulatory frameworks, enhanced federal data assets, and development of and incentives for standards adoption). Based on the comments across all functionalities, participants foresee the need for widespread cooperation to make data available and useful for research, while maintaining the privacy and security of patient health information.

Future Directions

In the section that follows, we consider the implications of this activity for the future of PCOR data infrastructure. Specifically, we discuss: 1) how the findings intersect with the OS-PCORTF Strategic Framework; and 2) areas in need of development to expand the capacity of PCOR data infrastructure.

Implications for ASPE's Strategic Framework

The priority challenges and improvements identified in this activity generally aligned with ASPE's existing strategic framework. Participants did not signal a need to expand the definitions of the five functionalities or significantly adjust ASPE's strategic framework. Overall, the activity served to underscore the importance and relevance of the framework and the work of the OS-PCORTF.

For example, the challenges emphasized the importance of disseminating existing standards, policies, and best practices. Suggested improvements called for tools and guidance to enhance interoperability and facilitate data linkages. Such improvements would render data from patients and their social context (e.g., PPI and SDOH) and publicly-funded data resources more valuable. Cataloging federal data assets would increase researchers' awareness and use of such resources. Improved ability to link these and other data would newly equip researchers to explore robust questions that require integration of claims, clinical, and patient-generated health data. This serves as one example of the ways stakeholder input from this activity aligns with ASPE's own priorities. Below, we discuss how current projects funded by the OS-PCORTF align with the stakeholder-generated challenges and improvements.

Alignment with OS-PCORTF Projects

It is important to recognize that ASPE has already funded OS-PCORTF projects that are intended to help address several of the top five identified challenges and improvements across functionalities. Select examples of alignment between projects and challenges and improvements are provided below in Table 2.[3] [4] This alignment is promising and further underscores the need for dissemination of OS-PCORTF-funded project work and the resulting products to the research community.

[3] Dullabh P, Dhopeshwarkar R, Heaney-Huls K, Hovey, L, Rajendran N, Moriarty E, Steiner C. Building the Data Capacity for Patient-Centered Outcomes Research: The 2018 Annual Report. Prepared under Contract No. HHSP233201600020I. https://aspe.hhs.gov/system/files/pdf/259016/2018PortfolioReport.pdf

[4] Dullabh P, Dhopeshwarkar R, Heaney-Huls K, Sanders E, Hovey, L, Rajendran N, Moriarty E, Sidi M. Building the Data Capacity for Patient-Centered Outcomes Research: The 2019 Annual Report. Prepared under Contract No. HHSP233201600020I. https://aspe.hhs.gov/system/files/pdf/259016/2019OS-PCORTFPortfolioReport.pdf

Table 2. Examples of OS-PCORTF Projects That Address Functionalities and Stakeholder Suggestions

Top Five Ideas	Project Title and Brief Description of Project Objectives
Functionality 1. Improvement #2: Further develop and disseminate standards, services, and policies to assure data quality and metadata descriptions for research and practice	**Standardization and Querying of Data Quality Metrics and Characteristics for Electronic Health Data (FDA)** Create and implement a metadata standard data capture and querying system for data quality and characteristics, data source and institutional characteristics, and "fitness for use."
Functionality 1. Improvement #4: Develop and implement core outcome sets to build connections across research, clinical practice, and quality improvement and improve the relevance of routinely recorded clinical data for research.	**Harmonization of Clinical Data Element Definitions for Outcome Measures in Registries (AHRQ)** Convene clinical topic-specific working groups to discuss current outcome measures and how their data definitions can be harmonized to promote the use of common definitions across systems. The resulting definitions are to be made publicly available to PCOR researchers and analysts. **Capstone for Outcomes Measures Harmonization Project (AHRQ)** Improve collection and use of outcomes measures by linking clinical data to two different registries and pilot testing the bidirectional exchange of data between the registries and clinical sites.
Functionality 2. Improvement #5. Establish best practices for the standardized collection of patient generated health data (PGHD), including patient-reported outcomes (PROs), SDOH, and other data. **Functionality 4.** Challenge #3: Address difficulties integrating PROs and patient-generated health data into electronic health records (EHRs).	**Advancing the Collection and Use of Patient-Reported Outcomes through Health Information Technology (AHRQ, ONC)** Develop technical tools for collecting and integrating standardized PRO data into EHRs or other health information technology systems.
Functionality 3. Challenge #5: Offer practical guidance for the harmonization of different common data models (CDMs).	**Harmonization of Various Common Data Models and Open Standards for Evidence Generation (FDA, NIH/NCI, NIH/National Center for Advancing Translational Sciences ((NCATS), NIH/NLM, ONC)** Build data infrastructure for conducting PCOR using data from routine clinical settings, including insurance billing claims, EHRs, and patient registries. Harmonize several existing common data models, potentially including PCORnet and other networks.

Top Five Ideas	Project Title and Brief Description of Project Objectives
Functionality 3. Improvement #5: Offer updated, public data sets that provide useful zip-level information on SDOH related neighborhood characteristics (food desert, food swamp, walkability, transportation, public parks and green spaces, safety) for use by clinicians and researchers.	**Enhancing Patient-Centered Outcomes Research (PCOR): Creating a National Small-Area Social Determinants of Health Data Platform** Develop a consolidated set of national standardized databases on valid and reliable SDOH factors at the small-area and other geographic levels, building on existing databases developed by federal agencies (e.g., AHRQ, the Health Resources and Services Administration (HRSA), CDC, ASPE, and NIH). Include SDOH data elements related income, employment, food, housing, environment, economics, education, safety, transportation, etc., in addition to health status and health care access and utilization.

Alignment with ASPE Milestones for Progress

The priority challenges and improvements are also closely aligned with the milestones ASPE has identified for the five functionalities of the OS-PCORTF Strategic Framework (**Appendix C**).[5] On a macro-level, success will be measured against the development and use of standards, services, policies and governance to support the data infrastructure functionalities.

Milestones for the Use of Clinical Data for Research. The top five challenges related to the use of clinical data for research all point to a need to reach the third milestone of *standards* that support secure, electronic query of structured data across clinical research and delivery systems, including standards for open-access. This will help researchers handle technical barriers to accessing clinical information from EHRs or other clinical information systems. Participants also highlighted the need to develop and disseminate metadata standards to ensure data quality and establish a common understanding of whether data meets "fitness for use" criteria for specific research purposes. This priority

Use of Clinical Data for Research Milestones

1. Establish services and tools to support data access, querying, and use, including privacy-preserving analytics and queries. These services and tools would be leveraged nationally and are not likely to be developed by the private sector.
2. Develop support services and tools that can be leveraged nationally and are not likely to be developed by the private sector; these tools would test the quality of unstructured and structured data to answer PCOR questions.
3. Develop standards that support secure, electronic query of structured data across clinical research and delivery systems, including standards for open-source access.
4. Develop and test metadata standards that describe data quality.
5. Create a policy framework for privacy-preserving access and querying of clinical data by researchers conducting PCOR, and policies that govern the use of the services that support data access, querying, and use.
6. Develop a policy framework for ensuring clinical data used for research is of "research grade."

[5] Dhopeshwarkar R, Dullabh P, Dungan R, et al. Building Data Capacity for Patient-Centered Outcomes Research Portfolio Highlights (2016 – 2019) Impact, Opportunities, and Case Studies. Washington, DC: Office of the Assistant Secretary for Planning and Evaluation. https://aspe.hhs.gov/system/files/pdf/259016/OS-PCORTFImpactReport508.pdf

aligns with the fourth milestone related to developing and testing metadata standards for data quality.

The improvements also align with key milestones for this functionality. Pursuing several of the improvements could support the development of *tools and support services* unlikely to be developed by private sector (the first milestone), but crucial for advancing the national data infrastructure. Such tools would offer new means of testing the quality of unstructured and structured data, including the SDOH and PPI noted by participants throughout the prioritization activity. Finally, creating policy-driven incentives to ensure adherence to data standards would facilitate secure, electronic query of data across clinical research and health care delivery systems.

Standardizing Data Collection Milestones

1. Support the development of a set of research common data elements (CDEs) in specific gap areas and support development of a governance structure for CDE harmonization.
2. Support the development of repositories/portals for CDEs, standards for utilizing CDEs for research, and services to allow researchers to easily utilize standardized components.
3. Support research and/or crowd-sourced methods to determine which of the standardized collection components and services are most valuable.
4. Create policies to promote the adoption and use of valuable standardized collection components and services.

Milestones for Standardizing Data Collection. While stakeholders did not specifically focus on the need for common data elements as described in the first milestone, several of the challenges relate to a need for more consistent adoption and use of data standards for specific data types—including common clinical variables and data not routinely collected within clinical settings such as SDOH and PPI. These challenges point to opportunities for policies and services that support researchers' use of available standards in data collection, cleaning and analysis as described in the third and fourth milestones.

The top five suggested improvements relate directly to the creation of policies to— promote the adoption and use of standards and services—the fourth milestone. The activity also highlighted a perceived gap in CDEs for patient-provided data and outcomes measures. The recommendations to address this gap align well with the first two milestones, which specifically address CDE development and harmonization.

Linking Clinical and Other Data for Research Milestones

1. Leverage existing standards and support the development and balloting of needed standards of patient data linkage.
2. Establish HHS policies that promote appropriate data-linking based on the frame work noted in the milestone above.
3. Create a policy framework to facilitate patient data linkage in accordance with existing laws.

Milestones for Linking Clinical and Other Data for Research. Many of the priority challenges emphasize, with different levels of specificity, barriers to data sharing and standardization that ultimately hinder data linkage efforts. Stakeholders called for more data sharing across sectors (e.g., ensuring capture and integration of social service data with health care data) and more efficient data linkage (e.g., addressing challenges with patient matching and record linkage variables). These challenges echo all three milestones focused on existing standards and policies that promote patient data linkage.

Suggested improvements echo some of the same major themes, while also offering guidance regarding how to make progress towards the three milestones. Some of the priority improvements suggested continued work to increase data quality and standardization, and others emphasized the need for policies that better enable the use and sharing of data. Overall, informants highlighted a need for improved policy frameworks (from the federal to the organizational level), or a series of standard operating procedures, that can inform data linkage and governance activities, which is a need the milestones also acknowledge.

Collection of Participant-Provided Information Milestones

1. Support the development of tools and services that can be leveraged nationally and are not likely to be developed by the private sector. These tools and services will facilitate the collection and exchange of PPI, including national services for electronic capture and management of PPI and release of data for PCOR.
2. Support the development of a core set of standards for the collection and integration of prevalent use cases of PPI for PCOR, by leveraging existing standards and filling gaps.
3. Create policies and share best practices for collection and integration of prevalent use cases of PPI for PCOR.

Milestones for the Collection of Participant Provided Information. Participant-generated challenges and improvements relating to collecting, standardizing, and integrating PPI data are in clear alignment with milestones for this functionality. In relation to the second milestone, participants specifically highlighted a need for standards to support the collection and use of data from wearables and other remote-monitoring devices. Their prioritized challenges and improvements also indicated a need for strategies to encourage clinician and patient adoption and use of PROs while balancing privacy and security concerns related to leveraging PROs in research, which is echoed across the three milestones.

Milestones for the Use of Enhanced Publicly-Funded Data Systems for Research. The challenges focused on addressing data access issues aligns with the first milestone on enhancing publicly funded data systems to facilitate their access and ease retrieval for research purposes. Additionally, one of the identified challenges highlights the need for a federated infrastructure to support learning where researchers can access publicly funded data without needing to move it.

Use of Enhanced Publicly-Funded Data Systems for Research Milestones

1. Support the enhancement of strategic publicly-funded data systems (including CMS data) to facilitate their access and use, and ease retrieval of data for research purposes.
2. Support the further development of key federally initiated data systems for research.

The informants' suggested improvements aim to make publicly-funded data systems more accessible to researchers. For example, a tool to catalog federal data assets may increase the ease of data retrieval for research. In addition to improving navigation of federal data assets, another suggestion encourages the development of a devoted space for cross-agency data. These suggestions align with the second milestone to support further development of key federal data systems for research.

Areas for Further Exploration

We developed this activity with the understanding that it would provide ASPE initial considerations and insights for further refinement during its strategic planning process. Consequently, our approach to the prioritization activity did not include follow-up with participants to clarify responses.

The activity revealed opportunities to further specify priorities across functionalities. It is important to note that participants were not explicitly asked to prioritize responses in one functionality area over those in another. ASPE could consider running a follow-on activity that would ask respondents to prioritize two batches of information: 1) the top five challenges across all functionalities, and 2) the top five improvements across all functionalities. This could provide an additional level of input as to the most pressing challenges and promising improvements. **Appendix D** provides process considerations for future ASPE strategic planning activities.

There is also an opportunity to explore results from this activity with federal partners and end users. To deepen the agency's understanding of the themes that arose, ASPE could consider posing additional questions to both federal partners and other end-users around specific areas of interest (Table 3).

Table 3. Additional Areas to Probe with Federal Partners and Other End-Users

Federal Partners	• How can ASPE collaborate with your agency and others to enhance the quality and consistency of data for PCOR over the next decade? • How can we work together to make publicly-funded data systems more user-friendly? • How would you describe your agency's vision for PCOR data infrastructure? • What are your priorities and/or insights when it comes to addressing a lack of standardized data definitions? • How are you addressing the standardization and integration of SDOH data? • How can ASPE better support cross-agency innovation and shared learning?
End Users	• What steps can ASPE take in collaboration with end-users to enhance the quality and consistency of data for PCOR over the next decade? • What improvements are needed to make publicly-funded data systems more user-friendly? • What advances are on the horizon for EHR data usability for research purposes? • What innovations are you seeing related to the standardization and integration of non-clinical data such as SDOH data? • What are the barriers to an uptake in the use of PROs among both care providers and patients? • What is your approach to addressing the issue(s) of underrepresentation of vulnerable populations in health research?

Finally, stakeholders often noted challenges and suggested improvements that portfolio projects are actively addressing. This points to additional opportunities to familiarize stakeholders with the work of the portfolio. Specifically, dissemination strategies could be developed to raise awareness of how existing projects have contributed to PCOR data infrastructure so far and the ways in which they address priority challenges.

Appendix A. Virtual Listening Session Participants

Name	Affiliation
Participants	
Alyce Adams	Kaiser Permanente Division of Research
Joseph Blumenthal	MedStar
Thomas Carton	Louisiana Public Health Institute
Vishal Chaudhry	Washington State Health Care Authority
Christopher Chen	Washington State Health Care Authority
Lillian Coral	Knight Foundation
Amy Costello	University of New Hampshire
Peter Ekhart	Illinois Public Health Institute
John Glaser	Cerner
Crispin Goytia-Vasquez	Icahn School of Medicine at Mount Sinai
Claudia Grossman	Patient-Centered Outcomes Research Institute
Amy Hawn Nelson	Actionable Intelligence for Social Policy at the University of Pennsylvania
Benson Hsu	Sanford Health
Joyce Hunter	Vulcan Enterprises
Abel Kho	Feinberg School of Medicine at Northwestern University
Michelle Leavy	Agency of Healthcare Research and Quality
Kristin Lyman	Louisiana Public Health Institute
Sean Manion	ConsenSys Health
Keith Marsolo	Duke University
Erika Martin	University at Albany
Rozalina McCoy	Mayo Clinic
Deven McGraw	Ciitizen
Eneida Mendonca	Regenstrief Institute, Inc.
Penny Mohr	Patient-Centered Outcomes Research Institute
Tim Pletcher	Michigan Health Information Network Shared Services
Maik Schutze	Commonwealth of Kentucky Cabinet for Health and Family Services
Art Sedrakyan	Weill Cornell Medicine
Nilay Shah	Mayo Clinic
Nirav Shah	Stanford University Department of Medicine
Stephanie Shimada	Veterans Affairs
Eric Sid	NIH's National Center for Advancing Translational Sciences
Julia Skapik	National Association of Community Health Centers
Rajan Sonik	AltaMed Institute for Health Equity
Alexander Turchin	Brigham and Women's Hospital
Theresa Walunas	Feinberg School of Medicine at Northwestern University
Neely Williams	Community Partners' Network, Inc.

Name	Affiliation
OS-PCORTF Staff	
Lauren Adams	AcademyHealth
Mandar Bodas	Office of the Assistant Secretary for Planning and Evaluation
Rina Dhopeshwarkar	NORC at the University of Chicago
Nancy De Lew	Office of the Assistant Secretary for Planning and Evaluation
Priyanka Desai	NORC at the University of Chicago
Prashila Dullabh	NORC at the University of Chicago
Rachel Dungan	AcademyHealth
Aldren Gonzales	Office of the Assistant Secretary for Planning and Evaluation
Priya Govil	AcademyHealth
Angel Han	AcademyHealth
Allison Isaacson	AcademyHealth
Susan Lumsden	Office of the Assistant Secretary for Planning and Evaluation
Lisa Simpson	AcademyHealth
Scott Smith	Office of the Assistant Secretary for Planning and Evaluation
Marcos Trevino	Office of the Assistant Secretary for Planning and Evaluation

Appendix B. List of Participant Responses, by Functionality

Using the Codigital platform, participants reviewed prescribed questions and generated the following responses.

Functionality 1. Use of Clinical Data for Research

What are the most pressing data infrastructure challenges related to the use of clinical data for research in the next 10 years?

1. Disconnect between the data necessary for many research questions (e.g., SDOH, PROs) and the data routinely recorded in EHRs, particularly the data recorded in structured form
2. Lack of interoperability across EHR software platforms, settings and sectors (e.g., between health care, social service and public health sectors
3. Issues with data quality, such as data inaccuracy or incompleteness, with electronic health record (EHR) data
4. Lack of alignment across standards efforts. OS, ONC, FDA, etc. all pushing efforts for slightly different purposes. Need to focus in order to increase adoption
5. Lack of standardized data definitions/elements and audits for interrater reliability in individual registries, research studies, and harmonization across registries/other data sources will impact the reliability/usability of data from these sources
6. Issues of data quality including duplicate data points for patients across a city that use multiple health systems for their care
7. Fragmented EHR market leading to a large number of data models, forming a barrier to data sharing
8. Inefficiencies, duplication, lack of standardization, and difficulty linking across multiple investments in large electronic data warehouses (e.g., Sentinel, PCORnet, CDC's Vaccine Safety Data link, NIH's Clinical Research Network, MHRN)
9. The interoperability of registries with EHRs (and other data sources) in order to reduce the burden of data collection and improve data quality will continue to be a challenge for the foreseeable future even with initiatives like FHIR
10. Existing resources linked by common data models (such as PCORnet CDM, OMOP and ACT/i2B2) are available but difficult to access because there are few friendly gateways that make the access process clear and easy
11. Trust in data and analysis quality is limited due to lack of transparent audit trail for all data back to the source and other contextual metadata (e.g., workflow, setting) that may alter interpretation
12. Lack of nationally representative cross-institutional datasets that combine data from complementary sources (e.g., claims and EHR data)
13. Lack of semantic data standards for research in many medical subfields prevent useful data merging

14. The lack of patient knowledge, and involvement in the process of procuring the data. Patients participating in the development and being informed, educated, and co-curators of the data by patients would enrich the source and outcomes of BIG DATA
15. Long delay between high quality, nationwide data generation (as opposed to single institution EHR studies) and the data being made available for research use, precluding timely research to inform practice
16. Existing publicly available clinical data resources (e.g., Centers for Medicare and Medicaid Services (CMS)) have a significant time lag impairing their usability for clinical research (e.g., for outcome ascertainment for pragmatic clinical trials)
17. Limited investment in and facilitation of nationally representative clinical cohorts that can be used for clinical research and quality improvement
18. High cost of many data sources, precluding investigators from being able to access it in the absence of a substantial grant or institutional investment
19. Access to available data by all qualified researchers including patients and advocates
20. Prohibitive (for academic research) cost of existing large clinical datasets especially if not part of a large NIH-funded research team
21. Lack of financial incentives for making commercially produced data available at reasonable cost to researchers
22. Lack of knowledge/understanding about new and emerging methods for extracting data (particularly from unstructured clinical notes) for research
23. Risk for re-identification of rich health data is greater than we thought. The HIPAA "safe harbor" standard is not adequate protection

What improvements are needed (e.g., policies, governance, standards, services, technology) in the use of clinical data for research over the next 10 years?

1. Create policy or financial incentives for EHR vendors to adopt consistent standards in how they structure key patient demographics and utilization data to facilitate comparisons across sites and systems.
2. Further development and dissemination of standards, services, and policies to assure data quality and metadata descriptions for research and practice
3. Rigorously assess investments in large electronic data warehouses (e.g., PCORnet, Sentinel, NIH CRN) to consider synergies, gaps, and avoiding wasteful redundancies for further building out of infrastructure.
4. Develop and implement core outcome sets to build connections across research, clinical practice, and quality improvement and improve the relevance of routinely recorded clinical data for research
5. Better incentives for adoption of standards, services, etc. within health systems and health plans
6. Improve dissemination and utilization of emerging methods for using and extracting data for research (e.g., natural language processing, artificial intelligence, machine learning)

8. Facilitate creation of public-private partnerships that could combine and make publicly available data from government (e.g., CMS or VA) and private sources
9. Facilitate creation of the market for [de-identified] clinical datasets to provide incentives for data creators to improve data quality, timeliness, and availability
10. There needs to be policies in place for the following: Data Quality Data use Structure Patients need to be engaged in this process because ultimately it is their personal healthcare data that will be made available
11. Provide resources to address non-technological challenges (e.g., human, legal, regulatory, reputational or organizational challenges) that are barriers to the use of clinical data for research
12. Convene clinical and research stakeholders to evaluate which existing standards for clinical data collection best meet the needs for researchers, to identify a standard that maximizes utility for both purposes
13. Convene stakeholders to agree on principles, procedures, and tools regarding privacy protection and minimizing re-identification risk for rich health data
14. Better applications of federated learning and decentralized AI to reduce the need to send/copy/transfer data, allowing analytics and machine learning to be deployed while data stays local
15. Reduce disparities between IRB and data privacy/access regulations across geographic jurisdictions to improve researchers' ability to utilize standardized and consistent nationwide data for research
16. More infrastructure for auditable, controlled access of data via decentralized ledgers to track any use or changes to data, enabling more researchers to accelerate new knowledge development
17. Further develop tools such as bulk FHIR to enable transfer of data for research purposes
18. Improve timeliness of availability of federal and state government (e.g., CMS, state Medicaid) data sources
19. Make non-CMS government clinical data (e.g., VA or TriCare) publicly available for research
20. Reduce financial barriers for smaller academic teams to acquire and use clinical data for research

Functionality 2. Standardized Collection of Standardized Clinical Data

What are the most pressing data infrastructure challenges related to the standardized collection of standardized clinical data in the next 10 years?

1. Creating data standards that EHR vendors adhere to and are required across EHR platforms and establishing infrastructure to test and ensure EHR vendors adhere to standards
2. Lack of investment in development/deployment of technical infrastructure to support social services, which generate SDOH data just as clinical services generate EHR and clinical claims data
3. Limited access to data not commonly collected in the course of clinical care in a study (e.g., social determinants of health data)
4. Lack of standardized methods for cleaning clinical data. Standardizing clinical data capture is an important but much larger task. Agreeing on standardized methods for cleaning and mapping clinical/SDOH data (some exist already) would be helpful
5. Varying levels of local implementation of standardized terminology for labs (LOINC), medications (RxNorm, CUI), diagnoses (ICD-10 and SNOMEDCT)
6. How to incorporate SDOH and PROs into the EMR for use
7. Lack of incentives to capture data in a high-quality fashion at the point of care. Better data supports better research, precision medicine, etc. We should pay for good data quality
8. Lack of standardized outcome measures for many condition areas
9. Limited systematic documentation and the lack of publicly available development / testing of metadata standards to support assessment of data quality and fitness of use
10. Many of the critical pieces of information for understanding care and outcomes are in clinical notes which are unstructured and very costly to translate into useful, structured data that can be easily mined
11. Difficulties incorporating social determinants of health (SDOH) and other data into an electronic health record (EHR) system
12. Lack of sufficient secondary demographic or treatment data available with primary data for subgroup analysis
13. Over-reliance on FHIR and other standards that lack sufficient semantic detail across medical sub-specialties to be effective for research
14. Lack of standardized measures for covariates that are useful to control for confounding and improve methods for causal inference in observational studies using EHR
15. Difficulties obtaining data or access to data, especially for patient-led research initiatives (i.e., where patients collect and contribute their data for research)
16. Current point-and-curse interface methods are unacceptable for capturing detailed structured clinical data during complex clinical workflows. A new revolution in interface technologies, not immediately obvious today, must replace methods
17. Validated clinical phenotypes for some diseases

18. The push for standardized data collection, and the need to meet regulatory requirements, have diminished the richness and clinical specificity of collected data
19. Limited guidance on expert determination for making de-identified datasets available in open formats

What improvements are needed (e.g., policies, governance, standards, services, technology) in the standardized collection of standardized clinical data over the next 10 years?

1. Federal policies and financial incentives to promote the adoption and use of standards across sectors
2. Standardize measurement of critical clinical covariates, such as obesity, smoking, substance abuse, medications
3. Integration of different data sources, including patient-provided data, and real-time data reconciliation to address 1) missing data and 2) conflicting data
4. Development and dissemination of standardized patient outcome measures that can be used in research and captured and used across clinical care settings (e.g., primary care + specialty)
5. Establish best practices for the standardized collection of patient-generated health data (PGHD), including patient-reported outcomes (PROs), social determinants of health, and other data
6. Improve the dissemination of standards and document uptake to demonstrate their value in the research and clinical practice settings
7. Facilitate the development of core outcome sets for PCOR in different disease areas
8. Filling gaps in existing standards (e.g., social determinants of health data) through stakeholder engagement and dissemination.
9. Develop and disseminate a common model for cleaning and mapping clinical data recorded in disparate ways that individual systems can use regardless of the nuances of how their individual systems record the information they capture
10. Deep, granular semantic data standards for each sub-specialty area or disease model (e.g., traumatic brain injury, autism, etc.)
11. Fewer standards that are more widely adopted. Focus on getting the core right, including usability, and then expand over time
12. Better provider training and less reliance on insurance coding standards
13. Collaborate with legal aid organizations to help with understanding social needs and their relative levels of severity. Legal aid organization are experts at SDOH and help coordinate systems of SDOH-related advocacy. Their knowledge will be key here
14. Establish best practices for expert determination on de-identification for making data publicly available in open formats
15. Services and technologies that allow patients to participate more fully in clinical research by facilitating data donation of their electronic health record (EHR) data

Functionality 3. Linking of Clinical and Other Data for Research

What are the most pressing data infrastructure challenges related to linking clinical and other data for research in the next 10 years?

1. Challenges with matching patients across data sets that come from different sectors (e.g., social service and health care) due to privacy constraints and immature master data management practice
2. Linkage to datasets that support understanding of environment exposures and the relationship of home and work locations to health
3. Access to data not commonly collected in the course of clinical care in a study (e.g., vital records, social determinants of health)
4. Decentralized systems to connect disparate datasets for deployment of remote automated differentiated analytics that deliver only combined derived data and keep data at its local source
5. Lack of practical guidance for the harmonization of different common data models
6. Lack of a legal/regulatory framework to facilitate the ethical and responsible integration of clinical and administrative data
7. Most data (EHR, claims) exists to facilitate billing and meet regulatory requirements, not to support patient health or research. Hence, needed data is either not included or is included in a format that is not as usable/useful as it could be
8. Linking clinical data with administrative data from social safety net programs that address SDOH (e.g., SNAP, SSI) could provide key insights, but the data and legal investments needed to achieve these linkages are not prioritized
9. Create legal templates, technical support, and other resources to facilitate linking healthcare data with administrative data from safety net programs that address SDOH (e.g., SNAP, SSI, public housing)
10. Lack of a Universal Medical Identifier (outlawed by HIPAA)
11. Difficulty linking to clinician data - making it difficult to control for clinical prescribing preferences, which are useful for observational CER
12. Data quality issues for electronic health data (EHR) and claims data
13. Advanced systems for keeping patient data in the hands (and control) of patients who can provide access to combined identifiable PHI and tangential behavioral data at the time of their choosing through advanced privacy cryptography (e.g., ZKP)

What improvements are needed (e.g., policies, governance, standards, services, technology) in linking clinical and other data for research over the next 10 years?

1. Standardization of core record linkage variables (names, addresses, phone numbers, other pseudo-identifiers)
2. Clarity of roles, responsibilities, and regulations related to cross-sector and cross-jurisdiction (e.g., states) data sharing

3. Work with states to break down barriers to obtaining identifiable Medicaid data for PCOR, with appropriate protections, to enable linkage with EHRs and survey data
4. Ability to capture longitudinal social determinants of health (SDOH) and share these data across settings
5. Updated, public datasets that provide useful zip-level information on SDOH related neighborhood characteristics (food desert, food swamp, walkability, transportation, public parks and green spaces, safety) for use by clinicians and researchers
6. Practical guidance that harmonizes common data models
7. Create policies to ensure that clinical data collected outside of traditional healthcare settings (e.g., minute clinics, pharmacies) meets certain minimum standards so that they can be used for more than just billing
8. Improve unique identifiers for clinicians to enable linkage to patient and health system level data for observational CER
9. Make publicly available datasets that include linked clinical and non-clinical data that is already in government possession (e.g., census or other survey data aggregated by census tract)
10. Create a reliable opt-out mechanism for individuals that would improve public acceptance of data sharing and integration between clinical and other data sources
11. Develop and maintain a repository that identifies state and other policy changes in critical areas of healthcare that might lend themselves to natural experiments
12. Better deployment of decentralized AI/machine learning to standardize data in conjunction with analysis
13. Improved tools and resources for privacy preserving record linkages

Functionality 4. Collection of Participant-Provided Information (PPI)

What are the most pressing data infrastructure challenges related to the collection of participant-provided information in the next 10 years?

1. Lack of standards for collecting patient provided information from medical devices, wearables etc.
2. Lack of guidance and standards for balancing privacy and security concerns for patient data in PCOR with enhanced access to such data
3. Difficulties integrating PROs and patient-generated health data into electronic health records (EHRs)
4. Inconsistent standards make comparisons of similar data collected from different devices (e.g., glucometers, BP meters) and mobile applications or wearables difficult
5. Difficulty collecting PROs especially in populations that have less access to or familiarity with the technology needed to report their outcomes. Lack of a clear ROI for patients to report and attrition in longer term PRO studies impact usability
6. How to incorporate this information into clinical practice. Are the data useful, do clinicians know how to interpret, etc.
7. Most data that is currently collected is driven by reimbursement pressures. There are no such pressures for PROs
8. Digital divide limits use of clinical tools and wearables and collection of PROs among diverse and medically vulnerable populations
9. Lack of detailed contextual information (metadata) that enables accurate interpretation of participant-provided information
10. Difficulties obtaining patient-generated health data from wearable electronic devices (e.g., technological barriers and lack of trust from patients)
11. Data collected from patients is not consistently displayed back to them or their providers in a manner that would support their patient engagement and self-management
12. Clinicians may not know what to do with patient-reported information, especially if the data raises concerns that are outside the clinician's ability to control or address
13. Making it possible for patients to provide data when many patients, especially from underserved areas and populations, have no/limited technology access, familiarity, and trust
14. Data should to be patient-controlled and self-sovereign
15. Integration of genomic data from persona genomics services and strategies for making this data useful to clinicians and researchers (with appropriate privacy protections)
16. Need to develop strategies to encourage adoption and use of PROs by clinicians and patients (even if the infrastructure is there, we still need clinicians and patients to use the PROs to have sufficient data for research)

What improvements are needed (e.g., policies, governance, standards, services, technology) in the collection of participant-provided information over the next 10 years?

1. Generate core list of key information (data and metadata) needed from wearables, devices, and mobile apps for researchers to be able to understand which data points are comparable and use them to study patient engagement and impact on outcomes
2. Disseminate technology standards to support consistent recording and extraction of PRO data in the EHR
3. Develop and disseminate standards to support the aggregation of and integration of medical device data with other data sets including electronic health record (EHR) data
4. Develop and disseminate standards to support the aggregation of and integration of data from wearable technology with other data sets including electronic health record (EHR) data
5. Need to develop strategies to encourage adoption and use of PROs by clinicians and patients (even if the infrastructure is there, we still need clinicians and patients to use the PROs to have sufficient data for research)
6. Identify with stakeholder input and recommend 1 core PRO to be collected on a routine basis - such as SF-12, or a single item health status measure. Engage with PCORI and NCQA to convene stakeholders and encourage health system participation
7. Disseminate best practices on establishing collaborative partnerships with patients on federally-funded PCOR data infrastructure projects
8. Guidance on how best to collect and display SDOH data in a manner helpful to clinicians and sensitive to patients
9. Services and technologies that allow diverse patients to participate more fully in clinical research by facilitating data donation of their electronic health record (EHR) data
10. Encourage developers to work together to ensure data from apps and devices interface with EHR/PHR so that providers/ patients can choose to import, share, visualize data trends and use PGD/PROs to support health management and shared decision making
11. Engage patients in learning how to make PGD and PRO data meaningful to them. How would they like to view, store, share their data?
12. Advances in behavioral economics methods that create a positive reinforcement loop for participants to remain engaged in data collection to prevent large dropout rates as interest wanes
13. Proper application of block chain/distributed ledger technology and advanced cryptography to allow for self-sovereignty for patients' own data and giving them control of access and use
14. Identify with stakeholder input and recommend one core patient-reported outcome to be collected on a routine basis - such as SF-12, or a single item health status measure

Functionality 5. Use of Enhanced Publicly-Funded Data Systems for Research

What are the most pressing data infrastructure challenges related to the use of enhanced publicly-funded data systems for research in the next 10 years?

1. Difficulties promoting and navigating cross-sector collaboration to bring data sets together at the local level (e.g., county or state)
2. A shared vision and centralized leadership to promote existing data assets and resources to leverage enhanced data systems
3. Lack of publicly available data assets to facilitate social determinants of health (SDOH) analysis
4. Limited access to data from regional health information exchanges that aggregate data from multiple payers
5. Need of an infrastructure to allow for federated learning across all publicly funded data where access is allowed without ever moving the data. "Use don't move" with deployed analytics and learning models will rapidly expand research
6. Limited access to Medicare data enclave, especially for linking data
7. Limited awareness of existing federal data assets
8. Two-year lag in national death index data
9. Lag and lack of richness of data on community context, e.g., from the Area Resource File, to help analyze social determinants of health
10. Cost of obtaining and maintaining these datasets
11. Lag in availability of Medicaid data, specifically the T-MSIS, hinders ability to examine social determinants of health
12. Lack of access to identifiable data for linkage from PDMP data for opioid research
13. Linkages/matching of datasets due to cost and lack of identifiers/minimal datasets (probabilistic matching has been difficult at best in many instances)
14. High cost of these data sources
15. Insufficient training on how to assess dataset fit-for-purpose (for those seeking to use government datasets in PCOR)
16. Four year lag in access to Medicare data

What improvements are needed (e.g., policies, governance, standards, services, technology) in the use of enhanced publicly-funded data systems for research over the next 10 years?

1. Increase ability to readily use, retrieve, link, and aggregate publicly-funded data (federal/state) for research
2. Develop cross-agency and public “sandboxes” to allow for innovation and exploration of cross-agency data products
3. Improve access to federal data assets related to public health surveillance to facilitate rapid analysis of clinical and other data for public health response

4. Improve awareness of and access to federal data assets through creation and maintenance of a single catalog of federal data assets
5. Increase availability and discoverability of de-identified open datasets that can be readily accessed by diverse users
6. Reduce regulatory and local IT infrastructure barriers to allow deployed decentralized AI to analyze data at rest with copying/moving
7. Develop a set of case studies or best practices around cross-sector collaboration around data linkage or data exchange
8. Federal policies to increase timeliness of NDI data
9. PCORI legislation states, The Secretary shall, with appropriate safeguards for privacy, make available to the Institute such data collected by the CMS under the programs under titles XVIII, XIX, and XXI. Make this happen for PCORnet and researchers
10. Lower the cost of access, especially for younger investigators or investigators from smaller institutions and without big grants
11. Federal policies to reduce latency of Medicare encounter data
12. Federal mandates to facilitate researcher access with appropriate protections to PDMP data for opioid research
13. Single portal for access to all data sources, resources, etc.
14. Federal policies to increase the speed and saliency of Medicaid data for the TMSIS

Appendix C. Strategic Framework Milestones by Functionality

Functionality	Milestones
Use of Clinical Data for Research	
	• Establish services and tools to support data access, querying, and use, including privacy-preserving analytics and queries. These services and tools would be leveraged nationally and are not likely to be developed by the private sector.
	• Develop support services and tools that can be leveraged nationally and are not likely to be developed by the private sector; these tools would test the quality of unstructured and structured data to answer PCOR questions.
	• Develop standards that support secure, electronic query of structured data across clinical research and delivery systems, including standards for open-source access.
	• Develop and test metadata standards that describe data quality.
	• Create a policy framework for privacy-preserving access and querying of clinical data by researchers conducting PCOR, and policies that govern the use of the services that support data access, querying, and use.
	• Develop a policy framework for ensuring clinical data used for research is of "research grade."
Standardized Collection of Standardized Clinical Data	
	• Support the development of a set of research CDEs in specific gap areas and support development of a governance structure for CDE harmonization
	• Support the development of repositories/portals for CDEs, standards for utilizing CDEs for research, and services to allow researchers to easily utilize standardized components.
	• Support research and/or crowd-sourced methods to determine which of the standardized collection components and services are most valuable.
	• Create policies to promote the adoption and use of valuable standardized collection components and services.
Linking Clinical and Other Data for Research	
	• Leverage existing standards, and support the development and balloting of needed standards for patient data linkage.
	• Establish HHS policies that promote appropriate data linking based on the framework noted in the milestone above.
	• Create a policy framework to facilitate patient data linkage in accordance with existing laws.
Collection of Participant-Provided Information	
	• Support the development of tools and services that can be leveraged nationally and are not likely to be developed by the private sector. These tools and services will facilitate the collection and exchange of PPI, including national services for electronic capture and management of PPI and release of data for PCOR.

Functionality	Milestones
• Support the development of a core set of standards for the collection and integration of prevalent use cases of PPI for PCOR, by leveraging existing standards and filling gaps.	
• Create policies and share best practices for collection and integration of prevalent use cases of PPI for PCOR.	
Use of Enhanced Publicly-Funded Data Systems for Research	
• Support the enhancement of strategic publicly-funded data systems (including CMS data) to facilitate their access and use, and ease retrieval of data for research purposes.	
• Support the further development of key federally-initiated data systems for research.	

Appendix D. Process Considerations for Future Strategic Planning Activities

The multi-part idea generation and prioritization effort was successful in soliciting and prioritizing feedback from a target audience. The activity not only generated key challenges and improvements for PCOR data infrastructure, but also highlighted potential areas for further exploration. Given the potential for additional stakeholder outreach, this section presents three process-based considerations that may be valuable for future efforts.

Use a phased approach that empowers participation and garners buy-in from participants. This activity's phased approach was successful in meaningfully engaging stakeholders. In advance of the virtual Listening Session, participants received a concise list of materials that provided an introduction to ASPE and its portfolio. The virtual Listening Session provided a forum for initial feedback and created buy-in for subsequent activities using the Codigital platform. By the time participants were invited to respond in Codigital, they were likely to be both familiar and comfortable with the questions asked in each domain.

We further divided the Codigital online activity into two distinct phases. Participants first provided their ideas and then prioritized them. Having two distinct phases encouraged participants to focus first on sharing and refining ideas, and then on defining priorities given their professional perspectives and expertise.

Offer different methods for participant engagement. We used several methods to engage participants throughout this activity, which allowed stakeholders to participate based on their availability and preferences. First, during the Listening Session participants were able provide their input verbally or in written form through the chat function. Second, the Codigital activity was structured to offer flexibility to participants. Participants could elect which domains to respond to, and which phases to participate in most actively, based on their own bandwidth and expertise. Affording this kind of flexibility can help to prevent feedback fatigue.

Use user-friendly technology to encourage participation. In lieu of person to person facilitation, we used the Codigital platform to guide idea generation and prioritization. Codigital provided user-friendly software that encouraged participation by providing results in real-time. Participants were largely able to navigate the platform without additional assistance from the project team. In addition, the technology allowed stakeholders to participate on their own time at multiple points in time, rather than having to rush to complete the full exercise in one sitting. The online mechanism facilitated input from diverse stakeholders with different contributions and preferences.