Physician and Physician Practice Research Database
Feasibility Report

AHRQ Physician and Physician Practice Research Database Feasibility Report

Prepared for:
U.S. DEPARTMENT OF
HEALTH AND HUMAN SERVICES
Agency for Healthcare Research and Quality
5600 Fishers Lane
Rockville, MD 20857
www.ahrq.gov

Contract Number: 75Q80120D00018

Prepared by:
NORC at the University of Chicago
Bethesda, MD

AHRQ Publication No. 24-0040
February 2024
Originally submitted August 15, 2022

Table of Contents

List of Exhibits and Tables

Executive Summary

The purpose of this feasibility report is to present the Agency for Healthcare Research and Quality (AHRQ) with a blueprint for developing a federally led data collection program on the U.S. physician workforce and physician practices that is primarily sourced through existing administrative data. The report discusses how a strategic, cost-effective, and timely option for such a program can emerge—one that successfully leverages these five key ingredients: data, conceptual design of the 3P-RD, data dissemination, cost, and partnerships. It provides recommendations for AHRQ's consideration on how to establish an expansive future iteration of the 3P-RD prototype that would regularly create and maintain these databases.

AHRQ's 3P-RD was motivated by the need for accurate and timely information on physician, physician practice and workforce behavior to support better strategic planning and federal funding allocations. To support that need, the team sought to determine the feasibility of establishing a program that can regularly release physician and physician practice information using existing administrative data resources. The effort involved the following tasks:

- Identifying data sources to use to create the physician and physician practice data files
- Developing construct definitions for physician and physician practice
- Developing the database architecture and then building the database by processing, standardizing, harmonizing, and linking disparate administrative data sources
- Exploring data release options that met AHRQ's goals and mandates while balancing the data use agreement (DUA) requirements of data owners whose data were being used for the 3P-RD
- Assessing AHRQ's role and its potential partner roles in building the 3P-RD

This report provides a roadmap for how a 3P-RD program can be enacted based on the feasibility study that NORC and its subcontractors conducted under this contract. It highlights key considerations for AHRQ based on our experience with planning and building the prototype physician and physician practice database for 13 states and how this feasibility study could be expanded into a federally led data development program.

1. Introduction

The Physician and Physician Practice Research Database from Administrative Data (3P-RD) initiative is one of two data development initiatives in the "AHRQ Innovations in Physician, Physician Practice, and Social Determinants of Health Data" project commissioned by the Agency for Healthcare Research and Quality (AHRQ) aimed at supporting research on existing and emerging issues related to healthcare policy and clinical practice research. AHRQ awarded this contract to NORC at the University of Chicago (NORC) and its partners from the American Academy of Family Physicians' Robert Graham Center (RGC) and the Kaiser Permanente Center for Health Research (KPCHR). The project began in September 2020, with a slated 23-month performance period. The overarching goal of the 3P-RD initiative was to further policy-relevant health services research by addressing gaps in current physician and physician practice data. The 3P-RD was a prototype that focused on building physician and physician practice files for 13 states to illustrate potential approaches to how a program version of this prototype could emerge as a proof of concept.

Motivation

AHRQ has a long history of creating research databases and disseminating them to the public to inform policy—e.g., to address pressing health care delivery policy issues, including the Medical Expenditure Panel Survey (MEPS), the Healthcare Cost Utilization Project (HCUP), and the Compendium of U.S. Health Systems. However, data gaps remain. Although an extensive amount of data is collected on health care providers, national data collection efforts conducted in the last decade have largely been sporadic, and highly targeted. The COVID-19 public health emergency (PHE) also amplified the need to have readily available data on physicians and practices that are actively delivering care to patient populations to inform national and state response efforts. The crisis has highlighted the nation's current lack of complete data on healthcare capacity that would effectively support national and state planning and response. Motivating this project was the need for accurate, timely information on physician, physician practice and workforce behavior to support better strategic planning and federal funding allocations.

Background

In 2014, AHRQ completed the Collecting Data on Physicians and Their Practices (CDPP) project, which provided the groundwork for future AHRQ efforts to address the need for data on physicians and their practices given the rapidly evolving health care environment.[1] Through this work, AHRQ identified a "strong need for dependable, comprehensive, accurate, and timely sources of information on physicians, physician practices, and physician practice behavior" for policymaking and research purposes. The CDPP project focused on primary collection of physician data through surveys and examined the landscape of existing data collection efforts,

[1] See DesRoches C, Rich E. Collecting Data on Physicians and Their Practices: Final Report (Prepared by Mathematica Policy Research for the Agency for Healthcare Research and Quality under Contract No. HHSP23320095642WC). August 2014.

identified gaps, and proposed a prototype survey instrument. In examining the available literature, the study found that although there was extensive data collection on physicians, data gaps made it difficult to characterize national trends. An important finding from the CDPP study was that multiple administrative data sources could potentially be linked to obtain comprehensive information on physicians and physician practices for conducting policy-relevant health services research.

Objectives of 3P-RD Initiative

Using administrative data, the 3P-RD initiative focused on collecting information on active physicians and physician practices. This prototype study was scoped to test the feasibility of creating a database for 13 states to address the following long-term objectives for AHRQ:

- Capture the universe of physicians and physician practices in the United States.
- Obtain characteristics of physicians and their practices that can be used for analytic and research purposes.
- Create research databases that are not only valuable as stand-alone products but that can also be linked to various data assets in the Department of Health and Human Services (HHS) such as AHRQ's other data assets like HCUP and the MEPS, the Centers for Medicare & Medicaid Services (CMS) administrative data, and Health Services Resources Administration's (HRSA) workforce data.
- Enable timely, accessible, policy-relevant descriptions, and analyses of 1) the current roles played by physicians and physician practices in the U.S. healthcare system, and 2) the relationship of such roles to levels, trends, and distributions of healthcare access, quality, and costs.
- Develop research databases appropriate for use by multiple stakeholders, including AHRQ staff members, other federal agencies, and external researchers.
- Develop and implement pragmatic, cost-effective strategies and tools for acquiring, disseminating, and using physician and physician practice data for conducting policy-relevant health services research.
- Establish clear roles for AHRQ and partners by leveraging ongoing efforts for synergies.

At the conclusion of the 3P-RD, NORC was tasked with reporting on how a viable and sustainable program could emerge based on the experiences of developing the 3P-RD prototype. This report captures the process for how such a program can come about. It provides a high-level discussion on what was done, insights and lessons learned from implementing this prototype study, and future considerations for how an expanded version of this program could be scaled to span nationally.

This current chapter, Chapter 1 presents the motivations for this work, the 3P-RD initiative's objectives, and a summary of the content of each chapter.

Chapter 2 discusses the current data landscape based on the findings of an environmental scan that the team conducted. We investigated what existing administrative data sources could be

leveraged in building the 3P-RD files. Our work was informed by a scan of the literature and key informant discussions. Our final decision for which data sources to use weighed the advantages and disadvantages of using each data source and was ultimately driven by what would best meet AHRQ's short- and long-term goals for the project. We present the strengths and limitations of the 3P-RD files that were constructed for states, along with opportunities for further data enhancements in future iterations.

Chapter 3 presents an overview of the construct definitions that we used to define physician and physician practice for the 3P-RD. We discuss how the definitions were framed for the prototype given the challenges with operationalizing those constructs using the data at hand. We provide a discussion of the existing limitations with some of the current definitions, what additional information would be needed to refine them, and the potential data sources that could be used to capture that information. In this chapter, we present recommendations for how future iterations of this project could expand certain constructs to account for additional elements.

Chapter 4 provides an overview of how the 3P-RD was constructed. It includes the key requirements and considerations that needed to be met in designing a data architecture and build for a database—one that is flexible and can effectively leverage the strengths of the data being sourced for the 3P-RD while also accounting for any data quality issues and variations in data availability across states. We discuss the key issues/challenges that need to be addressed for the 3P-RD construction. In addition, we indicate the processes that were successful during the prototype phase, offer suggestions for improvements, and identify new processes for future consideration. We elaborate on our data standardization and harmonization efforts, which represent the backbone of any work that relies on linking and combining disparate data sources. This was an especially important task when developing physician specialty standardization schemes and harmonizing the data to fit those schemes. We provide some reflections on data quality issues with the data sources that were used to build the files and we discuss potential benchmarks.

Chapter 5 provides an overview of data dissemination options for the physician and physician practice data. We discuss the significance of setting up a DUA that can effectively meet the data protection requirements of the database along with data file release options that would allow for greater access to data elements, but under stricter and more controlled environments.

Chapter 6 provides an overview of the role of AHRQ and its partners. AHRQ is pioneering what can be considered a very challenging yet critical national data resource to meet the needs of the ever-evolving healthcare landscape. Only a federal agency with access to the resources and budget necessary to scale this project to span nationally can lead this effort. We discuss how the 3P-RD has been recognized as a critical resource for various data owners. We discuss: 1) the importance of fostering strong and trusting state and federal partnerships in building the 3P-RD; and 2) the value in developing public-private partnerships with trusted private and/or nonprofit organizations—a partnership that would harness the agility of the private sector while retaining the oversight and protections provided by the government.

2. Data Landscape and Selection

Fulfilling the goals of the 3P-RD required a fundamental understanding of the research and its gaps, as well as the existing administrative data sources for developing comprehensive databases that can be used to address the immediate and longer-term public health efforts. One of the first steps for conducting this feasibility study consisted of researching what administrative data resources on physicians and their practices are available, the conditions for acquiring and using these data for AHRQ's internal purposes, and the possibility of creating from these data sources research databases for public release. In this chapter, we discuss what data sources were used for the prototype, along with the rationale for why those data were used. We determined that producing files that linked state medical board (SMB) licensure data to the National Plan & Provider Enumeration System (NPPES) and the Medicare Provider Enrollment, Chain, and Ownership System (PECOS) would form the core structure of the physician and physician practice state files. Claims files derived from CMS sources as well as state all-payer claims database (APCD) data, where available, would then supplement and enhance the core files. We present the advantages and challenges of using these various resources for the 3P-RD. We then present the data gaps (if any) along with alternative data sources for consideration in future iterations.

Key Requirements

When researching and selecting administrative data sources, the team operated with the understanding that there was a minimum set of key requirements that had to be met either through a single data source or through the combination of sources to achieve the intended goals of the 3P-RD. The data sources that would be used had to meet the following needs:

- Ability to capture the census of active physicians and physician practices in a state
- Ability to identify which physicians are actively delivering care to patients
- Ability to identify where the physicians are delivering care

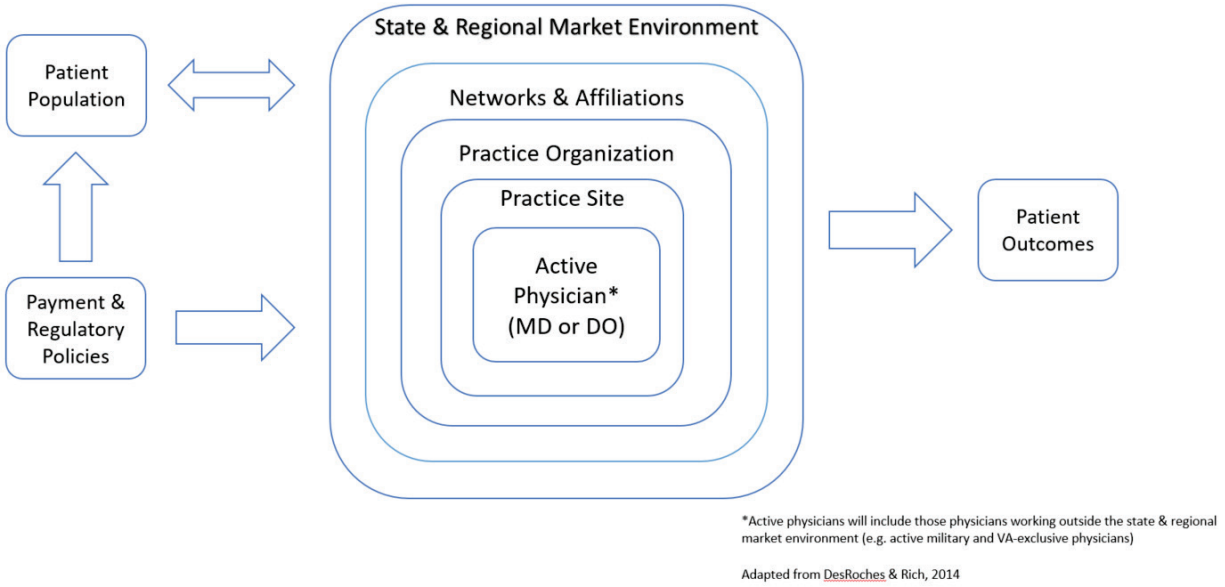
Insights from Prototype Phase

An assessment of the current data landscape began with an environmental scan that captured the current availability of administrative data sources that could be leveraged in the building of the 3P-RD. The conceptual framework depicted in Exhibit 2.1—adapted from AHRQ's CDPP study report[2] and used to guide the initial research— provided an identification and prioritization scheme for determining which data sources to use. The model situates individual physicians within multiple potential contexts that are important for policy-relevant health services research. Individual physicians are placed at the center of the matrix that is nested within layers of organizations that span practice sites, practice organizations, affiliated networks, and the market environment at the regional and state levels. Based on the relationships observed in care delivery,

[2] DesRoches CM, Rich EC. Collecting data on physicians and their practices. Final Report. Mathematica Policy Research; 2014.

this framework allows physicians to be nested in practice sites and within a larger practice organization, that may in turn be part of larger network of practices or a healthcare system. Some practices may comprise a single practice site. Information on these multiple contexts and relationships among them is critical to understanding provider supply and delivery of care within the healthcare system, specifically how policy and regulatory initiatives influence the behavior of physicians and their practices to impact patient outcomes, including access, quality, and cost of care.[3,4,5]

Exhibit 2.1. Conceptual Framework



Using this model, the team conducted a comprehensive search of the available data sources for physicians and physician practices, focusing on data accessibility and usability for developing a 3P-RD prototype that is useful for policy-making and research. We assessed the completeness of data on physicians and physician practices at the national and state levels. We then documented the breadth (e.g., data elements/characteristics) and quality of available information for physicians and physician practices. Table 2.1 summarizes the team's overall approach to evaluating the various data sources, including domains of inquiry and illustrative research questions regarding data sources and target states for the 3P-RD prototype. For each data source, we assessed its added value for capturing the census and characteristics of active physicians and physician practices. We evaluated the value added from each data source based

[3] Wilensky GR. Reforming Medicare's physician payment system. N Engl J Med. 2009;360(7):653-655.

[4] Casalino LP, Nicholson S, Gans DN, et al. What does it cost physician practices to interact with health insurance plans? A new way of looking at administrative costs—one key point of comparison in debating public and private health reform approaches. *Health Aff*. 2009;28(Suppl1):w533-w543

[5] Landon BE, Reschovsky J, Reed M, Blumenthal D. Personal, organizational, and market level influences on physicians' practice patterns: results of a national survey of primary care physicians. Med Care. 2001;39(8):889-905.

on several factors, including applicability to policy-making and research, accessibility and cost, types of information included, data quality and timeliness, and data harmonization and linkages.

Table 2.1. Organizing Approach for Data Landscape Review

Domain	Data Acquisition Evaluation
Purpose	• For what purpose was the data source originally developed? • How can it inform the development of the 3P-RD?
Accessibility of data	• Who owns the data and what is the process for requesting use? • Can data be included in a public release database? If not, what alternatives are possible (e.g., limited release for researchers, different levels of release based on user, de-identification or release of aggregate data)? • What is the cost and overall timeline for obtaining the data? Are there potential costs related to data cleaning or quality checks?
Types of information	• What data elements are available regarding physician and physician practice characteristics? • What types of practices and medical professionals are included in the dataset? • Does the dataset include medical practices at the national, state, regional, or local level?
Data quality & timeliness	• How complete are the data for capturing the census of active physicians and physician practices? • What elements are missing from the dataset? • What is the quality of the data? Are the data validated? • How recent are the data and how often are datasets updated?
Data harmonization & linkage	• What identifier data on physicians, other practitioners, and their practices' affiliations are included? • Can the data source be linked with HHS data sources to inform policy (e.g. AHRQ's MEPS and HCUP)? If so, using which linkage data elements? • Can the data source be linked to claims and other physician and physician practice level secondary data (e.g. CMS data sources like Medicare Care Compare)? • Are there parallel AHRQ efforts to link datasets that can be leveraged?
Applicability	• What AHRQ objectives can be informed by the available data? • What types of research questions can be informed by the data?
Strengths & limitations	• What are the overall strengths and limitations of the data for the 3P-RD prototype? • What are the caveats for using the data for the 3P-RD prototype?
States for database	• Which states are the most promising in terms of data quality, robustness, and timeliness? • Which states have the most complete information in available datasets? Do any states have unique information of value? How representative are states with respect to their geography, population, policy environment, and healthcare markets?

In addition to a scan of the literature, NORC met with subject matter experts (SMEs) and conducted a data owner outreach to guide the team's final decision on which data sources to

consider for building the 3P-RD. This outreach was especially important since a few data sources that were flagged in the environmental scan (such as APCD data) lacked sufficient information in the public domain to understand the quality and comprehensiveness of the data. Based on this review and our exchanges with data owners, the team focused on acquiring the data files that were determined as viable sources to pursue for the construction of the 3P-RD both for the short-term needs of the feasibility assessment and the long-term goals for developing a program.

Data Source Selection

The outcome of our data source research led us to select data sources that would provide the best information while balancing cost, accessibility and long-term viability for use. We relied on data sources that provided the necessary information for capturing the population of interest and were readily accessible— i.e., available either publicly or at a small fee with little to no data release requirements because they are in the public domain. These data sources were considered in the building of the core 3P-RD data files. This approach allowed the team to circumvent lengthy data acquisition exchanges that could risk the project's schedule and ability to release the data. We then identified another set of data sources that would be used to supplement and augment the 3P-RD. We discuss these various sets of data sources next.

Core data source files. The outcome of our research resulted in selecting SMB data, and NPPES and PECOS data files as the core elements for the physician and physician practice files. We found that there were several benefits in linking the SMB data to the publicly available NPPES and PECOS data files:

- The NPPES data file contains both the National Provider Identifier (NPI) and state licensing information. By linking the SMB file with the NPPES data file, an NPI can be linked to the SMB records.
- Having an NPI also allows for linkage to both PECOS and claims data.
- Another benefit of this linkage is to potentially capture providers from the NPPES and PECOS files who do not appear in the SMB state file.
- In addition, the NPPES and PECOS files contain information that is not captured by the SMB state file. This information can be used to capture additional information about the provider. Furthermore, information that is not well populated in the SMB state file can be augmented by the NPPES or PECOS files.

Linking the SMB data to NPPES and PECOS comprised the first step in creating the core physician file, allowing the team to identify all possible physicians in a state.

Supplemental data source files. We then identified APCD or CMS claims (claims) data elements as sources to enhance the core 3P-RD data files and found that claims offered certain advantages that would further enhance the core 3P-RD data files, although access to claims data can be challenging. Nonetheless, claims, particularly state APCD data, offer both a rich resource of information that we leveraged for this prototype and potential for future use for the program. States have been responding to the need for comprehensive APCDs to support health care and payment reform initiatives and increased transparency and information on the health care system. APCDs were initially created with the purpose of serving as a single source for claims

and enrollment data (where possible/available) across all sources of insurance coverage in that state. Although state APCDs vary in their degree of maturity and limitations set by federal and state requirements, they still present valuable information for individuals looking to understand health care utilization, spending, and costs.[6] For the 3P-RD, we sourced state APCD data from four of the 13 states that were selected for the study, with the intention of using those data to supplement the core physician and physician practice files that had been developed using SMB licensure data, NPPES, and PECOS. For non-APCD states, we used CMS claims data derived from Medicare Fee-for-Service (Medicare FFS), and Medicaid claims derived from the Transformed Medicaid Statistical Information System (T-MSIS). Table 2.2 lists the data sources we selected for the 3P-RD and provides the advantages and challenges that the team weighed when deciding on using them for the 3P-RD.

Table 2.2. Data Sources Selected for the 3P-RD

Data Sources	Advantages	Challenges
Individual state medical board (SMB) data files	• All states have state medical board data. • Contains accurate, up-to-date information on license status for physicians in the state. • Physicians are required to register with the SMB to practice medicine in the state. • Most states include dates of activation and expiration. • Data requests are straightforward and do not require a DUA.	• Data vary across states (e.g., variations in what variables are available, how data are stored, data formats, and the frequency with which data are refreshed). • Data may require extensive cleaning and processing to align with other administrative files. • Some states require data to be web-scraped, a time-intensive process. • Some states have restrictions on whether they allow their data to be included in the creation of a data product. • Not all state SMB data provide information to verify M.D. or Doctor of Osteopathy (D.O.) credentials. • Data do not indicate which physicians are actively providing care to patients.

[6] Costello A, Love D, Porter J, Peters A, Sullivan E, Informing health system change - use of all-payer claims databases, 2018. Available at: https://www.apcdcouncil.org/publication/informing-health-system-change-use-all-payer-claims-databases

Data Sources	Advantages	Challenges
National Plan and Provider Enumeration System (NPPES)	• Includes health providers (spanning more than just physicians) who can bill for their services[7],[8] (e.g., podiatrists, chiropractors, nurse practitioners, physician assistants, psychologists, and social workers) • Captures recent entries into the physician workforce. • Contains information for both organizations and individuals. • Contains a variety of information collected from each provider • Contains historical information (e.g., other license numbers) • File contains NPI (of high quality), making it easily linkable to any source with an NPI. • Contains license numbers (though of mixed quality) and name information, making it linkable to state-specific files. • Contains information related to specialty, via the taxonomy code which can be converted to a CMS specialty code using an online publicly available crosswalk. • Historically, NPPES provided only active NPIs, but in recent years CMS has added deactivated NPIs (identified via indicator) to the NPPES file. • Monthly files are released.	• There is no requirement to update the data when participants change location or specialty, retire, or die. • It only contains providers who have applied for an NPI. • License numbers vary in quality and formatting which results in challenges for linking disparate files that include the state license number. For example, a license number might include special characters (e.g., hyphens) in one file whereas the other file does not, thereby prohibiting linkage. • Information is supplied by the provider leading to: – Incomplete information (missingness) – Incorrect Information (outdated or miskey) – Information in different formats (e.g., provider may provide first and middle name in first name field, credentialling may be in list format with or without punctuation, etc.) – Providers appearing in the NPPES file may not be credentialed or hold an active license.

[7] Agency for Healthcare Research and Quality. The number of nurse practitioners and physician assistants practicing primary care in the United States. Available at: https://www.ahrq.gov/research/findings/factsheets/primary/pcwork2/index.html.

[8] Miller BF, Petterson S, Burke BT, Phillips RL, Green LA. Proximity of providers: Co-locating behavioral health and primary care and the prospects for an integrated workforce. *Am Psychol*. 2014;69(4):443-451.

Data Sources	Advantages	Challenges
Provider Enrollment Chain and Ownership System (PECOS)	• These data include variables allowing to link individuals to organizations and can therefore be used to create practice-level measures, including practice size and specialty composition. • These data also include all providers, not just physicians, which makes it possible to analyze teams within practices. • Because this enrollment system is linked to systems for enrolling in Medicaid,[9] PECOS files also include most physicians across all specialties, including pediatricians, who rarely see Medicare patients. • These data contain NPI making them easily linkable to other data sources with NPI. • These data contain enrollment state. • Information is harmonized with CMS coding schemes (specialty codes).	• Because it takes approximately two years to drop old enrollments from PECOS, associations between physicians and practices may need to be validated using current claims data. • Some information may be masked, such as the tax identifier number (TIN), which limits linking with other disparate data sources. • These data only contain Medicare providers who have registered with CMS to receive payment.
Claims data (APCD data And additional claims data [e.g., Medicare FFS, T-MSIS])	• Claims data provide information on patients and services physicians provide to patients. • Claims data include information on the provider type and physician specialty. • Claims data include ZIP Code information, indicating where the service might have been provided. • Claims data include data on the health system and the organization providing the service, indicating the location of care.	• Not all physicians with an active license provide care to patients; claims will only provide information on physicians whose NPI is used for billing services. • Physicians who provide care and submit a claim may have a license expire later in the year. • Not all physicians who provide care to patients submit a claim. • Data requests and DUAs can be lengthy and the process extensive, requiring multiple agencies and decision-makers to approve the request. • ZIP Code on claims may indicate the location of the billing office not the location of care.

[9] Centers for Medicare & Medicaid Services. Medicaid provider enrollment compendium (MPEC); 2018. Available at: https://www.medicaid.gov/sites/default/files/2019-12/mpec-7242018.pdf.

State Selection

As noted earlier, to assess the feasibility of this work, we selected 13 states for the prototype, with final selection based on an assessment of the SMB and APCD data, data acquisition timelines, AHRQ priorities, and variation both geographically and in state provider policies. Our assessment of the harmonized and standardized data from various data sources further informed the state selection process by providing insight into data quality and linkability of SMB data. To maximize geographic variation, inclusion of a frontier state was prioritized. The inclusion of a state bordering Arkansas (AR) was set as a priority to achieve provider policy comparison between neighboring states. The final selection for both the frontier state and AR-bordering state required additional analysis of the SMB-NPPES linked data. Based on these criteria, we chose the 13 states listed in Table 2.3 for the 3P-RD.

Table 2.3. Final State Selection

State	Reason for Inclusion
Arkansas	• Availability of APCD data that could be used for the prototype.
Arizona	• Population and geographic considerations.
California	• Geographic, population, and health policy considerations.
Colorado	• Availability of APCD data that could be used for the prototype.
Florida	• Has an APCD, although not available for use on the prototype.
Massachusetts	• Has an APCD, although not available for use on the prototype.
Maryland	• Availability of APCD data that could be used for the prototype.
Minnesota	• Has an APCD, although not available for use on the prototype.
Missouri	• AHRQ priorities
Montana	• Frontier state with high quality SMB data.
New York	• Interstate geographic variability between urban and rural areas
Texas	• Geographic, population, and health policy considerations.
Washington	• Availability of APCD data that could be used for the prototype

Future Considerations

For future iterations of the 3P-RD, we recommend building on the current data that were sourced for the prototype and incorporating additional data sources. Based on the lessons learned from the 3P-RD, future work should focus on continuing to build out knowledge of the data landscape and assessing and addressing data gaps. Below we present recommendations for AHRQ's consideration:

- Conduct an environmental scan periodically to stay informed on changes in the data landscape as additional data resources become available for use. State, federal, and private organizations continually develop data products to enhance analysis and provide transparency. Specifically, as more states develop and implement APCDs, there is an increased opportunity to partner with them and obtain state-wide claims data.

- Expand the number of states included in the 3P-RD. Obtaining SMB data is straightforward in most states, and the data can be used to build out the current core physician and physician practice data to include additional states.
- Work with federal partners to increase access to claims data. These federal resources will help fill data gaps in APCD data files. The Supreme Court ruling *in Gobeille vs. Liberty Mutual Insurance Company* limited the number of payers submitting claims to state APCDs. Self-insurers, including the Federal Employees Health Benefits program, do not need to submit claims to the APCD, resulting in a large data gap for some states. In addition, Veterans Affairs and TRICARE (military families) do not regularly submit claims to the APCD. Identifying and working with federal partners can fill current data gaps, thereby increasing the accuracy of identifying actively practicing physicians, patient panel characteristics, and volume of claims submitted.
- Work with commercial partners to fill the data gaps for non-APCD states. Many states do not have an APCD or do not provide APCD data to external researchers. CMS T-MSIS and Medicare files address part of the data gap for claims; however, physicians treating patients who have commercial insurance will have a remaining data gap. Identifying and working with partners (e.g., Blue Health Insurance, Optum) to obtain commercial claims will address data issues for states until APCD data are available.
- Work with states to obtain state Medicaid data. Currently there is an approximately two-year lag in the CMS T-MSIS data. Obtaining Medicaid data directly from a state will mitigate the lag in claims data, allowing licensing information and claims information to be more closely aligned.
- Leverage the HCUP data and/or partnerships with states for hospital discharge data. Hospital discharge data provide a wealth of information for hospital-based physicians and practices. Also, the data will address data gaps for patients without insurance, which can be critical in understanding policy questions related to the uninsured or underinsured populations.

3. Definition Construction

A critical component in the success of the 3P-RD prototype was developing constructs for how the project would define physician and physician practice (along with other key associated elements) that appropriately reflected and matched how current research and real-world applications were capturing those constructs. The team had to develop definitions that adequately captured concepts that were actionable and could be operationalized. This was particularly important when it came to defining physicians and physician practices. For this effort, the team relied on the conceptual model discussed earlier in Exhibit 2.1 to help guide the framework in which physicians operate. The team also reviewed existing resources and other efforts that already attempted to develop these definitions. Moreover, to be sure the definitions matched real-world applications, the team worked alongside SMEs and clinicians.

Developing definition constructs was foundational to our overall approach to the project, including how the data for the 3P-RD would be standardized and harmonized. The terms needed to meet recognized health care industry standards. This chapter discusses how terms were defined, starting with key requirements for variables essential to the 3P-RD. We identify areas where a variable definition was successful and areas that were challenging. Finally, we highlight key topics for AHRQ to consider in future work in the 3P-RD program.

Key Requirements

Although the definition of physician and physician practice could be expansive and would include the broad spectrum of the health care workforce along with a wide variety of places of practice, there was a minimum set of requirements when building appropriate constructs for the 3P-RD. These had to both meet the goals of the study and reflect the realities of the population and how that population was being captured and defined in the existing data and larger policy conversations. The following are the elements we needed to define and the minimum requirements for each:

- **Physician:** Doctor of Medicine (M.D.) or Doctor of Osteopathic Medicine (D.O.)
- **Status of a physician's license:** Active, expired, suspended
- **Active physician:** Delivering care to patients
- **Physician practice:** Where the physicians are delivering care (site of care not organization)
- **Practice size:** Identify the number of 3P-RD physicians and the total number of providers delivering care within identified practices
- **Physician specialty:** Identify primary and secondary specialty
- **Patient panel:** Patient characteristics of those with submitted claims
- **Services provided:** Identify the top procedures performed, either by representing specific procedure codes or identifying the top categories of procedures
- **Accepted payers:** Medicaid, Medicare FFS, Medicare Advantage, commercial insurance

Insights from Prototype Phase

The 3P-RD focuses on physicians only—i.e., for the purposes of this feasibility study, we chose not to include *all providers* (e.g., physician assistants, nurse practitioners, nurses) or *provider types* (e.g., hospitals and hospital systems, ambulatory surgery centers). As for physician practices, the 3P-RD definition captured how the physicians organize to deliver care. The 3P-RD definitions within the prototype were crafted to maintain adherence to the above-stated focus for physician and physician practice.

Although definitions were initially informed by the literature and existing data programs that were already capturing aspects of the population of interest, we further refined definitions through multiple reviews with both clinicians and SMEs, as well as through the data. We drew identifying characteristics for the physician and physician practices from information contained in the available data, definitions known and pervasive within health care research, and adherence to data use restriction requirements.

Definition Construction

The next set of tables represent key variables that were discussed and operationalized for the 3P-RD. Table 3.1 presents definitions for the main concepts of the 3P-RD, whereas Tables 3.2 and 3.3 focus on defining characteristics of physicians and physician practices.

Table 3.1. Operational Definitions of Key Variables for the 3P-RD

Variable	Description of Approach on 3P-RD	Operational Definition
Physicians	• Definition was aligned with the definition used by the American Medical Association (AMA).	• Doctor of Medicine (M.D.) or Doctor of Osteopathic Medicine (D.O.)
Physician practice	• Physician practices are identified using TIN or CCN (CMS certification number) and could comprise a solo physician or a physician group (organizational NPI). • Physician practices can be in a single site or multiple sites as identified by their site location address(es) or ZIP Code(s). • Physician practices include non-physician clinicians such as physician assistants, nurse practitioners, and others, identified by their NPI.	• Grouping of TIN-ORG NPI-SERVICE ZIP CODE

Variable	Description of Approach on 3P-RD	Operational Definition
State	• Physicians and physician practices can be assigned to states in one of two ways: 1. The location of their practice site is within a state. 2. Active delivery of care to patients within a state although their practice site might be in a neighboring state. • Preference was given to defining states using the first approach since states' data do not fully capture characteristics of physicians or physician practice sites that are outside a state.	• The geographic boundaries of a specific state will be used to define its physicians and physician practice sites.

Table 3.2. Definitions of Physician Characteristics for the 3P-RD

Variable	Definition
Active physicians	• Active physicians are identified at three levels: 1. Living physicians as identified by NPI 2. Living physicians holding active medical licenses with their state boards 3. Living physicians holding active licenses and actively engaged in patient care as observed on claims data.
Physician license status	• Identifies whether the physician has an active, expired, or suspended license.
Physician specialty	• The primary specialty is the specialty the physician most likely practices or is identified in source data as the primary specialty. • The secondary specialty is other listed specialties or identified as a secondary specialty in the source data. • Uses the provider specialty codes defined by CMS.
Accepted payers	• Flags indicating type of payers that the physician submitted claims for services rendered, such as commercial, Medicare FFS, Medicare Advantage, or Medicaid.
Services provided	• Identify from claims data the top procedure codes performed in the year
Patient panel	• Patient characteristics of those with submitted claims.

Table 3.3. Definitions of Physician Practice Characteristics for the 3P-RD

Variable	Definition
Physician practice ID	• Randomly assigned number for each identified physician practice (TIN-ORG NPI-SERVICE ZIP CODE)
Physician practice affiliation	• Physician practices can be owned by or financially affiliated with hospitals and health systems. We define health systems as entities comprising at least one hospital and one physician group providing comprehensive care and sharing common ownership or joint management.

Variable	Definition
Practice size	• The number of 3P-RD physicians and the number of non-3PRD providers associated with the physician practice ID
Number of 3P-RD physicians	• Count of the unique number of 3P-RD NPIs associated with a physician practice ID
Number of non-3P-RD providers	• Originates from claims data • Count of the unique number of non-3P-RD NPIs associated with a physician practice
Patient panel	• Patient characteristics of those with submitted claims
Rural vs. urban	• Identify each practice as rural or urban based on the U.S. Department of Agriculture's (USDA) rural-urban commuting area (RUCA) code • Uses the service ZIP Code and the USDA data.

Identified Successes and Challenges

In Table 3.4, we evaluate how successfully our definitions for the various constructs that we operationalized for the study captured the intended information. We also offer reflections on the challenges that we encountered when developing those definitions for further thought and consideration in future iterations of this project.

Table 3.4. Successful Definitions for the 3P-RD

Variable	Evaluation of Success	Identified Challenges
Physicians	• Identifying physicians using the definition of M.D. or D.O. is relatively straightforward in many data sources.	• Some data sources did not include the necessary information to identify whether a provider was an M.D. or D.O. Using claims as a method to 'fill the gap' was partially successful, although it lacked the ability to validate the determination. • Although the definition of a physician is widely accepted, it does not align with an evolving concept of a primary care provider. • Feedback from advisors and SMEs often resulted in inquiries about the reason for why other providers were not included although they provide care to patients regularly.

Variable	Evaluation of Success	Identified Challenges
Physician practice	• The definition was successful since most claims data sources included the necessary information to identify physician practices using the operational definition.	• Some APCDs do not include TIN in the data. However, we did have the PECOS Associate Control ID (PAC-ID), which is very similar to TIN. Using PAC-ID, we were able to identify practices. • It is unclear whether the ZIP Code in the claims data is the actual location of the service or the location of the billing office. This may cause inaccuracy of practice site assignments. • There are many one-offs within the data, possibly a result of errors in the billing information, inconsistency of data entry for billing, or other reasons. • The team resolved this issue by identifying the top five practice IDs most associated with each 3P-RD physician and including those on the physician file.
State	• The definition is successful.	• Data for physicians and/or practices along state lines may not represent the full picture if portions of the patient panel receive treatment in a different state. • 3P-RD files for all states will be required to accurately determine 'actively practicing', patient panel characteristics, and insurances accepted.
Active physicians	• The definition is successful if there are no gaps in claims data.	• Gaps in the claims data due to not all insurers submitting to the APCD impact the ability to identify whether a physician is actively delivering care.
Physician license status	• Partially successful definition	• Data varied greatly across states. It became difficult to identify classifications beyond active or inactive for most states. Therefore, the variable was limited to active, inactive, and unknown status. • An additional variable was added to include the original data from the source data file to provide granularity lost in the standardization process.

Variable	Evaluation of Success	Identified Challenges
Physician specialty	• Partially successful definition	• Many states provided open-text fields for physicians, allowing them the option to input multiple specialties in any order. Identification of primary and secondary specialty was not straightforward. • Based on feedback from two clinicians, the primary specialty was the most specialized specialty (e.g., cardiology) with the secondary specialty being the most general (e.g., internal medicine).
Accepted payers	• The definition is successful.	• This definition is limited by the claims data available.
Services provided	• The definition is successful.	• This definition is limited by the claims data available.
Patient panel	• The definition is successful.	• Patient panel definitions are limited by the claims data available. • Definitions must adhere to DUA restrictions.
Physician practice ID	• The definition is successful.	• Some APCDs did not include TIN in the data. However, we did have the PAC-ID, which is very similar to TIN, and were able to use that for the identification of practice IDs.
Practice size	• The definition successfully represented the number of 3P-RD physicians and other providers within a practice.	• Current definition and data sources do not allow for the inclusion of administrative staff.
Number of 3P-RD physicians	• The definition is successful.	• The current definition does not identify locum tenens physicians.
Number of non-3P-RD providers	• The definition is successful.	• The current definition does not indicate the number of non-3P-RD providers who are full-time and those who are part-time. • The current definition does not identify locum tenens physicians or other traveling providers.
Rural vs. urban	• The definition is successful.	• It is unclear whether the ZIP Code in the claims data is the actual location of the service or the location of the billing office. This may cause inaccuracies in rurality assignments.

Future Considerations

There are opportunities to refine and expand on how both physician and physician practice are being defined that would further enhance the program. Future work should build on lessons learned during the prototype phase, prioritizing a revisit of the definitions for physician and

physician practice. Further consideration and examination should also be given to addressing physicians who have practice sites in one state but serve patients in another state, especially as it relates to licensing and telemedicine. Below we suggest several variables for characteristics of physicians and physician practices to include in future iterations that may be helpful for researchers when analyzing physician workforce policies and issues.

Expand Providers

We recommend expanding the 3P-RD to include providers other than physicians (M.D. and D.O.). Through our discussions with SMEs and advisors, we found that they regularly mentioned the use of physician assistants and nurse practitioners to provide care to patients, especially in areas where the number of physicians is limited. Although the 3P-RD includes both physician assistants and nurse practitioners in the count of providers for the practice size, they are not included in the directory of the 3P-RD. Adding other provider types to the definition of "physician" in the 3P-RD would provide a more comprehensive understanding of the workforce and provider capacity.

Disciplinary Action and Medical Malpractice

The prototype includes information on disciplinary action and medical malpractice for physicians. The variable is derived from SMB data and varies by state. Currently, only two states in our study (California and New York) provided data on medical malpractice, whereas 11 states offered information on disciplinary history. The medical malpractice variables included in the prototype could be further expanded in the 3P-RD program through the standardization of medical malpractice variables across states. Likewise, the disciplinary action variable would also benefit from further refinement of the variable construction.

Disciplinary action is represented in two ways in the prototype: 1) The first variable simply indicates whether there is a history of disciplinary action. 2) The second variable provides the text data that were derived from the SMB data to provide detailed information for researchers if needed when conducting analyses. Future work for the 3P-RD program should consider harmonizing the text to provide greater granularity and better classification of disciplinary action. Like the provider specialty harmonization, the text provided varies by state and will take time and careful consideration to harmonize.

Other data sources for data on medical malpractice, such as the HRSA National Practitioner Data Bank (NPDB) should also be considered. However, due to privacy concerns, it may not be possible to include detailed level information from the HRSA NPDB. Aggregating more detailed malpractice information to a 'yes'/'no' indicator may satisfy privacy requirements for data release.

Insurance Acceptance Status

Both the physician and the physician practice data files contain variables that indicate Medicare FFS, Medicare Advantage, Medicaid, and commercial insurance acceptance. These variables are derived from the claims data and indicate acceptance of the insurance based on the evidence that a claim has been submitted for payment. Whether a physician or physician practice accepts a particular payer, regardless of a history of submitting a claim, is an area for future work. Most

states did not provide the information in the SMB data, so additional data sources are needed to address this gap.

Some state APCD eligibility data provide information on the product type of commercial insurance (e.g., Preferred Provider Organization (PPO), and Health Maintenance Organization (HMO) plans). Adding variables to indicate what type of commercial plan the physician and physician practice accept is another area of expansion.

Number of Patients Serve/Size of Panel

Future work should refine variables in the physician and physician practice data files that capture the number of patients served (i.e., size of patient panel). Building on work from the prototype phase, the next phase would focus on understanding which patients are attributed to a physician and/or physician practice and which patients are transient. Health care researchers use several definitions. Future work should consider whether: 1) the same definition should be used across all aspects of the 3P-RD or 2) the definition should change based on whether a physician or practice is hospital-based. Claims data coming from disparate sources—state APCD and CMS Medicare FFS—present a challenge for ensuring one patient is not inappropriately counted twice. Furthermore, the quality of T-MSIS data may present additional challenges in identifying patient attribution.

Size of Practice

We recommend refining the definition of the size of a physician practice. The current definition defines the practice size by counting the distinct providers (as identified by NPI) who have submitted a claim for a service provided at the practice site. However, this could be an overcount, as it may include providers who were present only temporarily. Locum tenens physicians travel to provide coverage during a staff physician's absence. Likewise, traveling physician assistants, nurse practitioners, and nurses can augment staff as needed. Modifying the definition of practice size has the potential to provide more accurate information. Several options can be considered when trying to capture more detailed information on practice size, such as: 1) capturing the use of traveling providers, 2) identifying the number of full-time staff members, and 3) identifying the number of administrative staff members.

Ownership

Since the data that were used to create the 3P-RD included TIN and PAC-ID, there is an opportunity to expand the 3P-RD physician practice data to include information on ownership status. There are three potential ways of accessing ownership status: 1) cross-linkages between TIN and other entities associated to the TIN (e.g., if the TIN is associated with a physician site that is also related to a hospital, then the physician site is likely owned by the hospital); 2) incorporating another secondary data source (e.g., data from the American Hospital Association, AHRQ Compendium of Health Systems); or 3) calling each physician site to ask about their characteristics. Although developing this measure would be valuable, the execution may be difficult, time-consuming, and costly, depending on available data.

Hospital-Based Practices

We recommend expanding the physician practice file to include information on the type of hospital-based practice. The prototype indicates which physician practices are hospital-based. Including additional information on the type of practice would provide researchers with information on workforce-related questions pertaining to emergency departments, outpatient centers, laboratory and pathology practices, radiology practices, and other specialties within a hospital setting.

Number of Nonphysician Medical Staff

Expanding the work of the prototype to include more granular data on nonphysician practitioners working in a physician practice is highly recommended. Future work can include indicator variables and/or counts of practitioners who are physician assistants, nurse practitioners, social workers, nurses, psychiatrists, and dietitians. Expanding the definition to include nonphysician medical staff members should be done in coordination with considerations for expanding the types of providers included in the 3P-RD.

Accountable Care Organization (ACO) Model Participation

Both the physician and physician practice databases can be enhanced to incorporate participation in ACO models by using data from CMS's Shared Savings Program (SSP) and Next Generation ACO (NGACO). CMS provides data for identifying physicians who participated in the SSP and NGACO[10] models. Indicator and dates of participation variables for physicians and practices associated with participating in an ACO model allow researchers to understand the impact of how physicians organize and practice medicine. Furthermore, the addition of ACO information indicates the level of risk a physician may be willing to take on.

[10] For more information, see https://resdac.org/cms-data?tid_1%5B1%5D=1&tid%5B6061%5D=6061.

4. Database Design and Data Development

The construction of the 3P-RD includes evaluating various aspects of the database design and mechanisms for data processing. This chapter provides a brief overview of the key requirements and considerations involved in designing a database and developing a data processing methodology that is flexible to effectively leverage the strengths of the disparate data sources. These requirements would also need to account for the inherent data quality issues, as well as variations in data availability across states. Finally, in this section, we discuss the processes that were successful for the prototype, provide suggestions for improvements, and identify new processes for future consideration.

Database Design

The database design for a 3P-RD program will need to accommodate multiple workstreams, including maintaining the initial prototype states and the inclusion of additional states, variables, and years. As a result, although database design aspects between the 3P-RD prototype and program may be similar, the final solution may be different. We present several key requirements of evaluating design options to ensure timeliness of database release, efficient and effective maintenance, and understandability for end-users.

Key Requirements

Transitioning the 3P-RD from the prototype phase to the 3P-RD program requires assessing the current database design while considering programmatic goals, timelines, and scope. Below is a list of key requirements for determining how best to design the database:

- **Replicability:** Replicability is an essential aspect of the database design and build process as it enables AHRQ to continue to expand and enhance the 3P-RD and reproduce the database on an iterative basis (e.g., quarterly or annually).
- **Scalability:** Scalability is essential for the database, as numerous sources are anticipated to be added to the 3P-RD over time as the database expands to include additional states, variables, and years. Scalability of the design allows for additional data sources to be incorporated into the processing and for data size to increase over time without relinquishing operational efficiencies and data quality.
- **Ease of linkage:** Critical aspects of possible database designs include linking tables and data storage to allow for easy linkage between the physician and physician practice data files and other key data sources.
- **Storage:** Database designs also differ in how they store data, and it is important to select one that efficiently stores data.
- **Flexibility:** Flexibility is a key consideration of the 3P-RD database design since the 3P-RD program will expand to include additional states and data sources.
- **End-user compatibility:** One key feature to consider while evaluating possible database designs is the capability for its end-users to easily retrieve, query, and analyze data within the 3P-RD.

Insights from 3P-RD Prototype Phase

For the 3P-RD prototype, we considered two database designs: a data mart and a relational database. Table 4.1 presents an evaluation of both options. Ultimately, the team determined that for the prototype an independent data mart containing flat files was the best solution, given the project's cost and timeline constraints. This offered the best solutions for the prototype—one that balanced operational efficiencies and total effort for setup and costs. It provided the team with some flexibility in dealing with unforeseen data challenges due to varying structures. A relational database would have required the team to revisit the original architecture. This would have added more time and cost to the project—both of which were already limited.

Table 4.1. Evaluation of Database Design Options for the 3P-RD Prototype

Evaluation Domain	Data Mart	Relational Database
Advantages	• A data mart structure with SAS datasets prioritizes time to identify and incorporate data from various sources vs. streamlining into one relational database. • A programmer would create code to transform the data into the specified format. Future iterations of the 3P-RD can use the code with minimal adaptations. • If additional data are received, a data mart's flexibility allows other variables to be easily added by adjusting the code to insert the variable into the SAS table.	• More efficient to store data compared to large data mart • Easier and faster to query data • May have additional security that can be implemented, making data vendors more comfortable with sharing data • Data validation and evaluation can be set according to predetermined rules, thereby automating quality assurance processes • Easily scalable, as the size of current data increases and additional data sources are added to the 3P-RD

Evaluation Domain	Data Mart	Relational Database
Disadvantages	• Less efficient method of storing and processing large administrative data files • Less efficient for querying data	• Requires, at minimum, a database administrator and operational staff to maintain over time • Due to the variety of incoming data sources, a database architect may be required to ensure proper functionality, efficiency, and table linkage of the database design. • Has an intense process for extracting, transforming, and loading (ETL) data. We anticipate data sources to vary across states, potentially increasing ETL's intensity and proving difficult to streamline into one relational database in the time limitations of the prototype project. • Adjustments may be required to the original design to accommodate unique value-added features. • Need to ensure that enough resources have been allocated to account for processing and storage (size of files) • The more files are created and stored on the database, the slower it may become.
Cost considerations	• Minimal cost since data are stored within a simple structure that does not require technical resources such as a database administrator or architect.	• High costs for initial setup and additional costs for maintenance • Need to staff for data developers and architects
Timeline considerations	• Minimal time to stand-up and maintain file structure	• Timeline for setup will vary and depend on the complexity of the data being stored (e.g., number of data sources, file structure, required relationships between data elements), the selected data architecture and the rate at which the data are expected to grow.
Prototype choice	• Selected	• Not selected for the prototype

Future Considerations

Although we selected a data mart model for the prototype, we anticipate that, in the long term, switching to a relational database will be better, especially when the number and size of data for

the 3P-RD are anticipated to increase. Given the structure of a relational database, it is easily scalable; maintains data integrity; automates quality assurance during the ETL process; and is highly efficient for users to access, manipulate, and query data. Although more personnel are needed to design and maintain a relational database, using information from the prototype phase (e.g., the number and type of disparate data sources, ETL processing techniques, interim and final data files, linkage requirements) provides strategic information for a database architect and an administrator to construct a relational database that would more efficiently receive, store, and maintain the 3P-RD files. Furthermore, a relational database would provide additional efficiencies for ETL processing, data transformation, and querying of the final 3P-RD to create public use files (PUFs) and restricted use files (RUFs) for external users.

Data Development

The data development for the 3P-RD includes the ETL process, standardizing and harmonizing, linking data files, and integrating value-added data elements. Several tools and methods, each with its own advantages and disadvantages, can be used to conduct the various aspects of data development.

Key Considerations

The team evaluated a range of available tools and techniques to determine which is the most appropriate for developing the 3P-RD. Below, we provide a list of key elements that the team needed to consider when developing a data development plan.

- **Software tools for processing:** Several software packages are available to use for data development (e.g., SAS, R, Python). It is important to identify the advantages and disadvantages of each.
- **Data transformation:** Data harmonization, standardization and transformation techniques and requirements will differ across disparate data files. Files need to be transformed prior to conducting linking and further data processing.
- **Identification variables:** It is critical to have variables available for linking disparate data files for data development and for the 3P-RD to link to other data sources. Linking variables will minimize processing time and increase usability of the 3P-RD for research.
- **Data availability and variability:** Disparate data files contain differing levels of provider information. Data processing and information available for the 3P-RD may not be consistent across all 50 states.
- **Benchmarking data:** To ensure the end product is of high quality and accurately reflects information for health-policy use, it is critical to benchmark the data to determine the level of quality and accuracy of the 3P-RD data files.

Insights from the Prototype Phase

Overall, the process that was implemented for the prototype is sustainable and can be expanded for the 3P-RD program, even if more states are added to the list. Relying on programming techniques that streamline and automate the processing of data can decrease processing time

and increase focus on refinement activities for the 3P-RD. Next, we discuss in more detail how we approached each key component and provide an evaluation of our decisions.

Software tools. Data processing for the prototype had to occur across many environments in a defined timeframe. Focusing our work to use only one software (e.g., SAS) allowed the team to develop programming code that streamlined processing and minimized duplicative code development. As a result, once coding logic was generated and quality checked, that code was then used across different work environments, such as on NORC's systems and CMS's Virtual Research Data Center (VRDC) when developing the physician data or the physician practice data. Table 4.2 presents our main takeaways when evaluating the success and challenges we faced in implementing our approach to software tool selection.

Table 4.2. Software Tools – Takeaways from the 3P-RD Prototype

	Takeaways
Success	• SAS is widely used across the healthcare field, including federal agencies. • SAS is proficient in the management and transformation of large data files. Due to the manner in which SAS processes data, data cleaning and processing can be done efficiently.
Challenge	• SAS requires a license and can be expensive if the organization does not already have an enterprise license. • SAS can be slow to query data, depending on the environment in which it resides. • More programmers are becoming increasingly proficient in other software packages, increasing the use of less efficient processing techniques and code within SAS. • Web-scraping and data encryption cannot always be accomplished using SAS. Other software packages may be required or more efficient.

Data harmonization and standardization. Streamlining and automating data standardization and harmonization across disparate data files increases overall programming efficiency. Most data harmonization and standardization conducted for the 3P-RD were rather straightforward (except the specialty harmonization, which we discuss next). Examples of these processes include automating how variable names and formats are harmonized across files, ensuring that coding is consistent across data files (e.g., M/F, 0/1, male/female), and aligning incoming data sources that contain the same information, even when source variables are differently named (e.g., license status and license type). Table 4.3 presents our main takeaways when evaluating the success and challenges we faced in implementing our approach to data harmonization and standardization.

Table 4.3. Data Harmonization and Standardization – Takeaways from the 3P-RD Prototype

	Takeaways
Success	• Automation of data processing reduces manual re-assignment.
Challenge	• How states captured and stored specialty, license status (e.g., active, suspended), and disciplinary information was not consistent. Extensive assessment of disparate data for the development of a standardized classification system and subsequent harmonization of disparate files are required by clinicians and SMEs (detailed information on specialty harmonization is addressed separately). • Validation of harmonized data sources

Specialty harmonization. For the 3P-RD prototype, specialty harmonization required significant time and resources to standardize the data. State SMB data, state APCD claims data, CMS Medicare FFS, and T-MSIS data often use different classification systems for physician specialty. The team needed to align specialty assignments across files to ensure proper linking across those files and allow end-user analysis to be consistent across states and data sources. As a result, the team transformed and standardized specialty information data into similar formats and harmonized those data into one classification system. Table 4.4 presents our main takeaways when evaluating the success and challenges we faced in implementing our approach to specialty harmonization.

Table 4.4. Specialty Harmonization – Takeaways from the 3P-RD Prototype

	Takeaways
Success	• Using the CMS Provider Specialty classification aligns the 3P-RD with other data sources. • CMS provides a list for Provider Specialty to Taxonomy that served as a crosswalk for data processing and harmonization. • Using multiple clinicians to review classified data improves accuracy of classifications. • Clinicians and SMEs provide insight into practical application of specialty descriptions. • The CMS Medicare FFS and T-MSIS data files provide information on the specialty that physicians submit for billing. Assessing specialty codes for billing by patient and claim volume provides validation of specialty harmonization logic.
Challenge	• The harmonization of provider specialty is a laborious process. • SMB data files may contain free text for specialty information resulting in complicated programming to clean the data prior to processing data to the standardized classification system. • Not all specialties could be assigned or classified due to ambiguous text.

Identification variables. Identification variables (IDs) can be used for linking disparate data files. The use of unique IDs is essential to streamline linking and to reduce processing times, as well as to identify individual physicians, other providers, and practice locations. Table 4.5 presents our main takeaways when evaluating the success and challenges we faced in using these variables.

Table 4.5. Identification Variables – Takeaways from the 3P-RD Prototype

	Takeaways
Success	• Most administrative files contain at least one type of ID, such as the NPI, state licensure number, or TIN. • IDs can be used for both identification and linking purposes. • Most data files contain an NPI, reducing the processing time and increasing accuracy of linkage between files. • State license numbers are present in professional association, state agency, and federal agency data files. • Provider names are included in many administrative provider data sources.
Challenge	• Not all files contain the same identification number, thus requiring additional processing. As a result, additional steps are required for processing or the sequence for processing files may need to be adjusted to ensure linking variables are present on all necessary files. • APCDs can have multiple IDs for members and providers, including NPI and payer-specific member/provider IDs. Multiple IDs within a state data file will increase processing time. • IDs may be used for multiple purposes—e.g., the NPI may be used for both the individual provider and the organizational provider. Therefore, additional variables may be required for identification (e.g., individual vs. entity) and data processing. • Provider and billing policy may allow multiple providers to bill under the same NPI, affecting the accurate identification of physicians and nonphysician providers. The volume of services performed, and patients seen by a physician or practice may also be inflated due to such a billing policy. • Other information such as state license numbers and provider names can be used for linking but may require additional data cleaning and matching to standardize data across disparate data files.

Data availability and variability. Provider information is not consistent across all disparate data files (Table 4.6). Claims data contain provider information related to billing, whereas SMB data include information specified by the state for licensing. In addition, states vary in the amount of information they will release for use and have varied requirements concerning whether the information can be used in an integrated data product such as the 3P-RD.

Table 4.6. Data Availability and Variability – Takeaways from the 3P-RD Prototype

	Takeaways
Success	• All provider data include at least one provider ID. Claims and federal data sources primarily use the NPI, with some containing license numbers. States primarily use the state license number, although some include the NPI. This provides the ability to link data. • State data uniformly contain the same information for providers, such as name, license number, license dates. • Some states (e.g., Florida, California) contain information about military affiliation and/or medical school location. • Providers in all states have sources for medical board and claims data, although the data source will vary by state. • Linking, standardization, and harmonization allow provider data to be aligned across states and data gaps to be filled. • Provider adjudication uses claims data to confirm information from the SMB to ensure that only validated physicians were included in the 3P-RD.
Challenge	• Not all SMBs release information for active providers, whereas other states include information on all providers who applied for a license. • Some SMBs include information on all providers who have applied for a license without clear indication of their provider type. As a result, it is necessary to validate providers identified in the SMB data as physicians (M.D., D.O.). The provider adjudication process would help fill the gap in SMB data. • Often state medical boards do not contain the NPI—the variable required for linking to NPPES—thereby challenging the ability to obtain the NPI associated with the state license. Since NPPES is manually entered by the physician, the license number is not standardized between the disparate data sources and may impact the linkage success rate across files. • Reasons for disciplinary action vary by state. Attempts to harmonize and standardize the information were not cost-effective due to the numerous options states allow. Furthermore, descriptions are vague and difficult to classify. As a result, the prototype left the information from the state data as is. • License status varied by state. Although license status (e.g., active, expired, suspended) was harmonized and standardized across states, we also included a text field so that the original data may also be represented. This allows individual researchers to reclassify this variable differently if preferred. • Race/ethnicity data have a high level of missingness in all data sources.

Benchmarking data. A critical step in data processing for the 3P-RD program is data benchmarking (Table 4.7). Identifying high-quality, appropriate data sources for benchmarking is necessary to ensure that 3P-RD files are being compared to similar standards. Benchmarking also demonstrates the accuracy of data linkage and areas of remaining data gaps.

Table 4.7. Benchmarking Data – Takeaways from the 3P-RD Prototype

	Takeaways
Success	• Use of the Association of American Medical Colleges (AAMC) workforce survey provided detailed information on the number of physicians practicing by state. • Reviewing statistical summary of data files highlights outliers within the data. • Reviewing state policy for billing and researching workflow for identified specialties (e.g., radiologists and pathologists) provided supporting documentation of reasonable distribution of claims and patients.
Challenge	• The 3P-RD includes the census of physicians, regardless of license status (e.g., active, expired) or actively providing services (e.g., active license but conducts medical research). Benchmarking against the AAMC workforce survey provides only an estimate for the census of physicians. • Identification of high-quality, readily available data sources is difficult. Data sources such as the AMA Masterfile and IQVIA OneKey data require a license that can be costly. • Benchmarking physician practices is difficult due to differences in practice definition between the 3P-RD and other data sources. Many rely on only the TIN and organizational NPI to identify the physician group, whereas the 3P-RD includes the ZIP Code of the service (as found on the claims).

Future Considerations

We recommend building on the processes and programming code developed for the 3P-RD prototype. Future focus should include enhancing programming and statistical techniques for specialty harmonization, provider adjudication, and race/ethnicity identification (Table 4.8).

Table 4.8. Evaluation of Database Design Options for the 3P-RD Prototype

Key Consideration	Future Consideration
Processing claims files	• Use a relational database to process the data to reduce processing times.
Specialty harmonization	• Use existing specialty harmonization as a foundation for future specialty harmonization efforts, including creation of new state-specific specialty files. • Use clinicians and SMEs to conduct quality control on specialty assignments and provide feedback for classification of new state data. • Build on existing code to improve methods of automation and to streamline processing.

Key Consideration	Future Consideration
Claims data	• Using the APCD for claims is preferred due to the initial processing that states conduct, including consolidating patient and provider IDs across submitters and providing information for commercial, state, and federal payers. • Preference should be given to state Medicaid claims, since T-MSIS files have already been processed and may be less granular on some domains. • For states without an APCD, discuss the option of obtaining state Medicaid data. • Consider using state hospital discharge data (HDD) to access information on the physician workforce in institutional settings. HDD also provides information on the uninsured population.
Data availability and variability	• As with specialty harmonization, use the existing programming code from the prototype as a foundation for future adjudication efforts. As the number and quality of data sources increase, expand on provider adjudication to ensure the full census of physicians is being captured. • Identify which variables that are currently not available in states may be collected in the future. • Conduct periodic environmental scans to identify additional data sources that may address data gaps in the SMB data. • The use of clinicians and SMEs may help when developing algorithms to determine provider type from available information. • The rate of missingness of race/ethnicity data needs to be addressed. Evaluating variables within the SMB data or techniques for imputing information may help to address the issue while state and federal agencies work on collecting higher-quality race/ethnicity information in the future.
Benchmarking	• Identifying a validated, high-quality data source to use for benchmarking is critical to verify active physicians and physician practices. • Validate and benchmark physician practice sites in the state.

5. Data Dissemination

Data dissemination for the 3P-RD includes PUFs, RUFs, and end-user documentation. Providing a rich resource for both AHRQ's internal purposes and external health care researchers requires careful determination of how data can be disseminated so that restrictions can be adhered to while balancing data utility. This can provide researchers the most granular information allowable without risking disclosures. RUFs and PUFs allow varying levels of information to be disseminated. Finally, end-user documentation of the 3P-RD is vital for users to fully understand and properly use the 3P-RD.

Key Requirements

- Data dissemination must adhere to DUA restrictions from relevant data providers, including data suppression rules and data dissemination requirements from data providers.
- AHRQ is required to adhere to the AHRQ Confidentiality Statute when releasing data. Data dissemination must comply with Section 944(c) of the Public Health Service Act (42 U.S.C. 299c-3(c)).
- The data files released must be useful for policy analysis. Therefore, files should provide information on physicians and physician practices with the most granular level of information to provide researchers the ability to understand and analyze the capacity of health care supply, how physicians organize, the types of services they provide, and the patients served.

Insights from Prototype Phase

Releasing the 3P-RD for researchers, both internal and external to AHRQ, requires attention to details related to DUAs, data owner concerns, level of granularity necessary for policy analysis, end-user experience, and data lifecycles. Several options are available to AHRQ for data release, each offering varying levels of accessibility and data granularity. Data release timelines will depend on the types of files released, timelines for source data, and requirements of DUAs. For the prototype, the team identified three types of files for release for both the physician and the physician practice data files. We consolidated end-user documentation into a primary document to ensure simplicity for end-users.

Data Dissemination Files

For the prototype, the team created two PUFs and two RUFs for each of the physician and physician practice files, with a corresponding codebook for each. We also prepared a data dictionary for both the PUFs and RUFs. Table 5.1 describes each file and provides key insights from the prototype work.

Table 5.1. 3P-RD Files for Data Dissemination

File	Description	Key Insights
Physician PUF	• Directory of physicians in the 3P-RD • Contains information derived only from publicly available data sources.	• Physician privacy concerns are rising. • Data owners are concerned about disclosure risk for physicians and practices that provide services to vulnerable populations.
Physician practice PUF	• Directory of physician practices in the 3P-RD • Contains information derived from publicly available sources, focusing on variables drawn from the specialty harmonization work	• May be easier to expand given the aggregated data • Data owners are concerned about disclosure risk for physicians and practices that provide services to vulnerable populations.
Geographic PUFs	• Versions of the physician PUF and physician practice PUF aggregated to the 3-digit ZIP Code. • Includes selected variables derived from claims data (e.g., minimums and maximums for patient age, claims per month and patients per month)	• Data owners less concerned about sharing data aggregated to the ZIP Code level. • Due diligence still requires assessing disclosure risk for physicians and practices that provide services to vulnerable populations.
Physician RUFs	• Directory of physicians in the 3P-RD • Contains information derived from publicly available data sources and where the data owner has given permission (e.g., variables derived from claims)	• Further discussions with data owners may allow for additional variables derived from claims data to be included. • Data disclosure will be required to ensure that privacy concerns for vulnerable patient populations are maintained.
Physician Practice RUFs	• Directory of physician practices in the 3P-RD. • Contains information derived from publicly available data sources and where the data owner has given permission. • Inclusion of hospital-based practice information, which relies on claims information.	

File	Description	Key Insights
Geographic RUFs	• Versions of the physician PUF and physician practice PUF aggregated to the ZIP Code • Includes selected variables derived from claims data (e.g., minimums and maximums for patient age, claims per month and patients per month)	• Able to provide data at the ZIP-5 level rather than ZIP-3. • Working with data owners may allow all or most variables to be released. • Not every variable should be aggregated up to the ZIP Code level. Aggregation may render some variables meaningless at the ZIP Code level. For example, the average number of claims submitted in a ZIP Code may be skewed due to outliers within the ZIP Code.
End-User documentation: data dictionary and codebook	• Can be provided both in Excel and Word/PDF formats • Documentation should contain information on variable construction and methodology. • Documentation should highlight key differences by state.	• Documentation can be streamlined using programming code, ensuring that the information accurately reflects the data files. • It is important to document for end-users the impacts of cell suppression on the data, especially for summary statistics across any of the files.

Timeline of Data Lifecycles

The data lifecycles vary greatly across the disparate data sources. SMB data tend to be updated annually, whereas CMS T-MSIS data lag by several years–e.g., the preliminary calendar year (CY) 2020 data were released in November 2021. Furthermore, data release cycles may not align with data update cycles. Although new NPPES data are released monthly, physicians are required to update information as needed. Therefore, although new data are released, the information they contain may not be up to date for a provider (although providers are required to update information within 30 days). As a result, the data lifecycles of source data do not easily align with one another.

The prototype opted to solve issues with conflicting timelines by pulling the most recent data from each data source. The 3P-RD prototype reflects the most recent data at the time of data delivery. All claims were aligned to a single year, so states with more advanced APCD data delivery schedules might have provided data for both 2019 and 2020, but only 2019 claims were used for the 3P-RD, as that was the year for which claims data were available across all data sources. We obtained the most recent SMB, NPPES, and PECOS data, reflecting information from 2021. The team included variables in the 3P-RD that indicated the year of data used for each data source. Therefore, researchers are able to incorporate date information into their evaluation of the data and discussion of findings.

Future Considerations

Future work for the 3P-RD program should include several topics of discussion for data dissemination, including file type, DUA requirements, data disclosure, and release schedule. Due to data release policy concerns within AHRQ and for data owners, identifying and clarifying key areas of concern early in the program development will be important. Drawing from lessons learned during the prototype, we recommend the following areas for consideration:

- Hold internal discussions with AHRQ legal experts to determine the level of information that can be released for physicians and physician practices.
- Identify variables of interest stemming from claims data early on and begin discussions with data owners to determine whether the information can be released under a PUF or RUF without additional review by the data owner.
- Evaluation of disclosure should be conducted for Medicaid populations, patients with rare diseases, and residents of rural areas. Data owners may want to be assured that patients who fall into these categories are not at risk for disclosure in a physician and physician practice database.
- Consider physician privacy concerns and develop approaches to mitigate the impact on 3P-RD data dissemination.
- Releasing data annually will reflect information on physicians with a current license and the most recent information about those actively providing care.
- Using the most recent claims available by source can provide the most accurate picture by state, although it may decrease the ability to compare providers across state lines.

6. Role of AHRQ and Partners

AHRQ's Role

AHRQ is responding to multiple federal priorities, including the Federal Data Strategy's call for the federal government to accelerate the use of data to deliver on mission, serve the public, and steward resources, while protecting security, privacy, and confidentiality. Through this effort, AHRQ is pioneering what can be considered a challenging yet critical national data resource to meet the needs of the ever-changing health care landscape. This effort can only be led by a federal agency with access to the resources and budget necessary to scale this project to span nationally. This feasibility study provided another powerful use case for how AHRQ can meaningfully engage states and other data owners in developing an infrastructure that can regularly create and disseminate files that span multiple states. Because of AHRQ's experience in working with and harnessing sensitive data (such as through the HCUP program), the agency has demonstrated its ability to be a trusted steward of the data.

For AHRQ, this prototype provided an opportunity to test how a program can be developed and what considerations need to be accounted for in sourcing and managing a program built on this experience. This prototype also offered AHRQ a glimpse into how to motivate various types of data owners to participate in this endeavor and how a sustainable partnership be established. It also allowed AHRQ to understand what challenges do various data owners face, and how can this project support and address some of those challenges.

This feasibility study also offered data owners a glimpse into how their data can be used, the intended relationship between them and AHRQ, and the opportunities for how this work could enhance their own products, including stating business cases to their own constituents and stakeholders.

Data Partners

Need for strong relationships with data owners and other stakeholders who have access to the range of data.

As described earlier, the 3P-RD served as an important use case for both AHRQ and various data owners to understand how administrative data can be leveraged to develop a national resource like this. Data owners often face a number of constraints that affect their ability to share their data. These constraints range from state legislative constraints to their organization's business model, and they may limit or prevent them entirely from participating in a data program meant to create PUFs. Through this scaled feasibility study, data owners are able now able to see how a viable model for engagement with the federal government could emerge. It also serves as a tool for them to advocate with their stakeholders for the short-term and long-term value of participating in this effort.

Key to all these partnerships is a long-term commitment and a clear vision that can be supported by sufficient resources. AHRQ's setup can reflect the HCUP setup—an already proven model that has effectively leveraged data from different types of partners across states.

Federal-Federal Partnership

AHRQ's engagement with CMS on this project offers a powerful example of how important a cross-agency federal partnership is to enhance the value of the work and to ensure its sustainability. For the prototype, AHRQ was able to leverage its existing relationship and DUA with CMS to help the team access CMS's claims files in the Chronic Conditions Data Warehouse (CCW) VRDC environment. The CMS files were important to supplementing the state files—particularly for states without an APCD—and served as an important resource to fill existing data gaps.

The AHRQ-CMS relationship also allowed for major efficiencies in multiple areas in the execution of the project:

- **Time efficiencies.** Allowing access to the files early in the project enabled the team to begin work.
- **Resource efficiencies.** By working with already well-curated federal data files. the team was able to minimize the resource and processing burden that would have otherwise been costly if the data had not already been processed and maintained.
- **Budget efficiencies.** CMS had priced out our data request with special consideration for AHRQ and provided access to their data at a reduced fee.
- **Data support efficiencies.** Through AHRQ's relationship with CMS, the team was able to connect with all the right resources in a timely manner and have a direct line of contact with CMS if any issues emerged while working with their data.

These various efficiencies allowed our team to stay within the contract's budget and schedule. The CMS data resources were critical to filling existing data gaps that otherwise would have been challenging to address in the files. The CMS-AHRQ data partnership example is one that AHRQ should continue to explore with other federal agencies that collect data that would help fill existing data gaps and further enhance the value of this program. AHRQ should explore how a partnership could be established with agencies in the U.S. Department of Veterans Affairs, the Office of Personnel Management, and the Department of Defense to help access Veterans Affairs (VA) and TRICARE data and data from the Federal Employees Health Benefits program to help fill data gaps that would otherwise be very challenging to address.

Federal-State Partnership

State APCDs have also often expressed concern regarding confidentiality and privacy. Many of them are mandated by state legislation to meet certain requirements regarding how to collect and release data. AHRQ can support some of these APCDs through these types of data-building resources that demonstrate to their legislative constituents the value of participating in this federally led program.

Although each state APCD has its own customized DUA and data release processes, navigating each unique setup when looking to leverage multiple state data for a single project can be very

challenging. According to a U.S. Department of Labor (DOL) report,[11] however, states have expressed interest in the possibility of creating a standard DUA, possibly similar AHRQ's HCUP. This could be an opportunity for AHRQ to negotiate with state APCDs to implement uniform agreements and processes. The value of this exchange would be particularly important for establishing a sustainable 3P-RD program.

Federal-Private Partnership

There are a few private data collection efforts that would be relevant to this program. These efforts span commercial payers and other proprietary datasets (e.g., Aetna and Blue Health Intelligence data, and data programs such as IBM® MarketScan® Research Databases program, and the Health Care Cost Institute's commercial claims dataset). However, there are several challenges in using these sources, such as costs to access data (e.g., costly licensing fees) and data dissemination limits (e.g., not allowing data to be used on any publicly disseminated files). There could also be concerns regarding how AHRQ's work may seem to duplicate their own efforts or may raise existential concerns about the potential for a reduction of their purpose.

Nonetheless, there are important opportunities for either linking 3P-RD physician and physician practice data with these data or using those sources for benchmarking the 3P-RD. AHRQ should consider building a long-term partnership to empower the value of these private datasets and the value in having their owners participate in this effort.

Program Partners/Contractors

In addition to establishing data partnerships, AHRQ should also consider program management and support from trusted third-party intermediaries—whether a private or nonprofit organization contractor. This partnership would harness the agility of the private sector while retaining government oversight and protections. The partnership would operate, maintain, and manage contracted support services necessary to develop, expand, and maintain a 3P-RD program. The contractor would implement the program for AHRQ by supporting the effort in such capacities as:

- Completing data applications and negotiating or updating memorandums of agreement or DUAs
- Purchasing data, recruiting additional data from existing partners, and establishing new partnerships when possible
- Processing data obtained from partner organizations and developing and maintaining the 3P-RD databases
- Coordinating the centralized distribution of 3P-RD data and handling activities such as reviewing applications to access restricted public-release 3P-RD databases.

[11] U.S. Department of Labor, SAPCDAC. State all payer claims databases advisory committee report with recommendations under Section 735 of the Employee Retirement Income Security Act of 1974. Available at: https://www.dol.gov/sites/dolgov/files/ebsa/about-ebsa/about-us/state-all-payer-claims-databases-advisory-committee/final-report-and-recommendations-2021.pdf.

Conclusion

The onset of the COVID-19 public health emergency highlighted the need to develop a program on the U.S. physician workforce and physician practices, and AHRQ swiftly stepped into the role of fulfilling that need. With the understanding of how difficult developing an entire program in the middle of a crisis can be, initiating this feasibility study was a shrewd example of how effective change can come about through careful planning. Based on the outcomes of this study, it is clear that building out the infrastructure for developing this program is possible and critical to meeting the policy and research needs of multiple stakeholders.

Ideas have power to shape public policy in innovative and meaningful ways, as demonstrated through all of AHRQ's current data programs. NORC appreciates the opportunity to have contributed to the foundations of an exciting and critical data development program. Our team is grateful for all the contributions of our subcontractors, all the SMEs who were consulted, and all the data owners we engaged with during the project, with a special note of appreciation to the various state APCD and CMS representatives. The input and support we received and the exchanges we have had were invaluable and added significantly to the quality of our work. We would also like to reiterate our appreciation for AHRQ's leadership, guidance, and vision throughout the project.