AHRQ
Agency for Healthcare
Research and Quality

AUGUST 2022

AHRQ Physician and Physician Practice Research Database

Methodology Report

Table of Contents

List of Tables and Figures

1. Introduction

The AHRQ Physician and Physician Practice Research Database (3P-RD) was developed to help address gaps in current physician and physician practice data at the state and market levels. The 3P-RD prototype contains information pertinent to supply-side policies and questions, focusing on active physicians' current roles in the healthcare system and the corresponding relationships within practices and facilities. The database harmonizes physician and physician practice data from administrative and other available data sources for 13 states and includes data elements related to characteristics of physicians and physician practices useful for informing policy-relevant health services research.

3P-RD States. The following states are represented in the final 3P-RD public-use files (PUFs): Arkansas, Arizona, California, Colorado, Florida, Massachusetts, Maryland, Minnesota, Missouri, Montana, New York, Texas, and Washington.

3P-RD Data Files. The outcome of this effort produced the following 4 sets of files for each of the 13 states:

- **A Physician Directory Public Use File** which provides information and characteristics for physicians within the state that have a license.
- **A Geographic Physician Public Use File** which provides summarized information and characteristics of physicians within the state. Physician information has been summarized at the 3-digit ZIP Code.
- **A Physician Practice Directory Public Use File** which provides information and characteristics of practices within the state.
- **A Geographic Physician Practice Public Use File** which provides summarized information and characteristics of practices within the state. Practice information has been summarized at the 3-digit ZIP Code.

Purpose of this document. The *AHRQ Physician and Physician Practice Research Database Methodology Report* serves as a resource for data users to reference when first working with the 3P-RD. It provides information on how the 3P-RD was developed including descriptions of the data cleaning and validation routines of the data sources that were used to build the files, how the files were then linked and harmonized and then used to build the final set of state-level files. Users are expected to refer to this document along with the 3P-RD codebooks.

In the next section we provide an overview of the documentation accompanying the 3P-RD files, and in subsequent sections, we discuss how the 3P-RD was constructed.

2. Documentation Overview

This section provides an overview of the methodology report, the codebooks and the data dictionary that have been made available for users of the AHRQ 3P-RD.

2.1 Methodology Report

This methodology report contains information for researchers about the structure of the 3P-RD files and the data sources that were used to create them. It includes a description of the data cleaning and harmonization procedures that were implemented on the source files. It then discusses how each state file was built for the thirteen states represented in the 3P-RD. The next set of sections in this report describe the following:

- **Construct Definitions and Data Sources** (section 3): This section describes the data sources that were used to build the 3P-RD. It provides the overall conceptual design and how the various data sources were used the create the files.
- **Data Processing** (section 4): This section describes the data cleaning and harmonization procedures that were implemented across the various data sources.
- **State-Specific Builds** (section 5): This section describes how physician and physician practice files were developed for each state.
- **Public Release Files** (section 6): This section discusses what and how the data was distilled into the current public release files that have been made available to users.

2.2 PUF Codebooks

A total of **four (4) codebooks** are available to users as PDF documents. Each codebook is associated with either one of the four types of physician and physician practice state-level public use files (PUFs): physician and physician practice directory files, and the physician and physician practice geographic files. Each codebook provides a list of available data elements and their characteristics. The codebooks include a variable label, variable name, input data source, variable type and length, and state-specific information.

2.3 PUF Data Dictionary

A data dictionary is also available to users as an Excel spreadsheet. It provides a list of available data elements and their characteristics for both the directory and geographic PUFs. It demonstrates how variables have been harmonized across the various data sources that were used when building the 3P-RD. The data dictionary contains four Excel sheets: a contents sheet describing the contents of the data dictionary, a legends sheet describing coding for payer prefix of variables derived from the claims data, a sheet for physician data elements and a sheet for physician practice data elements.

3. Construct Definitions and Data Sources

3.1 Construct Definitions

The 3P-RD focuses on physicians that are defined as Doctor of Medicine (M.D.) or Doctor of Osteopathic Medicine (D.O.). It currently does not include all providers (such as 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. While definitions were initially informed by the literature and existing data programs that were already capturing aspects of the population of interest, definitions were further refined through multiple reviews with both clinicians and subject-matter experts as well as through working with the data. The identifying characteristics for the physician and physician practices were drawn from information contained within the available data, definitions known and pervasive within healthcare research, and adherence to data use restriction requirements.

The next set of tables represent key variables that were operationalized for the 3P-RD. **Table 3.1** presents definitions for the main concepts of the 3P-RD, while **Tables 3.2** and **Table 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 that 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
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 are 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.
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.

3.2 Data Sources

The core structure of the physician and physician practice state files was created by linking state medical board (SMB) licensure data to the National Plan & Provider Enumeration System (NPPES) and the Medicare Provider Enrollment, Chain, and Ownership System (PECOS). Claims files derived from the Centers for Medicare & Medicaid Services (CMS) sources as well as state all-payers claims databases (APCDs) data (where available) supplemented and enhanced the core files.

Core Data Source Files. There were several benefits to linking the SMB data to the publicly available NPPES and PECOS data files such as:

- 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 associated with the SMB records.
- The files can be linked to both PECOS and claims data through NPI.
- Providers from the NPPES and PECOS files that do not appear in the SMB state file can now be captured through the SMB-NPPES-PECOS file linkage.
- 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. Further, 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 consisted of the first step in creating the Core physician file allowing the team to identify all possible physicians within a state.

Supplemental Data Source Files. APCD or CMS claims data elements were then identified as sources to enhance the core 3P-RD data files. Claims data, particularly state APCD data, offered a rich resource of information that was leveraged for the 3P-RD. State APCD data was available to use for 4 of the 13 states that were selected for the study. This data was used 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).

4. Data Processing

Data processing procedures were implemented in the construction of the 3P-RD. Data processing for the 3P-RD entailed importing acquired data, performing initial data cleaning and harmonization of data elements across disparate data sources, and validating the harmonization process using third-party data. A final adjudication of providers was then conducted to identify physicians (i.e., Medical Doctor or Doctor of Osteopathic Medicine) to create state-specific physician files. These files are identified as the Core physician files. Quality assurance checks were conducted across all data processing tasks to ensure the accuracy of data preparation for building the 3P-RD. A description of how the various data files were cleaned, validated, harmonized, and linked is provided next.

4.1 Data Cleaning, Standardization, and Validation

4.1.1 Purpose

The purpose of the data cleaning and validation routine was to apply generalized cleaning to targeted variables in the provider files. The provider files included the individual SMB, NPPES, and PECOS data files. Fields that were routinely cleaned included names for organizations and individual providers, date values, and address information which included address, city, state, and ZIP Codes.

4.1.2 Overall Process

Data cleaning and validation began with importing data into SAS. Once imported, the number of variables were verified against the information provided in the data documentation for each data source to confirm that variable names, data values and the number of variables were correctly imported into the SAS files. Variable frequency distributions were also run to verify the content included in each dataset. Once datasets were validated, the data was processed to clean the following fields: name, date, and address.

Several states required changes to SMB data processing. SMB data acquired through web-scraping or states that contain physician information across several files, required a different initial import process. For those states, their files were merged into one state SMB data file using state-specific logic. **Table 4.1** provides the state-specific import steps that were used for states which required additional data import and cleaning processing.

Table 4.1. State-specific SMB Data Processing

State	Number of Files/File Origin	Import Process Changes
California	18 data files received through SFTP	Merged the necessary files by license ID to create a clean SMB file with the information we needed
Florida	22 data files	Merged the necessary files by Professional code and license ID to create a clean SMB file with the information required to create the FL 3P-RD physician file.
Missouri	2 files: MD and DO	Combined the MD and DO files and then processed the combined file using standard processing.

State	Number of Files/File Origin	Import Process Changes
New York	Web Scraping	The New York data was web scrapped. The scraping process followed a recursive search and select approach from which A-Z was entered and searched in the last name field. Post-search, each physician link that appeared in the search was opened and information was collected.

4.1.3 Name Clean-up/Standardization Routine

For the routine cleaning of name fields, both general and field-specific cleaning and standardization criteria were developed. These criteria included changing the name fields to all uppercase, removing multiple blank spaces, and removing leading and trailing blank spaces from the name. No clean-up steps were taken for organization name fields due to their complexity (for example, some organization names contained symbols and numbers that were important to preserve in the name field). The following clean-up/standardization rules were applied to individual provider's first, middle, and last name fields:

1. **Accented letters:** Adjusted to their non-accented equivalents
2. **Brackets** *(commonly indicating a nickname):* Adjusted to parenthesis
3. **Double quotes** *(commonly indicating a nick name):* Adjusted to single quote
4. **Unknown name text** (*commonly identified as 'Unknown', 'Don't Know', 'No First Name*'): Removed
5. **Zeros:** Changed to the letter 'O' when found between two letters (e.g., R0BERT = ROBERT)
6. **Numeric values, control characters, and punctuation:** Removed (excluding periods and hyphens)
7. **Single quotes:** Updated to parentheses when surrounding a name
8. **Last name only:** Corrected and standardized special name text (this includes 'ST' transformed to 'SAINT' and corrections such as 'O BRIEN' to 'O'BRIEN')
9. **Prefix, suffix, credentials:** It is common for prefix, suffix, and provider type (credentials) to be separated into their own individual fields. Therefore, prefix, suffix, and provider type (credentials) were removed from their respective name fields.

4.1.4 Date Clean-up Routine

Date values were simply compared to a starting and ending point input by the user of the program. If a date fell outside of this range, it was updated to a null value. For example, it is common for a system to input a default value (typically, 01/01/1900). Improbable date values such as dates that occur in the far past or future (i.e., a 2017 date in a 2016 data file) were removed from the file.

4.1.5 Full address clean-up/standardization routine

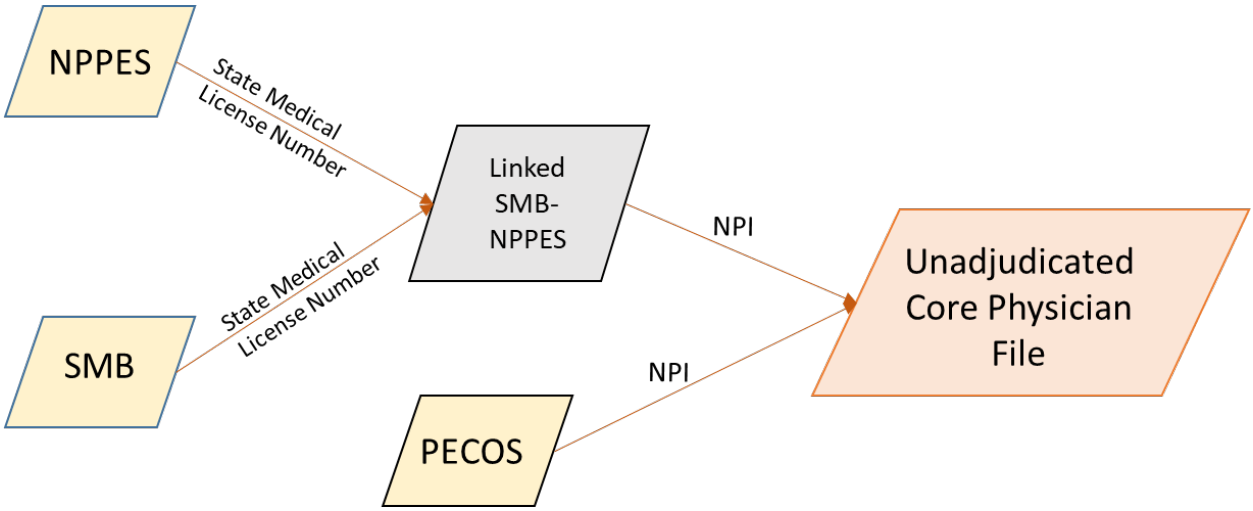
A description of the clean-up routine for address fields is provided below:

1. **Street address:** No changes were made other than setting all data to uppercase, removing leading and trailing blank values, and removing double space characters were applied to the street address fields.
2. **City:** No changes were made other than setting all data to uppercase, removing leading and trailing blank values, and removing double space characters were applied to the city fields.
3. **State:** State fields were reported in a variety of different formats across the various data sources. For example, state can be reported as a SSA, FIPS, or USPS alpha code. Part of data cleaning entailed standardizing the state values in the provider files to their corresponding USPS 2-digit alpha value.

4.2 SMB Linkage With NPPES and PECOS

Linking the SMB data to NPPES and PECOS was the first step in creating the Core physician files. **Exhibit 4.1** illustrates the process for how the files were linked to create the Core physician files.

Exhibit 4.1. Linking SMB, NPPES, and PECOS to Create the Core Physician File



4.2.1 Linking SMB to the NPPES File

The only way PECOS and SMB files can be linked was through NPI since the PECOS file only has NPI and the SMB data file only contains state license numbers. Linking the SMB files to the NPPES file was a necessary first step to pull in NPI information. The merge was separated into several steps: the first was to identify non-organization NPPES providers with a license number issued in the SMB state being processed, merge the SMB and NPPES data files (using license number as the join key), and finally validate the joined records using provider first and last name.

4.2.2 Maximizing Linking Algorithm

To maximize the number of joined pairs, three variants were created for the license numbers in each of the data files (SMB and NPPES):

- License numbers with all punctuation removed (e.g., ABC-0135 would become ABC0135)

- License number with all punctuation and alphabetical letters removed (e.g., ABC-0135 would become 0135)
- License number type converted to numeric type, removing all non-numeric values and leading zeros (e.g., ABC-0135 would become 135).

These variants are crucial since providers inconsistently reported their license numbers in the NPPES file. For example, some providers reported the full license number (punctuation, alphabetical characters, and numeric values) while others only reported the numeric values of their license number. The joined records resulting from each of the matches were appended into a single file, containing all joined SMB and NPPES records.

The final step in the linkage was to remove any record pairing that was deemed a non-match. Match status was assigned by comparing the provider's first and last names on the joined records. The Jaro-Winkler[1][2] string comparator, which assigns a similarity score between 0 (not similar) and 1 (exact match), was used to ascertain the similarity of the name information between the joined records. A manual review of a 5% simple random sample of linked records was used to determine the appropriate cutoff values of the Jaro-Winkler scores for first and last name. Records with a first name score greater than 0.85 (85%) and a last name score greater than 0.85 (85%) or a first name score greater than 0.8 (80%) and a last name score equal to 1 (100%) were deemed matches. Any linked record that did not satisfy either rule, were excluded from the final linked SMB to NPPES file.

4.2.3 Linking PECOS to the merged SMB+NPPES file

Several steps were taken for merging in the PECOS information. These steps included merging in PECOS information to records that matched on NPI, identifying non-organization practitioners in the PECOS file with an enrollment state or practice state equal to the SMB state being merged and appending to merged records, and comparing provider name information to confirm accurate matches.

4.3 Data Assessment and Harmonization

Due to the multiple source files that were used to develop the 3P-RD, all data from those sources were harmonized into a common data structure. This section documents specialty harmonization in detail.

4.3.1 Specialty Harmonization

An iterative process was implemented for harmonizing physician specialty codes and physician specialty descriptions across the three primary provider data sources when developing the 3P-RD Core Physician Research File. Harmonization of physician specialty code is a result of marked differences between data coming from an individual SMB and the CMS, specifically the NPPES and the PECOS. Each of the three data sources use different coding systems for specialty. NPPES lists the provider's

[1] Jaro M. Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. J Am Stat Assoc. 1987 Jan 01;406:414-420.

[2] Winkler W. String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. Proceedings of the Section on Survey Research Methods. American Statistical Association. 1990. 354-9.

taxonomy code, the PECOS data use the CMS physician specialty coding system, and the SMB data is state specific and lacks a uniform coding system. Therefore, it is essential to align the SMB data and NPPES coding structures to PECOS to assure accurate linkage between files and to standardize coding of physician specialty information across states. The Medicare Provider and Supplier Taxonomy Crosswalk dataset (Medicare crosswalk), which includes the Medicare provider taxonomy code and the corresponding Medicare specialty code (if available), was used to harmonize the NPPES data and assist with the SMB harmonization.

The iterative harmonization process followed four specific steps:

1. **Automatic assignment of specialty code.** Programming code to automatically assign a CMS physician specialty code to the SMB data systematically compares words found in the SMB specialty description to the Medicare crosswalk specialty descriptions. If a match between the SMB specialty description and the Medicare crosswalk specialty descriptions is present, the program assigns the corresponding CMS physician specialty code to the SMB data. For example, the SMB physician specialty description OTOLARYNGIC is similar to the Medicare crosswalk specialty Otolaryngology (code = 04) and is therefore applied to the output data.
2. **Review of programmatic assignment.** A manual review of all output generated by the code was completed to ensure that an accurate assignment was made. For records where the programming code did not assign a specialty code or the assigned specialty code was inaccurate, a manual investigation was conducted to identify the correct specialty code. First, a reviewer searched the relevant taxonomy code descriptions in the Medicare crosswalk to determine if the SMB description was a taxonomy code name. If a match existed between taxonomy name and SMB specialty description, the Medicare specialty code from the Medicare crosswalk was used. If a match did not exist, a reviewer conducted an internet search to determine the type of doctor that would match the SMB description. Finally, a reviewer categorized the remaining unassigned or ambiguous SMB records as OTHER.
3. **Review of manual assigned specialty codes by a clinician.** Two clinicians were consulted to review the manually assigned records as a method for validation. For example, SMB data includes the term radiology as a part of many different specialty descriptions (e.g., abdominal radiology, breast radiology, vascular and interventional radiology). The Medicare crosswalk includes both diagnostic radiology (code 30) and interventional radiology (code 94). Based on the clinician's review, Medicare specialty code 94 was assigned to the SMB specialty descriptions that specifies 'interventional radiology.' Otherwise, all other SMB specialties that include radiology were classified as 'diagnostic radiology.' Records that were unknown and could not be assigned were classified as OTHER. Any logical conflicts between the two clinical reviews were resolved through a discussion with the reviewing clinicians.
4. **Adjustment to the initial programming code to incorporate reviewed output.** Using the output generated from the programming code and updates from the manual review and the clinician review, a comprehensive, state-specific 3P-RD specialty code crosswalk was generated. The finalized comprehensive 3P-RD specialty crosswalk was used for the SMB data harmonization, in preparation for the creation of the adjudicated 3P-RD Core physician file for each state.

Primary vs. Secondary Specialty Variables

The 3P-RD Physician files included both primary and secondary specialty fields. The above process was applied for both the primary and secondary specialty variables. As a result of the manual and clinical review conducted by clinician and SME consultants on the project, the programming code was adjusted to accommodate several key analytical decisions for distinguishing primary and secondary specialty from a complex SMB specialty description.

1. **Emergency Medicine:** Frequently the specialty 'emergency medicine' was in combination with other specialties, such as sports medicine, toxicology, or cardiology. Based on the clinical review, 'emergency medicine' was identified as the primary specialty and the remaining specialty description as the secondary specialty. This approach was determined based on how that specialty is practiced within the Emergency Department—the physician is an emergency medicine physician with additional training and expertise.
2. **Pediatric Medicine, Internal Medicine, and Family Practice:** Pediatric medicine, internal medicine and family practice were assigned as secondary specialty while the other SMB specialty description was identified as the primary specialty. A few reasons informed this approach. First, these specialties may be a prerequisite for more specific certification or training. Second, other specialties, especially if the physician was board certified in another specialty, were more likely to be the primary focus for providing patient care. For example, 'pediatric cardiology' resulted in 'cardiology' as the primary specialty and 'pediatric medicine' as the secondary. Similarly, 'internal medicine, medical oncology' resulted in 'medical oncology' as the primary specialty and 'internal medicine' as the secondary specialty.

Specialty Categories

The 3P-RD Physician Practice files aggregate physician specialty information to identify top specialties represented in the practice. In addition, indicator variables are present to identify the practice as having primary care physicians (PCP), medical specialties, surgical specialties, obstetrics-gynecological specialties, psychiatric specialties, hospital-based specialties, and other. Grouping of physician specialties into these seven categories was completed in conjunction with the harmonization process (**Table 4.2**).

Table 4.2. Physician Specialty Categories.

Specialty Category	Specialty Description	
Primary Care Physician	• General practice • Family practice • Internal medicine • Osteopathic manipulative therapy	• Hospice and Palliative Care • Pediatric medicine • Geriatric medicine • Preventive medicine

Specialty Category	Specialty Description	
Medical	• Allergy/immunology • Cardiology • Dermatology • Gastroenterology • Neurology • Cardiac Electrophysiology • Pulmonary disease • Nephrology • Infectious disease • Endocrinology • Rheumatology • Addiction medicine	• Hematology • Hematology/oncology • Medical oncology • Sleep medicine • Interventional cardiology • Advanced heart failure and transplant cardiology • Medical toxicology • Hematopoietic cell transplantation and cellular therapy • Medical Genetics and Genomics • Micrographic Dermatologic Surgery • Adult Congenital Heart Disease
Surgical	• General surgery • Otolaryngology • Neurosurgery • Ophthalmology • Orthopedic surgery • Sports medicine • Plastic and reconstructive surgery • Colorectal surgery	• Thoracic surgery • Urology • Hand surgery • Peripheral vascular disease • Vascular surgery • Cardiac surgery • Surgical oncology
Obstetrics/Gynecology	• Obstetrics/gynecology	• Gynecologist/oncologist
Hospital-based	• Anesthesiology • Interventional Pain Management • Pathology • Physical medicine and rehabilitation • Diagnostic radiology	• Nuclear medicine • Pain Management • Critical care (intensivists) • Radiation oncology • Emergency medicine • Interventional radiology • Hospitalist
Psychiatric	• Psychiatry • Geriatric Psychiatry	• Neuropsychiatry
Other	• Unknown physician specialty	

4.3.2 Specialty Validation

The provider's identifiers in the merged harmonized datafile were used to pull specialty information from the CMS Medicare FFS and T-MSIS files, separately. Due to data quality issues, the T-MSIS Medicaid specialty information was unusable for the following states: CA, CO, MA, MD, MN, MO, NY, and TX (see **Table 4.3** for state-specific data quality issues). The top billed specialty for Medicare and Medicaid (where applicable) was populated based on claim volume. The specialty validation data was used as confirmation and any specialty information from the SMB/PECOS/NPPES harmonization process was not replaced.

Table 4.3. T-MSIS Provider Specialty Data Quality Issues.

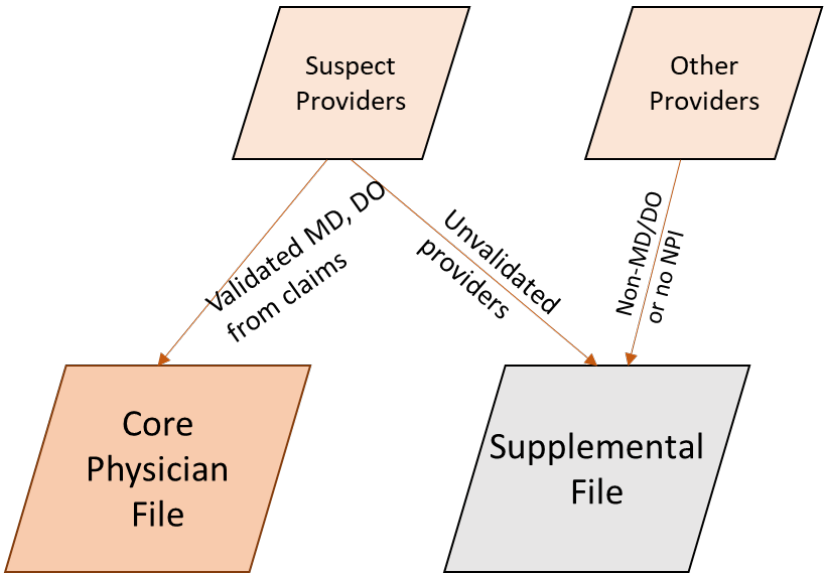
State	Data Quality Category	Description of Specific Issue
CA	Missing Values	98% of providers missing specialty information.
CO	Frequency of specialty codes	45% of providers with a nurse practitioner specialty code (50). 38% of providers with a general surgery specialty code (01).
MA	Frequency of specialty codes	31% of providers with a general surgery specialty code (01).
MD	Frequency of specialty codes	64% of providers with an “All other suppliers” specialty code (87).
MO	Missing Values	56% of providers missing specialty information.
NY	Missing Values Frequency of specialty codes	22% of providers missing specialty information. 73% of providers with a general surgery specialty code (01).
TX	Missing Values	26% of providers missing specialty information.

4.3.3 Provider Adjudication

During the initial processing of the Core physician files, providers were assigned to one of three files for further processing: core physicians, suspect physicians, and all other providers. Providers placed in the Core physician file were providers with credentialling data indicating either MD or DO. Providers moved to the suspect physician file included those with no credentialling information, providers where the specialty did not align with credential (i.e., Nurse Practitioner specialty but has MD credential), or if credential indicated non-MD/DO but specialty is typical of physician. The remainder were placed in the other provider file.

Validated specialty information was used to determine the final assignment of providers in the suspect file to either the final Core physician file or the supplemental file (**Exhibit 4.2**). Providers who can be determined from Medicare FFS and T-MSIS billing were moved to the Core physician file. Others are combined with the “all other” providers file to create a supplemental provider file for internal AHRQ assessment. Providers may appear more than once in the Core physician file as a result of the specialty validation. For example, a physician may be in the SMB data multiple times, with one record per specialty board certifications. Disparate data sources, such as the SMB and claims data, may have primary and secondary specialty information reversed. The validated specialty information was used to de-duplicate the physician file by selecting the record with the most specialized specialty in the primary spot. If no record can be logically selected, a record was selected at random. The supplemental file contains any provider without an NPI (i.e., not linked with NPPES/PECOS or providers with NPI’s that cannot be verified as MD/DO).

Exhibit 4.2. Provider Adjudication.



4.4 Processing APCD and CMS Claims Data

Initial quality checks were conducted on the claims files which included verifying imported files contained the same record counts and file size and those listed in control totals provided by the data vendor were conducted. Also, a visual inspection of a random sample of records was conducted to ensure data imported properly. Finally, discussions with the data vendor were conducted to understand caveats to key variables used for building the 3P-RD, such as rendering NPI, billing NPI, billing taxpayer identification number (TIN), ZIP Code, and submitter provider identifiers.

5. State-Specific Builds

There was some variation in the process for how state physician and physician practice files were built for the 3P-RD. An overview of the process is provided below, followed by descriptions of state-specific programming needs based on the uniqueness of each source data.

5.1 Overview

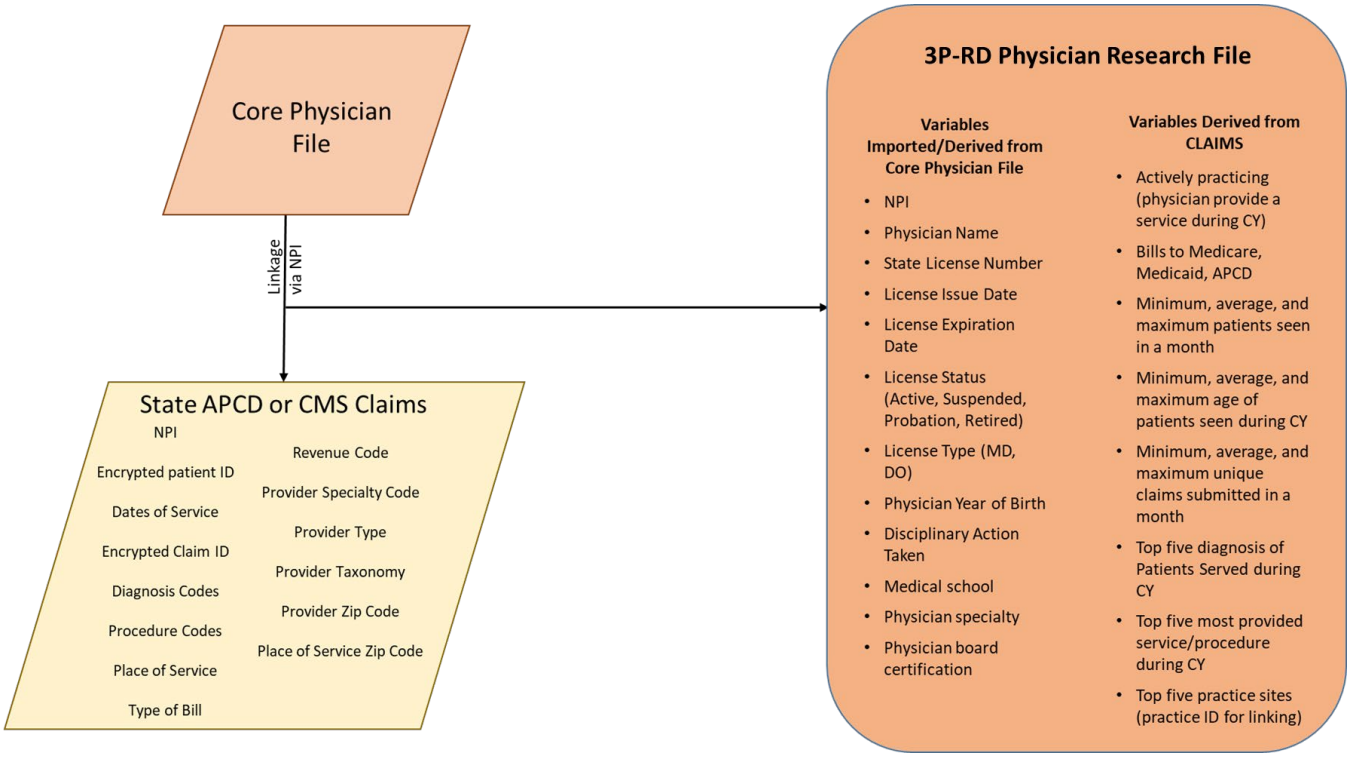
The 3P-RD PUFs include data elements from public data sources, as well as variables created from any transformation and aggregation of data from the claims data sources. Primarily, linking occurs between the claims data to the corresponding state Core physician file (the linked SM-NPPES-PECOS file described above – **Exhibit 4.1**). For states with APCD data elements, the CMS Medicare and Medicaid claims data was used for gaps within the APCD data, as necessary. The linkage between the claims data and Core physician file was done through the NPI.

5.1.1 Physician File

Building the 3P-RD Physician File included aggregating the claims data for each rendering NPI, linking the Core physician file, and performing initial quality checks on the file to verify that the aggregated physician characteristics are reasonable (**Exhibit 5.1**). The origin and processing of the physician characteristics depended on what data was available for each state. Physician characteristics include number of claims billed, payors billed, patient panel demographics, top procedures billed, and top diagnoses of patient panel. For Medicare FFS, the CMS files were used for all states. APCD states include data for commercial and Medicaid populations (except Maryland); for non-APCD states, T-MSIS data were used to develop Medicaid variables. Regardless of the claims data source, merging between the Core physician file and the claims data was done though the NPI as the unique merge key.

The same method was mostly applied for creating the physician-level claims variables across all states; however, specific coding and data transformation techniques were applied to individual states based on the source of the claims data. Different data structures and contents between state APCD files and CMS Medicare FFS and T-MSIS data required adjustments to processing. Once the physician variables were generated from the claims data, the data was merged with the final Core physician to create each state's 3P-RD Physician File. Finally, summary statistics were generated, and a visual inspection was performed to ensure data was reasonable (e.g., maximum number of claims submitted in one month).

Exhibit 5.1. Linking Core Physician File to Claims Data to Create 3P-RD Physician Files



5.1.2 Practice File

Building the 3P-RD Physician Practice File first began with identifying practice sites. A practice site was understood as being the location where the rendering physician provided services to the patient panel. The 3P-RD practice was defined as the combination of taxpayer identification number, organizational NPI, and ZIP Code. The purpose of using the combination of TIN, ORG NPI and ZIP Code to define a physician practice is to identify all the locations a physician may be practicing at, allowing for geographic analysis of services provided. Depending on the health system, several options existed when combining TIN and ORG NPI.

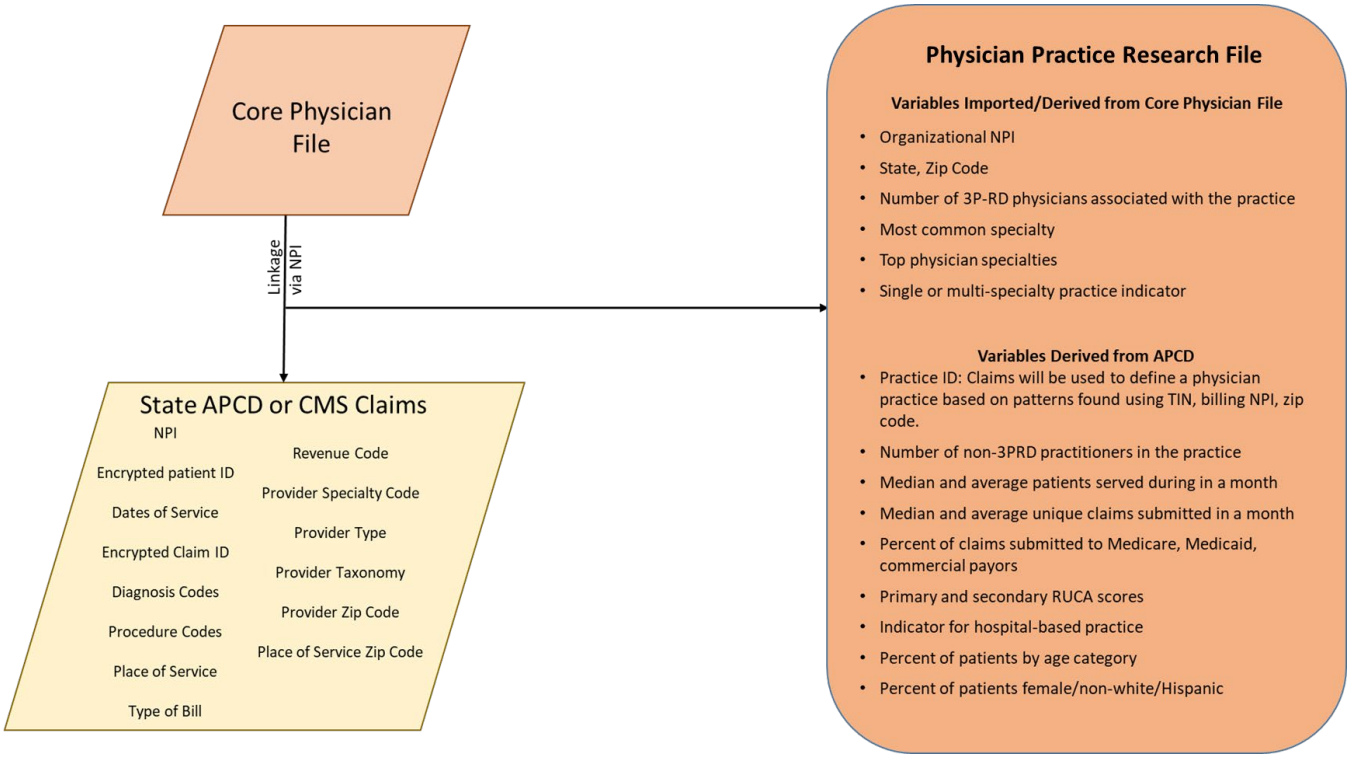
- Multiple ORG NPIs fall under one TIN
- Multiple TINs fall under one ORG NPI
- Multiple ZIP Codes are associated with both options above.

A practice ID was assigned to each combination of TIN-ORG NPI-ZIP Code identified within the claims data. After practice ID was assigned, the Core physician file was merged with claims files, and the data was aggregated to create the physician practice characteristics. Like the physician processing, the same method was applied for aggregating claims data across all states, however, specific coding and data transformation techniques were applied to individual states based on the source of the claims data. Processing adjustments were needed due to differences in data structures and TIN or ORG NPI information available between state APCD files and CMS Medicare FFS and T-MSIS data. After creating each state's 3P-RD Physician Practice File, summary statistics were generated, and a visual

inspection was performed to ensure data was reasonable (e.g., percent of patient panel female, number of 3P-RD physicians associated with the practice).

Using a combination of TIN-ORG NPI-service ZIP Code resulted in associating many practices with only one physician. While solo practices do exist, the definition resulted in a high frequency of solo practices. Upon review, some of the solo practices were a consequence of multiple ZIP Codes being used for a service location (e.g., practices located in cities with more than one ZIP Code). To identify the top practices associated with physicians, the number of patients served in a month and number of claims submitted in a month were analyzed for each 3P-RD physician. The top five (5) practice IDs for each 3P-RD physician were identified, and the practice IDs were included into the 3P-RD Physician File.

Exhibit 5.2. Linking Core Physician File to Claims Data to Create the 3P-RD Physician Practice Files



By linking via NPI, physicians were identified within each state having an active license that have actively provided services to patients during the reference year. Data elements from other data sources appear in the new linked file, including provider degree, disciplinary action, and other provider identifiers and demographics. The linked dataset also included physician and physician practice characteristic variables created from each state's specific claims, such as average number of patients seen in a month, specialty most associated with billing, primary billing location, and identifying relationships between physician practice patterns for the creation of unique physician practice sites (**Exhibit 5.2**).

5.2 Colorado Build

The Center for Improving Value in Health Care (CIVHC) provided a total of 30 CO APCD files. Files included four members and eligibility files, five provider files, and four medical claims files. In addition, 17 reference files were also delivered. Discussions with CIVHC were held to understand how the files related to each other to ensure that the correct member and provider identifiers were being used during processing. Specifically, the CO APCD data included multiple identifiers for each member and provider within the data (**Table 5.1**).

Table 5.1. CO APCD Identifiers

Variable	Description	Notes for Use
Member ID	• Payor-specific unique identifier	• A member may have more than one member ID. 0.04% missing in the 2019 data.
Member Composite ID	• CO APCD unique identifier	• A member should only have one member composite ID, it has a one–to-many relationships with the member ID. None missing in the 2019 data.
Service Provider ID	• Payor-specific unique provider identifier	• An individual provider can have more than one provider ID. 0.12% missing in the 2019 data.
Service Provider Composite ID	• NPPES-specific unique provider identifier	• A provider should only have one provider composite ID, it has a one-to-many relationship with the provider ID. 1.87% missing in the 2019 data.

The purpose for using composite identifiers was to try and deduplicate the non-composite identifiers. There was a one-to-many relationship of one composite ID with many non-composite IDs. When merging non-composite member/provider variables with the claims file, payor IDs needed to be included.

In addition to the identifiers, the processing for CO was adjusted to use the line of business code to identify the payor of the claims as Medicare Advantage, Commercial, and Medicaid. Also, a capitation flag variable was used to further differentiate between Medicaid Fee-for-Service and Medicaid Managed Care.

Prior to the processing of physician and physician practice variables, a 2019 CO APCD claim file was created by taking all claims with either a start or end date year in 2019. Any claim included in the 2019 file with a start or end date year not in 2019 was identified as a spillover claim. After generating the 2019 CO APCD claim file, the following information was added: provider information, member information, and header information which included the line of business code, the principal diagnosis code (ICD 10), and the capitation flag.

5.2.1 Physician File

Beyond the processing adjustments described above, no other changes were implemented to generate the 3P-RD CO physician file.

5.2.2 Practice File

The CO practice file was created using the CO APCD data. Since the data contained both commercial, Medicare Advantage and Medicaid data, it was more comprehensive for the identification of practice sites within CO rather than only using the Medicare FFS files (as done in non-APCD states). Once the 2019 CO APCD claim file was generated, each practice was identified by filtering to claims with non-null billing and servicing composite provider IDs and billing provider IDs. The provider ID was associated with the provider's information as it pertained to the claim and the provider composite ID is associated with the provider's information of the provider overall. Multiple provider IDs can be associated with provider composite IDs. The billing provider ID was used to identify the TIN for the billing provider; the billing composite provider ID was used to identify the organizational NPI; and the servicing provider composite ID was used to identify the provider's zip code, city, state, and individual NPI. The TIN, organizational NPI, and zip code were concatenated and assigned a randomly generated number. This became the physician practice ID. After the physician practice IDs were generated, an analytic file was created with the practice ID attached to each claim line.

5.3 Arkansas Build

The AR APCD files did not have a specified unique identifier that could be used to link across files. Instead, researchers need to create a unique key to identify unique individuals and link across the member eligibility and the medical files. Using AR documentation[3] in conjunction with discussions with the Arkansas Center for Health Improvement (ACHI), which oversees the AR APCD, the necessary identification variables needed to process and link the AR files for the 3P-RD were generated (**Table 5.2**).

Several changes from the standard processing overview were made for the processing of AR data. First, two member IDs are created. The SE_ID is the unique member ID that is used to link the member file with their corresponding claims; the SE_ID is the combination of the Entity ID and the enrollee ID and is specific to the payor (**Table 5.1**). The study_ID is a 'global' member ID that typically covers multiple SE_ID (i.e., study_ID is an ID that deduplicates the SE_ID; it is a one-to-many relationship of one study_ID with many SE_ID). When merging the SE_ID and provider variables to the claims file, the submitter ID (a unique number for each payor submitting claims to AR APCD) needs to be included in the processing. The submitter_ID is a categorical variable that indicates whether the submitting entity is a commercial payor, Medicaid, or Medicare. Second, the claim submitter category was used to separate out the Medicaid and Commercial claims. Third, Medicaid FFS and MCO claims cannot be further differentiated due to the lack of information on Medicaid capitated services. Lastly, the original variable names provided by AR were renamed following a more user-friendly naming convention for data processing. This eliminated the need to constantly reference the data dictionary to identify the variables needed to complete the analysis.)

As with the CO processing, a 2019 AR APCD claim file was created prior to generating the physician variables by selecting claims that started in/prior to and ended in/after 2019. Provider information was

[3] 2022 Data Attribute Supplement for Data Requesters. Retrieved from: https://www.arkansasapcd.net/Docs/282/

then merged into the claim file by NPI, submitter ID and submitter category, which was used to identify Medicaid and commercial claims.

Table 5.2. AR APCD Identifiers

Variable	Description	Creation Logic
SE_ID	• A unique member ID that is used to connect members to claims data within payers.	• SE_ID is created using both Entity ID* and Enrollee ID** • *Entity ID is a code that represents a payer within the claims data. • **Enrollee ID is a Member ID. • % missing = 0.0%
Study_ID	• A unique member ID that is used to connect members across payers.	• Study_ID is created using both APCD Unique ID* and Member Gender • *APCD Unique ID is a variable that is hashed securely to indicate a member across sources and data file types. • % missing =0.0%
PROV_National_Prvdr_ID	• A unique identification number for covered healthcare providers.	• No processing required to create variable; available in AR APCD) • % missing =1.7%
ELG_Submitter	• A code representing the entity submitting payments.	• No processing required to create variable; available in AR APCD) • % missing =0.0%
ELG_Submitter_Category	• A code representing payer type.	• No processing required to create variable; available in AR APCD) • % missing =0.0%

5.3.1 Physician File

No other adjustments to processing, beyond those described above, were made to generate the 3P-RD AR physician file.

5.3.2 Practice File

The AR practice file was created using the AR APCD data. Since the data contains both commercial and Medicaid data, it was more comprehensive for the identification of practice sites within AR rather than only using the Medicare FFS files (as done in non-APCD states). The AR Practice IDs were created following the physician practice definition as described above in the overview using the billing TIN, which was encrypted, servicing zip code and billing NPI.

5.4 Maryland Build

The MD APCD data was provided by the Maryland Health Care Commission (MHCC). The team collaborated with MHCC to import and understand the MD APCD data files. In total, there were 4 files per year for the MD APCD. Files included a member eligibility file, a prescription and pharmacy file, an

institution level file, and a professional level claims file. In addition, the team received user manuals which helped utilize the data. All the MD APCD analytic files were linkable through the unique patient identifier – PIDBDGP – and included several provider identifiers (**Table 5.3**).

Table 5.3. MD APCD Identifiers

Variable	Description	Notes for Use
PIDBDGP	• Unique patient identifier	• Available across all analytic files and used to link all analytic files together
PROVID	• Provider ID	• Servicing provider identifier
NP_SP_NPI_P	• Provider NPI	• Servicing provider identifier

Prior to physician variable processing, a 2019 MD APCD claim file was created by isolating the 2019 claims from the multiple years of data received. An initial assessment of the presence of spillover claims identified no spillover claims with a start or end date out of the year 2019 present in the data. After generation of the 2019 MD APCD claim file for data processing, patient race and ethnicity information only present on the member eligibility file was added.

5.4.1 Physician File

The MD APCD data did not contain MD Medicaid data. The processing for the MD physician file combined commercial and Medicare Advantage data from the MD APCD, Medicaid data from CMS T-MSIS, and Medicare FFS data from CMS Medicare data.

5.4.2 Practice File

The MD practice file was created using the MD APCD data. Since the data contains both commercial, Medicare Advantage and Medicare supplemental data, it was more comprehensive for the identification of practice sites within MD rather than only using the Medicare FFS files (as done in non-APCD states).

5.5 Washington Build

The team worked with WA Health Care Authority (HCA) and Onpoint Health Data to receive, import, and understand the WA APCD data files. A total of four (4) medical claims files, two (2) provider files, and two (2) files for members and eligibility were received. In addition, a list of reference files was received. Discussions with Onpoint Health Data and referencing the WA-APCD Data Dictionary facilitated understanding of how data files related to each other to make the most efficient linkages across files. Specifically, assuring the correct identifiers and linkage variables were used during processing since some files contain more than one ID variable (**Table 5.4**).

Table 5.4. WA APCD Identifiers

Variable	Description	Notes for use
internal_provider_id (rendering_internal_provider_id)	This field contains an ID that represents a unique provider.	This field is used to aggregate all records associated with a provider.
medical_claim_header_id	This field contains an ID that identifies a unique claim.	This field is used to link the medical claims header file to the medical crosswalk file.
medical_claim_service_line_id	This field contains an ID that identifies a unique service line of a submitted claim record.	This field is used to link the medical claims line file to the medical crosswalk file.
internal_member_id	This field contains an ID that represents a unique member.	This field is used to aggregate all records associated with a member.

A 2019 WA APCD claim file was created prior to generating the physician variables by selecting claims that started in/prior to and ended in/after 2019. Provider information and member eligibility information were then merged into claim file on rendering internal provider ID and internal member ID, respectively.

The product code in the claim file was used to identify the types of payers as Medicaid, Commercial and Medicare Advantage. Medicaid FFS and MCO were further differentiated using a capitation flag. Race was recoded from the WA APCD categories into the 3P-RD White, non-White, Hispanic and non-Hispanic groups.

5.5.1 Physician File

No other adjustments to processing, beyond those described above, were made to generate the 3P-RD WA physician file.

5.5.2 Practice File

The WA APCD data did not contain TIN, encrypted or unencrypted. To identify practice sites for the WA 3P-RD, the PECOS Associate Control ID (PAC-ID) was used. PACID has a nearly one-to-one correlation to TIN. PACID was added to the WA APCD files by linking the PECOS file to the WA APCD medical claims file using the Organizational NPI. The relationship between the PACID and WA organizational NPIs was assessed to determine if PACID could be used in lieu of TIN for the creation of the practice ID.

- 330,711 distinct organizational NPIs from the WA provider file.
- 121,671 distinct organizational NPIs that do not have a PACID
- 151,524 distinct PACIDs and 209,040 distinct organizational NPIs were linked between the PECOS file and the WA provider file
- 7,807 distinct PACIDs that are associated with 2 or more NPIs
  - Since multiple organizational NPIs can be associated with a TIN, this many-to-one relationship is not an issue for processing.

- 436 distinct organizational NPIs that are associated with 2 or more PAC IDs
  - One PACID was selected randomly as being the associated PACID to each organizational NPI. This is <1% of all organizational NPIs in the WA provider file.

5.6 Non-APCD States Build

APCD data was not available for the remaining nine 3P-RD states which include Arizona, California, Florida, Massachusetts, Maryland, Minnesota, Missouri, Montana, New York, and Texas. Though some states such as Florida, Massachusetts, and Minnesota have APCD databases, that data was not available to use for this set of 3P-RD files. For these states, CMS T-MSIS and Medicare FFS claims were used to develop their state 3P-RD files.

Since the conceptual model assumes access to and use of the APCD data, several changes were required to the process to accommodate data availability and structure of CMS data. Several process changes were made to T-MSIS and Medicare FFS process for the physician file. The primary change to data processing for T-MSIS and Medicare FFS data is the environment – all data processing occurred on CMS' Virtual Research Data Center, with the final Medicaid and Medicare physician and practice files being downloaded and then merged into Core physician files. Also, the methodology to identify a physician practice needed to change to accommodate Medicare FFS structure and variable inclusion.

5.6.1 T-MSIS Physician Processing

Upon review of the files, it was determined that the TAF Other Services File and the TAF Demographic and Eligibility File were needed from the TMSIS files. Unlike the APCD data, which provided data in an annual format, the TAF Other Services File is stored at the month-year level. To streamline and align processing to methods used for APCD data, all twelve months of data for 2019 were appended and the files were restricted to NPIs identified in the Core physician file. Only variables necessary for processing data were kept as a way to mitigate against long processing times. Variables included for processing are claim ID, beneficiary ID, the MSIS ID, state code, provider NPI, service dates, bill type code, place of service code, diagnosis codes, claim type code, type of service code, and procedure code. Since the TAF files have two beneficiary (member) IDs – Beneficiary ID and MSIS ID - the most appropriate ID was determined based on CMS guidance. If the beneficiary ID was present, then the member ID was defaulted to the Beneficiary ID. Otherwise, the member ID was created as the concatenation of MSIS ID and state. Demographic information for T-MSIS processing was retrieved from the TAF Demographic and Eligibility File and merged onto the claims data. Data for beneficiaries with claims but without eligibility during the calendar year were dropped from the file. Once generated, we followed the same processing as done for the APCD states.

After generating the T-MSIS Medicaid physician file, the data was submitted to CMS for review and exported from the VRDC. After export, the T-MSIS Medicaid physician file was linked to the Core physician file to create the 3P-RD Physician File for non-APCD states.

5.6.2 Medicare FFS Physician Processing

All state 3P-RD Physician Files include Medicare FFS variables. The processing of the Medicare FFS data focuses on the Medicare Carrier file, as it contains claims specific to individual physicians. Processing logic for Medicare FFS was aligned to APCD and T-MSIS processing. The logic employed to create the variables is consistent across all data files. Similar to T-MSIS data structure, Medicare FFS files are kept in month-year format. Due to the large size of Medicare FFS, processing was completed at the month level and then data was aggregated up to the year level. For example, the minimum, average and maximum patients seen in a month was determined by calculating the number of patients in each month, concatenating all monthly files together, and then determining the minimum, maximum and average from the 12 monthly variables. By constructing the code in a monthly processing manner, we were able to minimize processing time.

After generating the Medicare FFS physician file, the data was submitted to CMS for review and exported from the VRDC. After export, the state Medicare FFS physician files were linked to the Core physician file to create the 3P-RD Physician File for non-APCD states and enhance the 3P-RD Physician File for APCD states.

5.6.3 Medicare FFS Physician Practice Processing

For states without APCD data, Medicare FFS data was used to identify physician practices. The Medicare FFS Carrier file was used for processing to identify how and where physicians provided services to Medicare beneficiaries. Overall, the processing for practice files using the Medicare data aligned with logic used for APCD data. However, as in the physician file processing, processing was conducted for each month to mitigate processing times. Only variables required for the identification of practice ID and creation of practice characteristics were kept, such as tax identification number, provider zip code, organizational NPI, claim ID, line number, beneficiary ID, provider state code, provider NPI, place of service code. Once each month was processed, all months were concatenated to make an annual file. Records that did not have the appropriate format for TIN, zip codes, organizational NPIs, and physician NPIs were dropped from processing.

After generating the Medicare FFS physician file, the data was submitted to CMS for review and exported from the VRDC. After export, the state Medicare FFS physician files were linked to the Core physician file to create the 3P-RD Physician File for non-APCD states and enhance the 3P-RD Physician File for APCD states.

5.7 Suppression Policy

Each state file – physician and physician practice – required a level of suppression as outlined in the DUA governing the data used for the state's data. Data vendors required cell suppression to maintain privacy protections of health information. However, data suppression rules were not uniform across all data sources. This resulted in differences in how data could appear and in turn be released on the 3P-RD. **Table 5.5** describes how the data suppression rules were applied across the 3P-RD.

Table 5.5. Cell Suppression Logic

Data Source	Suppression Logic	Exceptions	Files
AR APCD	• No cell suppression. • The files contain provider level data without total counts. Therefore, cell suppression is not required.		• AR Medicaid, Commercial and Medicare Advantage data in the APCD.
CO APCD	• No cell suppression. • The files contain provider level data without total counts. Therefore, cell suppression is not required.		• CO Medicaid, Commercial and Medicare Advantage data in the APCD.
MD APCD	• No cell suppression. • DUA is held with AHRQ and therefore no cell suppression was required for delivery.		• MD Commercial and Medicare Advantage data in the APCD.
WA APCD	Suppress cells less than 11. Physician: • All minimum, maximum, and average variables. Practice: • All minimum, maximum, and average variables.	• Zero (0) • Physician Counts • Percentages • Age	• WA Medicaid, Commercial and Medicare Advantage data in the APCD.
CMS: Medicare FFS & T-MSIS data	Suppress all counts, minimum, median, and maximum values: • where total claims for the provider is less than 100. • values less than 11 regardless of total claim count.		• All states: Medicare FFS fields • Non-APCD states and Maryland: Medicaid fields

6. Public Release Files

PUF files were created for the 3P-RD. The first, the Directory PUF, is a directory of physicians and practices within the state and contains information derived from publicly available data sources, such as SMB, NPPES, and PECOS. For the geographic PUF files, the 3P-RD physician and practice files are be aggregated to the ZIP Code level (**Table 6.1**). Through the claims data, aggregated provider information related to physician characteristics and patient population statistics were derived. Adherence to the various requirements outlined in multiple data use agreements across various state and federal agencies were maintained by aggregating physician and practice characteristics.

Table 6.1. 3P-RD Physician and Practice PUF

	Physician	Practice
Directory PUF	• NPI • Physician Name • Top five practice sites (practice ID for linking) • State License Number • License Issue Date • License Expiration Date • License Status (Active, Suspended, Probation, Retired) • License Type (MD, DO) • Physician Year of Birth • Disciplinary Action Taken • Medical school • Physician specialty • Physician board certification	• Organizational NPI • State, Zip Code • Number of 3P-RD physicians associated with the practice • Most common specialty • Top physician specialties • Single or multi-specialty practice indicator • Variables Derived from APCD • Practice ID: Claims will be used to define a physician practice based on patterns found using TIN, billing NPI, zip code. • Number of non-3PRD practitioners in the practice
Geographic PUF	• State Zip Code • Count of unique physicians • Percent of physicians with an active state license • Count of physicians with MD • Count of physicians with DO • Total number of physicians that are actively practicing • Percent of physicians female/non-White/Hispanic • Top three most common medical schools • Top five most common physician specialties • Minimum/maximum number of claims per month across all 3P-RD physicians • Minimum/maximum age of patients across all 3P-RD physicians	• State Zip Code • Count of unique practice IDs • Average count of 3P-RD physicians associated with a practice in the ZIP code • Top three most common medical schools • Top five most common physician specialties • Percent of practices in zip code with OB/GYN specialty • Percent of practices in zip code with psychiatric specialty • Percent of practices in zip code with PCP focus • Percent of practices in zip code with medical specialties • Percent of practices in zip code with surgical specialties