



## Data Content Areas Descriptions

## Table of Contents

|                                                      |    |
|------------------------------------------------------|----|
| Table of Contents .....                              | 2  |
| Overview .....                                       | 3  |
| Census .....                                         | 4  |
| Death (Death and Cause of Death) .....               | 6  |
| Laboratory Result.....                               | 7  |
| Patient Information (Enrollment) .....               | 8  |
| Patient Information (Demographics) .....             | 9  |
| Patient Information (Languages) .....                | 10 |
| Patient Reported Outcomes (PRO): .....               | 11 |
| Pharmacy (Pharmacy and EverNDC): .....               | 12 |
| Social History.....                                  | 14 |
| Tumor.....                                           | 15 |
| Utilization (Encounter, Diagnosis, Procedures) ..... | 17 |
| Utilization (Facilities).....                        | 19 |
| Utilization (Providers) .....                        | 20 |
| Vital Signs .....                                    | 22 |

## Overview

The Virtual Data Warehouse (VDW) tables are grouped into different content areas. Each content area has 1 or more tables that are linked together to give the complete information. For example, the Utilization content area has several tables that provide data on the individual patient encounters, including data on any diagnoses or procedures that occurred at those encounters.

This document provides a description of each content area, ways the data can be used, the key variables and important information you should know when using these data.

If you have any questions, please contact your organization's Site Data Manager or any of the HCSRN VOC members.

## Census

### Description:

The Census subject area geocodes consumer address information to geocodes and provides access to standardized area level sociodemographic information from the U.S. Census.

- CENSUS\_LOCATION: Contains one record per MRN location period **and** GEOCODE\_BOUNDARY\_YEAR.
- CENSUS\_DEMOGRAPHICS\_ACS: Contains one record per 5-year release of the ACS.
- CENSUS\_DEMOGRAPHICS\_DECENNIAL: Contains one record per Decennial release. Questions revolve around race, ethnicity, and housing stock.

### Uses:

Examples of how Census data can be useful to researchers:

- Population Level Estimates: Social Determinants of Health (SDoH), specifically socioeconomic status (SES) (e.g., "We estimate that ~15% of the population of interest has income below poverty.").
- Context effects: living in a neighborhood affects your geographic access or exposure to something (food desert, play space). (e.g., "Areas with poverty rates above 20% are 1.5x more likely to experience Y.")
- Composition effects: living with similar/dissimilar people affects behavior (social capital, "community").

### Key Variables:

- LOC\_START
- LOC\_END
- GEOCODE\_BOUNDARY\_YEAR – FK to ACS and Decennial
- GEOCODE – FK to ACS and Decennial
- CENSUS\_YEAR – Not a key, but must be specified
- MATCH\_STRENGTH – dropping zip code matches from analysis is common practice

### Important to know:

- Census Location Update schedules differ by site – Sites have different update schedules for geocoding. While some sites might choose to perform just-in-time geocoding, others might update the file on an annual basis.
- GEOCODE\_BOUNDARY\_YEAR – The year the census boundaries were drawn. Boundaries are redrawn at every Decennial Census. We would expect to see 2010, and 2020 boundaries available for VDW sites. The shapes formed by the boundaries are also referred to as polygons. Use this in your joins between CENSUS\_LOCATION and other tables. If specified ACS CENSUS\_YEAR >=2020 with GEOCODE\_BOUNDARY\_YEAR=2010, the result returned will be missing many records.
- GEOLEVEL – Probably best to keep tract only matches. Matches at zip centroids will present tracts at the zip centroid. Although you could use values from the tract, they tend to differ significantly from the entire zip code. Instead, we suggest using HUD [crosswalks](#) to weight the

- CENSUS\_YEAR
  - ACS – This represents the release year of the ACS file. CENSUS\_YEAR=2012 means that the statistics were generated from each month between 2008 and 2012 (5 years). Choose the CENSUS\_YEAR that aligns most with your study. Studies that span longer periods of time might require different solutions or further explanation around the analysis.
  - DECENNIAL – This represents the release year and the GEOCODE\_BOUNDARY\_YEAR in the Decennial file.
- GEOCODE (n.) (syn. GEOID) - The unique identifier of a geography or polygon. VDW keeps geocodes at the tract level. Hierarchical strings that have the following structure:
  1. State (first 2 characters) – permanent boundaries
  2. County (next 3 characters) – permanent boundaries
  3. Tract (next 6) - Census Tracts are small, relatively permanent statistical subdivisions of a county or statistically equivalent entity that can be updated by local participants prior to each decennial census. Tracts generally have a population between 1,200 and 8,000 people, with an optimum size of 4,000 people. A census tract usually covers a contiguous area; however, the spatial size of census tracts varies widely depending on the density of settlement.

## Death (Death and Cause of Death)

### Description:

The DEATH table contains one record per person in the DEMOGRAPHICS table for whom there is some belief that the person may be deceased. The CAUSE OF DEATH table contains one record per person per known cause of death, and links to the DEATH table by person identifier. When sources of death provide conflicting information, site data managers make local determinations as to which source to use, collating information when possible, and reflecting their confidence in the observation overall using the CONFIDENCE variable.

### Uses:

Examples of how Death data may be useful to researchers:

- Studies of rates of death linked to healthcare utilization data
- Examining longitudinal factors for mortality
- Inclusion or exclusion from a population based on vital status
- Linkage to Cause of Death data
- Outcomes analysis for study cohorts

### Key Variables:

- Person ID
- Date of death
- Confidence
- Source(s) of death data
- Cause of Death
- Type of Cause of death

### Important to know:

- Sites may have additional death data available that requires DUAs from their state(s) for access. Contact the Site Data Manager for information.
- The table has an indicator for whether the death data is imputed, and if so, how it was imputed. An example of when this might occur is if a site sources a death solely from a claim received indicating a discharge status of “expired”. The death date may be imputed from the discharge date of the claim.
- The list of sources of death information is in descending order of reliability indicating all sources that report this person’s death. For example: If State, tumor registry, and a local source all indicate a person’s death and the state is considered most reliable and the local source the least reliable, then the list of sources would be “STO”.
- The presence of a record in the death data does not confirm that an individual is deceased, nor does it mean there will be no other data for the individual beyond the death date. Charts may indicate phone calls or clinical documentation that may show in the utilization table, and enrollment may continue beyond the death date (or month end), depending on an employer’s rules for continuing benefits for dependents. A combination of the CONFIDENCE variable and other elements within the VDW may need to be used to truly confirm death, particularly in the case of “Poor” or “Fair” confidence.

- The types of Causes of Death include Immediate/Primary, Underlying, Contributory, and Other.

## Laboratory Result

### Description:

The Laboratory Result table holds results, not orders, of laboratory tests performed on patients. If a test is not resulted for whatever reason (e.g., specimen not sufficient, patient did not show), then that test should not appear in the table. Tests are generally resulted by a site's own laboratory, but results can be 'hand-entered' when 'walked in' by the patient from outside sources. The table includes most chemistries desired by projects, some hematology and coagulation, opioid urine toxicity, hepatitis C and HIV microbiology, influenza, and COVID-19 results. Most tests included in the lab table are determined by the HCSRN Lab workgroup. Sites may include additional tests. A "best practice" is to prefix any additional test-type with the site's acronym.

### Uses:

Examples of how Lab Result data can be useful to researchers:

- Identifying cohorts with lab result values above or below certain study determined values.
- Assess the existence or progression of disease.
- Determine opioid exposure.

### Key Variables:

- Person ID
- Type of lab test
- Lab test result
- Result unit
- Normal result range
- Abnormal indicator flag
- Dates test was ordered, gathered, resulted
- Specimen source
- LOINC (used to pull results into the VDW, may offer insight for investigators familiar with them)

### Important to know:

- Always use the result unit with any quantitative value. Test results with different units can be held in the same Test-Type.
- Although standardization of results for qualitative values is encouraged this work takes time and is continuous at all sites. Any new data pull of lab data into the VDW may bring a qualitative result value that was not seen earlier. Projects may need to perform standardization of qualitative result values.
- Not all types of tests have been added at every site. Ask your research partners questions about data availability early in the planning stage of a project. Adding a lab test is not easy nor is it quick.
- Don't assume that data held by other research partners looks like the data at your site. Perform basic comparison QA between sites.
- Results should be pulled from the VDW for a project by Test-Type, not by LOINC.

## Patient Information (Enrollment)

### Description:

The Enrollment table documents periods over which the people in the table were associated with the implementing site, such that we should expect to see at least some of their encounters, pharmacy fills, etc. in the other VDW file implementations. That person/site association can take a few different forms:

- For sites that have an insurance function (e.g., HMOs, Insurers) the person could be covered by the implementing site.
- For sites that have a patient-care function (e.g., HMOs, care providers) the person could be the site's patient.
- For sites that have both a patient-care function *and* high market dominance in their geographic area, the person may be fairly guaranteed to get any medical care from the implementing site.

### Uses:

The primary use for Enrollment data is to *define populations* for research projects—generally populations *at risk* for various exposures or outcomes. This is often necessary to calculate rates of illnesses, say, or to marshal sets of people who have *not* had a particular exposure, to serve as control subjects.

### Key Variables:

- Person ID.
- Start and end dates of the association
- Primary type of the association—insured, patient or geographic
- Where known, expected completeness of data capture
- Provider and clinic primarily responsible for the person's care

### Important to know:

- Because not every HCSR member site has an insurance function, at some sites the enrollment data is inferred/constructed from patient contact data. This is of course an inexact science.
- Regardless of the association type, it is possible that capture of particular types of data (e.g., encounters, lab tests, or other data in the rest of the VDW), may be known to be incomplete. Where that is known, the implementing site's expectation is documented in the capture fields.

## Patient Information (Demographics)

### Description:

The Demographics table documents static and slowly changing attributes of the people found in the implementing site's VDW files (e.g., date of birth, sex at birth, and race/ethnicity). There is one record per person, and the variables document the last-known information related to each person.

### Uses:

Demographics serves as a look-up for any and all people ever associated with the implementing site over all time. It is useful any time you need to know e.g., the age of a person at the time of a given event, or their current value of sex, race or ethnicity.

### Key Variables:

- Person ID
- Date of birth
- Sex (assigned at birth)
- Gender Identity

### Important to know:

Not everyone in the Demographics table:

- Has a current association with (e.g., is a current enrollee of) the implementing site.
- Is currently alive.
- Has complete data for all variables (e.g. they may not have known date of birth, known sex, etc).

## Patient Information (Languages)

### Description:

The Patient Languages table holds information on what languages people in the implementing site's VDW can speak and read.

### Uses:

This table is useful for developing cohorts who can be communicated with via specific languages, and possibly also for identifying people with particular national origins (though that would of course be inexact).

### Key Variables:

- Person ID
- Language
- How the indicated person uses this language (spoken, written or both)
- Whether this is the person's primary language

### Important to Know:

Only people with some level of known language appear in this table.

## Patient Reported Outcomes (PRO):

### Description:

The Patient Reported Outcomes (PRO) data are collected from patient self-administered questionnaires or interview-based questionnaires completed by health care team members. PRO data may be collected during office visits, via electronic data capture, or over-the-phone, etc. The PRO data is contained in three tables:

- PRO\_TYPES: Contains high level information about survey types. Each record contains a unique identifier, PRO\_ID, defined by distinct survey type/subtype combinations. For example, the PRO type 'PHQ' has multiple subtypes (PHQ-9, PHQ-2, etc.) associated with it.
- PRO\_SURVEYS: Contains all distinct survey question text for all PROs. There may be multiple versions of the questions as they are often revised over time.
- PRO\_SURVEYS\_REPONSES: Contain patient level response data linked to survey questions. Other data include survey date and time, mode of data collection and a link to the VDW encounter table.

### Uses:

Examples of how PRO Personal Health Questionnaire (PHQ) data can be useful to researchers:

- Studies of how depression severity relates to diabetes, CVD, dementia, etc.
- Targeted interventions based on depression severity
- Using PRO EHR data to evaluate system-level mental health programs, and improve studies of depression treatment and prevention

### Key Variables:

- Survey name
- Question text
- Question version
- Person ID
- Response date and time
- Response text
- Survey mode
- Encounter ID

### Important to know:

- New PRO surveys will be added as needed. The HCSRN currently has implemented the following surveys into the VDW PRO tables:
  - Patient Health Questionnaire (PHQ2, PHQ9)
  - Generalized Anxiety Disorder survey (GAD-7, GAD-2)
  - Columbia Suicide Severity Rating Scale (C-SSRS)
  - Alcohol Use Disorders Identification Test (AUDIT-10, AUDIT-2, AUDIT-C)
  - Edinburgh Postnatal Depression Score (EPDS)
  - Other surveys are available among the CESR sites.
- The wording of questions may change over time (e.g., due to phrasing revisions, grammatical corrections, etc.). Original question text is stored as administered (question\_text\_orig) and the question version number is noted.

- Each response to a survey question gets a dedicated record in the PRO\_SURVEY\_RESPONSES table, so any summary data must be grouped as needed (by survey, date, MRN, etc.) to avoid overcounts.

## Pharmacy (Pharmacy and EverNDC):

### Description:

The PHARMACY table contains data on medications dispensed in the outpatient setting. Dispenses in inpatient settings are excluded, along with incomplete or unfilled medication orders. Information on the drug dispensed is available in the EVERNDC companion lookup table. Compound drugs may be represented differently across sites. If a drug of interest is a compound drug, users may want to check with individual sites on their handling of these dispenses. The EVERNDC table is a lookup table containing all NDC values ever included in the PHARMACY table at each site. That is, pharmacy dispenses occurring at site A will have at least one corresponding record in site A's EVERNDC table but not necessarily any records in site B's EVERNDC table.

### Uses:

Examples of how Pharmacy data may be useful to researchers:

- Studies involving the medication use process, including medication adherence, linked to healthcare utilization data.
- Inclusion or exclusion in a population based on medication use.
- Targeted interventions based on usage of specific types of drugs.
- Usage of medications over time

### Key Variables:

- Person ID
- Date of dispense
- NDC
- Days Supply
- Units dispensed
- Prescribing Provider
- Generic name
- Brand name
- Unit of Measure
- Strength

### Important to know:

- A link between Pharmacy and Medication Orders is available at over half of the sites. Contact individual sites for availability if interested.
- Site specific values for non-standard items are allowed in the NDC column. This includes but is not limited to compound drugs, stock items, repackaged drugs, and study medications. Some site specific NDCs are identifiable via a starting set of characters, such as 11111 or 99999, or a term, and should be identifiable via the NDC\_SITE\_SPECIFIC\_FLAG variable.
- For multi-site projects, a locally created list of NDCs rarely captures all NDCs used at all participating HCSR sites. A two-step approach that uses the EverNDC table will help compile a

complete list of all NDCs for the studied drug. The lead analyst of a project can first write a distributed program to query the EverNDC tables across sites. They may use generic names, brand names, AHFS codes (if available), or GPI codes (if available) to identify the NDCs in use at other sites. They can then collate the NDCs identified from all participating sites to compile a final list of NDCs for the study.

- Number of dispenses and of members who used each drug during a specific period can be obtained by querying the VDW Pharmacy file using the NDCs identified from the two-step approach described above.
- Unit of Measure and Strength appear as reported in source data without standardization or cleaning except for storing values in all upper case.

## Social History

### Description:

The SOCIAL HISTORY table contains behavioral measures taken by health professionals during clinic visits, over the telephone, or via questionnaires. These measures include the use of tobacco, alcohol, and illegal drugs as well as sexual partners and contraceptive use.

### Uses:

Examples of how Social History data may be useful to researchers:

- Studies of how alcohol, illegal drug, or tobacco use relate to utilization or specific diagnoses.
- Inclusion or exclusion in a population based on a behavioral measure.
- Studies related to specific sexual partners or safe/unsafe sex.
- Linkage to healthcare utilization data.

### Key Variables:

- Person ID
- Date of member contact
- Tobacco use and types
- Alcohol use and type
- Illicit drug use and type
- Sexual partner types
- Birth control use
- Encounter ID

### Important to know:

- Because of the range of tobacco products available and the variability in the level of detail recorded, precision and capture of tobacco related variables may vary both across and within sites. Consistency of response is also not guaranteed over time.
- Social History measures may carry special privacy concerns. Beyond the nature of these variables, the use of free text fields in any content area warrants additional care as they may contain personal health information (PHI). Users are encouraged to consult with privacy experts and/or experienced users before extracting data if they are unfamiliar with these additional data sensitivities. QA testers should also review all free text fields included in a QA and remove any PHI before sharing QA results.
- Data on marijuana use is currently contained as free text in a variable. However, relevant information may also be contained in a medical center's chart notes or "problem lists," which would not then be transferred into the Social History table.
- The utilization linking variable is populated only when there is an encounter that can be definitively matched to the social history record.

# Tumor

## Description:

The Tumor content area contains tumor registry data including tumor and patient characteristics as well as treatment information. The data are collected from in-house tumor registries, hospital registries or SEER registries. The vast majority of data correspond directly to NAACCR data items, but there are a few created summary variables and translated dates in the Tumor Main table. Data items corresponding to NAACCR items are identified by their NAACCR XML tag.

Tumor data are distributed over two tables:

- Tumor Main: A wide table containing 49 variables, where each row represents one tumor. In most cases, this table should contain sufficient information to identify the initial study cohort.
- Tumor Supplementary: A long, narrow table with 3 fields, where each row contains one data item associated with a tumor. These are typically less frequently used items that provide additional detail to the information contained in the main table. Information can be linked to data from the main table using the tumor ID.

## Uses:

Naturally, Tumor data are most frequently used in cancer related projects (outcomes analysis, treatment patterns and effectiveness, incidence reporting, etc.). Sometimes, tumor data are also used to identify covariates or exclusions for other types of studies.

## Key Variables:

### Main table

- Person ID
- Tumor ID
- Diagnosis date
- Primary Site
- Histology
- Grade
- Behavior
- AJCC Stage
- SEER Summary Stage
- Treatment summary variables

### Supplementary table

- Tumor ID
- Attribute (NAACCR XML Tag)
- Value

## Important to know:

- The content of the Tumor tables closely follows NAACCR standards, which are updated frequently. As a result, the way similar information is stored in the Tumor table will change over time. For example, site specific disease indicators, tumor size, or some treatment information will be stored in different fields depending on the year of diagnosis. For some key information, such as staging or grade, the Main table will provide a created summary variable that remains

valid across time, but all content of the Supplementary is directly affected by NAACCR changes. Users will therefore have to consult both the VDW and the NAACCR documentation to ensure that they are able to retrieve all pertinent information.

- The Supplementary table layout is not intuitive to use, but the workgroup has provided a macro that greatly facilitates the task.
- Some dates that are completely missing in the Main table may be found as incomplete dates in the Supplementary table.
- The data dictionary contains a list of NAACCR items that all sites are expected to populate to the extent possible. However, many items may not be available at all sites. Sites may also choose to include additional items not on the list. Users should consult the most recent QA report to find out which items are available at each site.

## Utilization (Encounter, Diagnosis, Procedures)

### Description:

The Utilization content area gives insight to a member/patient's interactions with our health system. This content can tell us what kind of interaction (outpatient, inpatient, telephone, email, etc.), what the outcome was (diagnoses made), and what services were provided to the patient (procedures, level of service, etc.).

- The **ENCOUNTER** table contains medically related interactions between a patient and a medical provider regardless of care setting or type of encounter. Content is characterized as outpatient encounters, virtual care, hospital stays, institutional stays, and other encounters. Information including location, provider, date of admission and discharge can be found in this table.
- The **DIAGNOSIS** table contains all ICD-9 and ICD-10 diagnoses captured for each encounter in the ENCOUNTER table, as well as the diagnosing provider.
- The **PROCEDURES** table contains CPT-4, HCPCS and ICD-9/10 procedure codes captured in the system for each encounter. Evaluation and management (level of service) codes and revenue codes are also stored in this table.

### *Examples of how Utilization data can be useful to researchers:*

- Defining analytic cohorts or outcomes by encounter type, such as inpatient and ED use
- Identifying patient recruitment populations based on care settings or department specialty
- Studying patterns of care (post-exposure follow-up, virtual care outcomes, etc.)
- Determining disease state or comorbidity burden
- Understanding treatment patterns or use of innovative procedures.
- Performing cost burden analysis based on standardized costing methods (DRG and length of stay, Average Weighted Cost, etc.)

### Key Variables:

- Encounter ID
- Person ID
- Encounter Type (inpatient, same day surgery, ambulatory office, video, etc.)
- Provider (Visit/Attending, Diagnosing and Performing)
- Location
- Department
- Date (admission/visit, discharge, procedure)
- Discharge disposition and status
- Diagnostic Relational Grouper (DRG) code
- Diagnosis Code and Type of Diagnosis
- Procedure Code and Type or Procedure
- Sourcing (EMR, Claims, etc.)

## Important to know:

### *Linkages between utilization tables and other content:*

- The encounter identifier (ENC\_ID) uniquely identifies each encounter and may be used to link diagnoses, procedures and other content (e.g., vitals, medication orders, patient questionnaires, etc.) to a specific encounter.
- The diagnoses guidelines state that diagnoses are only included if there is a patient/provider interaction that results in a record in the ENCOUNTER table, to ensure they are valid diagnoses. In some cases, diagnoses may be attached to lab orders or other clinical documentation and therefore these diagnoses would not be found in the VDW.
- Procedures are included whether or not an encounter can be linked.
- Data on all services provided by the health plan, whether or not they were provided to a member (enrollee) of the plan should be included. This means that the utilization tables may need to be checked against the enrollment table to identify study populations. Users should not assume that people found in the Utilization tables are members of the health plan.

### *Multiple or “duplicate” encounters on same day:*

- Encounters with the same provider on the same day are allowed if that is the truth in the source data and have unique encounter identifiers within the source system. This may occur between encounters of different types (e.g., telephone call followed by an ambulatory office visit) or encounters of the same type (e.g., an emergency room encounter in the morning and another later in the day).
- In general, the intention of the encounter table is to describe all significant interactions between patients and medical providers. Depending on the source system used, there may be differing levels of granularity within a site’s data. For example:
  - Sites that own their own hospitals may be able to separate out emergency room, Observation bed, and inpatient admissions into separate encounters within the same facility, while claims-based sites will likely have these visits bundled into a single encounter. In these cases, revenue codes may be used to determine ER charges, Observation bed nights, etc.
  - Claims-based sites, or sites using billing systems to determine utilization, may not be able to discern whether a telephone visit was actually scheduled. Evaluation and management codes may be used to help determine “significance” of a telephone call.

### *General inclusion/exclusions:*

- Sites currently may have differing local rules on inclusion/exclusion from the VDW, that can affect volumes of care. For example, several VDW sites exclude encounters with no diagnosis or procedure attached, to help mitigate space issues. This could potentially make studies that look at provider burden based on volume or “significant” vs “non-significant” care more difficult.
- Admitting diagnoses and diagnoses sourced from the problem list are excluded from the DIAGNOSES table.
- Some sites may be using reporting systems to source data that limit the number of diagnoses and procedures that are attached to an encounter. This is typically found in claims systems or legacy systems.

## Utilization (Facilities)

### Description:

The FACILITY content is comprised as a reference table describing the identity, relationship, and location for a given facility code coming out of an administrative source system, linkable through the FACILITY\_CODE variable.

Examples of how Facility data can be useful to researchers:

- Clustering or adjusting analyses by facility
- Identifying facilities for intervention studies
- Studying health care barriers or patient behaviors through “distance from clinic”
- Assisting in identifying core partnering hospitals for claims-based sites
- Understanding pal data gaps by facility (services offered or not offered)

### Key Variables:

- Facility ID
- Relationship (Owned Facility vs. External)
- Relationship History
- Physical address
- Latitude/Longitude
- Address type (Clinical location or Billing Vendor)

### Important to know:

- Most sites will have usable data for facilities owned by their site’s health care system. While much care has been taken to identify details on facilities, not every facility will have useable information.
- Latitude and longitude are highly encouraged to be populated for owned facilities, but not required for external facilities.
- Sometimes it may be difficult for a site to discern between a vendor/billing provider and a clinical facility. The ADDRESS\_FACILITY\_TYPE may be a best guess to support research. Sites are encouraged to find and use the service location over a billing provider where possible. If a facility from an encounter is unable to be linked to the FACILITY table, it may be worth trying to map to the PROVIDER table, as it may represent a billing or service *provider* rather than a place of service or *facility*.
- Future improvements to the facility table are needed to help identify facility types. Local variables may be available to help discern between ambulatory clinics, inpatient hospitals, psychiatric hospitals, skilled nursing facilities, nursing homes, etc.
- Relationship history indicates whether the facility was once external and now owned, or vice versa. Local expertise may be needed to understand the timeframes at which facilities changed ownership.

## Utilization (Providers)

### Description:

The PROVIDER content tables contain details about the providers found within other content areas within the VDW. Two tables are available for use:

- The **PROVIDER** data contains demographic and training information that may influence how they practice or interact with patients. Each provider is represented by one unique record.
- The **PROVIDER\_TAXONOMIES** table (*available in fall 2023*) provides insight into a provider's training and licensing through standardized CMS taxonomies indicating provider types and specialties. A provider may have multiple taxonomies, over time, or in which they are currently licensed.

Examples of how Provider data can be useful to researchers:

- Clustering analyses by provider
- Identifying providers for intervention studies
- Studies on care practices and health disparities
- Cohort identification by care or diagnoses provided by specific provider types

### Key Variables:

- Provider ID
- Birth year
- Gender
- Race/Ethnicity
- Graduation year
- Provider Types
- Provider specialties and subspecialties

### Important to know:

- Most sites will have usable data for providers employed by their site's health care system. While much care has been taken to identify details on providers, not every provider will have useable information, particularly historically.
- A "provider" may be:
  - an individual person,
  - a billing or service provider, meaning that they are organizations, provider practice groups, or facilities, or
  - a "resource", meaning a float pool resource, or even a room booked for a specific service. A taxonomy may be still be attached if there is enough specificity to determine a proper taxonomy that would normally be assigned to a person.

The ENTITY\_TYPE may be used to discern between these different types.

- It is recommended that sites make every effort to ensure providers can be linked from other content to a provider record in the PROVIDER and PROVIDER\_TAXONOMIES tables, however there will likely a level of missingness that each site is willing to accept, particularly when related to one-off claims or unidentifiable values from legacy data sources.

- Many providers have multiple taxonomies and can be licensed in more than one state. All taxonomies are kept supporting studies where there is interest in whether a provider was ever trained or formerly practicing with a given specialty. Due to the inability to understand the timeframes during which a provider practiced a specific taxonomy, there is a variable called SITE\_CURRENT\_RECOMMENDED to help identify the taxonomy most likely representing how a provider is currently practicing at a site. It considers the state in which the taxonomy was licensed as well as the most recent taxonomy entered.
- The NPI registry provides a good starting point for provider-to-taxonomy linkages, but it is important to note that it is not necessarily the most accurate. Some providers who obtained NPIs before the taxonomies were fully built out may simply be listed as general practice physicians due to a default in the system. It is recommended to sites to either keep local provider type information and/or map their local provider types and specialties to the standardized taxonomy (in addition to using the NPI registry) to look for conflicts and flag the SITE\_CURRENT\_RECOMMENDED flag accordingly.
- The provider ID is unique to a provider as defined by the site. Some individuals may have multiple provider IDs or even NPIs over time. The different NPIs may in some cases represent different eras of a provider's training (e.g., as an RN and then as an MD), but some providers will maintain the same NPI over time. If a site may choose to combine multiple provider IDs into a master provider ID in their VDW, OR they may choose to leave them separate for granularity and traceability back to the source – particularly in the case of multiple NPI numbers. It is important to know and have these discussions within study teams to understand how a site manages this.

## Vital Signs

### Description:

Vital Signs includes physiological (i.e., temperature, pulse, respiratory rate, blood pressure) and anthropometric (i.e., height, weight, BMI, infant head circumference) measures taken by health professionals during encounters. It also includes these measures reported by a patient in a Virtual Care Encounter and measures collected through an approved medical device.

### Uses:

Examples of how Vital Signs data can be useful to researchers:

- Studies tracking pregnancy weight over time.
- Targeted interventions based on blood pressure, BMI, or infections.
- Identifying low birth weight infants.
- Research on preeclampsia or heart rate abnormalities.
- Linkage to healthcare utilization data.

### Key Variables:

- Person ID
- Measure date and time
- Blood pressure (also encounter type at which blood pressure was taken and position)
- Height and weight
- For infants only: head circumference
- Encounter ID

### Important to know:

- Body Mass Index (BMI) is not a cleaned variable on the VITAL SIGNS table. Users are recommended to calculate BMI at the time of actual need making use of the VDW Standard Macros.
- The RAW variables' primary purpose is for capturing categories or ranges that may be recorded in the system (e.g., for BMI\_RAW values like "overweight", or "25.0-25.9" or for DIASTOLIC\_RAW values like "HYPERTENSION" or "100-109").
- Utilization linking variable is populated only when there is an encounter that can be definitively matched to the vital record.
- Some sites use Celsius, while most others use Fahrenheit.