PEPR Data Repository Specifications

Overview

The Pediatric Patient-Reported Outcomes in Chronic Disease (PEPR) Data Model is based on the PEDSnet Common Data Model (CDM) which is based in structure on the OMOP Common Data Model.

The PEDSnet CDM is an expansion of the OMOP CDM to accomodate requirements of both the PCORnet Common Data Model and the primary research cohorts established in PEDSnet. The PEPR Common Data Model is an expansion of the PEDSnet CDM to incorporate patient reported outcomes (PROs) that are not captured in a standard fashion in the OMOP or PEDSnet CDMs.

This document provides ETL processing assumptions and conventions developed by the PEPR consortium data contributors.

PEPR Data Standards and Interoperability Policies:

For details on the PEPR Consortium data network architecture, data access and data archiving, refer to the *PEPR Data Sharing Policy, Version 3.0.*

- concept IDs are taken from OMOP 5 vocabularies for PEDSnet CDM v3.7
- concept IDs specific to the PEPR consortium that are not part of the OMOP 5 vocabulary are generated at the the PEPR DCC at CHOP and are submitted to be included in the OMOP vocabulary
- The PEPR consortium does not utilize licensed vocabularies
- Some source fields may be considered sensitive by data sites. Many of these fields are used to generate an ID field. Sites are free to obfuscate or not provide source values that are used to create ID variables. Sites must maintain a mapping from the ID variable back to the original site-specific value for local re-identification tasks but are not required to submit source values.
  - The CHOP DCC will not release source values nor will they post source values on Harvard Dataverse
  - Sites choosing to send data to CHOP can include source values in the data submission
  - Sites choosing to directly upload to Harvard Dataverse are not expected to upload source values
  - Source value obfuscation techniques may include replacing the real source value with a random number, an encrypted derivative value/string, or some other site-specific algorithm
- The PCORnet CDM has specific definitions for null values (as seen below). For the PEPR CDM, please use the following logic to determine the correct concept value to use for **source_concept_id** fields where there are null values in `*_source_value`

Null Name	Definition	PEPR `*_concept_id`	`*_source_concept_id`	`*_source_value`
NULL	A data field is not present in the source system (i.e. there is no way to ascertain the value of the data field from the data that the site already has). Note this is not a 'NULL' string but the NULL value			
NI	'No information' A data field is present in the source system, but the source value is null or blank	44814650	0	value as in source (leave NULL)
UN	'Unknown' A data field is present in the source system, but the source value explicitly denotes an unknown value	44814653	0	value as in source (however the source system defines an unknown value)
OT	'Other' A data field is present in the source system, but the source value cannot be mapped to the CDM	44814649	0	value as in source

- Guidelines for populating `*_source_value`
  - When possible, provide "human readable" values rather than coded value e.g. for `gender_source_value`, the source value at your site may be `1` for Female and `2` for Male. Please provide the label value of `Female` and `Male` as the `gender_source_value`

Data Extraction Guide

Please use the table headings in the following tables as a guide in extracting and submitting data. All **network required** fields must be submitted even if not submitting data in a field (see missing value specifications above). The following table contains definitions for interpreting the table specifications contained in this document:

Field	NOT Null Constraint	Network Requirement	Data Type	Description	PEPR Conventions
Column name as it should appear in the .csv sent to the DCC or uploaded to Harvard Dataverse. Note capitalizations and underscores	Yes if it cannot be NULL and No if it can be NULL	Is this field a network requirement? (Yes or No) Note that a field may be required in the data submission, but need not be populated. In this case, the field will have a Network Requirement of 'yes' and a NOT Null Constraint of 'no'. Source values are not to be uploaded to Harvard Dataverse but should be kept locally for DQA and reverse mapping.	Format of the data in this field (e.g. Integer, Varchar)	Brief definition of field	PEPR-specific definitions of field - may contain value sets for field contents

PERSON

The person domain contains records that uniquely identify each patient in the source data. Each person has associated demographic attributes, which are assumed to be constant for the patient throughout the course of their periods of observation (or over the course of their PRO collection period). All other patient-related data domains have a foreign-key reference to the person domain.

Field	NOT Null Constraint	Network Requirement	Data Type	Description	PEPR Conventions
person_id	Yes	Yes	Integer	A unique identifier for each individual. This is created by each contributing site	Centers generate their own ids, unique within their own center, but can overlap with other centers in the PEPR consortium. For each center, should be consistent across the condition_occurrence, observation, pro_occurrence and visit_occurrence tables
gender_concept_id	Yes	Yes	Integer	A foreign key that refers to a standard concept identifier in the vocabulary for the gender of the person.	**Ambiguous:** concept_id = 44814664 **Female:** concept_id = 8532 **Male:** concept_id = 8507 **No Information:** concept_id = 44814650 (Vocabulary_id='PCORNet') **Unknown:** concept_id = 44814653 **Other:** concept_id = 44814649
year_of_birth	Yes	Yes	Integer	The year of birth of the person in YYYY format	
month_of_birth	Yes	Yes	Integer	The month of birth of the person in MM (e.g. if May 1, `05`)	
day_of_birth	Yes	Yes	Integer	The day of the month of birth of the person in DD (e.g. if May 1, `01`)	

Field	NOT Null Constraint	Network Requirement	Data Type	Description	PEPR Conventions
race_concept_id	Yes	Yes	Integer	A foreign key that refers to a standard concept identifier in the Vocabulary for the race of the person	*Predefined values:* **American Indian/Alaska Native:** concept_id = 8657 **Asian:** concept_id = 8515 **Black or African American:** concept_id = 8516 **Native Hawaiian or Other Pacific Islander:** concept_id = 8557 **White:** concept_id = 8527 **Multiple Race:** concept_id = 44814659 (vocabulary_id='PCORNet') **Refuse to answer:** concept_id = 44814660 (vocabulary_id='PCORNet') **No Information:** concept_id = 44814650 vocabulary_id='PCORNet') **Unknown:** concept_id = 44814653 **Other:** concept_id = 44814649
ethnicity_concept_id	Yes	Yes	Integer	A foreign key that refers to the standard concept identifier in the Vocabulary for the ethnicity of the person	*Predefined values:* **Hispanic:** concept_id = 38003563 **Not Hispanic:** concept_id = 38003564 **No Information:** concept_id = 44814650 (vocabulary_id='PCORNet') **Unknown:** concept_id = 44814653 (vocabulary_id='PCORNet') **Other:** concept_id = 44814649 (vocabulary_id='PCORNet')
location_id	Yes	Yes	Integer	The value of the first three digits of the zip code of residence for the patient	The value of the Zip 3, formatted as three digits, e.g. `123`
care_site_id	Yes	Yes	Integer	A foreign key to the care site table where details of the hospital at which the participant was included are captured.	Should indicate the hospital at which the patient was recruited.

Field	NOT Null Constraint	Network Requirement	Data Type	Description	PEPR Conventions
gender_source_value	No	No	Varchar	The source code for the gender of the person as it appears in the source data	The person's gender is mapped to a standard gender concept in the Vocabulary; the original value is stored here for reference. See `gender_concept_id`
race_source_value	No	No	Varchar	The source code for the race of the person as it appears in the source data	The person race is mapped to a standard race concept in the Vocabulary and the original value is stored here for reference. For patients with multiple races (i.e. biracial), race is considered a single concept, meaning there is only one race slot. If there are multiple races in the source system, concatenate all races into one source value, and use the concept_id for Multiple Race.
ethnicity_source_value	No	No	Varchar	The source code for the ethnicity of the person as it appears in the source data	The person ethnicity is mapped to a standard ethnicity concept in the Vocabulary and the original code is, stored here for reference.
site	Yes	Yes		The name of the study/cohort for which the patient is included in the data repository	Should be the same for all patients involved in a single study

Details of categorical definitions:
Ethnicity:

For PEPR, a person with Hispanic ethnicity is defined as "A person of Cuban, Mexican, Puerto Rican, South or Central American, or other Spanish culture or origin, regardless of race."

Race: **American Indian or Alaska Native:** a person having origins in any of the original peoples of North and South America (including Central America), and who maintains tribal affiliation or community attachment **Asian:** A person having origins in any of the original peoples of the Far East, Southeast Asia, or the Indian subcontinent including, for example, Cambodia, China, India, Japan, Korea, Malaysia, Pakistan, the Philippine Islands, Thailand, and Vietnam. **Black or African American:** A person having origins in any of the black racial groups of Africa. **Native Hawaiian or Other Pacific Islander:** A person having origins in any of the original peoples of Hawaii, Guam, Samoa, or other Pacific Islands. **White:** A person having origins in any of the original peoples of Europe, the Middle East, or North Africa. For patients with multiple races (i.e. biracial), race is considered a single concept, meaning there is one race slot. If there are multiple races in the source system, concatenate all races into one `race_source_value` (see below) and use `concept_id` code as 'Multiple Race'

CARE_SITE

The care site domain contains a list of uniquely identified physical or organizational units where healthcare delivery is practiced (offices, hospitals, clinics, etc.).

For PEPR, the care site table should include all hospitals at which patients were recruited or have visits recorded (in the visit_occurrence table) over the course of their inclusion in the study.

Field	NOT Null Constraint	Network Requirement	Data Type	Description	PEPR Conventions
care_site_id	Yes	Yes	Integer	A unique identifier for each defined location of care within an organization. Here, an organization is defined as an institution that collects its patient information in one source system	IDs must be consistent across the visit_occurrence (indicating site at which patient filled out survey) and person (indicating site at which patient was recruited) tables.
care_site_name	Yes	Yes	Varchar	Description of the care site	Name of the care site
place_of_service_concept_id	Yes	Provide When Available	Integer	A foreign key that refers to a place of service concept identifier in the Vocabulary	**Urgent Care Facility:** concept_id= 8782 **Rural Health Clinic:** concept_id= 8761 **Outpatient:** (Examples: Hospital Dialysis, HOD, Day Hospital, Day Medicine) concept_id= 8756 **Office:** concept_id= 8940 **Inpatient Psychiatric Facility:** concept_id= 8971 **Inpatient Hospital:** concept_id= 8717 **Independent Clinic:** concept_id= 8716 **Emergency Room - Hospital:** concept_id= 8870 **Other Inpatient Care:** concept_id= 8892 **Unknown:** concept_id= 44814653 **Other:** concept_id= 44814649 **No information:** concept_id= 44814650

Field	NOT Null Constraint	Network Requirement	Data Type	Description	PEPR Conventions
care_site_source_value	No	No	Varchar	The identifier for the organization in the source data, stored here for referece.	If care site source values are deemed sensitive by your organization, insert a pseudo-identifier (random number, encrypted identifier) into the field. Sites electing to obfuscate care site_source_values will keep the mapping between the value in this field and the original clear text location source value. This value is only used for site-level re-identification for study recruitment and for data quality review. For EPIC EHRs, map care_site_id to Clarity Department. Sites may consider using the care_site_id field value in this table as the pseudo-identifier as long as a local mapping from care_site_id to the real site identifier is maintained.
place_of_service_source_value	No	No	Varchar	The source code for the place of service as it appears in the source data, stored here for reference.	

CONDITION_OCCURRENCE

The condition occurrence domain captures records of a disease or a medical condition based on diagnoses, signs and/or symptoms observed by a provider or reported by a patient. The PEPR condition_occurrence table contains conditions of two types: (1) a participant's cohort-qualifying (i.e. primary) condition and (2) conditions derived from the comorbidity checklist (i.e. secondary condition). The condition_type_concept_id explains the provenance of the conditions.

Co-morbidity checklist: Participants were evaluated on a yes/no/unknown/not ascertained scale for the following conditions.

- Asthma
- Obesity
- Autism
- ADHD
- History of low birth weight
- Atopic Dermatitis/Eczema

If a study is not including the condition checklist in the minimum required dataset, each condition from the condition checklist should be included for each patient in the cohort with a condition_type_concept_id of 44814650 (No information)

Field	NOT Null Constraint	Network Requirement	Data Type	Description	PEPR Conventions
condition_occurrence_id	Yes	Yes	Integer	A unique identifier for each condition occurrence event. (Primary key)	Each institution uses its own unique sequencing to populate this field
person_id	Yes	Yes	Integer	A foreign key identifier to the person who is experiencing the condition. The demographic details of that person are stored in the person table	
condition_concept_id	Yes	Yes	Integer	A foreign key that refers to a standard condition concept identifier in the Vocabulary.	**DCC will send a list of acceptable concept_ids based on all cohort-level conditions and co-morbidities**
condition_source_value	No	No	Varchar	The source code for the condition as it appears in the source data. This code is mapped to a standard condition concept in the Vocabulary and the original code is stored here for reference.	
visit_occurrence_id	No	Yes	Integer	A foreign key to the visit in the visit table during which the condition was determined (if available)	Can include a reference to the visit_occurrence table to identify the wave during which the condition was recorded

Field	NOT Null Constraint	Network Requirement	Data Type	Description	PEPR Conventions
condition_type_concept_id	Yes	Yes	Integer	A foreign key that refers to a standard condition_type identifier in the Vocabulary and is used to identify the source of the condition	Potential values of condition type: **Primary Condition:** concept_id = 44786627**Secondary Condition:** concept_id = 44786629 Where primary condition is the cohort-qualifying condition and secondary condition is any co-morbidity indicated as 'yes' on the comorbitidy checklist**No diagnosis:** condition marked as No on the condition checklist - concept_id = 4011458 **Patient not asked:** condition checklist was not filled out for the patient - concept_id = 46273465 **No information:** condition checklist was obtained for the patient, but the result will not be represented in the PEPR minimum dataset - concept_id = 44814650

OBSERVATION

The observation domain supports capture of data not represented by other domains such as structured measurements. The observation table in the PEPR data model contains the Area Deprivation Index (ADI), determined based on the patient's location as reported at baseline. ADI national percentile is represented in the value_as_number column with the unit_concept_id describing the unit (national percentile) of the ADI.

Field	NOT Null Constraint	Network Requirement	Data Type	Description	PEPR Conventions
observation_id	Yes	Yes	Integer	A unique identifier for each observation	Sites generate unique sequential values for this field
person_id	Yes	Yes	Integer	A foreign key identifier to the person about whom the observation was recorded. The demographic details of that person are stored in the person table	Should be consistent across all tables

Field	NOT Null Constraint	Network Requirement	Data Type	Description	PEPR Conventions
observation_concept_id	Yes	Yes	Integer	A foreign key to the standard observation concept identifier in the Vocabulary	concept_id representing ADI to be generated by CHOP **DCC** will need to let sites know the concept_id for Area Deprivation Index
observation_date	Yes	Yes	Date	The date of the baseline observation for the patient	Date shifted according to PEPR Data Sharing Plan
value_as_number	Yes	Yes	Integer	The observation result stored as a number. Only applicable to observations where result is stored as a numeric value.	All ADI should be represented as value_as_number. Required unless otherwise discussed with DCC
unit_concept_id	Yes	Yes	Integer	Indicates the unit of the value_as_number as a standard concept_id.	**DCC** will send a vocabulary with the concept_ids for national percentile
visit_occurrence_id	No	Yes	Integer	A foreign key to the visit in the visit table during which the observation was recorded.	The visit should indicate the the observation was recorded at baseline
observation_source_value	No	No	Varchar	The observation code as it appears in the source data. This code is mapped to a standard concept in the Vocabulary and the original code is stored here for reference.	

Field	NOT Null Constraint	Network Requirement	Data Type	Description	PEPR Conventions
unit_source_value	No	No	Varchar	The source code for the unit as it appears in the source data. This code is mapped to a standard unit concept in the Vocabulary and the original code is stored here for reference.	Indicates the unit of the value_as_number

Note: For PEPR CDM v1 only the ADI is represented in the observation table and should be represented as value_as_number unless otherwise discussed with the DCC

PRO_OCCURRENCE

The pro occurrence domain captures information that is derived from a structured survey or similar instrument.

Field	NOT Null Constraint	Network Requirement	Data Type	Description	PEPR Conventions
pro_id	Yes	Yes	Integer	Unique identifier for each PRO response (Primary Key)	
person_id	Yes	Yes	Integer	Foreign key to the person table	
pro_concept_id	Yes	Yes	Integer	A foreign key referring to a standard concept ID in the Standardized Vocabularies for the specific question or item used to obtain this PRO; this value may encode the instrument as well	will specify either an individual item whose response is captured, or a composite score for all or part of an instrument. This concept will also typically include information about the instrument from which the item was drawn. However, where necessary for disambiguation or for convenience in analyses, the instrument_concept_id can provide (possibly redundant) specification of the instrument.

Field	NOT Null Constraint	Network Requirement	Data Type	Description	PEPR Conventions
instrument_concept_id	No	Yes	Integer	A foreign key referring to a standard concept ID in the Standardized Vocabularies for each instrument	Standard Vocabulary will be provided by CHOP
instrument_version	No	Yes	Varchar	Version of instrument	
pro_type_concept_id	Yes	Yes	Integer	A foreign key referring to a standard concept ID in the Standardized Vocabularies for the administration method and mode	Provides information about how the PRO was obtained, including method (e.g.self-administered or interviewer-administered) and mode (e.g.paper and pen, simple computer-based survey, CAT)
respondent_type_concept_id	No	Yes	Integer	A foreign key to a standard concept ID in the Standardized Vocabularies indicating the respondent type (e.g. patient, parent, proxy) to the PRO items	provides information about the person who directly provided the PRO (e.g.patient, parent, proxy)

Field	NOT Null Constraint	Network Requirement	Data Type	Description	PEPR Conventions
value_as_concept_id	No	No	Integer	A foreign key to a standard concept ID in the Standardized Vocabularies indicating the response to items for qualitative items	A standardized representation of the PRO response is captured in value_as_number (typically for numeric or Likert scales), or value_as_concept_id (typically for qualitative scales, Yes/Noresponses). Both may be used if a particular categorical response corresponds to a specific raw score (e.g.Likert scale of Never-Rarely-Sometimes-Often-Always corresponding to 1-5 score). If the response can only be represented as text, both of these fieldsare set to NULL, and analyses use value_source_value.
value_as_number	No	No	Numeric	Raw score for quantitative items	A standardized representation of the PRO response is captured in value_as_number (typically for numeric or Likert scales), or value_as_concept_id (typically for qualitative scales, Yes/Noresponses). Both may be used if a particular categorical response corresponds to a specific raw score (e.g.Likert scale of Never-Rarely-Sometimes-Often-Always corresponding to 1-5 score). If the response can only be represented as text, both of these fieldsare set to NULL, and analyses use value_source_value.
theta	No	No	Numeric		
scaled_score	No	No	Numeric		The method for obtaining scaled_score is determined by the instrument (e.g.T-score for PROMIS items)
standard_error	No	No	Numeric		
pro_response_date	No	Yes	Date	Date on which the PRO was obtained	Shifted according to PEPR PRO date shifting convention. Date and time data elements should follow ISO 8601 conventions.

Field	NOT Null Constraint	Network Requirement	Data Type	Description	PEPR Conventions
pro_response_time	No	No	Time	The time in which the PRO was obtained. (To accommodate all temporal analyses, a data type of datetime or timestamp can be used, combining measurement_date and measurement_time)	Date and time data elements should follow ISO 8601 conventions.
instrument_source_value	No	No	Varchar	Source value for the instrument name	
pro_source_value	No	No	Varchar	Source value for the item ID	
admin_method_source_value	No	No	Varchar	Source value for the method the respondent completed measure (self-administered vs interview-administered)	
admin_mode_source_value	No	No	Varchar	Source value for the mode the respondent completed measure (paper/telephone/computer/CAT)	
respondent_source_value	No	No	Varchar	Source value indicating respondent type	
value_source_value	No	No	Varchar	Source value for item response	

CHOP will process source values to create standardized vocabulary for the following concept_id data elements:

- pro_concept_id
- instrument_concept_id
- pro_type_concept_id
- respondent_type_concept_id
- value_as_concept_id

Example of table that each center will submit to DCC (note that the pro_occurrence table should be a single table but is broken into separate lines below due to page space limitations in this document):

pro_id	person_id	pro_ concept_id	instrument_ concept_id	instrument_version	pro_type_ concept_id
1234	1			2.0	

respondent_type_concept_id	value_as_concept_id	value_as_number	theta	scaled_score
		1		

standard_error	pro_response_date	pro_response_time	visit_occurence_id	instrument_source_value
	2018-01-01	1:00:00	1	PROMIS Parent Proxy Asthma Impact

pro_ source_value	admin_method_ source_value	admin_mode_ source_value	respondent_ source_value	value_ source_value
2220R2r	interview-administered	telephone	Parent Proxy	Never

VISIT_OCCURRENCE

The visit occurrence domain captures information about the time at which the patient reported outcome was recorded.

Field	NOT Null Constraint	Network Requirement	Data Type	Description	PEPR Conventions
visit_occurrence_id	Yes	Yes	Integer	A unique identifier for each each person's survey completion	Primary key (must be unique)
person_id	Yes	Yes	Integer	A foreign key identifier to the person for whom the visit (survey completion) is recorded. The demographic details of that person are stored in the person table.	
visit_start_date	Yes	Yes	Date	The start date of the visit (or, the date that the survey was completed)	Date should be shifted as according to the PEPR conventions
care_site_id	No	Yes	Integer	A foreign key to the care site in the care_site table for which the visit (or survey) was recorded	

Field	NOT Null Constraint	Network Requirement	Data Type	Description	PEPR Conventions
visit_concept_id	No	Yes	Integer	A foreign key that refers to a place of service concept id in the vocabulary	**Inpatient Hospital Stay:** (e.g. Hospital Admissions) concept_id = 9201 **Ambulatory Visit:** (e.g. Office Visits or Appointments) concept_id = 9202 **Emergency Department:** concept_id = 9203**Long Term Care Visit:** concept_id = 42898160 **Other ambulatory Visit:** (e.g. Telemedicine) concept_id = 44814711 **Non-Acute Institutional Stay:** concept_id = 44814710 **Emergency Department Admit to Inpatient Hospital Stay (If sites are unable to split the encounter):** concept_id = 2000000048 **Observation Stay:** concept_id = 2000000088 **Administrative Visit:** (e.g. Professional Billing or Hospital Abstractions) concept_id = 2000000104 **Unknown:** concept_id = 44814653 **Other:** concept_id = 44814649 **No information:** concept_id = 44814650

List of fields for which the DCC will need a comprehensive list of potential values from each site in order to generate/publish a vocabulary:

- condition_concept_id: the 6 items on the comorbidity checklist + the cohort-qualifying condition
- pro_concept_id: all item IDs (i.e. 2220R2r)
- instrument_concept_id: all instruments (i.e. PROMIS Parent Proxy Asthma Impact)
- pro_type_concept_id: all admin methods (i.e. self-administered, interview-administered) and all admin modes (i.e. paper, telephone, computer, CAT)
- respondent_type_concept_id: all respondent types (i.e. patient, parent, proxy)
- value_as_concept_id: all PRO responses that are not represented as numbers (i.e. Yes, No, Never, Rarely)