

## PCORnet Data Checks v1

These data checks were informed by Mini-Sentinel's data characterization programs ([http://www.minisentinel.org/data\\_activities/distributed\\_db\\_and\\_data](http://www.minisentinel.org/data_activities/distributed_db_and_data)) and by data characterization activities during Phase 1 of PCORnet. Data checks will be expanded and refined over time. Data checks 1.01-1.04 are assessed via PCORnet's CDM Diagnostic Query Package; all others are evaluated via the Data Characterization Query Package.

Category	Check	Description
Data Model Conformance	DC 1.01	DataMart is missing one or more required tables.
	DC 1.02	DataMart has one or more tables which do not conform to CDM specifications.
	DC 1.03	DataMart is missing one or more required fields.
	DC 1.04	DataMart has one or more fields which do not conform to CDM specifications for data type, length, or name.
	DC 1.05	DataMart has errors in the primary key definition for populated tables.
	DC 1.06	DataMart has one or more fields which contain records with values outside of CDM specifications.
	DC 1.07	DataMart has missing values for fields which are required to be populated.
Data Plausibility	DC 2.01	DataMart has more than 5% of records with future dates in any date field. Future dates may be attributable to data entry errors in the source data or ETL errors such as including scheduled appointments in the Encounter table.
	DC 2.02	DataMart has more than 20% of records in the lowest or highest categories of age, height, weight, diastolic blood pressure, and/or systolic blood pressure. A high number of records in these categories may indicate incorrect measurement units.
Data Completeness	DC 3.01	DataMart's average number of diagnoses records per encounter (diagnosis records/encounters in the Encounter table) is less than 1.0 for ambulatory (AV), inpatient (IP), emergency department (ED), or ED to inpatient (EI) encounters.
	DC 3.02	DataMart's average number of procedure records per encounter (procedure records/encounters in the Encounter table) is less than 1.0 for ambulatory (AV), inpatient (IP), emergency department (ED), or ED to inpatient (EI) encounters.
	DC 3.03	DataMart's rate of missing or unknown data values exceeds 5% for any of the following fields which are required for most research studies: BIRTH_DATE, SEX, DX_TYPE, PX_TYPE, and VITAL_SOURCE. Data missingness may be a legitimate source data constraint or an ETL issue.
	DC 3.04	DataMart's rate of missing or unknown data values exceeds 15% for fields which may be required for some research studies: RACE, DISCHARGE_DISPOSITION (IP/IS/EI encounters only), and PDX (IP/IS/EI encounters only). Data missingness may be a legitimate source data constraint or an ETL issue.
Other Analysis Considerations	DC 4.01	DataMart's DIAGNOSIS table has a minimum ADMIT_DATE after January 2010. DataMarts should include data that can be well curated. When possible, DataMarts should include historical data from no later than 2010 to the present.
	DC 4.02	DataMart's PROCEDURES table has a minimum ADMIT_DATE after January 2010. DataMarts should include data that can be well curated. When possible, DataMarts should include historical data from no later than 2010 to the present.
	DC 4.03	DataMart's VITAL table has a minimum MEASURE_DATE after January 2010. DataMarts should include data that can be well curated. When possible, DataMarts should include historical data from no later than 2010 to the present.
	DC 4.04	DataMart does not include all of the following encounter types: ambulatory (AV), inpatient (IP or EI), and emergency department (ED or EI) encounters. This complement of encounter types is not required but may be important for some research studies.
	DC 4.05	DataMart has obfuscated or imputed one or more dates. Imputed or obfuscated dates are permissible but are important to consider when interpreting results.