pcornet

The National Patient-Centered Clinical Research Network

Work Plan:
Data Curation Query Package

v5.13
July 6, 2020

TABLE OF CONTENTS

I. Purpose and Scope

The purpose of the Data Curation Query Package v5.13 is to characterize the data in PCORnet Common Data Model (CDM) v5.1. This package examines all 22 tables. The package consists of the Potential Code Errors query, a Data Curation query and an Empirical Data Curation Report which summarizes key information from the query output and evaluates the results against PCORnet's Data Check v8. Output tables will be produced by running SAS programs against static local DataMarts in PCORnet CDM v5.1 with SAS data types.

Query results will be used by the PCORnet Coordinating Center's Distributed Research Network Operations Center (DRN OC) to ensure a foundational level of data quality across the networks. Approved results may be used to provide initial feasibility estimates for prep-to-research queries, inform study planning activities, and to create DataMart-level, HPRN/CRN-level or network-level reports. Data aggregated at the network level may be shared publicly. DataMart-level results may be published within PCORnet and data from this query can be used to inform the PCORnet Coordinating Center of data availability and fitness for use in response to PCORnet queries, to enable the Coordinating Center to share high-level counts of data variables with requestors, to present PCORnet-level Aggregate Data on public-facing websites, (e.g., PCORnet.org) in manuscripts consistent with the guidelines of the International Committee of Medical Journal Editors ("**ICMJE**"), and for use in PCORnet marketing materials.

To provide the DRN OC with additional insight into the query results, the ETL Annotated Data Dictionary (ETL ADD) must be updated prior to submitting the response to this query. The ETL ADD is stored in a REDCap® database.

Low Cell Count Threshold

Users may specify a low cell count threshold that establishes the minimum number of observations required to protect against possible identification of subject data. Query results greater than zero and less than the threshold will be changed to BT (below threshold) and treated as missing. For example, if a DataMart sets a low cell count threshold of 5, cell counts between 1 and 4 will be changed to BT. The low cell count threshold applies to all query results except for descriptive statistics. The low cell count threshold treatment for each query is shown in Section IV.

The default low cell count threshold value is set to zero (0) in accordance with the CRN Scope of Work. If this value is changed it will be highlighted in orange in the Empirical Data Curation report.

Potential Code Errors Report

The query package produces a Potential Code Errors report which identifies exceptions to the expected code length or content for ICD9/ICD10 diagnosis codes; ICD9/ICD10-PCS and CPT/HCPCS procedure codes; LOINC codes; RXNORM_CUI codes; and NDC codes by applying the heuristics described in Section II.

Data Curation tables

The query package produces up to 248 query output tables depending on how many CDM tables are populated and how the program is executed. Information about each output table is provided in Section IV.

Empirical Data Curation Report

The query package produces an Empirical Data Curation (EDC) report. The EDC Report summarizes key information from the data curation query ouput tables and identifies exceptions to the PCORnet Data Checks. The table of contents is shown in Section V.

Lookback Date

The lookback date is the earliest date included in the query results. The lookback date will be calculated by subtracting the lookback period from the Query Response Date which is the date the data curation query (specifically, the cross-table portion) was run. The default lookback period is 20 years (240 months). **The lookback value should not be modified unless you are instructed to do so**.

This date restriction does not apply to the DEMOGRAPHIC, DEATH, HARVEST, DEATH_CAUSE,PROVIDER or LDS_ADDRESS_HISTORY tables. The dates used to select records in the other tables are ENROLLMENT.ENR_START_DATE, ENCOUNTER.ADMIT_DATE, DIAGNOSIS.ADMIT_DATE, PROCEDURES.ADMIT_DATE, VITAL.MEASURE_DATE, DISPENSING.DISPENSE_DATE, LAB_RESULT_CM.RESULT_DATE, CONDITION.REPORT_DATE, PRO_CM.PRO_DATE, PRESCRIBING.RX_ORDER_DATE, PCORNET_TRIAL.TRIAL_ENROLL_DATE, MED_ADMIN.MEDADMIN_START_DATE, OBS_CLIN.OBSCLIN_DATE, OBS_GEN.OBSGEN_DATE and IMMUNIZATION.VX_RECORD_DATE. Since some of these variables do not have to be populated, records with missing dates will also be included in the data curation query results.

Questions about this query package should be sent to Laura Qualls (laura.qualls@duke.edu).

II. Potential Code Errors Query

The purpose of the Potential Code Errors Program is to help network partners identify exceptions to the expected formats for selected codes. Output tables will be produced by running SAS programs against static local DataMarts in PCORnet CDM 5.1 format with SAS data types.

This program identifies exceptions to the expected code length or content for 8 coding terminologies used in one or more CDM tables. Heuristics are conservative to allow for all potential implementations (e.g. current LOINC codes are 5+ digits, but the program allows for the shorter deprecated codes; ICD10 procedure codes must be 7 alphanumeric chara). These heuristics will **not** identify all erroneous codes, and will only evaluate codes which are classified as one of the qualifying terminologies (e.g. a ICD9 diagnosis code labeled as a SNOMED-CT code type will not be evaluated). The CDM specifications (available on pcornet.org) provide guidance on addressing potential errors (General Implementation Guidance issue #5). The heuristics and CDM tables for each terminology are shown in the table below.

Terminology	**Unexpected length (after removing decimals if applicable)**	**Unexpected string**	**Unexpected alphabetical character**	**Unexpected numeric character**	**CDM tables**

ICD9 diagnosis	Not 3-5	000.x	Any alphabetical character other than E or V	No numeric characters	CONDITION;
					DIAGNOSIS;
					OBS_GEN

ICD10 diagnosis	Not 3-7	000.x or 999.x	First character is not alphabetical	No numeric characters	CONDITION;
					DIAGNOSIS;
					OBS_GEN

CPT/HCPCPS	Less than 5	00000x or 99999x	n/a	No numeric characters	IMMUNIZATION;
					OBS_GEN;
					PROCEDURES;

ICD9 procedures	Not 3-4	00.00	Any alphabetical character	n/a [c]	OBS_GEN; PROCEDURES
ICD10 procedures	Not 3-7 [a]	0000000 or 9999999	n/a	n/a	OBS_GEN; PROCEDURES

NDC	Not 11	00000000000 or 99999999999	Any alphabetical character[2]	n/a [c]	DISPENSING;
					IMMUNIZATION;
					MED_ADMIN;
					OBS_GEN;
					PROCEDURES

RXNORM_CUI	Not 2-7	n/a	Any alphabetical character	n/a [c]	IMMUNIZATION;
					MED_ADMIN;
					OBS_GEN;
					PRESCRIBING

LOINC	Not 3-7 [b]	No hyphen in the penultimate position	Any alphabetical character	n/a [c]	LAB_RESULT_CM;
					OBS_CLIN;
					OBS_GEN

a. Must be 7 digits for billing but this allows for shorter codes which may appear in electronic health record data.
b. Current LOINC codes are 5+ digits but this allows for shorter deprecated codes.
c. Redundant with the unexpected alphabetical character rule.

The data dictionary for the tables produced for by the PCORnet Code Errors query is as follows:

code_summary

Note: This dataset will include 1 row for each observed combination of table and code type.

Field	Description
TABLE	DIAGNOSIS, DISPENSING, LAB_RESULT_CM, MED_ADMIN, OBS_GEN, OBS_CLIN, PRESCRIBING, or PROCEDURES
CODE_TYPE	09,10, CH, ND, LC, or RX
BAD RECORDS	Count of potentially bad records
TOTAL RECORDS	Count of total records
PCT	The percent of records which are potentially bad records.

bad_dx; bad_px; bad_disp; bad_pres; bad_lab; bad_medadmin; bad_obsclin; bad_obsgen

Note: Tables will only be created if the table contains at least 1 record for the relevant code types.

Field	Description
[Varies]	Pseudoidentifer for the table, e.g. encounterID
CODE_TYPE	09,10, CH, ND, LC, or RX
CODE	The code. This field is renamed from the table-specific fields, e.g. DX, PX, LAB_LOINC, and MEDADMIN_CODE.
CODE_CLEAN	Uppercase code which discards decimals, dashes, commas, spaces and trailing blanks
CODE_LENGTH	Length of code_clean
ANYALPHA	The position of the first alphabetical character; 0 if there are no alphabetic characters
ANYDIGIT	The position of the first numeric character; 0 if there are no numeric characters
TABLE	DIAGNOSIS, DISPENSING, LAB_RESULT_CM, MED_ADMIN, OBS_GEN, OBS_CLIN, PRESCRIBING, or PROCEDURES
UNEXP_LENGTH	Error indicator. Yes=1; No=0; null=not applicable.
UNEXP_ALPHA	Error indicator. Yes=1; No=0; null=not applicable.
UNEXP_STRING	Error indicator. Yes=1; No=0; null=not applicable.
UNEXP_NUMERIC	Error indicator. Yes=1; No=0; null=not applicable.

III. Data Curation Query Definitions

The definitions for variables included in the query output are as follows:

- ADMIT_DATE Mismatch: These fields are replicated from the ENCOUNTER table to the PROCEDURES and DIAGNOSES table. The number of mismatched records is the number of records in PROCEDURES or DIAGNOSIS where these fields do not match the value in the ENCOUNTER table.
- ALL_N or RECORD_N or N: Count of records with non-missing values for the specified field.
- DATASET: CDM table name
- DISTINCT_N: Count of records with unique values for the specified field.
- DISTINCT_ENC_ID_N: Count of records with unique values for ENCOUNTERID.
- DISTINCT_PATID_N: Count of records with unique values for PATID.
- DISTINCT_VISIT_N: Count of unique visits in the ENCOUNTER table. Visits are a concatenation of PATID + PROVIDER_ID + ENC_TYPE + ADMIT_DT.
- ELIG_RECORD_N: Count of records in the ENCOUNTER table where PATID, PROVIDER_ID, ENC_TYPE, and ADMIT_DT are all populated.
- ENC_TYPE Mismatch: These fields are replicated from the ENCOUNTER table to the PROCEDURES and DIAGNOSES table. The number of mismatched records is the number of records in PROCEDURES or DIAGNOSIS where these fields do not match the value in the ENCOUNTER table.
- EXP_SPECIMEN_SOURCE: The expected specimen source based on the values established by LOINC®.
- KNOWN_TEST: Total number of records where LAB_LOINC is not null.
- KNOWN_TEST_RESULT: Total number of records where (1) LAB_LOINC is not null and (2) RESULT_NUM is not null and RESULT_MODIFIER is not in (null, NI, UN, OT) or (3) RESULT_QUAL is in ("BORDERLINE", "POSITIVE", "NEGATIVE" or "UNDETERMINED")
- KNOWN_TEST_RESULT_NUM: Total number of records where the test and result are known, as follows: (1) LAB_LOINC is not null and (2) RESULT_NUM is not null and (3) RESULT_MODIFIER is not in (null, NI, UN, OT).
- KNOWN_TEST_RESULT_NUM_SOURCE: Total number of records where the test and result are known, as follows: (1) LAB_LOINC is not null and (2) RESULT_NUM is not null and (3) RESULT_MODIFIER is not in (null, NI, UN, OT) and (4) SPECIMEN_SOURCE is not in (null, NI, UN, OT, UNK_SUB, SMPLS, SPECIMEN)
- KNOWN_TEST_RESULT_NUM_UNIT: Total number of records where the test and result are known, as follows: (1) LAB_LOINC is not null and (2) RESULT_NUM is not null and (3) RESULT_MODIFIER is not in (null, NI, UN, OT) and (4) RESULT_UNIT is not in (null, NI, UN, OT)
- KNOWN_TEST_RESULT_NUM_SRCE_UNIT: Total number of records where the test and result are known, as follows: (1) LAB_LOINC is not null and (2) RESULT_NUM is not null and (3) RESULT_MODIFIER is not in (null, NI, UN, OT) and (4) SPECIMEN_SOURCE is not in (null, NI, UN, OT, UNK_SUB, SMPLS, SPECIMEN) and (5) RESULT_UNIT is not in (null, NI, UN, OT)
- KNOWN_TEST_NUM_RESULT_RANGE: Total number of records where the test, numeric result, and normal range are all known, as follows: (1) LAB_LOINC is not null and (2) RESULT_NUM is not null and (3) RESULT_MODIFIER is not in (null, NI, UN, OT) and (4) one of the following is true: (4a) NORM_MODIFIER_LOW='EQ' and NORM_MODIFIER_HIGH='EQ' and NORM_RANGE_LOW is not null and NORM_RANGE_HIGH is not null or (4b) NORM_MODIFIER_LOW in ('GT','GE') and NORM_MODIFER_HIGH='NO' and NORM_RANGE_LOW is not null and NORM_RANGE_HIGH is null or (4c) NORM_MODIFIER_HIGH in ('LE','LT') and NORM_MODIFIER_LOW='NO' and NORM_RANGE_HIGH is not null and NORM_RANGE_LOW is null.
- NMISS or NULL_N: Count of records with null or missing values for the specified field.

- ENCOUNTERID Orphan: An ENCOUNTERID which is not in the ENCOUNTER table and appears in any other table.
- PATID Orphan: A PATID which is not in the DEMOGRAPHIC table and appears in any other table.
- PROVIDERID orphan: A PROVIDERID which is not in the PROVIDER table and appears in any other table.
- RECORD_PCT: The percent of all records. Will be blank for rows with values of 0 or BT (below threshold).
- RECORD_N_RXCUI: Count of records with non-missing values for RXNORM_CUI.
- RECORD_N_LOINC: Count of records with non-missing values for LOINC.
- RESPONSE_DATE: Date the query package was run (ie, SAS system date).
- QUERY_PACKAGE: Query package name.
- RXNORM_CUI_TTY_TIER: The term type (TTY) that the RXNORM_CUI is mapped to. Tier 1: RXNORM_CUI_TTY in ('SCD','SBD','BPCK','GPCK'). Tier 2: RXNORM_CUI_TTY in ('SBDF','SCDF','SBDG','SCDG','SBDC','BN','MIN'). Tier 3: RXNORM_CUI_TTY in ('SCDC', 'PIN','IN'). Tier 4: RXNORM_CUI_TTY in ('DF','DFG'). NULL or missing=RXNORM_CUI_TTY='NULL or missing'.
- STAT: Descriptive statistic (e.g. minimum, maximum, median).
- TAG: CDM field name
- VALID_N: Number of records in a valid format. Used for fields without a prespecified value set.
- VISIT: As stated in the PCORnet Common Data Model, for the Encounter table, "each record will generally reflect a unique combination of PATID, ADMIT_DATE, PROVIDERID, and ENC_TYPE". Thus, a visit is a concatenation of PATID + ADMIT_DATE+ PROVIDERID + ENC_TYPE.

IV. Data Curation Query Output Tables

For table shells of each dataset, please refer to the Technical Specifications available on the Data Curation page on iMeet.

ID	PCORnet Table(s)	Output table	Output table description
1	CONDITION	cond_l3_condition	CONDITION frequency
2	CONDITION	cond_l3_n	Counts PATID, ENCOUNTERID, and CONDITIONID
3	CONDITION	cond_l3_rdate_y	REPORT_DATE year frequency
4	CONDITION	cond_l3_rdate_ym	REPORT_DATE year month frequency
5	CONDITION	cond_l3_source	CONDITION_SOURCE frequency
6	CONDITION	cond_l3_status	CONDITION_STATUS frequency
7	CONDITION	cond_l3_type	CONDITION_TYPE frequency
8	DEATH	death_l3_date_y	DEATH_DATE year frequency
9	DEATH	death_l3_date_ym	DEATH_DATE year month frequency
10	DEATH	death_l3_impute	DEATH_DATE_IMPUTE frequency
11	DEATH	death_l3_match	DEATH_MATCH_CONFIDENCE frequency
12	DEATH	death_l3_n	Counts non-missing, distinct, and missing PATID and DEATHID
13	DEATH	death_l3_source	DEATH_SOURCE frequency
14	DEATH	death_l3_source_ym	DEATH_SOURCE and DEATH_DATE year month crosstab
15	DEATH_CAUSE	deathc_l3_code	DEATH_CAUSE_CODE frequency
16	DEATH_CAUSE	deathc_l3_conf	DEATH_CAUSE_CONFIDENCE frequency
17	DEATH_CAUSE	deathc_l3_n	Counts PATID, DEATH_CAUSE, and DEATHCID
18	DEATH_CAUSE	deathc_l3_source	DEATH_CAUSE_SOURCE frequency
19	DEATH_CAUSE	deathc_l3_type	DEATH_CAUSE_TYPE frequency
20	DEMOGRAPHIC	dem_l3_ageyrsdist1	Descriptive statistics for age. Age is calculated as current age or age at death if death date is known. If multiple death records exist, the earlier death date is used.
21	DEMOGRAPHIC	dem_l3_ageyrsdist2	Age group frequency. Age is calculated as current age or age at death if death date is known. If multiple death records exist, the earlier death date is used.
22	DEMOGRAPHIC	dem_l3_genderdist	GENDER_IDENTITY frequency
23	DEMOGRAPHIC	dem_l3_orientdist	SEXUAL_ORIENTATION FREQUENCY
24	DEMOGRAPHIC	dem_l3_hispdist	HISPANIC frequency
25	DEMOGRAPHIC	dem_l3_n	Counts non-missing, distinct, and missing PATID
26	DEMOGRAPHIC	dem_l3_patpreflang	PAT_PREF_LANGUAGE_SPOKEN frequency

ID	PCORnet Table(s)	Output table	Output table description
27	DEMOGRAPHIC	dem_l3_racedist	RACE frequency
28	DEMOGRAPHIC	dem_l3_sexdist	SEX frequency
29	DIAGNOSIS	dia_l3_adate_y	ADMIT_DATE year frequency
30	DIAGNOSIS	dia_l3_adate_ym	ADMIT_DATE year month frequency
31	DIAGNOSIS	dia_l3_dx	DX frequency
32	DIAGNOSIS	dia_l3_dxtype	DX_TYPE frequency
33	DIAGNOSIS	dia_l3_dx_dxtype	DX and DX_TYPE crosstab
34	DIAGNOSIS	dia_l3_dxpoa	DX_POA frequency
35	DIAGNOSIS	dia_l3_dxsource	DX_SOURCE frequency
36	DIAGNOSIS	dia_l3_dxtype_adate_y	DX_TYPE and ADMIT_DATE year crosstab
37	DIAGNOSIS	dia_l3_dxtype_dxsource	DX_TYPE and DX_SOURCE crosstab
38	DIAGNOSIS	dia_l3_dxtype_enctype	DX_TYPE and ENC_TYPE crosstab
39	DIAGNOSIS	dia_l3_enctype	ENC_TYPE frequency
40	DIAGNOSIS	dia_l3_enctype_adate_ym	ENC_TYPE and ADMIT_DATE year month crosstab
41	DIAGNOSIS	dia_l3_n	Counts PATID, ENCOUNTERID, and DIAGNOSISID
42	DIAGNOSIS	dia_l3_origin	DX_ORIGIN frequency
43	DIAGNOSIS	dia_l3_pdx	PDX frequency
44	DIAGNOSIS	dia_l3_pdx_enctype	PDX and ENC_TYPE crosstab
45	DIAGNOSIS	dia_l3_pdxgrp_enctype	PDX group and ENC_TYPE crosstab
46	DIAGNOSIS	dia_l3_dxdate_y	DX_DATE year frequency
47	DIAGNOSIS	dia_l3_dxdate_ym	DX_DATE year month frequency
48	DIAGNOSIS	dia_l3_dcgroup	Count of PATID by data curation diagnosis group
49	DISPENSING	disp_l3_ndc	NDC frequency
50	DISPENSING	disp_l3_ddate_y	DISPENSE_DATE year frequency
51	DISPENSING	disp_l3_ddate_ym	DISPENSE_DATE year month frequency
52	DISPENSING	disp_l3_dispamt_dist	Descriptive statistics for DISPENSE_AMT
53	DISPENSING	disp_l3_dose_dist	Descriptive statistics for DISPENSE_DOSE_DISP
54	DISPENSING	disp_l3_doseunit	DISPENSE_DOSE_DISP_UNIT frequency
55	DISPENSING	disp_l3_route	DISPENSE_ROUTE frequency
56	DISPENSING	disp_l3_n	Counts non-missing, distinct, and missing PATID, DISPENSINGID. PRESCRIBINGID, NDC, and valid NDCs. Valid NDCs are 11 digits with no dashes, ie. HIPAA format.
57	DISPENSING	disp_l3_supdist2	Record count by category of RX_DAYS_SUPP
58	DISPENSING	disp_l3_source	DISPENSE_SOURCE frequency
59	ENCOUNTER	enc_l3_adate_y	ADMIT_DATE year frequency
60	ENCOUNTER	enc_l3_adate_ym	ADMIT_DATE year month frequency
61	ENCOUNTER	enc_l3_admsrc	ADMITTING_SOURCE frequency
62	ENCOUNTER	enc_l3_dash2	Counts the number of patients with any AV, ED, IP, EI, or OS encounter

ID	PCORnet Table(s)	Output table	Output table description
			record with a populated ADMIT_DATE during the designated period prior to the maximum ADMIT_DATE. If the maximum ADMIT_DATE is in the future, the current date is used instead.
63	ENCOUNTER	enc_l3_ddate_y	DISCHARGE_DATE year frequency
64	ENCOUNTER	enc_l3_ddate_ym	DISCHARGE_DATE year month frequency
65	ENCOUNTER	enc_l3_disdisp	DISCHARGE_DISPOSITION frequency
66	ENCOUNTER	enc_l3_disstat	DISCHARGE_STATUS frequency
67	ENCOUNTER	enc_l3_drg_type	DRG_TYPE frequency
68	ENCOUNTER	enc_l3_enctype	ENC_TYPE frequency. (*Note:* Visits are a concatenation of PATID + PROVIDER_ID + ENC_TYPE + ADMIT_DT. ELIG_RECORD_N is a count of records where all fields used to define a visit are populated)
69	ENCOUNTER	enc_l3_enctype_adate_y	ENC_TYPE and ADMIT_DATE year month crosstab
70	ENCOUNTER	enc_l3_enctype_adate_ym	ENC_TYPE and ADMIT_DATE year month crosstab
71	ENCOUNTER	enc_l3_enctype_admsrc	ENC_TYPE by ADMITTING_SOURCE crosstab
72	ENCOUNTER	enc_l3_enctype_ddate_ym	ENC_TYPE and DISCHARGE_DATE year month crosstab
73	ENCOUNTER	enc_l3_enctype_disdisp	ENC_TYPE and DISCHARGE_DISPOSITION crosstab
74	ENCOUNTER	enc_l3_enctype_disstat	ENC_TYPE and DISCHARGE_STATUS crosstab
75	ENCOUNTER	enc_l3_enctype_drg	ENC_TYPE and DRG_TYPE crosstab
76	ENCOUNTER	enc_l3_n	Counts non-missing, distinct, and missing PATID, ENCOUNTERID, and PROVIDERID, and FACILITYID
77	ENCOUNTER	enc_l3_payertype1	PAYER_TYPE_PRIMARY frequency
78	ENCOUNTER	enc_l3_payertype2	PAYER_TYPE_SECONDARY frequency
79	ENCOUNTER	enc_l3_facilitytype	FACILITY_TYPE frequency
80	ENCOUNTER	enc_l3_facilityloc	FACILITY_LOCATION frequency
81	ENCOUNTER	enc_l3_facilitytype_facilityloc	FACILITY_TYPE and FACILITY_LOCATION(first 3 digits of a zip code) crosstab
82	ENROLLMENT	enr_l3_basedist	ENR_BASIS frequency
83	ENROLLMENT	enr_l3_n	Counts non-missing, distinct, and missing PATID, ENR_START_DATE, and ENROLLID (combination of

ID	PCORnet Table(s)	Output table	Output table description
			PATID, ENR_START_DATE, and ENR_BASIS)
84	ENROLLMENT	enr_l3_chart	CHART frequency
85	LAB_RESULT_CM	lab_l3_abn	ABN_IND frequency
86	LAB_RESULT_CM	lab_l3_dcgroup	Frequency by DC_LAB_GROUP
87	LAB_RESULT_CM	lab_l3_high	NORM_MODIFIER_HIGH frequency
88	LAB_RESULT_CM	lab_l3_loc	RESULT_LOC frequency
89	LAB_RESULT_CM	lab_l3_loinc	LAB_LOINC frequency
90	LAB_RESULT_CM	lab_l3_loinc_result_num	RESULT_NUM descriptive statistics by LAB_LOINC code
91	LAB_RESULT_CM	lab_l3_loinc_source	LAB_LOINC and SPECIMEN_SOURCE crosstab for a subset of LOINC codes
92	LAB_RESULT_CM	lab_l3_low	NORM_MODIFIER_LOW frequency
93	LAB_RESULT_CM	lab_l3_mod	RESULT_MODIFIER frequency
94	LAB_RESULT_CM	lab_l3_n	Counts non-missing, distinct, and missing PATID, LAB_RESULT_CM_ID, and ENCOUNTERID
95	LAB_RESULT_CM	lab_l3_priority	PRIORITY frequency
96	LAB_RESULT_CM	lab_l3_px_pxtype	LAB_PX and LAB_PXTYPE crosstab
97	LAB_RESULT_CM	lab_l3_px_type	LAB_PX_TYPE frequency
98	LAB_RESULT_CM	lab_l3_qual	RESULT_QUAL frequency
99	LAB_RESULT_CM	lab_l3_raw_name	RAW_LAB_NAME frequency
100	LAB_RESULT_CM	lab_l3_rdate_y	RESULT_DATE year frequency
101	LAB_RESULT_CM	lab_l3_rdate_ym	RESULT_DATE year month frequency
102	LAB_RESULT_CM	lab_l3_recordc	Frequency of records with varying levels of completeness across variables
103	LAB_RESULT_CM	lab_l3_snomed	RESULT_SNOMED frequency
104	LAB_RESULT_CM	lab_l3_source	SPECIMEN_SOURCE frequency
105	LAB_RESULT_CM	lab_l3_unit	RESULT_UNIT frequency
106	LAB_RESULT_CM	lab_l3_loinc_result_num_5y	RESULT_NUM descriptive statistics by LAB_LOINC code in the 5 years of the lookback period
107	LAB_RESULT_CM	lab_l3_loinc_source_5y	LAB_LOINC and SPECIMEN_SOURCE crosstab for a subset of LOINC codes in the 5 years of lookback period

ID	PCORnet Table(s)	Output table	Output table description
108	LAB_RESULT_CM	lab_l3_n_5y	Counts PATID, LAB_RESULT_CM_ID, and ENCOUNTERID in the 5 years of lookback period
109	LAB_RESULT_CM	lab_l3_recordc_5y	Frequency of records with varying levels of completeness across variables in the 5 years of lookback period
110	LAB_RESULT_CM	lab_l3_rsource	LAB_RESULT_SOURCE frequency
111	LAB_RESULT_CM	lab_l3_lsource	LAB_LOINC_SOURCE frequency
112	LAB_RESULT_CM	lab_l3_loinc_unit	LAB_LOINC and RESULT_UNIT crosstab
113	MED_ADMIN	medadm_l3_doseadm	Descriptive statistics for MEDADMIN_DOSE_ADM
114	MED_ADMIN	medadm_l3_doseadmunit	MEDADMIN_DOSE_ADMIN_UNIT frequency
115	MED_ADMIN	medadm_l3_n	Counts MEDADMINID and PATID
116	MED_ADMIN	medadm_l3_route	MEDADMIN_ROUTE frequency
117	MED_ADMIN	medadm_l3_source	MEDADMIN_SOURCE frequency
118	MED_ADMIN	medadm_l3_type	MEDADMIN_TYPE frequency
119	MED_ADMIN	medadm_l3_sdate_y	MEDADMIN_ START_DATE year frequency
120	MED_ADMIN	medadm_l3_sdate_ym	MEDADMIN_ START_DATE year month frequency
121	MED_ADMIN	medadm_l3_code_type	MEDADMIN_TYPE and MEDADMIN_CODE crosstab
122	MULTIPLE	datamart_all	DataMart metadata including variable names, variable lengths, data types and number of observations. Used to assess conformance to the required SAS structure for the PCORnet Common Data Model (CDM) v3.1.
123	MULTIPLE	elapsed_all	Displays the query start time, query end time, and query run time for each table created by the data_curation_all program, the cumulative run time for the program and the dataset loading time. Will only be present if the 'all' option is used.
124	MULTIPLE	elapsed_main	Displays the query start time, query end time, and query run time for each table created by the

ID	PCORnet Table(s)	Output table	Output table description
			data_curation_main program, the cumulative run time for the program and the dataset loading time. Will only be present if the programs are run separately.
125	MULTIPLE	elapsed_lab	Displays the query start time, query end time, and query run time for each table created by the data_curation_lab program, the cumulative run time for the program and the dataset loading time. Will only be present if the programs are run separately.
126	MULTIPLE	elapsed_xtbl	Displays the query start time, query end time, and query run time for each table created by the data_curation_xtbl program, the cumulative run time for the program and the dataset loading time. The DATAMART_ALL table is not included because it is just a print. Will only be present if the programs are run separately.
127	MULTIPLE	xtbl_l3_dash1	Counts the number of patients with any VITAL record with a populated MEASURE_DATE and a diagnosis record with a populated ADMIT_DATE and DX during the designated period prior to the maximum DIAGNOSIS.ADMIT_DATE. If the maximum ADMIT_DATE is in the future, the current date is substituted.
128	MULTIPLE	xtbl_l3_dash2	Counts the number of patients with any VITAL record with a populated MEASURE_DATE and a DIAGNOSIS record with a populated DX and ADMIT_DATE and either a PRESCRIBING record with a populated RXNORM_CUI and RX_START_DATE *or* a DISPENSING record with a populated DISPENSE_DATE and NDC during the designated period of time prior to the maximum DIAGNOSIS.ADMIT_DATE. If the maximum ADMIT_DATE is in the future, the current date is substituted.
129	MULTIPLE	xtbl_l3_dash3	Counts the number of patients with any VITAL record with a populated MEASURE_DATE and a DIAGNOSIS record with a populated DX and ADMIT_DATE and either (a PRESCRIBING record with a populated RXNORM_CUI and RX_ORDER_DATE *or* a

ID	PCORnet Table(s)	Output table	Output table description
			DISPENSING record with a populated DISPENSE_DATE and NDC) and a LAB_RESULT_CM record and RESULT_DATE during the designated period of time prior to the maximum DIAGNOSIS.ADMIT_DATE. If the maximum ADMIT_DATE is in the future, the current date is substituted.
130	MULTIPLE	xtbl_l3_date_logic	Identifies illogical relationships between BIRTH_DATE, DEATH_DATE, and key dates in other tables
131	MULTIPLE	xtbl_l3_dates	Descriptive statistics and counts of records with future dates or dates prior to January 2010 for all date fields.
132	MULTIPLE	xtbl_l3_lab_enctype	# of records and patients with lab records by encounter type.
133	MULTIPLE	xtbl_l3_metadata	HARVEST fields; maximum refresh date; query package; response date; low cell count threshold; operating system; SAS version and packages; SAS datastore (data or views); and query run time. There should only be 1 record in this table. The DATAMARTID and REFRESH_MAX fields are used extensively throughout the query package.
134	MULTIPLE	xtbl_l3_mismatch	Counts the number of records where there is a mismatch between a parent and child table. These checks include ENCOUNTERIDs that are not in the ENCOUNTER table; PATIDs that are not in the DEMOGRAPHIC table; PROVIDERIDs that are not in the PROVIDER table; and discordance in the fields that are replicated from the ENCOUNTER table to the PROCEDURES and DIAGNOSIS tables.
135	MULTIPLE	xtbl_l3_non_unique	Identify encounters which are associated with more than 1 patient (PATID) in the same table
136	MULTIPLE	xtbl_l3_pres_enctype	# of records and patients with prescribing records by encounter type.
137	MULTIPLE	xtbl_l3_times	Descriptive statistics for all time fields.
138	MULTIPLE	xtbl_l3_race_enc	# of records and patients by RACE among patients with at least 1 encounter after 2009(from 2010)
139	OBS_CLIN	obsclin_l3_n	Counts OBSCLINID and PATID
140	OBS_CLIN	obsclin_l3_code_type	OBSCLIN_TYPE and OBSCLIN_CODE crosstab

ID	PCORnet Table(s)	Output table	Output table description
141	OBS_CLIN	obsclin_l3_mod	OBSCLIN_RESULT_MODIFIER frequency
142	OBS_CLIN	obsclin_l3_qual	OBSCLIN_RESULT_QUAL frequency
143	OBS_CLIN	obsclin_l3_runit	OBSCLIN_RESULT_UNIT frequency
144	OBS_CLIN	obsclin_l3_type	OBSCLIN_TYPE frequency
145	OBS_CLIN	obsclin_l3_source	OBSCLIN_SOURCE frequency
146	OBS_GEN	obsgen_l3_mod	OBSGEN_RESULT_MODIFIER frequency
147	OBS_GEN	obsgen_l3_tmod	OBSGEN_TABLE_MODIFIER frequency
148	OBS_GEN	obsgen_l3_n	Counts OBSGENID, PATID, ENCOUNTERID and OBSGEN_PROVIDERID
149	OBS_GEN	obsgen_l3_qual	OBSGEN_RESULT_QUAL frequency
150	OBS_GEN	obsgen_l3_runit	OBSGEN_RESULT_UNIT frequency
151	OBS_GEN	obsgen_l3_type	OBSGEN_TYPE frequency
152	OBS_GEN	obsgen_l3_code_type	OBSGEN_TYPE and OBSGEN_CODE crosstab
153	OBS_GEN	obsgen_l3_source	OBSGEN_SOURCE frequency
154	PCORNET_TRIAL	trial_l3_n	Counts PATID, TRIALID, PARTICIPANTID, and TRIAL_KEY
155	PRESCRIBING	pres_l3_basis	RX_BASIS frequency
156	PRESCRIBING	pres_l3_dispaswrtn	RX_DISPENSE_AS_WRITTEN frequency
157	PRESCRIBING	pres_l3_freq	RX_FREQUENCY frequency
158	PRESCRIBING	pres_l3_n	Counts non-missing, distinct, and missing PATID, PRESCRIBINGID, ENCOUNTERID, and RX_PROVIDERID
159	PRESCRIBING	pres_l3_odate_y	RX_ORDER_DATE year frequency
160	PRESCRIBING	pres_l3_odate_ym	RX_ORDER_DATE year month frequency
161	PRESCRIBING	pres_l3_prnflag	RX_PRN_FLAG frequency
162	PRESCRIBING	pres_l3_rxcui	RXCUI frequency
163	PRESCRIBING	pres_l3_rxcui_rxsup	Descriptive statistics for RX_DAYS_SUPPLY by RXNORM_CUI
164	PRESCRIBING	pres_l3_rxcui_tier	RXNORM_CUI frequency by tier of term type
165	PRESCRIBING	pres_l3_rxdoseform	RX_DOSE_FORM frequency
166	PRESCRIBING	pres_l3_rxdoseodr_dist	Descriptive statistics for RX_DOSE_ORDERED
167	PRESCRIBING	pres_l3_rxdoseodrunit	RX_DOSE_ORDERED_UNIT frequency
168	PRESCRIBING	pres_l3_rxqty_dist	Descriptive statistics for RX_QUANTITY

ID	PCORnet Table(s)	Output table	Output table description
169	PRESCRIBING	pres_l3_rxrefill_dist	Descriptive statistics for RX_REFILLS
170	PRESCRIBING	pres_l3_route	RX_ROUTE frequency
171	PRESCRIBING	pres_l3_source	RX_SOURCE frequency
172	PRESCRIBING	pres_l3_rawrxmed	RAW_RX_MED_NAME frequency
173	PRESCRIBING	pres_l3_supdist2	Record count by category of RX_DAYS_SUPPLY
174	PRESCRIBING	pres_l3_rxcui_5y	RXNORM_CUI frequency and term type information in the 5 years of lookback period
175	PRESCRIBING	pres_l3_rxcui_tier_5y	RXNORM_CUI frequency by tier of term type in the 5 years of lookback period
176	PRO_CM	procm_l3_cat	PRO_CAT frequency
177	PRO_CM	procm_l3_itemfullname	PRO_ITEM_FULLNAME frequency
178	PRO_CM	procm_l3_loinc	PRO_LOINC frequency
179	PRO_CM	procm_l3_itemnm	PRO_ITEM_NAME frequency
180	PRO_CM	procm_l3_measure_fullname	PRO_MEASURE_FULLNAME frequency
181	PRO_CM	procm_l3_measurenm	PRO_MEASURE_NAME frequency
182	PRO_CM	procm_l3_method	PRO_METHOD frequency
183	PRO_CM	procm_l3_mode	PRO_MODE frequency
184	PRO_CM	procm_l3_n	Counts PRO_CM_ID, PATID, and ENCOUNTERID
185	PRO_CM	procm_l3_pdate_y	PRO_DATE year frequency
186	PRO_CM	procm_l3_pdate_ym	PRO_DATE year month frequency
187	PRO_CM	procm_l3_type	PRO_TYPE FREQUENCY
188	PRO_CM	procm_l3_source	PRO_SOURCE frequency
189	PROCEDURES	pro_l3_adate_y	ADMIT_DATE year frequency
190	PROCEDURES	pro_l3_adate_ym	ADMIT_DATE year month frequency
191	PROCEDURES	pro_l3_enctype	ENC_TYPE frequency

ID	PCORnet Table(s)	Output table	Output table description
192	PROCEDURES	pro_l3_enctype_adate_ym	ENC_TYPE and ADMIT_DATE year month crosstab
193	PROCEDURES	pro_l3_n	Counts non-missing, distinct, and missing PATID, ENCOUNTERID, and PROCEDURESID
194	PROCEDURES	pro_l3_ppx	PPX FREQUENCY
195	PROCEDURES	pro_l3_px	PX frequency
196	PROCEDURES	pro_l3_pxtype	PX_TYPE frequency
197	PROCEDURES	pro_l3_px_pxtype	PX and PX_TYPE crosstab
198	PROCEDURES	pro_l3_pxdate_y	PX_DATE year frequency
199	PROCEDURES	pro_l3_pxsource	PX_SOURCE frequency
200	PROCEDURES	pro_l3_pxtype_adate_y	PX_TYPE and ADMIT_DATE year crosstab
201	PROCEDURES	pro_l3_pxtype_enctype	PX_TYPE and ENC_TYPE crosstab
202	PROCEDURES	pro_l3_dcgroup	Count of PATID by data curation procedures group
203	PROVIDER	prov_l3_n	Counts PROVIDERID and PROVIDER_NPI
204	PROVIDER	prov_l3_npiflag	PROVIDER_NPI_FLAG frequency
205	PROVIDER	prov_l3_specialty	PROVIDER_SPECIALTY_PRIMARY frequency
206	PROVIDER	prov_l3_specialty_group	PROVIDER_SPECIALTY_PRIMARY group frequency
207	PROVIDER	prov_l3_sex	PROVIDER_SEX frequency
208	VITAL	vit_l3_bmi	BMI frequency
209	VITAL	vit_l3_bp_position_type	BP_POSITION_TYPE frequency
210	VITAL	vit_l3_dash1	Counts the number of patients with any vital record with a populated MEASURE_DATE during the designated period prior to the maximum MEASURE_DATE. If the maximum MEASURE_DATE is in the future, the current date is substituted.
211	VITAL	vit_l3_diastolic	DIASTOLIC frequency
212	VITAL	vit_l3_ht	HT frequency
213	VITAL	vit_l3_ht_dist	Descriptive statistics for HT
214	VITAL	vit_l3_mdate_y	MEASURE_DATE year frequency
215	VITAL	vit_l3_mdate_ym	MEASURE_DATE year month frequency

ID	PCORnet Table(s)	Output table	Output table description
216	VITAL	vit_l3_n	Counts non-missing, distinct, and missing PATID, ENCOUNTERID, and VITALID
217	VITAL	vit_l3_smoking	SMOKING frequency
218	VITAL	vit_l3_systolic	SYSTOLIC frequency
219	VITAL	vit_l3_tobacco	TOBACCO frequency
220	VITAL	vit_l3_tobacco_type	TOBACCO_TYPE frequency
221	VITAL	vit_l3_vital_source	VITAL_SOURCE frequency
222	VITAL	vit_l3_wt	WT frequency
223	VITAL	vit_l3_wt_dist	Descriptive statistics for WT
224	LDS_ADDRESS_HISTORY	ldsadrs_l3_n	Counts for PATID, ADDRESSID
225	LDS_ADDRESS_HISTORY	ldsadrs_l3_adrsuse	ADDRESS_USE frequency
226	LDS_ADDRESS_HISTORY	ldsadrs_l3_adrstype	ADDRESS_TYPE frequency
227	LDS_ADDRESS_HISTORY	ldsadrs_l3_adrspref	ADDRESS_PREFERRED frequency
228	LDS_ADDRESS_HISTORY	ldsadrs_l3_adrscity	ADDRESS_CITY frequency
229	LDS_ADDRESS_HISTORY	ldsadrs_l3_adrsstate	ADDRESS_STATE frequency
230	LDS_ADDRESS_HISTORY	ldsadrs_l3_adrszip5	ADDRESS_ZIP5 frequency
231	LDS_ADDRESS_HISTORY	ldsadrs_l3_adrszip9	ADDRESS_ZIP9 frequency
232	IMMUNIZATION	immune_l3_n	Counts for PATID, IMMUNIZATIONID, ENCOUNTERID, PROCEDURESID, VX_PROVIDERID
233	IMMUNIZATION	immune_l3_rdate_y	VX_RECORD_DATE year frequency
234	IMMUNIZATION	immune_l3_rdate_ym	VX_RECORD_DATE year month frequency
235	IMMUNIZATION	immune_l3_adate_y	VX_ADMIN_DATE year frequency
236	IMMUNIZATION	immune_l3_adate_ym	VX_ADMIN_DATE year month frequency
237	IMMUNIZATION	immune_l3_codetype	VX_CODE_TYPE frequency
238	IMMUNIZATION	immune_l3_code_codetype	VX_CODE and VX_CODETYPE crosstab
239	IMMUNIZATION	immune_l3_status	VX_STATUS frequency
240	IMMUNIZATION	immune_l3_statusreason	VX_STATUS_REASON frequency
241	IMMUNIZATION	immune_l3_source	VX_SOURCE frequency
242	IMMUNIZATION	immune_l3_dose_dist	VX_DOSE descriptive statistics
243	IMMUNIZATION	immune_l3_doseunit	VX_DOSE_UNIT frequency

ID	PCORnet Table(s)	Output table	Output table description
244	IMMUNIZATION	immune_l3_route	VX_ROUTE frequency
245	IMMUNIZATION	immune_l3_bodysite	VX_BODY_SITE frequency
246	IMMUNIZATION	immune_l3_manufacturer	VX_MANUFACTURER frequency
247	IMMUNIZATION	immune_l3_lotnum	VX_LOT_NUM frequency
248	HASH_TOKEN	hash_l3_n	Count for PATIDs
249	HASH_TOKEN	hash_l3_token_availability	Count for PATIDs with all possible combination of tokens

V. Empirical Data Curation Report

The data from all data curation query output tables except for the *elapsed* datasets is compiled into a normalized dataset. The Empirical Data Curation (EDC) Report is produced from this dataset. The EDC Report summarizes key information from the query ouput tables and identifies exceptions to the PCORnet Data Checks. The report includes a table of contents, a data check exception summary, and up to 52 tables and charts, depending upon the number of CDM tables which are populated. The table of contents is below.

Section	Table	Table Description	Data Check(s)
n/a	n/a	Data Check Exception Summary	n/a
Section I: Descriptive Information	Table IA	Demographic Summary	n/a
	Table IB	Potential Pools of Patients	3.04, 3.05
	Table IC	Height, Weight, and Body Mass Index (BMI)	n/a
	Table ID	Records, Patients, Encounters, and Date Ranges by Table	n/a
	Table IE	Records Per Table by Encounter Type	n/a
	Table IF	Date Obfuscation or Imputation	n/a
	Table IG	Lab Results For Selected Lab Tests	3.13
	Table IH	Patients with Selected Diagnoses	n/a
	Table II	Patients with Selected Procedures	n/a
	Chart IA	Trend in Vital Measures by Measurement Date, Past 5 Years	n/a
	Chart IB	Trend in Encounters by Admit Date and Encounter Type, Past 5 Years	n/a
	Chart IC	Trend in Institutional Encounters by Discharge Date and Encounter Type, Past 5 Years	n/a
	Chart ID	Trend in Laboratory Results by Result Date, Past 5 Years	n/a
	Chart IE	Trend in Prescribed Medications by Rx Order Date, Past 5 Years	n/a
	Chart IF	Trend in Dispensed Medications by Dispense Date, Past 5 Years	n/a
	Chart IG	Trend in Administered Medications by Start Date, Past 5 Years	n/a
	Chart IH	Trend in Condition Records by Report Date, Past 5 Years	n/a
	Chart II	Trend in Death Records by Death Date and Source, Past 5 Years	n/a
	Chart IJ	Trend in Immunization Records by Vx Record Date, Past 5 Years	n/a
Section II: Data Model Conformance	Table IIA	Primary Key Errors	1.05
	Table IIB	Values Outside of Common Data Model (CDM) Specifications	1.06
	Table IIC	Non-Permissible Missing Values	1.07
	Table IID	Diagnostic Errors	1.01, 1.02, 1.03, 1.04
	Table IIE	Orphan Records, Replication Errors and Encounter Duplication	1.08, 1.09,1.10, 1.11, 1.12,1.14
	Table IIF	Potential Code Errors	1.13
Section III: Data Plausibility	Table IIIA	Future Dates	2.01
	Table IIIB	Records with Extreme Values	2.02
	Table IIIC	Illogical Dates	2.03
	Table IIID	Encounters Per Visit and Per Patient	2.04
	Table IIIE	Laboratory Result Specimen Source Discrepancies	2.05

Section	Table	Table Description	Data Check(s)
	Table IIIF	Quantitative Lab Result Outliers, Selected Tests	2.06
	Table IIIG	Monthly Record Volume Outliers, Selected Domains	2.08
	Chart IIIA	Monthly Record Volume Outliers, Encounters	2.08
	Chart IIIB	Monthly Record Volume Outliers, Diagnoses	2.08
	Chart IIIC	Monthly Record Volume Outliers, Procedures	2.08
	Chart IIID	Monthly Record Volume Outliers, Vitals	2.08
	Chart IIIE	Monthly Record Volume Outliers, Prescribing	2.08
	Chart IIIF	Monthly Record Volume Outliers, Labs	2.08
Section IV: Data Completeness and Plausibility	Table IVA	Diagnosis Records Per Encounter, Overall and by Encounter Type	3.01
	Chart IVA	Diagnosis Records Per Encounter by Admit Date and Encounter Type, Past 5 Years	n/a
	Table IVB	Procedure Records Per Encounter, Overall and by Encounter Type	3.02
	Chart IVB	Procedure Records Per Encounter by Admit Date and Encounter Type, Past 5 Years	n/a
	Table IVC	Missing or Unknown Values, Required Tables	3.03
	Table IVD	Missing or Unknown Values, Optional Tables	3.03
	Table IVE	Principal Diagnoses for Institutional Encounters	2.07, 3.06
	Table IVF	Data Latency and Completeness of Encounter, Diagnosis and Procedure Data, Past 2 Years	3.07
	Table IVG	Data Latency and Completeness of Vital, Prescription, and Lab Data, Past 2 Years	3.11
	Table IVH	RXNORM Term Type Mapping, Overall and Past 5 Years	3.08
	Table IVI	Laboratory Result Data Completeness, Overall and Past 5 Years	3.09, 3.10, 3.12
	Table IVI_Ref	Laboratory Result Data Completeness Definitions	n/a
Section V: Data Persistence	Table VA	Changes in Tables	4.01
	Table VB	Changes in Selected Encounter Types and Domains	4.02
	Table VC	Changes in Selected Code Types	4.03

VI. Program Package File Structure

Each request package distributed by PCORnet's DRN OC contains several sub-folders to organize program inputs and outputs. The subfolders must reside within an outer folder labeled with the query name designated in the DRN Query Tool. The subfolders are as follows:

- ***dmlocal***: Contains output generated by the request that should be saved locally but not returned to DRN OC. Output may be used locally or to facilitate follow-up queries.
- ***drnoc***: Contains output generated by the request that should be returned to the DRN OC via the PCORnet DRN Query Tool. These tables consist of aggregate data/output and transfer the minimum required to answer the analytic question.
- ***sasprograms***: Contains the master SAS program that must be edited and then executed locally.
- ***infolder****:* Contains all input programs and files needed to execute the request. These are created for each request by the DRN OC Data Curation team; the contents of this folder should not be edited.

VII. Files Included in Query Request

The following files are included in the Zip file distributed with the query request.

Cycle 8 Data Curation Query Package Checklist.pdf
Data Curation Query Package v5.13 Work Plan.pdf

infolder

1. data_curation_query_base.sas
2. data_curation_query_lab.sas
3. data_curation_query_main.sas
4. data_curation_print.sas
5. data_curation_query_xtbl.sas
6. dc_reference.cpt. This file includes 17 SAS datasets. Five (5) datasets are created for this query: lab_loinc_ref, lab_dcgroup_ref, rxnorm_cui_ref, dx_dcgroup_ref, and px_dcgroup_ref. Ten (10) datasets are derived from the Valuesets tab of the parseable file (2020-06-17-PCORnet-Common-Data-Model-v5dot1-parseable.xlsx): facility_type, pat_pref_language_spoken, payer_type_, provider_specialty_primary, _qual, _route, rx_dose_form, state, specimen_source, _unit, vx_body_site and vx_manufacturer.
7. edc_prep.sas
8. edc_reference.cpt. This file includes 12 SAS datasets: dc_summary, dc_tables, footers, headers, lab_volume_ref, missingness, pediatric_datamarts, required_structure, specimen_source_category, tbl_ivi_ref, toc and q2_stat_dlg_loinc.
9. edc_report.sas
10. edc_template.sas
11. normalization.sas
12. potential_code_errors.sas

sas_programs

1. 01_run_code_errors.sas
2. 02_run_queries.sas
3. 03_run_edc_prep.sas
4. 04_run_edc_report.sas

VIII. Output Files

Local files (*dmlocal* folder). DMID=DataMart ID; DATE=response date

Produced by	File description
pcornet_code_errors.sas	code_summary (SAS dataset and csv file) *Up to 10 error files, if relevant code types are present:* bad_condition (SAS dataset and csv file) bad_dx (SAS dataset and csv file) bad_px (SAS dataset and csv file) bad_pres (SAS dataset and csv file) bad_lab (SAS dataset and csv file) bad_disp (SAS dataset and csv file) bad_medadmin (SAS dataset and csv file) bad_obsclin (SAS dataset and csv file) bad_obsgen (SAS dataset and csv file) bad_immunization (SAS dataset and csv file)

data_curation_query_base.sas;
data_curation_query_main.sas;
data_curation_query_lab.sas;
data_curation_query_xtbl.sas
	Up to 248 output tables (SAS datasets and csv files; see section IV) and set.log (contains the output results of the PROC SETINIT procedure. The set.log information is used to populate XTBL_L3_METADATA.

normalization.sas	[DMID]_[DATE]_dc_norm.sas7dat

Files to be returned to the DRN OC (*drnoc* folder). DMID=DataMart ID; DATE=response date.

File name	Program produced by	File description
[DMID]_[DATE]_potential_code_errors.log	potential_code_errors.sas	The SAS log file for the program. Must be checked for errors and warnings.
[DMID]_[DATE]_Potential_Code_Errors.pdf	potential_code_errors.sas	The report produced by the program.
[DMID]_[DATE]_code_error_summary.cpt	potential_code_errors.sas	A SAS transport file containing the code error summary dataset produced by the program

If all data curation queries are run at once
[DMID]_[DATE]_data_curation_all.cpt
or
If data curation queries are run separately
[DMID]_[DATE]_data_curation_main.cpt
[DMID]_[DATE]_data_curation_lab.cpt
[DMID]_[DATE]_data_curation_xtbl.cpt
	data_curation_query_base.sas;
	data_curation_query_main.sas;
	data_curation_query_lab.sas;
	data_curation_query_xtbl.sas
		A SAS transport file containing all the SAS datasets produced by the program(s).

If all data curation queries are run at once
[DMID]_[DATE]_data_curation_all.pdf
or
If data curation queries are run separately
[DMID]_[DATE]_data_curation_main.pdf
[DMID]_[DATE]_data_curation_lab.pdf
[DMID]_[DATE]_data_curation_xtbl.pdf
	data_curation_query_base.sas;
	data_curation_query_main.sas;
	data_curation_query_lab.sas;
	data_curation_query_xtbl.sas;
	data_curation_print.sas
		A PDF containing a partial print of the output tables for the benefit of non-programmers. For ease of readibility, it excludes the first three columns of the table (DataMartID, Response Date, and Query Package), and large tables

File name	Program produced by	File description
		are limited to the 100 most frequent observations. Empty tables are not printed.

If all data curation queries are run at once
[DMID]_[DATE]_data_curation_all.log
[DMID]_[DATE]_data_curation_base.log
or
If data curation queries are run separately
[DMID]_[DATE]_data_curation_base.log
[DMID]_[DATE]_data_curation_main.log
[DMID]_[DATE]_data_curation_lab.log
[DMID]_[DATE]_data_curation_xtbl.log
	data_curation_query_base.sas;
	data_curation_query_main.sas;
	data_curation_query_lab.sas;
	data_curation_query_xtbl.sas
		The SAS log files for the programs. Must be checked for errors and warnings.

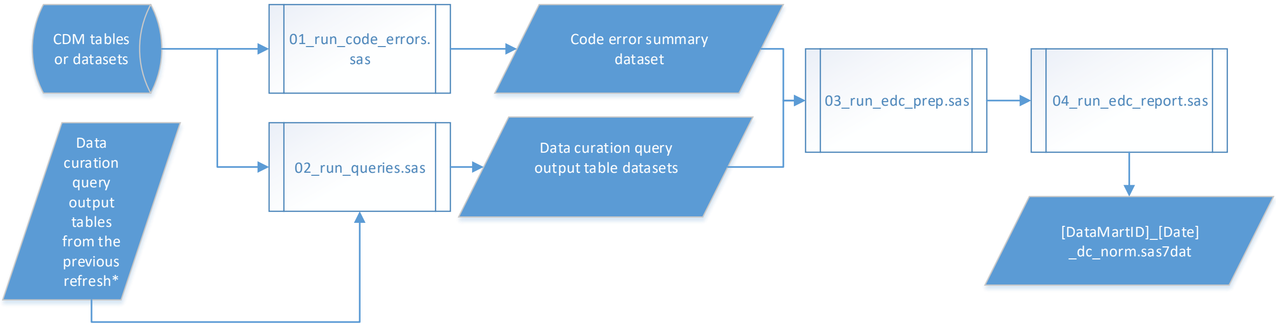
[DMID]_[DATE]_data_curation_progress_report.rtf	data_curation_query_base.sas;
	data_curation_query_main.sas;
	data_curation_query_lab.sas;
	data_curation_query_xtbl.sas
		A rtf file containing table names and their processing time

[DMID]_[DATE]_dc_norm.cpt	normalization.sas	A SAS transport file containing a normalized version of all data curation query output tables except the *elapsed* datasets.
[DMID]_[DATE]_normalization.log	normalization.sas	The SAS log file for the program. Must be checked for errors and warnings.
[DMID]_[DATE]_ EDCRPT.log	edc_report.sas	The SAS log file for the program. Must be checked for errors and warnings.
[DMID]_[DATE]_EDCRPT.pdf	edc_report.sas	The report produced by the program.

IX. Query Input and Output Diagram

The diagram below ilustrates how the query package uses information from the CDM tables and prior data curation results to produce the datasets in the *dmlocal* folder.

Required *dmlocal* inputs and outputs


CDM tables or datasets
01_run_code_errors.sas
Code error summary dataset
Data curation query output tables from the previous refresh*
02_run_queries.sas
Data curation query output table datasets
03_run_edc_prep.sas
04_run_edc_report.sas
[DataMartID]_[Date]_dc_norm.sas7dat

*These must match the EDC Report that was approved by the Coordinating Center

X. Responding to the Query Package

1) Prepare for the query as instructed in the Query Package Checklist.
2) Go to the DataMart Client and open the query package. Extract the contents, save them locally as described in Sections VI, and create the *drnoc* and *dmlocal* folders.
3) If the CDM data is stored in database tables, do the following. Otherwise proceed to Step 4.
   a) Consider compressing large tables to improve query response time.
   b) Modify the user inputs to use appropriate SAS/ACCESS options on a LIBNAME statement so that the program knows where to find the database tables. The examples below show connection information for an Oracle database; connecting to other database systems may require different connection information.
      (1) In the *sasprograms* folder, open **01_run_code_errors.sas, 02_run_queries.sas** and **04_run_edc_report.sas** and edit the `dpath` variable to include the appropriate database connection information. Be sure to use the `%str()` function to mask the embedded equal signs. For example: `%let dpath = %str(oracle user=”myuserid” orapw=mypasswd path=mydbname schema=myschema);`
      (2) In the *infolder f*older, open the **data_curation_query_base.sas** program edit the `libname pcordata` statement on Line 32 to remove the quotation marks, as: `libname pcordata &dpath;`
4) Open all programs in the *sasprograms* folder and modify the directory paths and inputs as instructed below. For reasons of compatibility and standardization, directory paths must meet the following criteria:

- DO use forward slashes (e.g. /) which are always compatible on both UNIX and WINDOWS.
- DO use end of path separators (e.g. /xyz/ and not /xyz) which are assumed by many programs.
- DO use beginning of path separators (e.g. /xyz) on UNIX.
- DO NOT use beginning of path separators on WINDOWS (e.g. P:/xyz not /P:/xyz).
- DO NOT surround directory paths with quotes (e.g. /xyz/ not "/xyz/").

   a) After `%let dpath=`, provide the directory path where your PCORnet CDM SAS data is located.
   b) After `%let qpath=`, provide the outer folder where the required folders were created.
   c) In the **02_run_queries.sas** program, populate the following user inputs.
      i) After `%let threshold=`, leave the default value of 0.
      ii) After `%let _grp,` provide one of the query group process option: `all`, `main`, `lab`, or `xtbl`.
         (1) *To run the programs all at once*: Select “`all`” to run the data curation query programs as a batch; this option is recommended if you are not an experienced SAS user and for the final submission to the DRNOC.
         (2) *To run the programs sequentially:* This is recommended for partners who have long run times and want to be able to remediate issues which only affect certain tables more easily. To do so, select one of the 3 options (`main`, `lab`, or `xtbl`) for the initial run and then repeat with the remaining options as instructed in Step 6.
      iii) After `%let lookback=`, leave the default value of 20.
   d) In the **04_run_edc_report.sas** program, populate the following user inputs:
      i) After `%let ppath=`, provide the outer folder containing the most recently approved query results (i.e. results for the previous DataMart refresh).
      ii) After `%let ets_installed=`, change the default value if needed as instructed in the program.
5) Open the **01_run_code_errors.sas** program. Run the program and review the log and output as instructed in the Query Package Checklist.
6) Open the **02_run_queries.sas** program. Run the program, either 1 time if you selected `%let _grp=all`, or 3 times in the sequence you desire (e.g. first with `lab`, next with `xtbl,` and finally with `main`). As it processes each query program, the program will print results to a PDF file, create

permanent SAS datasets for each output table, and import all permanent SAS datasets into a SAS transport file. Review the logs and output (see section IV and section VIII) as instructed in the Query Package Checklist. You may wish to review the output tables which could contain required data check exceptions before proceeding (e.g. from `xtbl`, review XTBL_L3_MISMATCH and XTBL_L3_NONUNIQUE). If you are working in Windows and executing the queries sequentially, you will need to close all open applications (e.g. PC SAS and Microsoft Word) before running the next program. Otherwise, you will get an error message from SAS like "`Fatal ODS error has occurred. Unable to continue processing this output destination`" and "`File is in use`". You can monitor the query progress by checking the [DMID]_[DATE]_data_curation_progress_report.rtf document in the *drnoc* folder. Depending on your SAS processing environment, you may also see the same information in the SAS OUTPUT window or RESULTS window.

7) Open and run the **03_run_edc_prep.sas** program. All data curation datasets must be present before proceeding with the EDC portion of this package. To ensure that this is the case, review the output in the result window. You should see a statement that says "`No datasets are missing`". If a dataset is missing, it will be listed in the output. If there is no output, confirm that you entered the correct information after `%let qpath=`. If necessary, rectify problems by returning to the **02_run_queries.sas** program to create the missing datasets.
8) Open and run the **04_run_edc_report.sas** program. This program will first call the **normalization.sas** program to create a dataset which combines all the data curation query output tables (`[DMID]_[DATE]_dc_norm.sas7bdat`). It will then call **edc_report.sas** to create the Empirical Data Curation (EDC) report from the normalized dataset and the code_error_summary dataset and print results to a PDF file. Review the logs and output as instructed in the Query Package Checklist.
9) If you need to modify your CDM data after running the queries, follow these guidelines for re-running the **02_run_queries.sas** programs, and then rerun the **03_run_edc_prep.sas** and **04_run_edc_report.sas** programs.
   a) You must re-run the `main` program unless the only change you made is to a field in the HARVEST table that is not a REFRESH date or the DATAMARTID.
   b) You must re-run the `xtbl` program if any of the following changes occurred: records were added or deleted, dates were changed, identifiers were changed, RACE or ENC_TYPE were changed, or anything in the HARVEST table was changed.
   c) You must re-run the `lab` program if you make any changes to the LAB_RESULT_CM table.
10) Update the online ETL Annotated Dictionary as instructed in the Query Package Checklist.
11) If desired, verify the contents of the cpt files by using a `proc cimport` statement, as shown in the example below:

```
libname outlib 'F:/pcornet/myproject/';
%let infile= 'F:/pcornet/myproject/T1D3_20151101_data_curation.cpt';
proc cimport infile=&infile library=outlib;
run;
```

12) Return the files in the *drnoc* folder (see section VIII) and a signed Query Package Checklist. If there is more than one version of any of the files in the *drnoc* folder, archive and/or delete the earlier versions and only return the ones with the most recent date (i.e., those reflecting the final results). Zip the contents of the drnoc folder into a file with your datamartid (e.g. [DATAMARTID]_DRNOC.zip).
13) Retain all output from the final run in the *dmloca*l folder for use in subsequent data curation queries as shown in the Query Input and Output Diagram (see Section IX).

XI. Version History

Date	Version	Description
Feb 3, 2016	v3.00	Original release.
Mar 17, 2016	v3.01	Corrected truncation of some query results by increasing field lengths. In VITAL_L3_HT, height categories of "<0" and "0-10" were both displaying as "0-10" due to a precision issue with PROC FORMAT/PROC MEANS; this was corrected. In PRO_L3_PXDATE_Y was incorrectly labeled ADMIT_DATE; this was corrected to PX_DATE. Updated all documentation and code to v3.01.
Nov 7, 2016	v3.02	Added queries of DEATH, DISPENSING, LAB_RESULT_CM, and PRESCRIBING (35 queries). Added 7 cross-table queries. Revised 14 queries (retained backwards compatibility). Revised the low cell count threshold logic to conform to PCORnet's new minimum bin size policy. Added the Empirical Data Curation Report.
Nov 18, 2016	v3.03	Eliminated the need for the SAS ACCESS/Interface to PC Files module. Resolves the following warning: "WARNING: In a call to the CATS function, the buffer allocated for the result was not long enough to contain the concatenation of all the arguments."
Mar 21, 2017	v3.04	Modified the program so that optional variables which are 100% missing will not cause errors or omissions. In ENC_L3_ENCTYPE, corrected the calculations for ELIG_RECORD_N and UNIQUE_VISIT_N. In XTBL_L3_DASH2 and XTBL_L3_DASH3, changed the logic to use PRESCRIBING. RX_ORDER_DATE instead of RX_START_DATE. In XTBL_L3_DASH3, changed the logic to not require LAB_RESULT_CM.LAB_NAME to be populated. In Empirical Data Curation (EDC) Table IIE, corrected the highlighting and added the PRESCRIBING table for orphan ENCOUNTERIDS. In Table IIIB, corrected the percentage calculations. In EDC Table IVD, corrected the "% of encounters without a principal diagnosis" calculation.
Jul 5, 2017	V3.10	Modified queries to conform to CDM v3.1. Added queries of the CONDITION, PCORNET_TRIAL, DEATH_CAUSE, and PRO_CM tables. Added 12 queries pertaining to previously characterized tables. Revised 31 queries. Incorporated PCORnet Data Checks v3.
Sept 18, 2017	V3.11	In the Data Curation query, corrected an omission in the "enc_l3_enctype_disdisp" query. In EDC Table IIB, added RX_QUANTITY_UNIT and corrected calculation for PX_TYPE. In EDC Table IVC, added DX_ORIGIN. In EDC Table IVE, corrected the percentage calculation.
Nov 20, 2017	V3.12	Incorporated the PCORnet Code Errors v3 program. In the Data Curation query, added 13 queries pertaining to previously characterized tables; revised 3 queries; and deprecated 6 queries. In the Empirical Data Curation report, incorporated PCORnet Data Checks v4, added 1 table, and revised 14 tables.
June 8, 2018	V4.10	Modified existing queries to conform to CDM v4.1. Incorporated the PCORnet Code Errors v5 program. In the Data Curation query, added 40 queries (24 pertaining to previously characterized tables; 16 for tables new to CDM v4.1); revised 27 queries; and deprecated 2 queries. In the Empirical Data Curation report, incorporated PCORnet Data Checks v5, added 1 table and 2 charts, and revised 15 tables.
June 29, 2018	V4.11	Corrected the DIAGNOSIS and PROCEDURES information in Table ID. Added additional DATE_MGMT fields to Table IIB.
Oct 8, 2018	V4.12	Corrected minor bugs in v4.11. Split the data curation program into 3 programs. Separated the data curation and code errors "run" programs. In the Data Curation program, modified the lookback logic to remove the date restriction from the DEATH table and include records with non-missing dates. Added an Empirical Data Curation (EDC) preparation program. In the Empirical Data Curation program, updated the reference files to reflect Cycle 5 results and to exclude LOINC codes which have no variation from Data Check 2.06.
Dec 20, 2018	V4.13	Full data curation for OBS_CLIN and OBS_GEN tables. For the data curation query, added query progress check report; updated Value Set Reference File to v1.5; and added dia_l3_dxtype and pro_l3_pxtype. Updated Potential Code Errors to v6 which incorporates OBS_CLIN, OBS_GEN, and

Date	Version	Description
		MED_ADMIN. In the Empirical Data Curation report, incorporated PCORnet Data Checks v6, added 4 tables and 1 chart, revised 11 tables and charts and switched to PDF format.
Mar 19, 2019	V4.14	Corrected minor bugs in v4.13 affecting PRO_L3_PXTYPE and Data Check 2.07.Updated the Data Curation Lab Group reference file to v3.1 and the network-wide results displayed in the EDC report to the most recently available data.
Jun 17, 2019	V4.15	Set the low cell count threshold to 0. Added Section IX: Query Input and Output Diagram. Added information about the new CDM Value Set Conformance Query to Section X: Responding to the Query Package.
Oct 7, 2019	V5.10	Additions and revisions to support CDM v5.1 and PCORnet Data Checks v7. In the Data Curation query, added 42 queries (17 pertaining to previously characterized tables; 25 for tables new to CDM v5.1); revised 30 queries; and updated the reference files for Data Curation Lab Groups, RXCUIs, and lab outliers. In the Empirical Data Curation report, revised 21 tables and added 4 tables and 7 charts.
Jan 6, 2020	V5.11	Corrected minor bugs in V5.10. Updated the Data Check programming so that (a) Data Check 2.08 (monthly outliers) will work for DataMarts that have SAS_ETS, and (b) in Data Check 2.06 (lab outliers) the pediatric lab reference range will be applied to pediatric DataMarts created in October 2019.
Apr 7, 2020	V5.12	In the Data Curation program, added 1 new query and revised 31 queries. In DIA_L3_DASH1, ENC_L3_DASH1, ENC_L3_DASH2, XTBL_L3_DASH1, XTBL_L3_DASH2, XTBL_L3_DASH3, and VIT_L3_DASH1, the logic was changed to use SAS system date to calculate a consistent timespan for all DataMarts. Updated Empirical Data Curation programs to incorporate PCORnet Data Checks v8, to suppress printing of Tables IIB and IIC if all fields conform to specifications, and to allow the user to designate if SAS_ETS is installed. Removed Data Curation table shells from the WorkPlan since these are available in the technical specifications posted on iMeet.
July 6, 2020	V5.13	In the Data Curation program, removed 10 query output tables that were no longer needed and incorporated a new parseable file (2020-06-17-PCORnet-Common-Data-Model-v5dot1-parseable.xlsx). In the Empirical Data Curation programs, fixed a defect in Data Check 3.13 that was failing to flag some exceptions for lab tests where the lab volume percent was below threshold but above 0, and modified Data Check 2.05 to use a lookup table.