Work Plan:
Data Curation Query Package

V6.15
Dec 6, 2024

TABLE OF CONTENTS

I. Purpose and Scope

The purpose of the Data Curation Query Package v6.15 is to characterize the data in PCORnet Common Data Model (CDM) v6.1. This package examines all 23 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 Checks v17. Output tables will be produced by running SAS programs against static local DataMarts in PCORnet CDM v6.1 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, 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.

Potential Code Errors Report

The query package produces a Potential Code Errors report which (a) 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 and (b) identifies LOINC codes that are in the incorrect CDM table based on the LOINC CLASSTYPE attribute. The expected CLASSTYPE for each CDM table is included in the CDM specification and listed in Section II.

Data Curation Tables

The query package produces up to 275 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 10 years (120 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, LDS_ADDRESS_HISTORY, HASH_TOKEN, or LAB_HISTORY tables. The dates used to select records in the other tables are ENROLLMENT.ENR_START_DATE, ENROLLMENT.ENR_END_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_START_DATE, OBS_GEN.OBSGEN_START_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), Keven Kunz (keven.kunz@duke.edu), and Yan Xu (yan.xu2@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 6.1 format with SAS data types.

This program has two components. First, it 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). 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 code 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
CPT/HCPCPS	Less than 5	00000x or 99999x	n/a	No numeric characters	IMMUNIZATION; OBS_GEN; PROCEDURES;
CVX	Not 2-3	n/a	Any alphabetical character	n/a [b]	IMMUNIZATION
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
ICD9 procedures	Not 3-4	00.00	Any alphabetical character	n/a [b]	OBS_GEN; PROCEDURES
ICD10 procedures	Not 7	0000000 or 9999999	n/a	n/a	OBS_GEN; PROCEDURES
LOINC	Not 3-8 [a]	No hyphen in the penultimate position	Any alphabetical character	n/a [b]	LAB_RESULT_CM; LAB_HISTORY; OBS_CLIN; OBS_GEN
NDC	Not 11	00000000000 or 99999999999	Any alphabetical character[2]	n/a [b]	DISPENSING; IMMUNIZATION; MED_ADMIN; OBS_GEN; PROCEDURES
RXCUI	Not 2-7	n/a	Any alphabetical character	n/a [b]	IMMUNIZATION; MED_ADMIN; OBS_GEN; PRESCRIBING

a. Current LOINC codes are 5+ digits but this length heuristic allows for shorter deprecated codes.
b. Redundant with the unexpected alphabetical character rule.

The program also identifies LOINC codes that are in the incorrect CDM table based on the LOINC CLASSTYPE attribute. This section contains content from LOINC® (http://loinc.org). LOINC is copyright © Regenstrief Institute, Inc. and the Logical Observation Identifiers Names and Codes (LOINC) Committee

and is available at no cost under the license at http://loinc.org/license. LOINC® is a registered United States trademark of Regenstrief Institute, Inc. This program uses LOINC release 2.78 and excludes six LOINC codes that could be potentially be stored in more than one CDM table. The expected CLASSTYPE for each CDM table is specified in the CDM specification and listed below.

CDM table	Expected CLASSTYPE(s)
LAB_RESULT_CM	1 (laboratory class)
LAB_HISTORY	1 (laboratory class)
OBS_CLIN	2 (clinical class)
OBS_GEN	2 (clinical class), 3 (Claims attachments) or 4 (Survey)

III. Query Variable Definitions

For a definition of each of the variables in the query package, please refer to the PCORnet Data Curation Query Package v6.15 Specifications, PCORnet Potential Code Errors v13 Specifications, and the PCORnet Empirical Data Curation v6.15 Specifications available on the Data Curation page on iMeet.

IV. Query Output Tables

For a full list of output tables and sample table shells, please refer to the PCORnet Data Curation Query Package v6.15 Specifications, PCORnet Potential Code Errors v13 Specifications, and the PCORnet Empirical Data Curation v6.15 Specifications available on the Data Curation page on iMeet.

V. Empirical Data Curation (EDC) Report

The data from all data curation query output tables, except for the *elapsed* and *datamart_all* datasets, are 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 54 tables and charts, depending upon the number of CDM tables populated. The table of contents is shown below. Note: the sequencing will skip letters for deprecated tables or charts for consistency between query packages.

Section	Table	Table Description	Data Check(s)
Data Check Summary	DC Summary	Data Check Exception Summary	n/a
Section I: Descriptive Information	Table IA	Demographic Summary	n/a
	Table IB	Potential Pools of Patients	2.09, 3.04, 3.05
	Table IC	Height, Weight, and Body Mass Index (BMI)	n/a
	Table ID	Records, Refresh Dates, Patients, Encounters, and Date Ranges by Table	1.18
	Table IE	Records Per Table by Encounter Type	n/a
	Table IF	Date Obfuscation or Imputation	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
	Chart IK	Trend in Clinical Observation Records by Start Date, Past 5 Years	n/a
	Chart IL	Trend in General Observation Records by Start 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, 1.17
	Table IIE	Orphan Records, Replication Errors, Encounter Duplication and Hash Token Duplication	1.08, 1.09,1.10, 1.11, 1.12,1.14, 1.15, 1.19
	Table IIF	Potential Code Errors and Misplaced Codes	1.13, 1,16
	Table IIG	LOINC Panel Codes	1.20
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

Section	Table	Table Description	Data Check(s)
	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
	Chart IIIG	Monthly Record Volume Outliers, Med Admin	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	3.08, 3.15
	Table IVI	Laboratory and Clinical Result Data Completeness	3.09, 3.10, 3.12, 3.16, 3.17
	Table IVI_Ref	Laboratory and Clinical Observation Result Data Completeness Definitions	n/a
	Table IVJ	Data Latency and Completeness of Medication Administration, Dispensing and Clinical Observation Data, Past 2 Years	3.14
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. 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. These tables consist of aggregate data/output and transfer the minimum required to answer the analytic question.
- ***infolder****:* Contains all input SAS programs and reference datasets (cpt 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.
- ***sasprograms***: Contains the SAS programs that must be edited and then executed locally.

VII. Files Included in Query Request

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

Subfolder	Type	Files
n/a	PDF and Word Document	Cycle 17 Data Curation Query Package Checklist.docx Data Curation Query Package v6.15 Work Plan.pdf Cycle 17 Additional Approval Criteria and Exemption Request Form.docx
/infolder	SAS programs	data_curation_base.sas data_curation_print.sas data_curation_tables.sas edc_prep.sas edc_report.sas edc_template.sas normalization.sas potential_code_errors.sas
	Reference datasets	loinc.cpt: Contains loinc.sas7dat dc_reference.cpt: See DC_Reference.cpt below edc_reference.cpt: See EDC_Reference.cpt below
/sasprograms	SAS programs	01_run_potential_code_errors.sas 02_run_queries.sas 03_run_edc_prep.sas 04_run_edc_report.sas

dc_reference.cpt

Dataset	Description
cdm_version facility_type pat_pref_language_spoken payer_type provider_specialty_primary qual route dose_form state specimen_source unit vx_body_site vx_manufacturer	List of allowed values. Derived from the Valuesets tab of the current version of the parseable file (2023_04_03_PCORnet_Common_Data_Model_v6dot1_parseable.xlsx). The parseable file is available on iMeet. Used in EDC Table IIB and Data Check 1.06.
rxnorm_cui_ref	List of RXNORM_CUI codes and their Term Type Tier mapping. Used in EDC Table IVH and Data Check 2.08.
dc_tables	List of output tables that are used by the EDC programs. This includes all datasets produced by the data curation query programs, except for the elapsed_ datasets and the code_summary dataset produced by the potential code errors program. This dataset includes the dcpart_macro_value, which is used to determine which data curation output tables are part1 and part2. This reference file is used by the data curation programs (to determine which ones are part1 and part2) and by 03_run_edc_prep.sas (to determine if all expected datasets are present).
address_rank	List of all possible combinations of ADDRESS_USE and ADDRESS_TYPE. The dataset assigns a rank to the values in each field (ADDRESS_USE_RANK and ADDRESS_TYPE_RANK) and to the combined values (ADDRESS_RANK). This dataset is used for creating the xtbl_l3_zip5_1y and xtbl_zip5_5y datasets.

edc_reference.cpt

Dataset	Description
dc_summary	Data check descriptions and network results
footers	Text for the EDC report footers
headers	Text for the EDC report headers
required_structure	Contains a list of the expected tables and fields (name, data type, and length) in the CDM. Derived from the Fields tab of the current parseable file. Used in EDC Table IID and Data Checks 1.01, 1,02, 1.03, and 1.04.
tbl_ivi_ref	Text for Table IVI_Ref
Toc	Text for the Table of Contents to be printed in the EDC report
Toc_part1	Text for the Table of Contents in part 1 of the EDC report when running the package in a modular fashion

VIII. Output Files

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

Produced by	File description
01_run_potential_code_errors.sas • potential_code_errors.sas	code_summary (SAS dataset and csv file) elapsed (SAS dataset) misplaced_loincs (SAS dataset and csv file) *Up to 11 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_lab_hist (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)
02_run_queries.sas • data_curation_base.sas; • data_curation_tables.sas	Up to 275 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.
03_run_edc_prep.sas	No files are created. After running the program, you should see a statement in the results window that shows "`No datasets are missing`", or a list of the missing datasets.
04_run_edc_report.sas • normalization.sas • edc_report.sas	[DMID]_[DATE]_dc_norm.sas7dat required_structure.csv

Files to be returned to the DRN OC (*drnoc* folder). DMID=DataMart ID; DATE=response date. Log files must be checked for errors and warnings.

File name	Program produced by	File description
[DMID]_[DATE]_potential_code_errors.log	potential_code_errors.sas	The SAS log file for the program.
[DMID]_[DATE]_Potential_Code_Errors.pdf	potential_code_errors.sas	The report produced by the program.
[DMID]_[DATE]_code_summary.cpt	potential_code_errors.sas	A SAS transport file containing the code_summary and elapsed dataset produced by the program
If data curation queries are run non-modularly [DMID]_[DATE]_data_curation_all.cpt **or** *If data curation queries are run modularly* [DMID]_[DATE]_data_curation _dcpart1.cpt [DMID]_[DATE]_data_curation_dcpart2.cpt	data_curation_base.sas; data_curation_tables.sas	A SAS transport file containing all the SAS datasets produced by the program(s).
If data curation queries are run non-modularly [DMID]_[DATE]_data_curation_all.pdf **or** *If data curation queries are run modularly* [DMID]_[DATE]_data_curation_dcpart1.pdf [DMID]_[DATE]_data_curation_dcpart2.pdf	data_curation_base.sas; data_curation_tables.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 are limited to the 100 most frequent observations. Empty tables are not printed.
If data curation queries are run non-modularly [DMID]_[DATE]_data_curation_all.log [DMID]_[DATE]_data_curation_base.log **or** *If data curation queries are run modularly* [DMID]_[DATE]_data_curation _base.log [DMID]_[DATE]_data_curation_dcpart1.log [DMID]_[DATE]_data_curation_dcpart2.log	data_curation_base.sas; data_curation_tables.sas	The SAS log files for the programs.
[DMID]_[DATE]_data_curation_progress _report.rtf	data_curation_base.sas; data_curation_tables.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* and *datamart_all* datasets.
[DMID]_[DATE]_normalization.log	normalization.sas	The SAS log file for the program.
If data curation queries are run non-modularly [DMID]_[DATE]_ EDCRPT_all.log **or** *If data curation queries are run modularly* [DMID]_[DATE]_EDCRPT_dcpart1.log [DMID]_[DATE]_EDCRPT_all.log	edc_report.sas	The SAS log file for the program.

File name	Program produced by	File description
If data curation queries are run non-modularly [DMID]_[DATE]_EDCRPT_all.pdf **or** *If data curation queries are run modularly* [DMID]_[DATE]_EDCRPT_dcpart1.pdf [DMID]_[DATE]_EDCRPT_all.pdf	edc_report.sas	The report produced by the program.

IX. Submission File Inventory

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 all files into a compressed file with your datamartid (e.g. [DATAMARTID]_DRNOC.zip).

Run option: Non-Modular (submit all contents of the /drnoc folder and 1-2 forms, for a total of 13-14 files):

1. [DATAMARTID]_[DATE]_data_curation_progress_report.rtf
2. [DATAMARTID]_[DATE]_data_curation_base.log
3. [DATAMARTID]_[DATE]_data_curation_all.cpt
4. [DATAMARTID]_[DATE]_data_curation_all.log
5. [DATAMARTID]_[DATE]_data_curation_all.pdf
6. [DATAMARTID]_[DATE]_EDCRPT_all.pdf
7. [DATAMARTID]_[DATE]_EDCRPT_all.log
8. [DATAMARTID]_[DATE]_dc_norm.cpt
9. [DATAMARTID]_[DATE]_normalization.log
10. [DATAMARTID]_[DATE]_potential_code_errors.log
11. [DATAMARTID]_[DATE]_potential_code_errors.pdf
12. [DATAMARTID]_[DATE]_code_summary.cpt
13. [Cycle #] Data Curation Query Package Checklist
14. [Cycle #] Additional Approval Criteria and Exemption Request Form (if applicable)

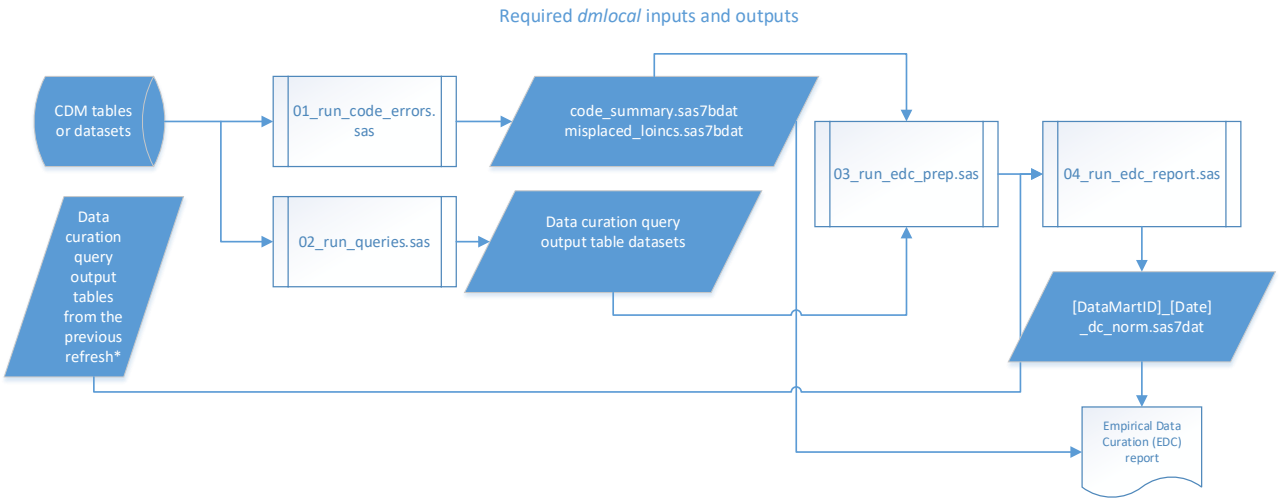
Run option: Modular (submit all contents of the /drnoc folder and 1-2 forms, for a total of 18-19 files):

1. [DATAMARTID]_[DATE]_data_curation_progress_report.rtf
2. [DATAMARTID]_[DATE]_data_curation_base.log
3. [DATAMARTID]_[DATE]_data_curation_dcpart1.cpt
4. [DATAMARTID]_[DATE]_data_curation_dcpart1.log
5. [DATAMARTID]_[DATE]_data_curation_dcpart1.pdf
6. [DATAMARTID]_[DATE]_data_curation_dcpart2.cpt
7. [DATAMARTID]_[DATE]_data_curation_dcpart2.log
8. [DATAMARTID]_[DATE]_data_curation_dcpart2.pdf
9. [DATAMARTID]_[DATE]_EDCRPT_all.pdf
10. [DATAMARTID]_[DATE]_EDCRPT_all.log
11. [DATAMARTID]_[DATE]_EDCRPT_part1.pdf
12. [DATAMARTID]_[DATE]_EDCRPT_part1.log
13. [DATAMARTID]_[DATE]_dc_norm.cpt
14. [DATAMARTID]_[DATE]_normalization.log
15. [DATAMARTID]_[DATE]_potential_code_errors.log
16. [DATAMARTID]_[DATE]_potential_code_errors.pdf
17. [DATAMARTID]_[DATE]_code_summary.cpt
18. [Cycle #] Data Curation Query Package Checklist
19. [Cycle #] Additional Approval Criteria and Exemption Request Form (if applicable)

X. Query Workflow Diagrams

Figure 1 illustrates how the query package uses information from the CDM tables and prior data curation results to produce the datasets in the *dmlocal* and *drnoc* folder.

Figure 1: Required *dmlocal* inputs and outputs



The figures below illustrate how to use the modular (Figure 2) or non-modular (Figure 3) approach to running the query package. The modular approach allows partners to identify and correct exceptions to required data checks before running the remainder of the data curation query programs. *Note:* Exceptions to Data Check 1.13 (More than 5% of ICD, CPT, LOINC, RXCUI, or NDC codes that do not conform to the expected length or content) will require re-running the Potential Code Errors program, regardless of which approach is used.

Figure 2: Data Curation Query Package Workflow, Modular Option

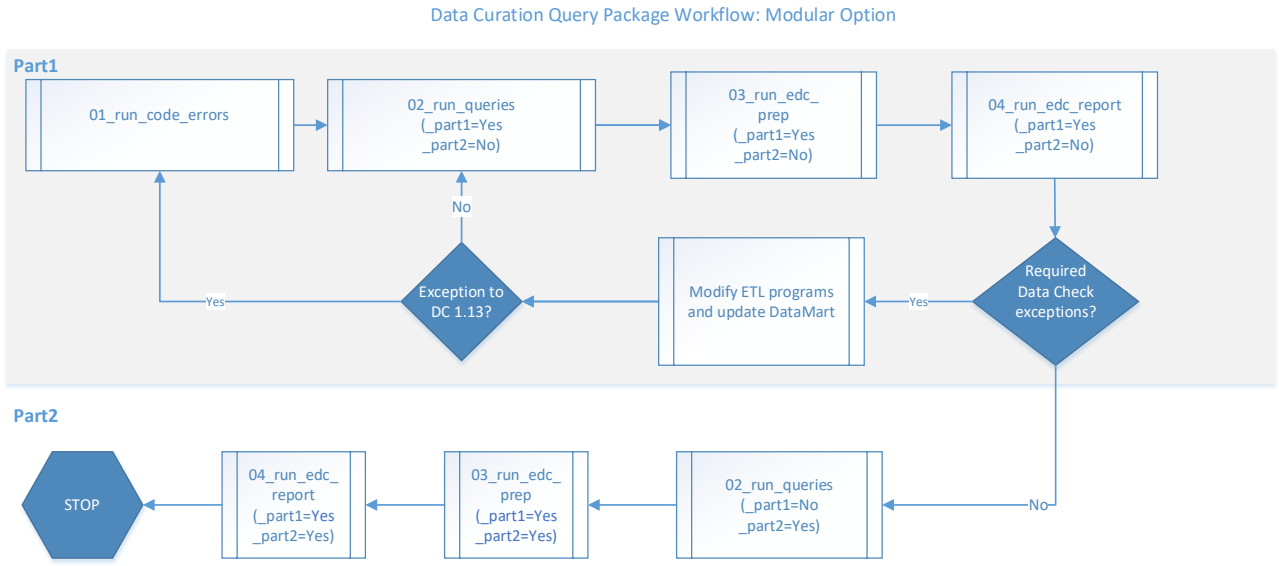
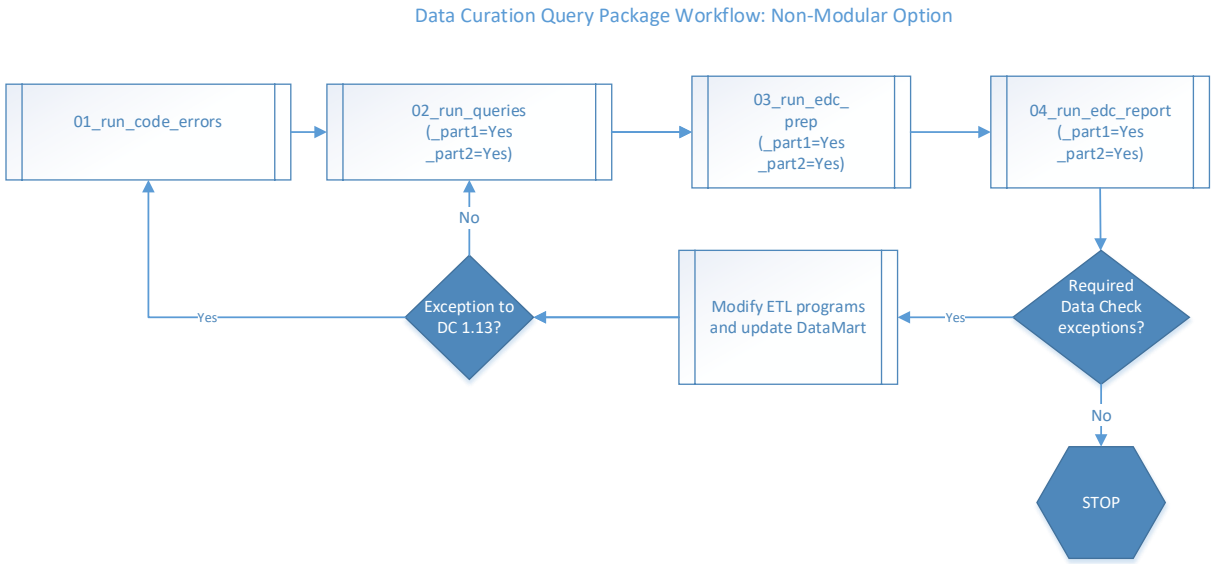


Figure 3: Data Curation Query Package Workflow, Non-Modular Option



XI. Responding to the Query Package

1) Prepare for the query as instructed in the Query Package Checklist.
2) Open the query package. Extract the contents, save them locally as described in Section VI, and create the *drnoc* and *dmlocal* folders.
3) Populate the directory paths 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) After `%let ppath=`, provide the outer folder containing the most recently approved query results (i.e. results for the previous DataMart refresh). If this is the DataMart's first refresh and the DataMart does not have results from a previous DataMart refresh, provide the same value that was provided for `qpath`. Failure to point to the correct ppath will result in the need for a resubmission.
   d) If the CDM data are stored in database tables, 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.
4) Open the **01_run_potential_code_errors.sas** program.
   a) Assign directory paths as described in Step 3.
   b) Run the program.
   c) Review the log and output as instructed in the Query Package Checklist.
5) Open the **02_run_queries.sas** program.
   a) Assign directory paths as described in Step 3. After `%let ets_installed=`, change the default value of Yes to No if applicable (see additional instructions in the program). If you are running the program modularly, modify `_part1` and `_part2` as shown in Figure 2.
   b) Run the program. 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. 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.
   c) Review the output and logs as instructed in the Query Package Checklist.
   d) If you are working in Windows and need to re-run the program, first close all open applications (e.g. PC SAS and Microsoft Word). 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`".
6) Open the **03_run_edc_prep.sas** program.
   a) Assign directory paths as described in Step 3. If you are running the program modularly, modify the `_part1` and `_part2` as shown in Figure 2.
   b) Run the program and 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 datasets are missing, re-run the **01_run_potential_code_errors.sas** and/or **02_run_queries.sas** programs to create the missing datasets before proceeding.
7) Open the **04_run_edc_report.sas** program.
   a) Assign directory paths as described in Step 3. If you are running the program modularly, modify the `_part1` and `_part2` as shown in Figure 2.

b) Run the program. This program will first call the **normalization.sas** program to create a dataset which combines all the data curation and code errors 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 `[DMID]_[DATE]_dc_norm.sas7bdat` dataset and the EDC reference datasets and prints results to a PDF file. If the program is run modularly (see Figure 2), part 1 will produce a portion of the EDC report mainly associated with the required data checks, while part 2 will produce the full EDC report that will be returned to the Coordinating Center.

c) Review the logs and output as instructed in the Query Package Checklist.

d) In investigating any Data Check exceptions, you may wish to review the reference datasets used by the data checks (see Section VII). You may access these by opening the `edc_reference.cpt` and `dc_reference.cpt` files, as shown in the example below.

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

8) Update the online ETL Annotated Dictionary as instructed in the Query Package Checklist.

9) 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_all.cpt';
proc cimport infile=&infile library=outlib;
run;
```

10) Submit the files as instructed in Section IX: Submission File Inventory.

11) IMPORTANT: You must retain all output in the *dmlocal* folder for use in subsequent data curation queries as shown in Figure 1.

XII. 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 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.

Date	Version	Description
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 Work Plan 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.
Dec 21, 2020	V6.00	Updated for the CDM v6.0 specifications and a new parseable file (2020_10_22_PCORnet_Common_Data_Model_v6dot0_parseable). In the Data Curation program, added 20 query output tables and revised 37 query output tables. In the Empirical Data Curation program, updated existing tables and data checks to support CDM v6.0, added Data Checks 1.15 and 1.16, revised the logic for Data Checks 3.01, 3.02, 3.07 and 3.09, and added charts IK and IL.
Apr 5, 2021	V6.01	In the Data Curation program, added 9 query output tables and added a new query group process option, `obsclin`.
Apr 7, 2021	V6.02	Fixed a bug in the Empirical Data Curation portion of the package that was causing duplicate rows in Table IVI and errors on the Data Check Summary page for Data Checks 3.10 and 3.12.
Jul 6, 2021	V6.03	Updated for a new parseable file (2021_04_12_PCORnet_Common_Data_Model_v6dot0_parseable). In the Data Curation program, added 6 new query output tables and revised 13 query output tables. In the Empirical Data Curation program, added 1 new table; revised the logic for Data Checks 2.07, 2.08, 3.02, 3.03, 3.06, and 4.02; and added Data Checks 1.17, 3.14, and 3.15.
Jan 3, 2022	V6.04	Updated for a new parseable file (2021_11_29_PCORnet_Common_Data_Model_v6dot0_parseable). In the Data Curation program, revised the default lookback period to 10 years, revised the lookback date for ENROLLMENT to consider either ENR_START_DATE or ENR_END_DATE, added 2 new query output tables, and revised 9 query output tables. In the Empirical Data Curation program, fixed the calculations for Data Checks 3.06 and 2.07 in Table IVE; fixed the calculation for PRESCRIBING Tier 1 brand in Table IVH; incorporated PCORnet Data Checks v11 by revising Data Checks 1.09 and 3.03 and adding Data Checks 1.18 and 1.19.
Jul 1, 2022	V6.05	In the Data Curation program, added 9 new query output tables and revised 9 query output tables. In the Empirical Data Curation program, fixed an error in Table IVE that affected Data Check 3.06 and incorporated PCORnet Data Checks v12.
Jan 03, 2023	V6.06	In the Data Curation program, revised 3 query output tables and revised the logic for calculating Length of Stay. In the Empirical Data Curation program, incorporated PCORnet Data Checks v13 and fixed errors in Data Check 1.19 and 2.07.
July 03, 2023	V6.10	Updated package for CDM v6.1. In the Data Curation program, added 1 query output table and revised 11 output tables, including fixing an error in the dia_l3_pdxgrp_enctype table that is used for Data

Date	Version	Description
		Check 2.07. In the Empirical Data Curation program, incorporated PCORnet Data Checks v14, updated the Data Check Summary to display exceptions to the required Data Check 1.17 in red font, updated Table IB (changes to potential pool calculations and addition of Data Check 2.09), updated Table IIA and IIC (changes to HASH_TOKEN table), updated Table IVI (addition of OBS_CLIN data), and added CSV versions of 4 of the reference datasets (required_structure; lab_volume_ref; specimen_source_category; and q2_stat_dlg_loinc) to the *dmlocal* folder.
December 15, 2023	V6.11	Modified entire package to create the ability to run the query into two parts (part 1 and part 2). This updated format made the query group processing option (%let grp) obsolete. All programs associated with running the query package in parts (obsclin, lab, xtbl, main) have been consolidated into a single program (data_curation_tables.sas). The reference dataset obsclin_code was removed. Added 2 additional figures to Section IX demonstrating the use of the part1 and part2 macros. Updated the majority of the reference datasets in the Data Curation program, deprecated 13 query output tables, added 2 query output tables, and revised 2 query output tables.
January 10, 2024	V6.12	Fixed a bug in the data_curation_tables program that was causing problems for partners who ran the query package modularly and closed their SAS session before running part2. Removed duplicates from the lab_loinc_ref reference table (used in lab_l3_loinc_source, lab_l3_loinc_source_5y, and Data Curation 2.05).
March 11, 2024	V6.13	Updates made to the data_curation_tables program to improve query run time for date logic.
June 5, 2024	V6.14	In the Potential Code Errors portion, updated to LOINC v2.76; fixed a bug in the allowed LOINC length for OBS_GEN, OBS_CLIN, and LAB_HISTORY; added an elapsed table; and corrected the expected length for ICD procedure codes in the workplan. In the Data Curation portion, deprecated 5 tables, revised 3 tables, updated the rxnorm_cui reference dataset, and removed the option to suppress low cell counts. In the Empirical Data Curation portion, deprecated 5 tables; deprecated 3 data checks; revised 1 data check; added 2 data checks; and altered the formatting of the charts and data check exceptions to be color-blind friendly.
Dec 6, 2024	V6.15	In the Potential Code Errors portion, updated to LOINC v2.78. In the Data Curation portion, deprecated 8 tables (labhist_l3_rhigh_dist, labhist_l3_rlow_dist, lab_l3_loinc_result_num_5y, lab_l3_n_5y, lab_l3_recordc_5y, medadm_l3_rxcui_tier_5y, pres_l3_rxcui_5y, pres_l3_rxcui_tier_5y), added 3 new tables (xtbl_l3_zip5_1y, xtbl_l3_zip5_5y, obsclin_l3_loinc), and updated the rxnorm_cui reference dataset. In the Empirical Data Curation portion, revised 2 data checks (2.08 and 3.07); added 1 data check (1.20).