TriNetX

DATA DICTIONARY

September 2022

(Applies to datasets generated as of September 1, 2022)

Table of Contents

Revision History

Effective Date	Change(s)

October 2, 2020	Encounter Table:
	• New data elements
	  ○ start_date_derived_by_TriNetX
	  ○ end_date_derived_by_TriNetX
	• New description
	  ○ derived_by_TriNetX

October 29, 2020	All Files:
	• New data element
	  ○ Source ID
	Diagnosis File:
	• New data element
	  ○ Principal Diagnosis Indicator
	Procedure File:
	• SNOMED License Agreement below for SNOMED data coming in the future
	Medication File:
	• New file name
	  ○ Previous: medication.csv
	  ○ Current: medication_ingredient.csv

November 12, 2020	Patient File:
	• New data element
	  ○ Marital status
	Diagnosis File:
	• Changed ordering of columns
	  ○ Principal Diagnosis Indicator is now after the Code column
	Procedure File:
	• New data element
	  ○ Principal Procedure Indicator
	Medication Ingredient File:
	• New data element
	  ○ Unique ID
	New Medication Drug File:
	• New file
	  ○ medication_drug.csv is a new file that provides drug level information with RxDrug and NDC codes

November 24, 2020	All files: • Source ID field ○ TriNetX-NLP option removed Medication Drug File:

	• Change to Code System ○ Code system options are RxNorm and NDC

January 14, 2021	Patient File:
	• New data element
	  ○ Month/Year of Death
	• Removed data element
	  ○ Age at Death

January 21, 2021	Patient File: • New data element ○ Reason YOB Missing
February 18, 2021	New Manifest file
March 4, 2021	Length column added to all tables
March 18, 2021	Patient File: • New data element ○ Patient Regional Location
April 16, 2021	Encounter File: • Removed Type options of FIELD and INPATIENT ACUTE

July 21, 2021	Diagnosis File:
	• New Field
	  ○ Admitting Diagnosis
	  ○ Patient Reason for Visit

August 18, 2021	Patient Demographics • Removed Postal Code Diagnosis • Added reason_for_visit
September 1, 2021	Page Header update
September 23, 2021	Updated the code system description within Medication Drug table to include the following additional language: NDC codes are eleven digits in length. Important reminder to not forget the issue with leading zeros in MS Excel
November 23, 2021	Diagnosis File: • Updated admitting_diagnois from Boolean to VARCHAR and updated description.

December 22, 2021	TriNetX Data Overview Updated the language under 'What are the characteristics of the HCOs that provide TriNetX with data?'
January 10, 2022	Lab Results • Added units_of_measure Vitals • Added units_of_measure • Updated field text_val to text_value
January 27, 2022	Medication Drug Table • Added Note Diagnosis • Updated 'reason_for_visit' from Boolean to VARCHAR • Updated description 'reason_for_visit' • Updated description 'principal_diagnosis_indicator' Procedure • Updated description 'principal_diagnosis_indicator' Standardized Terminology • Updated code_system sample data to ICD Lab Result • Updated description for code_system TriNetX Data Overview • Updated language and sentence structure
February 7, 2022	All tables • Added brief sentence on tables relationships with .csv files in dataset zip file. Medication Ingredient • Added Note Medication Drug • Added Note
September 1, 2022	Patient demographics • Added death_date_source_id to data dictionary spreadsheet to properly reflect existing output

	Diagnosis • Updated length of principal_procedure_indicator in data dictionary to properly reflect upper limits of the field Procedure • Updated length of principal_procedure_indicator in data dictionary to properly reflect upper limits of the field Patient cohort • Added source_id to data dictionary to properly reflect existing output Standardized terminology • Updated length of code_description and unit in data dictionary to properly reflect upper limits of the fields

TriNetX Data Overview

TriNetX datasets provide researchers access to de-identified patient data from networks of healthcare organizations (HCO) and other data providers.

Below are a set of questions and answers that describe the overall characteristics of TriNetX data.

What kind of data comes in a TriNetX dataset?

TriNetX datasets are comprised of clinical patient data such as demographics, diagnoses, procedures, labs, and medications. This is commonly referred to as real-world data (RWD).

The data in TriNetX datasets are:

- Primarily from HCOs electronic medical record (EMR) systems
- Collected for the primary purpose of providing care to patients

The data in TriNetX datasets are not:

- Claims data, data primarily collected for the purposes of billing
- Data collected for the purposes of randomized clinical trials

Where does the data in a TriNetX dataset originate?

Data in TriNetX datasets comes from HCOs and other data providers. The data these entities provide primarily comes from:

- EMR systems
    - Structured data
    - Unstructured data processed by Natural Language Processing (NLP) technology
- Cancer registries
- Other sources (e.g., genomic data from third party genomic testing labs)

What are the characteristics of the HCOs that provide TriNetX with data?

Most of the HCOs are large academic medical institutions with both inpatient and outpatient facilities. Most of these HCOs are adult acute-care hospitals with multiple facilities and locations. Most HCOs are currently located within the United States with the remaining spread across the globe.

HCOs provide TriNetX with both inpatient and outpatient data. The data they provide is representative of the entire patient population at the HCO. Most HCOs provide an average of seven years of historical data.

How is data transformed from its original source?

TriNetX typically receives data from HCOs and other data providers in one of two ways:

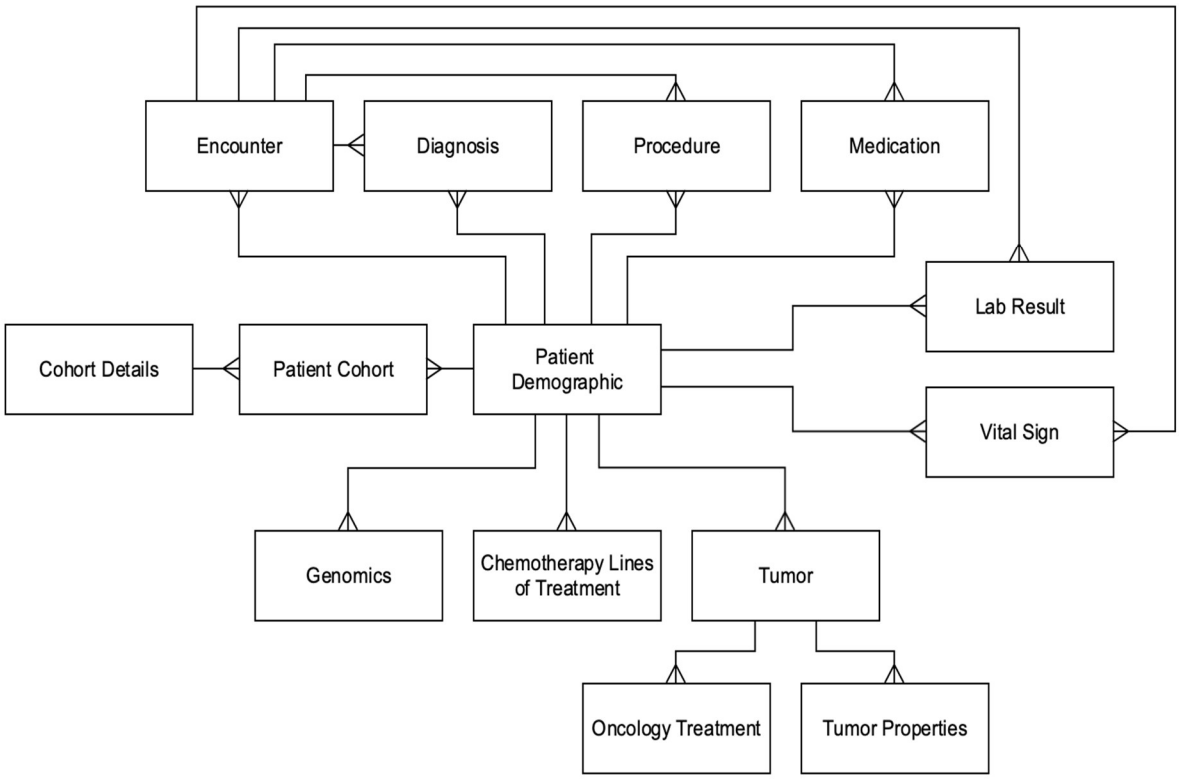
1. TriNetX ingests data directly from an HCO's research repository (e.g., i2b2) into the TriNetX environment
2. An HCO or data provider sends TriNetX data extracts in the form of CSV files

TriNetX maps the data to a standard and controlled set of clinical terminologies. The data is then transformed into a proprietary data schema. This transformation process includes an extensive data quality assessment that includes 'data cleaning' that rejects records that do not meet the TriNetX quality standards.

How fresh (up to date) is the data?

One of the distinguishing characteristics of the TriNetX dataset is that it is continuously refreshed. HCOs and other data providers update their data at various times with over 80% refreshing in 1, 2, or 4-week frequency intervals. The average lag time for an HCO's source data refresh is one month.

Data Tables Relationship



Patient Demographic Table

The Patient Demographics Table provides details on variables found in the patient.csv file.

Data Element	Data Type	Length	Sample Data	Description
patient_id	VARCHAR	200	123456789	The unique ID for the patient (de-identified). New unique ID's are generated for each dataset.
sex	VARCHAR	50	M	The biological sex of the patient. Possible values are M, F, Unknown.
race	VARCHAR	180	White	The race of the patient. Possible values are American Indian or Alaska Native, Asian, Black or African American, Native Hawaiian or Other Pacific Islander, White, Unknown.
ethnicity	VARCHAR	180	Hispanic	The ethnicity (cultural background) of the patient. Possible values are Hispanic or Latino, Not Hispanic or Latino, Unknown.
marital_status	VARCHAR	180	Married	The marital status for a patient. Possibilities: Married, Single, Unknown
year_of_birth	BIGINT	4	1958	The birth year of the patient. May be blank if the birth year occurred more than 90 years before the year the dataset was created.
reason_yob_missing	VARCHAR	50	Protected	The reason a Year of Birth is missing. Possibilities: Protected, Present
month_year_death	BIGINT	6	201905	The month and year of a patient's death. This field is rounded to the following month to protect patient privacy.
death_date_source_id	VARCHAR	200	EHR, Death Registry	Describes the source of the death date and is only present if a patient is deceased
patient_regional_location	VARCHAR	100	Middle Atlantic	The location of the patient. The geographic granularity is dependent on the cohort. For U.S. patients, there are four levels of granularity: 1. United States of America 2. East or West 3. Census Region 4. Census Division. For patients outside of the U.S., the value = ex-US.
source_id	VARCHAR	50	HCO	The data source for this each patient record. Data source option is HCO.

Encounter Table

The Encounter Table provides details on variables found in the encounter.csv file.

Data Element	Data Type	Length	Sample Data	Description
encounter_id	VARCHAR	200	987654321	The unique ID for the encounter (de-identified).
patient_id	VARCHAR	200	123456789	The unique ID for the patient (de-identified). New unique ID's are generated for each dataset.
start_date	DATETIME (YYYYMMDD)	8	20110315	The date the encounter began.
end_date	DATETIME (YYYYMMDD)	8	20110318	The date the encounter ended.
type	VARCHAR	50	AMB	The care setting of the encounter. Possible values are Ambulatory (AMB), Emergency (EMER), Home Health (HH), Inpatient Encounter (IMP), Inpatient Non-acute (NONAC), Observation (OBSENC), Pre-admission (PRENC), Short Stay (SS), Virtual (VR). These values are based on HL7 v3 Value Set ActEncounterCode.
start_date_derived_by_TriNetX	BOOLEAN	1	T	Flag that indicates whether the encounter start date was derived by TriNetX. Possible values are T for TRUE and F for FALSE.
end_date_derived_by_TriNetX	BOOLEAN	1	T	Flag that indicates whether the encounter end date was derived by TriNetX. Possible values are T for TRUE and F for FALSE.
derived_by_TriNetX	BOOLEAN	1	T	Flag that indicates whether the encounter start date, end date or the encounter itself was derived by TriNetX. Possible values are T for TRUE and F for FALSE.
source_id	VARCHAR	200	HCO	The data source and data type. Data source options are TriNetX and HCO. Data Type option is NLP. Possibilities: TriNetX, HCO, HCO-NLP.

Diagnosis Table

The Diagnosis Table provides details on variables found in the diagnosis.csv file.

Data Element	Data Type	Length	Sample Data	Description
patient_id	VARCHAR	200	123456789	The unique ID for the patient (de-identified). New unique ID's are generated for each dataset.
encounter_id	VARCHAR	200	987654321	The unique ID for the encounter (de-identified).
code_system	VARCHAR	50	ICD-10-CM	The name of the code system in which this diagnosis is coded. Possible code systems are ICD-9-CM, ICD-10-CM.
code	VARCHAR	100	E11	The diagnosis code.
principal_diagnosis_indicator	VARCHAR	10	P	Indicates whether the diagnosis code was the P (Primary) or a S (Secondary) reason for an encounter. If not known listed as Unknown
admitting_diagnosis	VARCHAR	1	T	Indicates whether this diagnosis code was the admitting diagnosis. Possible values are T for TRUE, F for FALSE, U for Unknown
reason_for_visit	VARCHAR	1	F	Indicates whether the diagnosis was the patient's reason for visit. Possible values are T for TRUE, F for False, and U for Unknown
date	DATETIME (YYYYMMDD)	8	20110315	The date the diagnosis was recorded.
derived_by_TriNetX	BOOLEAN	1	T	Flag that indicates whether the diagnosis was derived by TriNetX. Possible values are T for TRUE and F for FALSE.
source_id	VARCHAR	200	HCO	The data source and data type. Data source options are TriNetX and HCO. Data Type option is NLP. Possibilities: TriNetX, HCO, HCO-NLP.

Procedure Table

The Procedure Table provides details on variables found in the procedure.csv file.

Data Element	Data Type	Length	Sample Data	Description
patient_id	VARCHAR	200	123456789	The unique ID for the patient (de-identified). New unique ID's are generated for each dataset.
encounter_id	VARCHAR	200	987654321	The unique ID for the encounter (de-identified).
code_system	VARCHAR	50	ICD-10-PCS	The name of the code system in which this procedure is coded. Possible code systems are ICD-9-CM, ICD-10-PCS, CPT, HCPCS.
code	VARCHAR	100	03CJ0ZZ	The procedure code.
principal_procedure_indicator	VARCHAR	10	P	Indicates whether the diagnosis code was the P (Primary) or a S (Secondary) reason for an encounter. If not known listed as Unknown
date	DATETIME (YYYYMMDD)	8	20150314	The date the procedure was recorded.
derived_by_TriNetX	BOOLEAN	1	T	Flag that indicates whether the procedure was derived by TriNetX. Possible values are T for TRUE and F for FALSE.
source_id	VARCHAR	180	HCO	The data source and data type. Data source options are TriNetX and HCO. Data Type option is NLP. Possibilities: TriNetX, HCO, HCO-NLP.

Medication Ingredient Table

The Medication Ingredient Table provides details on variables found in the medication_ingredient.csv file.

Note:

- The medication drug and medication ingredient tables are unique in that they are the only two tables with a unique_id.
- The medication drug table is a subset of the medication ingredient table.

Data Element	Data Type	Length	Sample Data	Description
patient_id	VARCHAR	200	123456789	The unique ID for the patient (de-identified). New unique ID's are generated for each dataset.
encounter_id	VARCHAR	200	987654321	The unique ID for the encounter (de-identified).
unique_id	VARCHAR	200	452378345	The unique ID for the medication drug which can be used to link ingredient(s) to a single drug in the medication_drug.csv file.
code_system	VARCHAR	50	RxNorm	The name of the code system in which this medication is coded. The code system is RxNorm.
code	VARCHAR	100	26225	The medication code.
start_date	DATETIME (YYYYMMDD)	8	20120914	The date the medication order, prescription, or administration was recorded.
route	VARCHAR	200	Oral Product	The route of administration. Possible values are Drug implant, Inhalant, Injectable, Intraperitoneal, Nasal, Ophthalmic, Oral, Otic, Rectal, Topical, Urethral, Vaginal, Unknown.
brand	VARCHAR	200	Zofran	The medication brand.
strength	VARCHAR	200	4 mg	The medication strength.
derived_by_TriNetX	BOOLEAN	1	T	Flag that indicates whether the medication was derived by TriNetX. Possible values are T for TRUE and F for FALSE.
source_id	VARCHAR	200	HCO	The data source and data type. Data source options are TriNetX and HCO. Data Type option is NLP. Possibilities: TriNetX, HCO, HCO-NLP.

Medication Drug Table

The Medication Drug Table provides details on variables found in the medication_drug.csv file.

Note:

- The medication drug and medication ingredient tables are unique in that they are the only two tables with a unique_id.
- The medication drug table is a subset of the medication ingredient table.

***Important to Note**:

The NDC and RxNorm codes that are included here within the Medication Drug Table are not within the Standardized Terminology table below because of size constraints. For assistance with codes we recommend the following link: RXNav (here).

Data Element	Data Type	Length	Sample Data	Description
patient_id	VARCHAR	200	123456789	The unique ID for the patient (de-identified). New unique ID's are generated for each dataset.
encounter_id	VARCHAR	200	987654321	The unique ID for the encounter (de-identified).
unique_id	VARCHAR	200	452378345	The unique ID for the medication drug which can be used to link drug to ingredient(s) in the medication_ingredient.csv file.
code_system	VARCHAR	50	RxNorm	The name of the code system in which this medication is coded. Possibilities: RxNorm, NDC NDC codes are eleven digits in length. Important reminder to not forget the issue with leading zeros in MS Excel.
code	VARCHAR	100	26225	The medication code.
start_date	DATETIME (YYYYMMDD)	8	20120914	The date the medication order, prescription, or administration was recorded.
route	VARCHAR	200	Oral Product	The route of administration. Possible values are Drug implant, Inhalant, Injectable, Intraperitoneal, Nasal, Ophthalmic, Oral, Otic, Rectal, Topical, Urethral, Vaginal, Unknown.
brand	VARCHAR	200	Zofran	The medication brand.
strength	VARCHAR	200	4 mg	The medication strength.
quantity_dispensed	VARCHAR	50	60	The quantity dispensed for a drug.
days_supply	VARCHAR	50	30	The days' supply for a drug.
derived_by_TriNetX	BOOLEAN	1	T	Flag that indicates whether the medication was derived by TriNetX. Possible values are T for TRUE and F for FALSE.
source_id	VARCHAR	200	HCO	The data source and data type. Data source options are TriNetX and HCO. Data Type option is NLP. Possibilities: TriNetX, HCO, HCO-NLP.

Lab Result Table

The Lab Result Table provides details on variables found in the lab_result.csv file.

Data Element	Data Type	Length	Sample Data	Description
patient_id	VARCHAR	200	123456789	The unique ID for the patient (de-identified). New unique ID's are generated for each dataset.
encounter_id	VARCHAR	200	987654321	The unique ID for the encounter (de-identified).
code_system	VARCHAR	50	LOINC	The name of the code system in which this lab observation is coded. The code system is LOINC or TNX.
code	VARCHAR	100	2885-2	The code representing the lab test.
date	DATETIME (YYYYMMDD)	8	20120914	The date the test result was recorded.
lab_result_num_val	DECIMAL	50	7	The lab result for numeric results.
lab_result_text_val	VARCHAR	100	Positive	The lab result for text results.
units_of_measure	VARCHAR	40	mg/dL	The lab result units of measure for numeric results.
derived_by_TriNetX	BOOLEAN	1	T	Flag that indicates whether the lab result was derived by TriNetX. Possible values are T for TRUE and F for FALSE.
source_id	VARCHAR	200	HCO	The data source and data type. Data source options are TriNetX and HCO. Data Type option is NLP. Possibilities: TriNetX, HCO, HCO-NLP.

Vital Sign Table

The Vital Sign Table provides details on variables found in the vitals_signs.csv file.

Data Element	Data Type	Length	Sample Data	Description
patient_id	VARCHAR	200	123456789	The unique ID for the patient (de-identified). New unique ID's are generated for each dataset.
encounter_id	VARCHAR	200	987654321	The unique ID for the encounter (de-identified).
code_system	VARCHAR	50	LOINC	The name of the code system in which this vital sign is coded. The code system is LOINC or TNX.
code	VARCHAR	100	8302-2	The code representing the vital sign.
date	DATETIME (YYYYMMDD)	80	20120914	The date the vital sign was recorded.
value	VARCHAR	72	72	The value of this vital sign.
text_value	VARCHAR	1020	Positive	The value for text results.
units_of_measure	VARCHAR	40		The lab result units of measure for numeric results.
derived_by_TriNetX	BOOLEAN	1	T	Flag that indicates whether the vital sign was derived by TriNetX. Possible values are T for TRUE and F for FALSE.
source_id	VARCHAR	200	HCO	The data source and data type. Data source options are TriNetX and HCO. Data Type option is NLP. Possibilities: TriNetX, HCO, HCO-NLP.

Tumor Properties Table

The Tumor Properties Table provides details on variables found in the tumor_properties.csv file.

Data Element	Data Type	Length	Sample Data	Description
patient_id	VARCHAR	200	123456789	The unique ID for the patient (de-identified). New unique ID's are generated for each dataset.
diagnosis_date	DATETIME (YYYYMMDD)	8	20120914	The date of the original primary cancer diagnosis.
observation_date	DATETIME (YYYYMMDD)	8	20121114	The date the property was recorded.
tumor_site_code_system	VARCHAR	50	ICD-O	The name of the code system in which the tumor site is coded. The code system is ICD-O.
tumor_site_code	VARCHAR	100	C50	The tumor site code.
morphology_code_system	VARCHAR	50	ICD-O	The name of the code system in which morphology is coded. The code system is ICD-O.
morphology_code	VARCHAR	100	8500/3	The morphology code.
tumor_property_code_system	VARCHAR	50	TriNetX – Tumor Property	The name of the code system in which the tumor property is coded. The code system is TriNetX. This is a code system created by TriNetX for oncology specific factors.
tumor_property_code	VARCHAR	100	CSF07-Colon|060	The code that indicates the type of tumor property.
derived_by_TriNetX	BOOLEAN	1	T	Flag that indicates whether the tumor property was derived by TriNetX. Possible values are T for TRUE and F for FALSE.
source_id	VARCHAR	200	HCO	The data source and data type. Data source options are TriNetX and HCO. Data Type option is NLP. Possibilities: TriNetX, HCO, HCO-NLP.

Oncology Treatment Table

The Oncology Treatment Table provides details on variables found in the oncology_treatment.csv file.

Data Element	Data Type	Length	Sample Data	Description
patient_id	VARCHAR	200	123456789	The unique ID for the patient (de-identified). New unique ID's are generated for each dataset.
diagnosis_date	DATETIME (YYYYMMDD)	8	20120914	The date of the primary cancer diagnosis.
tumor_site_code_system	VARCHAR	50	ICD-O	The name of the code system in which the tumor site is coded. The code system is ICD-O.
tumor_site_code	VARCHAR	100	C50	The tumor site code.
morphology_code_system	VARCHAR	50	ICD-O	The name of the code system in which morphology is coded. The code system is ICD-O.
morphology_code	VARCHAR	100	8500/3	The morphology code.
oncology_treatment_start_date	DATETIME (YYYYMMDD)	8	20121001	The start date of the course of oncology treatment.
oncology_treatment_code_system	VARCHAR	50	TriNetX – Oncology Treatment	The name of the code system in which the oncology treatment is coded. The code system is TriNetX. This is a code system created by TriNetX for oncology treatment.
oncology_treatment_code	VARCHAR	100	1390|1	The code for the oncology treatment.
derived_by_TriNetX	BOOLEAN	1	T	Flag that indicates whether the oncology treatment was derived by TriNetX. Possible values are T for TRUE and F for FALSE.
source_id	VARCHAR	200	HCO	The data source and data type. Data source options are TriNetX and HCO. Data Type option is NLP. Possibilities: TriNetX, HCO, HCO-NLP.

Tumor Table

The Tumor Table provides details on variables found in the tumor.csv file.

Data Element	Data Type	Length	Sample Data	Description
patient_id	VARCHAR	200	123456789	The unique ID for the patient (de-identified). New unique ID's are generated for each dataset.
diagnosis_date	DATETIME (YYYYMMDD)	8	20120914	The date of the original primary cancer diagnosis.
observation_date	DATETIME (YYYYMMDD)	8	20121114	The date the property was recorded.
tumor_site_code_system	VARCHAR	50	ICD-O	The name of the code system in which the tumor site is coded. The code system is ICD-O.
tumor_site_code	VARCHAR	100	C50	The tumor site code.
morphology_code_system	VARCHAR	50	ICD-O	The name of the code system in which morphology is coded. The code system is ICD-O.
morphology_code	VARCHAR	100	8500/3	The morphology code.
stage_code_system	VARCHAR	50	TriNetX – Oncology Stage	The name of the code system in which the tumor stage is coded. The code system is TriNetX. This is a code system created by TriNetX for tumor stages.
stage_code	VARCHAR	100	2b	The code for the tumor stage.
tumor_size	VARCHAR	50	T2	The code for the size of tumor.
number_of_lymph_nodes	VARCHAR	50	N1	The code for the degree of spread to regional lymph nodes.
metastatic	VARCHAR	50	M0	The code for the presence of metastasis.
derived_by_TriNetX	BOOLEAN	1	T	Flag that indicates whether the tumor entry was derived by TriNetX. Possible values are T for TRUE and F for FALSE.
source_id	VARCHAR	200	HCO	The data source and data type. Data source options are TriNetX and HCO. Data Type option is NLP. Possibilities: TriNetX, HCO, HCO-NLP.

Chemotherapy Lines of Treatment Table

The Chemotherapy Lines of Treatment Table provides details on variables found in the chemo_lines.csv file.

Data Element	Data Type	Length	Sample Data	Description
patient_id	VARCHAR	200	123456789	The unique ID for the patient (de-identified). New unique ID's are generated for each dataset.
start_date	DATETIME (YYYYMMDD)	8	20150314	The date the chemotherapy line of treatment was determined to start.
line	BIGINT	2	1	The sequential order of chemotherapy regimens. Possible values are 1, 2, 3, 4, or 5 with 1 = the first regimen and 5 = the last regimen. These lines are derived by TriNetX.
derived_by_TriNetX	BOOLEAN	1	T	Flag that indicates whether the chemotherapy line of treatment was derived by TriNetX. Possible values are T for TRUE and F for FALSE.
source_id	VARCHAR	200	HCO	The data source and data type. Data source options are TriNetX and HCO. Data Type option is NLP. Possibilities: TriNetX, HCO, HCO-NLP.

Genomics Table

The Genomics Table provides details on variables found in the genomic.csv file.

Data Element	Data Type	Length	Sample Data	Description
patient_id	VARCHAR	200	123456789	The unique ID for the patient (de-identified). New unique ID's are generated for each dataset.
code_system	VARCHAR	50	HGVS	The name of the code system in which genomic data is coded. The syntax of the code conforms to HGVS.
code	VARCHAR	100	BRAF p.V600E c.1799T>A	Variant description.
test_date	DATETIME (YYYYMMDD)	8	20120914	The date the genetic test was recorded.
derived_by_TriNetX	BOOLEAN	1	T	Flag that indicates whether the genomic data was derived by TriNetX. Possible values are T for TRUE and F for FALSE.
source_id	VARCHAR	200	HCO	The data source and data type. Data source options are TriNetX and HCO. Data Type option is NLP. Possibilities: TriNetX, HCO, HCO-NLP.

Cohort Details Table

The Cohort Details Table provides details on variables found in the cohort_details.csv file.

Data Element	Data Type	Length	Sample Data	Description
cohort_name	VARCHAR	500	Diabetes women aged 18-45	The name of the cohort included in the dataset.
cohort_number	BIGINT	2	1	The number of the cohort included in the dataset.
total_patient_records	BIGINT	10	20,000	The total number of patient records in the cohort in the dataset.

Dataset Details Table

The Dataset Details Table provides details on variables found in the dataset_details.csv file.

Data Element	Data Type	Length	Sample Data	Description
total_number_unique_patients	BIGINT	10	19,000	The total number of unique patient records across multiple cohorts in the dataset. A patient's record could be in a single cohort multiple times if the patient visited more than one HCO that contributed data to a cohort.
total_number_HCOs	BIGINT	4	7	The total number of healthcare organizations contributing data to the dataset.
date_created	DATETIME (YYYYMMDD)	8	20180316	The date the dataset was created.

Patient Cohort Table

The Patient Cohort Table provides details on variables found in the patient_cohort.csv file.

Data Element	Data Type	Length	Sample Data	Description
patient_id	VARCHAR	200	123456789	The unique ID for the patient (de-identified). New unique ID's are generated for each dataset.
cohort_name	VARCHAR	500	Diabetes women aged 18-45	The name of the cohort in which the patient's record is included.
cohort_number	BIGINT	2	1	The number of the cohort in which the patient's record is included.
source_id	VARCHAR	50	EHR	The data source and data type.

Standardized Terminology Table

The Standardized Terminology Table provides details on variables found in the standardized_terminology.csv file.

Data Element	Data Type	Length	Sample Data	Description
code_system	VARCHAR	50	ICD	The name of the code system in which the data element is coded.
code	VARCHAR	100	1191	The code for the data element.
code_description	VARCHAR	1500	Aspirin	The textual description of the data element.
path	VARCHAR	300	N0000010574/ N0000029132/ N0000029133/ N0000029135/ 1191	The terms the data element is mapped to and the path in which those terms exist.
unit	VARCHAR	40	inches	The unit of measurement for a code value. This field only applies to codes in the Lab Result table and the Vital Sign table.

Manifest Table

The Manifest Table provides details on variables found in the manifest.csv file.

Data Element	Data Type	Length	Sample Data	Description
file	VARCHAR	50	diagnosis.csv	The name of a file included in the dataset.
column_count	BIGINT	2	8	The number of columns in the file.
row_count	BIGINT	15	10,000	The number of rows in the file.
unique_patient_count	VARCHAR	10	456,789	The number of unique patients in the file. If there are no patient ID's in the file, the value will be -.