Research Dataset FAQ

The document supplements the TriNetX Research Data Dictionary. Together, our dictionary and FAQ provide a complete guide to the source, structure, interpretation, and use of our datasets.

TriNetX Research Data Dictionary gives you...	Dataset FAQ gives you...
• A list of the data tables that make up a dataset • Table relationships • The description and type (e.g. VARCHAR, BOOLEAN, DATETIME) of all data elements, organized by table	• A description of the data sources • Explanations for missing or discrepant data • Our methods for imputing data (e.g. deriving GFR, LVEF, etc.) • Steps for licensing and downloading datasets

FAQ by Topic

Data Sources

Where does the data come from?

As of August 2021, 52 healthcare organizations (HCOs) from the United States and Asia-Pacific contribute patient data to Dataworks. Electronic health record (EHR) data forms the foundation of our datasets, but a growing number of additional sources supplement this EHR data.

For example, in the United States, cancer registries and surgical pathology reports supply oncology data, while obituaries and public Social Security records provide mortality data. Among HCOs outside of the United States, EHR provides most oncology data.

Datasets contain:

- values from structured EHR fields (e.g. demographics; date-indexed encounters, diagnoses, procedures, and medications; genetic variants)
- facts and narratives from free text (e.g. medications identified through NLP)
- death dates from mortality registries
- tumor morphology and size data from tumor registries and surgical pathology reports

The HCOs providing this data encompass everything from freestanding acute care hospitals to networks of outpatient clinics. HCOs contributing to our datasets include:

- More than 100 community hospitals
- More than 500 outpatient clinics
- More than 10 U.S. academic medical centers
- Multiple premier U.S. pediatric hospitals
- A center for cardiac care and cardiothoracic surgery

You may define a dataset by clinical facts (e.g. all patients with a first-recorded diagnosis of melanoma since 2016), but not by the source HCO or country.

Back to question list

How much data do you have?

Data from **73 million patients** is available to license and download.

Depending on the length of their record and pattern of care, each patient represents anywhere from dozens to thousands of “facts”, e.g.

- diagnoses, medication orders, lab results, procedures, or clinical events tied to a specific date
- demographic attributes like race or sex
- genetic variants
- tumor size and staging

Our 70M patients represent 25 **billion facts**.

Back to question list

How can I access data from the EU?

TriNetX can help you analyze EU patients in different ways, depending on your needs.

Ask a question about outcomes, cohort differences, or treatment pathways on our self-service platform at live.trinetx.com. You’ll get statistically robust answers you can download as data tables and graphs.

We can also conduct chart reviews and observational studies at HCOs throughout Europe, conditional upon IRB/EC approval.

As a matter of institutional policy or culture, the European organizations who provide aggregate counts to the TriNetX platform do not yet allow end users to download patient-level data.

Back to question list

Data Format and Structure

What kind of files (and how many) will I get?

A dataset is made up of 18 comma separated value (CSV) files, including a terminology table and manifest table, which you'll download from our platform in a single compressed folder. Each file is a machine- and human-readable data table you can import into your preferred application.

	A	B	C	D	E	F	G
1	patient_id	encounter_id	code_system	code	date	lab_result_num_val	derived_by_TriNetX
92	de674da2db18a49f4c85b990096d59e6a4981adb	dff7d07cdeec33bbddd9f41e44fbb84d7659d25f	LOINC	1751-7	20141009	3.4	F
93	de674da2db18a49f4c85b990096d59e6a4981adb	c0278ce72f566717eaca9f01b8cd36568049d546	LOINC	1751-7	20141106	3.5	F
94	de674da2db18a49f4c85b990096d59e6a4981adb	d36986a58f0095ef95266627d319fdf449f1d588	LOINC	1751-7	20141204	3.9	F
95	de674da2db18a49f4c85b990096d59e6a4981adb	30e668602a4a27d037f99bd8bfb03904ce374cb2	LOINC	1751-7	20150107	3.7	F
96	de674da2db18a49f4c85b990096d59e6a4981adb	0573a3cf8725020554365afa706a0b0080a95f8d	LOINC	1751-7	20150204	3.9	F
97	de674da2db18a49f4c85b990096d59e6a4981adb	1e925b09a698c773c06849515bfb203d0f55a7fc	LOINC	1751-7	20150327	3.8	F
98	de674da2db18a49f4c85b990096d59e6a4981adb	cb64b72c54b1a117f38c25315d98eab61677828c	LOINC	1751-7	20150408	3.1	F
99	de674da2db18a49f4c85b990096d59e6a4981adb	3eac595f7a91581abfc450bae6fc35ff92f48d51	LOINC	1751-7	20150415	3.2	F
100	de674da2db18a49f4c85b990096d59e6a4981adb	903296fb8c0a33bf2917f93847435daa7b1b258d	LOINC	1751-7	20150422	3.1	F

Sample rows from the lab_values csv. Each patient_id and encounter_id is a synthetic identifier. Another CSV in this dataset, named standardized_terminologies, provides the description for the LOINC code indicated here. 1751-7 is Albumin [g/dL] in Serum or Plasma.

Our data dictionary provides a table-by-table breakdown of the data elements and their type.

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.
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.
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.

A page from the Data Dictionary, explaining each column in the lab_results table

[Back to question list]

What data elements are included?

Our data dictionary provides an exhaustive list and description of data elements you may find in your dataset.

Because data availability can vary by HCO, no single patient is guaranteed to have values for every data element represented within the dataset. For example, you may find that only a portion of the patients in your dataset have anatomic location codes assigned to their tumor, while only a (different) portion have genetic variants recorded. Even common demographic data, such as race, may be 'Unknown' for a patient. However, by requiring a known value for a data element in your query (e.g. M or F for sex), you can limit the number of these 'Unknown' values.

Back to question list

What terminologies do the sources use, and how do you harmonize them?

Our HCOs employ a wide array of terminologies in their EMR systems; more than two dozen across all the data domains (e.g., diagnoses, medications) represented in our datasets. While some terminologies, like ICD-10-CM and LOINC, are common, others may be unique to a geographic region, a health system, or even to a specific hospital department.

TriNetX's informatics team, in close collaboration with each HCO, builds and verifies data maps that translate information from a *source* (i.e., the HCO's EMR) to the *target* (i.e., the TriNetX terminologies used on our platform and in our datasets). For example, these maps transform the provider's medication codes (e.g., Medi-Span, Multum, NDC, ATC, etc.) to RxNorm medication codes.

This informatics team also practices ongoing quality assurance, using tools to identify anomalies and outliers that may indicate mappings in need of change.

Source and Master Terminologies by Data Domain

Domain	Example Source Terminologies	Target TriNetX Terminologies
Demographics	Various, uncoded, HL7	HL7
Diagnosis	ICD-9-CM, SNOMED CT, ICD-10-CM, ICD-10	ICD-10-CM
Procedure	Non-standard terminologies unique to organization, ICD-9-CM, SNOMED CT, ICD-10-PCS, CPT, HCPCS	ICD-10-PCS, CPT, HCPCS, SNOMED CT
Medication	WHODrug, NDC, RxNorm	RxNorm, with OMOP extension for medications not approved in the U.S.
Lab	Non-standard terminologies unique to lab, SNOMED CT, LOINC	LOINC
Oncology	Various	ICD-O
Genomics	Various	HGVS
Vital Signs	Various	LOINC
Visit Types	Various	HL7

Back to question list

Data Quality

Do you delete or clean any values in the data you receive from HCOs?

From the data HCOs provide to us, we preserve all *non-identifiable* data.

Non-identifiable data cannot be traced back to a specific person, and so excludes names, addresses, phone numbers, medical record numbers, and other uniquely identifying pieces of information.

For the clinical data that remains, TriNetX uses HIPAA's expert determination method to ensure that such data is de-identified. You can download our expert privacy attestations from this page.

The expert determination method allows us to include observation dates and patient years of birth in our datasets. This method also requires us to remove facts (e.g. a rare diagnosis, in combination with a rarely obtained age) whose inclusion would increase the risk that a patient could be re-identified. These occasions are themselves very rare. If removing data for privacy reasons poses any risk of biasing analysis, we will discuss the concerns and alternatives with you.

Note that we do not clean or correct any data that is nonsensical (e.g. a height of 3 inches), inconsistent with other facts about a patient (e.g. a perinatal procedure for a patient whose sex is M), or logically impossible (e.g. an encounter end date that precedes its start date). We preserve these anomalies unchanged so that you may handle them in the way that best suits your analysis.

Back to question list

Do you derive values from data you receive from HCOs?

In some cases, TriNetX derives a clinical value or date from data received directly from the HCO; that is, we infer an encounter, make a calculation on given data, or derive a value for a patient that did not itself come from the EHR. A row that contains a derived fact always contains a 'T' (for True) in the column labeled 'derived_by_TriNetX.'

Below are situations where TriNetX derives a value.

A patient's record includes an observation whose encounter_id does not appear in the encounter table

We derive as many encounters as there are unique encounter_ids within the observation tables (i.e. diagnosis, medication, procedure, lab result, vital signs) that are not already included in the encounter table. If two or more observations carry the same encounter_id, we derive a single encounter to accommodate those observations. We use the date of the earliest observation as the encounter's start date, and the date of the latest observation as the encounter's end date. We do not derive an encounter type.

Example #1

Patient ABC shows a protein level (LOINC 27298-9) on 20200301 (1 March 2020) and a diagnosis of nephritis (ICD-10-CM N00.9) on 20200302 (2 March 2020). The encounter_ids for the lab result and diagnosis differ from one another. Neither is represented in the encounter table.

TriNetX derives two encounters (both of type 'Unknown') within the encounter table, one for the lab result and one for the diagnosis, using the encounter_ids found in the lab result table and diagnosis table. The encounter for the lab result will show a start date and end date of 20200301. The encounter for the diagnosis will show a start date and end date of 20200302.

For both encounters, all three "derived_by" values – start_date_derived_by_TriNetX, end_date_derived_by_TriNetX, and derived_by_TriNetX – carry a T.

Example #2

Patient XYZ shows a test for Streptococcus pyogenes Ag in throat (LOINC 78012-2) on 20200401 (1 April 2020), a diagnosis of Streptococcal pharyngitis (ICD-10-CM J02.0) on 20200402 (2 April 2020), and a medication order for Amoxicillin (RxNorm 723) on 20200403 (1 April 2020). All three observations are associated with the same encounter_id, which is not represented in the encounter table.

TriNetX creates just one encounter (of type 'Unknown') with a start date of 20200401 and an end date of 20200403. All three "derived_by" values – start_date_derived_by_TriNetX, end_date_derived_by_TriNetX, and derived_by_TriNetX – carry a T.

An HCO-provided encounter is missing a start date or an end date

If an HCO provides us with an encounter without a start date, we assign the earliest start date from all observations associated with the same encounter_id. We then set start_date_derived_by_TriNetX to T. We also set derived_by_TriNetX to T.

If an HCO-supplied encounter is missing an end date, we assign the latest end date from all observations associated with the same encounter_id. We then set end_date_derived_by_TriNetX (and derived_by_TriNetX) to T.

Glomerular filteration rate (GFR)

For every value of serum, plasma, or blood creatinine level provided in the lab results table, we derive at least one GFR, provided that:

- the patient's sex is known, AND
    - the patient is an adult (>=18 years) whose age in years is known, OR
    - the patient is less than 18 years old and his or her record provides a height recorded within one year of the creatinine measurement date

See *How do you derive GFR?*

Oncology diagnoses (from tumor properties)

We use an ICD-O to ICD-10-CM mapping recommended by the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) Program to derive diagnoses from ICD-O data provided by the HCO. You can consult that mapping here. We flag these diagnoses as derived by TriNetX.

Back to question list

What do the values that appear for source_id (TriNetX, HCO, or HCO-NLP) mean?

Every observation table (e.g. Diagnosis, Procedure, Lab Result) includes a column labeled source_id. Each row in these tables will show either TriNetX, HCO, or HCO-NLP in that column.

source_id value	Meaning	Example
TriNetX	Fact is derived by TriNetX	An eGFR value
HCO	Fact is taken directly from a structured field in the HCO's EHR	A diagnostic code
HCO-NLP	Fact is abstracted via natural language processing from clinical notes stored by the HCO	A tumor stage

Why are there occasional mismatches between encounter dates and associated observation dates?

Occasionally, an observation date falls outside of the range given by the associated encounter's start and end dates.

Example:

In the lab table, patient ABC shows a lab test for albumin [mass/volume] in serum or plasma (LOINC 1751-7) dated 20200501 (1 May 2020). The test has an associated encounter_id of 789. In the encounter table, encounter_id 789 carries a start date and end date of 20200430 (30 April 2020). The HCO has provided both dates.

The difference in dates originates with the HCO and may not reflect an error. In the case above, the lab could have started with a blood draw performed on 30 April 2020, with results recorded the next day.

TriNetX does not alter either of the HCO-supplied dates for these inconsistencies.

Back to question list

What does it mean when an encounter start_date is after the end_date?

This discrepancy also comes from the HCO. We preserve it unchanged in the dataset.

Back to question list

Why do some encounters appear multiple times with different dates?

Duplicate encounter_ids originate with the HCO. The HCO may not define encounters in a way that requires a single start date. Data extraction and/or transformation processes at the HCO might also associate multiple start or end dates for a single encounter.

Note that any duplicated id's will always pertain to the same patient.

Back to question list

Are dates in the dataset actual, or are they shifted to protect patient privacy?

More than half of our sites give us actual observation dates. HCOs that shift dates typically do so in the range of plus or minus 7 to 365 days, applying the same shift to all observations so that the duration *between* any two observations is preserved. We do not indicate anywhere in our dataset if a patient represents a date-shifting HCO.

If you are conducting an incidence or prevalence analysis for a specific date range, it is possible that date shifting could skew the results. However, relative changes over time or differences between groups remain informative; for example, a lower incidence of treatment in one age group compared to another, or a greater prevalence of a disease among female patients as compared to male patients.

Back to question list

Why do some encounters have no facts?

Very rarely, the encounter table will contain an (HCO-provided) encounter_id not referenced in any other table. In these cases, TriNetX has removed the observation (or observations) associated with this encounter as part of our de-identification logic. Click here to download our list of sensitive terms from ICD-9, ICD-10, HCPCS, LOINC, and SNOMED.

Before assuming an encounter_id listed in the encounter table has no associated observations, be sure to consult every other table that references an enounter_id: diagnoses, lab results, procedure, medication, and vital sign. Some encounters contain only medication details or only lab results.

Back to question list

Why do so many encounters have an Unknown type?

HCOs do not always provide an encounter type. If they do, it may not match one of the 10 values we support for this dataset element, which are based on HL7 v3 Value Set ActEncounterCode.

- Ambulatory (AMB)
- Emergency (EMER)
- Field (FLD)
- Home Health (HH)
- Inpatient Encounter (IMP)
- Inpatient Non-acute (NONAC)
- Observation (OBSENC)
- Pre-admission (PRENC)
- Short Stay (SS)
- Virtual (VR)

If the value supplied by the HCO does not match one of the values above, or the HCO does not specify any value, we assign 'Unknown' as the encounter type.

Back to question list

How do you derive glomerular filtration rate (GFR)?

For every value of serum, plasma, or blood creatinine level provided in the lab results table, we derive at least one GFR, provided that:

- the patient's sex is known, AND
  - the patient is an adult (>=18 years) whose age in years is known, OR
  - the patient is less than 18 years old and his or her record provides a height recorded within one year of the creatinine measurement date

Exactly which rates we derive, and the formulas we use to derive them, depend on the types of creatinine values given, as well as the patient's age, sex, and race.

Derived GFRs

Code (LOINC)	Description	Formula	Derived for
62238-1	CKD-EPI GFR for adults	$141 \times \min(S_{cr}/\kappa, 1)^{\alpha} \times \max(S_{cr}/\kappa, 1)^{-1.209} \times 0.993^{Age} \times 1.018$ [if female] × 1.159 [if black] where: S_{cr} is serum creatinine in mg/dL (LOINC code 2160-0), κ is 0.7 for females and 0.9 for males, α is -0.329 for females and -0.411 for males, min indicates the minimum of S_{cr}/κ or 1, and max indicates the maximum of S_{cr}/κ or 1 Source: Levey AS, Stevens LA, Schmid CH, Zhang YL, Castro AF, 3rd, Feldman HI, et al. A new equation to estimate glomerular filtration rate. Ann Intern Med. 2009;150(9):604-12.	Adult patients of known sex, age, and race, with creatinine in **serum/plasma** (LOINC 2160-0) given
33914-3	MDRD GFR in serum/plasma for adults	$175 \times (S_{cr})^{-1.154} \times (Age)^{-0.203}$ × [0.742 if female] × [1.212 if black] where: S_{cr} is serum creatinine in mg/dL (LOINC code 2160-0), κ is 0.7 for females and 0.9 for males, α is -0.329 for females and -0.411 for males,	

		min indicates the minimum of S_{cr}/κ or 1, and max indicates the maximum of S_{cr}/κ or 1 Source: Levey AS, Coresh J, Greene T, Stevens LA, Zhang YL, Hendriksen S, Kusek JW, Van Lente F; Chronic Kidney Disease Epidemiology Collaboration. Using standardized serum creatinine values in the modification of diet in renal disease study equation for estimating glomerular filtration rate. Ann Intern Med. 2006 Aug 15;145(4):247-54.	
76633-7	MDRD GFR in blood for adults	Same as MDRD GFR in serum/plasma for adults	Adult patients of known age, sex, and race, with creatinine in **blood** (LOINC 38483-4) given
48643-1	MDRD GFR in serum/plasma for adults of unknown race, using black as race	Same as MDRD GFR in serum/plasma for adults, with the final x 1.212 term *included*	Adult patients of known age and sex, but **unknown** race, with creatinine in **serum/plasma** (LOINC 2160-0) given
48643-3	GFR in serum/plasma for adults of unknown race (MDRD), using non-black as race	Same as MDRD GFR in serum/plasma for adults, with the final x 1.212 term *excluded*	
50384-7	Schwartz GFR for children	(0.41 × Height in cm) / Creatinine in mg/dL	Pediatric patients with either creatine in serum/plasma or blood (LOINC 2160-0 or LOINC 38483-4), with a height recorded within one year of the creatinine measurement date

The HCO may also provide a GFR, for the same patient on the same date, that does not match the TriNetX-derived value. There is no single explanation for why these values may differ. The HCO's lab may use a slightly different formula, or may report out any value greater than the lower limit of normal as equal to that lower limit, so that all

GFR values greater than or equal to 90 mL/min/1.73m^2 as appear in the dataset as exactly 90.

Back to question list

How do you report a left ventricular ejection fraction (LVEF) that's given as a range?

Occasionally, an HCO will report a LVEF as a range. We convert these ranges to a single numeric value by taking the mean of the lower and upper bounds. For example, if an HCO reports an LVEF of 40% - 60%, the lab value table in our dataset will show 50.

Back to question list

The codes indicated in the Medication Drugs table do not correspond to any codes in the Standardized Terminology table. How can I map them to a drug name?

Two tables provide medication data: medication_ingredient.csv and medication_drug.csv

All codes in the **Medication Ingredient** table refer to drug ingredients. They are always RxNorm codes, referring to concepts of the type IN in RxNorm terminology. RxNorm ingredient codes are included in the Standardized Terminology table of your dataset.

All codes in the **Medication Drugs** table refer to *drugs*. Drug-level codes do not appear in the Standardized Terminology table of your dataset. To translate a drug code to its corresponding name, you will need to consult external sources. We can recommend RxNorm and FDA's "National Drug Code Directory." RxNorm contains NDC codes in addition RxNorm codes. The NDC resource contains only NDC codes.

RxNorm

User interface for looking up one code at a time	https://mor.nlm.nih.gov/RxNav/
User interface for processing many codes in a batch	https://mor.nlm.nih.gov/RxMix/
RxNorm API documentation	https://rxnav.nlm.nih.gov/APIsOverview.html
RxNorm file downloads*	https://www.nlm.nih.gov/research/umls/rxnorm/docs/rxnormfiles.html

Note that for RxNorm file downloads, you will first need to obtain sign in credentials from Unified Medical Language Systems (UMLS). These credentials are free. To create an account, visit: https://uts.nlm.nih.gov/uts/signup-login

NDC
https://www.fda.gov/drugs/drug-approvals-and-databases/national-drug-code-directory

Back to question list

Is there always a one-to-one correspondence between the medication ingredients listed for a patient encounter and the drug for that same patient encounter?

There is not.

In many cases, HCOs contributing to a dataset have provided medication data at the level of ingredients and not the drug (i.e., a combination of generic or brand name, strength, form, administration, etc.). In these cases, a patient may have data in medication_ingredient.csv but not in medication_drug.csv. Be sure to review both tables if you want to capture all medication data for a particular patient or patient encounter.

On the other hand, if an HCO does supply medication information at the drug level, your dataset will contain its ingredients on individual rows within the Medication Ingredient table, associated with the same patient id and encounter id. The drug will also appear in the Medication Drug table.

Back to question list

Why are medication rows duplicated?

Depending on how the HCO has structured medication details in their EHR, a drug's name may occupy one row of the medication table, while its route, brand, or strength may occupy one or more subsequent rows. Duplicate rows may also represent two or more administrations of the same drug on the same day.

Back to question list

Getting Started

Can I define my dataset (e.g., by diagnosis, medication, age of patients, date range, etc.)?

Yes. Defining a dataset begins on our Query Builder page, where you can apply as many terms (e.g. diagnoses, medications, procedures), qualifiers (e.g. cancer stages, medication brands), temporal conditions (e.g. before or after, within the last year), and logical operators ('must have', 'cannot have') as you wish.

Our Account Managers and Clinical Scientists can help you design a query that will generate the appropriate cohort for your needs.

Back to question list

What are the authorized uses?

TriNetX data may be used only for healthcare research, defined as follows:

Healthcare research refers to the branch of healthcare science that determines the safety and effectiveness (efficacy) of medications, devices, diagnostic products and treatment regimens intended for human use. These may be used for prevention, treatment, diagnosis or for relieving symptoms of a disease or medical condition. Research includes health economics and outcomes research.

Back to question list

How do I request a dataset if I'm a TriNetX subscriber?

You will need:

- access to a network from which datasets may be generated (as of September 2020, these include Dataworks, Diamond, and Brazil)
- the Procure datasets permission enabled

If you do not have Procure datasets enabled, you can ask your organization administrator to enable it on your account, or share the study containing your query with someone in your organization who already has it enabled.

To submit a dataset request:

1. Using query builder, create one or more queries on the Research network to define cohorts to include in the dataset.
2. Click the Request Dataset button.
3. Select one or more cohorts.
4. Click the Request button.
5. Review the summary of your selections.
6. If you have the procure permission enabled, enter a message to include with your request and click Submit. If you do not have the procure permission enabled, select a Procurer and enter a message to the Procurer. Click Share with Procurer.
7. If you have the procure permission enabled, when you click Submit your study will be shared with TriNetX. A TriNetX team member will reach out to you to begin the dataset licensing process. If you do not have the procure permission enabled your study will be shared with the procurer you selected who can then send the dataset request to TriNetX on your behalf.

Back to question list

I'm not a subscriber. Can I still license a dataset?

Yes. If you've been speaking to a TriNetX representative, he or she can quickly put you in touch with a query designer who will help you define your dataset. Alternatively, email join@trinetx.com with a summary of your data needs.

Back to question list