STARR-OMOP-Documentation-Technical-Specifications

STARR-OMOP Documentation

Technical Specifications

November 2025

Table of Contents

Summary

The technical document is updated only if there is a change in ETL resulting in a qualitative difference in the data. The underlying database is refreshed more frequently.

OMOP Version	V 5.3.1, https://github.com/OHDSI/CommonDataModel/releases/tag/v5.3.1
Themis rules	https://github.com/OHDSI/Themis
High level publication	https://arxiv.org/abs/2003.10534; rationale, technologies, analytics
User Guide	https://docs.google.com/document/d/1dTjCEvvU8sMd8CuaVx94YFoPTsW_xjOMoF9-xl5zHik/; getting access, etc.
Data Sources	We maintain the source information but otherwise combine data from: ● Children's Hospital Clarity ● Adult Hospital Clarity
Data Filters	There is data in Epic Hyperspace that is not available in Clarity due to technical reasons. There is data in Clarity that is not available for research use due to regulatory reasons. Finally, there is data that is filtered out if the OMOP quality checks fail. Even with all these filters, we have data from 2.5+ million patients.
Data types in OMOP	Aside from typical EMR data that you expect in OMOP CDM, the database contains: ● Clinical text (in NOTE table) and flowsheets in the observation table, inside JSON strings. ● Text mining output (in NOTE_NLP table)
Data sets	Pre-IRB datasets: ● STARR-OMOP-confidential (fka STARR-OMOP-deid) ● STARR-OMOP-confidential-1pcent (for query development and testing) (fka STARR-OMOP-deid-1pcent) Post-IRB dataset (currently only accessible via Research Informatics Center consultation): ● STARR-OMOP
Stanford OHDSI ATLAS	OHDSI ATLAS is on top of the STARR-OMOP-confidential (fka STARR-OMOP deid) dataset, where the clinical text and flowsheets are redacted. The underlying dataset is referred to as STARR-OMOP-confidential-lite (fka STARR-OMOP-deid-lite). ATLAS does not support these data fields. For more information, visit the ATLAS user guide.

More information about STARR	https://starr.sites.stanford.edu/; for training, other products etc.

Introduction

Research IT at Stanford Medicine brings data from the two Hospitals, normalizes it for research use, and makes it accessible to the research community. STARR-OMOP CDW launched in 2019 is part of the STARR portfolio. It is a database where Epic Clarity data from the Hospitals is converted to OMOP CDM. Note, a) not all Stanford patients or their encounters will have a corresponding entry in the research data warehouse and, b) OHDSI has a well-defined ETL process as well as stringent quality criteria, and some data may be excluded if it does not meet the inclusion standards. For more information about STARR-OMOP, please review our manuscript on the pre-print server.

Data Guide

1.1 Data Details

1.1.1 OMOP-CDM Schema

Please refer to the manuscript, “Supplementary Material, Section 5: Data transformation from EHR to OMOP CDM”. The transformation process from Clarity to the OMOP CDM can be summarized in five processes: ETL specifications, ETL code, mappings, data quality, and data release. To design our ETL, we currently use OMOP-CDM v 5.3.1. The OHDSI consortium OMOP CDM GitHub repository specifies a list of all the tables present in the 5.3.1 schema and the explanation for each one of the columns. The repository also contains a guide to populate each of the tables, THEMIS rules, which contain high-level specifications for ETL.

Among all the schema, the following are considered clinical event tables: CONDITION_OCCURRENCE, PROCEDURE_OCCURRENCE, DRUG_EXPOSURE, DEVICE_EXPOSURE, MEASUREMENT, AND OBSERVATION.

Other nuances of the STARR OMOP dataset are documented here.

11.1.2 Known Additional Sources

In this section, we mention probable sources for data that are not currently brought in into the OMOP- CDM. In the following, we will go table by table and mention probable sources or problems.

- DEATH
    - Cause_concept_id: Not populated in any of the records. A record is created in the death table for each record in the EPIC Clarity patient table where the death date is populated. There is no cause of death information in the EPIC Clarity Patient table, and we have not identified another table from which the cause of death can be obtained.
- DEVICE_EXPOSURE
    - Sources for blood product data have not been identified yet. Blood products will be stored in device_exposure according to conventions for populating the OMOP-CDM.
- PROVIDER
    - care_site_id: Not populated in 92% of the records.

1.1.3 Some Known Issues

In this section, we describe some known issues. The issues described are particular to the requirements of the OMOP-CDM implementation and do not constitute an issue with the source data itself.

- DRUG_EXPOSURE: Multi-day infusions are currently stored as a set of separate one-day records, where drug_exposure_start_date and drug_exposure_end_date have the same value, and there is no logical link between a set of such records. A future enhancement would collapse the set into a single record with drug_exposure_start_date set to the first day of the infusion and drug_exposure_end_date set to the last day of the infusion.
- DRUG_EXPOSURE: missing drugs with order_status = NULL
- DRUG_EXPOSURE: end dates are not properly populated not seeing an end date for outpatient prescriptions in many cases
- MEASUREMENT
    - operator_concept_id: is set to 0 in all records. However, there are free text measurement values that contain operators, and therefore there is an opportunity to extract those operators and populate operator_concept_id with their equivalent standard concepts.
- OBSERVATION
    - The flowsheet data has been incorporated into the observation table as JSON strings. Rows in the observation table where concept_id = 2000006253 contain these JSON strings.

1.2 ETL Specifications

The process of Extraction Load and Transformation was made in partnership with Odysseus Data Services. The specifications focus on the necessary rules to bring data from EPIC Clarity into the CDM. This specification and the corresponding SQL code developed is available upon request.

Please contact Research-IT if you are in need of those resources by emailing starr-omop-support@stanford.edu

1.3 Population of NOTE_NLP table: Annotations from Clinical Text

Please refer to our manuscript, Supplementary Material, “Section 7: Processing clinical text to identify known medical concepts”.
Please Note**: As of August 2022; we have discontinued generating this table. You can find the last populated NLP table in the dataset:
som-rit-phi-starr-trek.starr_omop_cdm5_deid_2022_08_10.note_nlp
Every with STARR OMOP confidential access in BQ should be able to access it.

1.4 De-identification of structured and unstructured data

Detailed information of the de-identification method can be found in our manuscript, Supplementary Material, “Section 6: De-identification methods”.

Here we will only focus on the operations performed to PHI found inside the free text that must be synchronized with the operations performed in PHI columns. Additionally, the de-identification pipeline supports a variety of “transformation” rules (aka privacy operations) that can be prescribed at runtime. PHI column in CDM can be categorized in three types: high risk identifier (e.g. person_id), low risk identifier (e.g. observation_period_id) and dates (e.g. visit_start_date). For high and low-risk identifiers we define three possible privacy operations:

- Removal (**del**): in this case, the column is entirely deleted from the dataset.
- Deletion of content (**del_cnt**): in this case, the contents of the column are deleted, but the column itself is preserved. This is particularly useful for OMOP-CDM since deleting a required column will “break” the schema.
- Substitution with a random identifier (**sub_rand**): the identifier in this column is substituted with a random identifier. Since the mapping is one-to-one the full list of originals and new identifiers must be stored. This is the preferred method for high_risk identifiers.
- Substitution by addition with a random number (**add_rand**): a random number is added to the identifier in this column. Since it is an addition, only one random number generated is stored. This is the preferred method for low-risk identifiers.

For dates there we define the following operations:

- Removal (**del**): in this case, the column is entirely deleted from the dataset.
- Deletion of content (**del_cnt**): in this case, the contents of the column are deleted, but the column itself is preserved. This is particularly useful for OMOP-CDM since deleting a required column will “break” the schema.

- Truncation (**trunc**): in this case, the date is truncated by preserving year only.
- Patient-age- at-event (**age_evt**): In this method, the age of the patient at a given event is given instead of the actual date. This is the less preferable method given that it is easier to reconstruct patient birthdate.
- Jittering (**jitt_date**): A random number is added to the date. This random number is unique per patient and must be applied to every event for the given patient. This preserves timeline which is highly desirable for longitudinal studies. This is the preferred method at Stanford.

The types of PHI found in the content of free text columns are the same as the ones described above: high and low-risk identifiers and dates. However, the operations here vary for high and low-risk identifiers. The operations for dates are the same. The possible operations for high and low-risk identifiers found in free text columns are:

- Masking (**msk**): this method substitutes the identifiers with a mask that only refers to the HIPAA type (e.g. [NAME]). This is the least desirable method given that legibility is impacted. This is important for chart review. Also, the performance of NLP tools is impacted.
- Generic replacement (**gen_rep**): here the identifier is substituted by a credible generic non-random replacement. For example, every time an SSN is found, this is replaced with 999-99-9999. This is the preferable model for high-risk identifiers.
- Surrogate (**surrogate**): here the identifier is replaced by a credible random substitute. For example, the name John Smith will be replaced by Dan Brown. This option is only available at the moment for names and locations.

For the OMOP-confidential (fka deid) dataset:

- Policy for patients with age > 90**: Once the patient turns 90 all the events associated with the patient are removed.
- **In collaboration with Stanford UPO**, we apply time jitter of plus or minus 30 days, never zero. Currently, we use the same jitter for the date of birth, date of death and the hospital events.

In collaboration with UPO, we have implemented a special processing for zipcodes in OMOP-confidential that has been arrived at in collaboration with Stanford UPO. When the zipcode (to be precise, if the ZCTA) has >20,000 people living in them, we are allowed to present the five digit ZCTA in the PHI scrubbed data. In most cases, the zip5 and ZCTA are geographically close but in a small number of cases, they are geographically dissimilar. For all practical research purposes, if zip5 works, then ZCTA works. We are currently using the 2020 census. For the rest of the patients, we retain the first three digits (***00) and the remaining are set to zero.

Note further that the zipcode represents a postal delivery route and can change over the course of a decade. The zip+4 changes even more frequently. The five-digit zipcode changes infrequently. These zipcode changes happen due to post office openings, closures, and boundary changes. So, in a small number of patients, the older five-digit zipcode stored in EHR may not represent the zipcode corresponding to the street address.

Also note that for social determinants of health studies, the ZCAT is often too large an area and too socio-economically diverse. A smaller, more homogenous sub-division is likely to be a census tract. But the census tract often has a very small population (<4000). Census tract data is deemed a Limited Data Set and is therefore not available in OMOP-confidential. Also, note that zip+4 is an identifier. In some circumstances, the 9-digit number can point to a single address.

Release Notes

3.1 Alpha Release (Aug 2019)

1.1. OMOP datasets were generated EPIC Clarity data from:
- 1.1.1. July 7, 2019, for LPCH and
- 1.1.2. July 3, 2019, for SHC.

1.2. Known Issues in this version of the ETL code:
- 1.2.1. Multi-day infusions are currently stored as a set of separate one-day records, where drug_exposure_start_date and drug_exposure_end_date have the same value, and there is no logical link between a set of such records.
- 1.2.2. *operator_concept_id** is set to 0 in all records. However, there are free text measurement values that contain operators, which can be extracted and used to populate operator_concept_id with their equivalent standard concepts.
- 1.2.3. There are records in the EPIC Clarity™ ORDER_RESULT table where ord_num_value=`9999999` and ord_value has useful data. Further analysis is needed in order to define a rule for including these records in the ETL that populates this table.
- 1.2.4. Codes from Billing data are not included in the Procedure and Diagnosis Tables.

3.2 December 2019 Release

1.3. OMOP datasets were generated using EPIC Clarity data from:
- 1.3.1. Nov 4, 2019, for LPCH and
- 1.3.2. Sept 29, 2019, for SHC.

1.4. This dataset includes the following improvements to the ETL code:
- 1.4.1. Flowsheet data is now included in the OBSERVATIONS table as JSON strings.

1.4.2. Diagnosis and Procedure Codes from billing data are incorporated into the CONDITION OCCURRENCE and PROCEDURE OCCURENCE tables.
1.4.3. LOINC code mappings (which had been done by the Shah lab) have been integrated into the MEASUREMENTS table
1.4.4. The PERSON_ID column is now stable between releases: i.e. If you build a cohort using a STARR OMOP de-id dataset; and rebuild it in future releases, the PERSON_IDS in your cohort will remain stable.
1.4.5. A timestamp Issue has been resolved: All dates are now in DateTime format and comply with the BigQuery interpretation of Datetime.(https://cloud.google.com/bigquery/docs/reference/standard-sql/data-types#datetime-type)

3.3 March 2020 Release

1.5. New OMOP datasets were generated using EPIC Clarity data from:
1.5.1. Feb 23, 2020, for LPCH and
1.5.2. Feb 22, 2020, for SHC.
1.6. This dataset includes the following improvements and enhancements to the OMOP ETL since Dec 2019
1.6.1. Canceled Labs have been filtered out and are no longer included
1.6.2. A large number of categorical Lab values have been mapped to their appropriate concepts_ids
1.6.3. Labs in the Measurement table containing value_as_number = 9999999.0 have been replaced by value_as_number = NULL

3.4 May 2020 Release

1.7. New OMOP datasets were generated using the following (more recent) data snapshots from Clarity:
1.7.1. May 23, 2020, for LPCH and
1.7.2. May 23, 2020, for SHC.
1.8. This version of the dataset includes the following significant improvements and enhancements to the OMOP ETL since March 2020 resulting in a more complete VISIT_OCCURRENCE table.
1.8.1. Visit_concept_id has been re-mapped resulting in more granular visit_concept_ids. The visit_concept_id now includes:
1.8.1.1. Laboratory Visit
1.8.1.2. Outpatient Visit
1.8.1.3. Inpatient Visit
1.8.1.4. Emergency Room and Inpatient Visit

1.8.1.5. Emergency Room Visit
1.8.1.6. Telehealth
1.8.1.7. No matching Concept- these are typically historical encounters, scans, documentation, letters, etc

Note that the earlier mapping only included Outpatient visits, Inpatient visits, and Emergency Room Visits. The table below shows the improved visit_concept_id counts in this OMOP dataset.

Row	concept_name	visit_concept_id	Counts
1	Telehealth	5083	179120
2	Laboratory Visit	32036	676261
3	Outpatient Visit	9202	7386164
4	Emergency Room Visit	9203	689852
5	No matching concept	0	51406577
6	Emergency Room and Inpatient Visit	262	173781
7	Inpatient Visit	9201	405552
8	Office Visit	581477	8445802

Table 1: Visit_concept_id counts (OMOP data from 5-23-2020):

1.8.2. If a hospital admission/discharge time was available, then that was used to populate visit start/end datetime. Else, effective datetime was used. Furthermore, we have now brought in the time component to the visit end/start datetime fields where available.
1.8.3. We are now filtering out (removing) visit dates before 2000-01-01 and greater than 6 months into the future. (Note: the source data often contains future dates corresponding to future scheduled visits).
1.8.4. Date information for the hospital admission time has been brought in from the Clarity *pat_enc_hsp* table wherever available.
1.8.5. Please note that this is *Phase 1* of the re-mapping. Furthermore:
1.8.5.1. For patients not discharged yet, the visit end date is set to the visit start date. This is because the CDM expects a value. For Covid19 cases, this will result in a handful of patients with erroneous discharge dates until they are actually discharged. This is because visit_end_date is required in the CDM conventions.

1.8.5.2. There can be inaccuracies due to mapping assumptions or variance in data entry by the thousands of providers.
1.8.5.3. Observation status was categorized as an inpatient, e-visits as telehealth, even though not typical of definitions in the clinic.

1.9. This version includes improvements to the MEASUREMENT table. Now, value_as_concept_id is populated for approximately 10% of the categorical lab values. This includes mapping for the results of the COVID test. This enables the use of public COVID cohorts on ATLAS.

3.5 June 2020 Release

1.10. New OMOP datasets were generated using EPIC Clarity data from:
1.10.1. June 20, 2020, for LPCH and
1.10.2. June 20, 2020, for SHC.
1.11. The VISIT_DETAIL table has been populated for the first time.
1.11.1. The data for this table is primarily sourced from the Clarity ADT (Admit Discharge Transfers) table.
1.11.2. Currently, this table only includes mapping for SHC. LPCH records are in the table but are not yet mapped and will be available soon.
1.11.3. We are aware that not every inpatient visit in visit_occurrence has a corresponding record in visit_detail. This discrepancy is being investigated.

3.6 July 2020 Release

1.12. New OMOP datasets were generated using the following (more recent) data snapshots from Clarity:
1.12.1. July, 4 2020, for LPCH and
1.12.2. July 4, 2020, for SHC.
1.13. Furthermore, this version of the dataset includes the following improvements and enhancements to the OMOP ETL since June 2020.
1.13.1. New annotations have been brought into the Note_nlp table - resulting in more granular annotations. This has resulted in a 12 fold increase in the size of the note_nlp table.

3.7 September 2020 Release

1.14. New OMOP datasets were generated using the following (more recent) data snapshots from Clarity:
1.14.1. Sept 5 2020, for LPCH and
1.14.2. Sept 5, 2020, for SHC.
1.15. The vocabulary being used for the note_nlp annotations has been further improved:

1.15.1. Common words that are not individually clinically informative have been removed. The jupyter notebook that generated the final vocabulary is available at: https://github.com/susom/starr-public/blob/master/notebooks/Creation_of_Vocabulary_for_STARR_MINER.ipynb
This has resulted in a meaningful reduction of the size of the note_nlp table- only affecting words that were removed.
1.15.2. Vocabulary version being used is:
1.15.2.1. OHDSI Vocabulary Version: v 5.0 31-MAR-20
1.15.2.2. Stanford Vocabulary Version: vocabulary_20200717

3.8 November 2020 Release

1.16. New OMOP datasets were generated using the following (more recent) data snapshots from Clarity:
1.16.1. Oct 24, 2020, for LPCH and
1.16.2. Oct 24, 2020, for SHC.
1.17. The location data has been incorporated and the LOCATION table has been populated in the identified OMOP. However, not all the location data is in.
1.18. The temperature data has been brought in from Flowsheets and is now included in the Measurement table. However, please note that the temperature data from flowsheets to the measurement table was incorporated; we have **not yet r**emoved the corresponding json records from the observation table.

3.9 December 2020 Release

1.19. New OMOP datasets were generated using the following (more recent) data snapshots from Clarity:
1.19.1. Dec 5, 2020, for LPCH and
1.19.2. Dect 5, 2020, for SHC.
1.20. De-identified zip code data has been incorporated into the LOCATION table in the starr_omop_cdm5_confidential_lite_latest(fka deid) OMOP dataset. However, not all the location/zip code data is in.
1.21. The logic used to keep/crop or de identify the zip codes based on the guidelines from HHS is the following:
1.21.1. The **"+4"** extension in the zipcodes which includes the hyphen and four digits after the five digits of the primary zip code are **always dropped.**
1.21.2. If a zip code in the census dataset has 3 or 4 or 5 digits, and it has more than 20,000 population, we keep the full zipcode(i.e 3-5 digits).

1.21.3. If a zip code has more than three digits, and if the zip code belongs to a group (grouped by first three digits) that has a combined population less than 20k, we keep the first 3 digits and set the remaining two digits to zero.
1.21.4. For all other cases, we remove the zip code completely.

3.10 March 2021 Release

1.22. New OMOP datasets were generated using the following (more recent) data snapshots from Clarity:
1.22.1. March 6, 2021, for LPCH and
1.22.2. March 6, 2022, for SHC.
1.23. We released and incorporated and updated the vocabulary used for generating STARR OMOP. This new vocabulary version corresponds to:
1.23.1. OHDSI Vocabulary Version: v5.0 06-NOV-20 and
1.23.2. Stanford Vocabulary Version: vocabulary_20210304
1.23.3. As a result of using the new vocabulary, we have :
1.23.3.1. New mappings for visits
1.23.3.2. new drugs in (e.g. redesmivir)
1.23.3.3. More codes/standard concepts that have been included in the OHDSI vocabulary as of November 2020.
1.24. We have now incorporated the following ~27 fields from flowsheet data into the Measurement table. This information is/has been available in the Observation table as json strings with observation_concept_id="2000006253"

	Concept_id	Loinc Code	Concept Name
1.	3005424	8277-6	Body surface area
2.	3020891	8310-5	Body temperature
3.	3025315	29463-7	Body weight
4.	21490675	60985-9	Central venous pressure (CVP)
5.	3012888	8462-4	Diastolic blood pressure
6.	21490565	60802-6	Dynamic plateau pressure
7.	3032652	35088-4	Glasgow coma scale
8.	3027018	8867-4	Heart rate
9.	3005629	3151-8	Inhaled oxygen flow rate
10.	21490581	60826-5	Lung compliance
11.	42527086	60949-5	Mean airway pressure
12.	3027598	8478-0	Mean blood pressure
13.	21490566	60804-2	Minimum alveolar concentration (MAC) for anesthesia.XXX Anesthetic agent.XX

14.	3045410	33425-0	Minute volume setting Ventilator
15.	21490615	60860-4	Nitrous oxide [VFr/PPres] Gas delivery system
16.	3024882	19994-3	Oxygen/Inspired gas setting [Volume Fraction] Ventilator
17.	21490855	76248-4	PEEP Respiratory system --on ventilator
18.	3036453	38214-3	Pain severity [Score] Visual analog score
19.	3011557	19931-5	Peak inspiratory gas flow setting Ventilator
20.	3025809	8634-8	Q-T interval
21.	3026258	8636-3	Q-T interval corrected
22.	3024171	9279-1	Respiratory rate
23.	21490553	60782-0	Sevoflurane gas delivered during case [Volume] from Gas delivery system
24.	3004249	8480-6	Systolic blood pressure
25.	3025853	20140-0	Volume expired
26.	3036277	8302-2	Body height

3.11 February 2022 Release

1.25. The condition_type_concept_id in the condition_occurrence table now includes the "Primary Diagnosis"

3.12 July 2022 Release

1.26. The payer_plan_period table is now populated in the latest refresh of STARR-OMOP-confidential (fka deid). This table captures details of the period of time that a Person is continuously enrolled under a specific health Plan benefit structure from a given Payer. The payer_concept_id represents the organization who reimburses the provider which administers care to the Person and has been mapped to the standard_concept_id (Accepted Concepts.). The payer_source_value contains payer_ SOPT Desc | SOPT ID | Epic Financial Class Name | Epic Financial Class ID.

Note: SOPT is the payer type standard that provides a mechanism for consistent reporting of payer data to public health agencies for health care services and research

1.27. We are now using the LOINC mappings (for lab measurements) provided to us by the hospital instead of our own custom mappings for lab LOINC codes. **Please**

review your current lab queries. We recommend using the Atlas tool to generate a set of LOINC codes for your labs of interest.

3.13 August 2022 Release

1.28. The full confidential (fka deid) OMOP OBSERVATION table contains flowsheets as jsons (See Release Dec 2019) . In the deidentification process, strictly numeric values in the value_as_string were getting nullified- we have now fixed this so that if the measured flowsheet value is numeric; the actual value is coming through. This was a bug fix and will allow researchers to actually access/use the values. Note that values containing characters like “>” etc will continue to be nullified.

1.29. The lab measurement_source_value column in the measurement table has now been modified to contain a pipe-delimited string of 3 values:

- **LOINC-code|base_name|component_name.** For example, 20584-9|WBC|WHITE BLOOD CELLS (WBC).
- In the past, labs only contained the LOINC code. The base name is a grouper with which researchers are more familiar.

3.14 May Release (May 2023)

1.30 **Oncology Data**: STARR-OMOP has been populated with tumor data from SHC's Epic Beacon. The tumor registry data will also be included in a future release. In summary, the integration of data from the Beacon module adds data for ~56,000 patients across ~20 years. Since 2016, there has been an average of ~6700 patients/year with data in Beacon. With this integration, researchers can search by histology, topography, clinical stage, pathologic stage, clinical/path TNM scores, and grade. Future plans include integration of laterality, tumor size, hormone receptor statuses, Gleason score, recurrence data, etc.

As of May 2023, the tumor data included are:

- Tumor histology and topography using the ICD-O-3 vocabulary. An example of the format is '8140/3-C34.9' where 8140/3 indicates the histology if available and C34.9 specifies the topography/site. The SEER ICD-O-3 site/histology validation list can be found here: https://seer.cancer.gov/icd-o-3/sitetype.icdo3.20220429.pdf
  The Athena vocabulary id is 'ICDO3'.
- The condition start date of the tumor is derived from the staging date. If the tumor was staged both clinically and pathologically but on different dates, then STARR-OMOP will contain two separate rows. One row will represent the clinical condition date and code, while another row will represent the

pathologic condition date and code. If the staging date is not available, the noted date from the problem list is used. If the noted date is null, then the contact date from the Epic Beacon summary table is used to populate the condition start date.

- Basic tumor descriptives are located in the MEASUREMENT table. The sample code below will show you how to link the ICDO3 CONDITION_OCCURRENCE data to the respective tumor data in MEASUREMENT using the modifier_of_event_id variable. The following tumor metadata have been populated using the 'Cancer Modifier' vocabulary in Athena, which the Oncology Working Group has decided will be the standard vocabulary for these variables. Other tumor variables will continue to be represented with the 'NAACCR' vocabulary.

Note: If you want to query all clinical T1* scores (T1, T1a, T1b, etc. of all AJCC editions), the CONCEPT_RELATIONSHIP table where the relationship_id 'subsumes' is not quite complete yet. You'll need to perform string matching on the concept_code '%clinical T1%'.

- More information can be found in the OHDSI community sites: https://github.com/OHDSI/OncologyWG
- A jupyter notebook with sample queries to explore the oncology data has been added to STARR gitlab and is available under a new notebook titled Tutorial5_Oncology_queries.ipynb

Following table presents a snapshot in time showing the distribution of all tumor data:

concept_name	concept_code	num_distinct_persons
Neoplasm defined only by topography: Breast, NOS	NULL-C50.9	9014
Neoplasm defined only by topography: Prostate gland	NULL-C61.9	6089
Neoplasm defined only by topography: Lung, NOS	NULL-C34.9	3763
Neoplasm defined only by topography: Bladder, NOS	NULL-C67.9	1829
Neoplasm defined only by topography: Colon, NOS	NULL-C18.9	1551
Neoplasm defined only by topography: Rectum, NOS	NULL-C20.9	1373
Neoplasm defined only by topography: Thyroid gland	NULL-C73.9	1278
Neoplasm defined only by topography: Endometrium	NULL-C54.1	1147
Neoplasm defined only by topography: Pancreas, NOS	NULL-C25.9	1040
Infiltrating duct carcinoma, NOS, of breast, NOS	8500/3-C50.9	1003
Neoplasm defined only by topography: Ovary	NULL-C56.9	948
Neoplasm defined only by topography: Connective, Subcutaneous and othe	NULL-C49.9	711
Neoplasm defined only by topography: Upper lobe, lung	NULL-C34.1	678
Neoplasm defined only by topography: Kidney, NOS	NULL-C64.9	626
Kidney, NOS	C64.9	564
Neoplasm defined only by topography: Upper-outer quadrant of breast	NULL-C50.4	562
Neoplasm defined only by topography: Cervix uteri	NULL-C53.9	454
Neoplasm defined only by topography: Stomach, NOS	NULL-C16.9	447
Neoplasm defined only by topography: Liver	NULL-C22.0	430
Adenocarcinoma, NOS, of lung, NOS	8140/3-C34.9	425
Neoplasm defined only by topography: Lower lobe, lung	NULL-C34.3	416
Neoplasm defined only by topography: Esophagus, NOS	NULL-C15.9	393
Neoplasm defined only by topography: Overlapping lesion of lung	NULL-C34.8	374
Neoplasm defined only by topography: Testis, NOS	NULL-C62.9	368
Neoplasm defined only by topography: Trachea	NULL-C33.9	362
Neoplasm defined only by topography: Base of tongue, NOS	NULL-C01.9	346
Adenocarcinoma, NOS, of prostate gland	8140/3-C61.9	338
Neoplasm defined only by topography: Tonsil, NOS	NULL-C09.9	330
Neoplasm defined only by topography: Sigmoid colon	NULL-C18.7	328
Neoplasm defined only by topography: Uterus, NOS	NULL-C55.9	308
Neoplasm defined only by topography: Overlapping lesion of bladder	NULL-C67.8	302
Neoplasm defined only by topography: intrahepatic bile duct	NULL-C22.1	266
Neoplasm defined only by topography: Oropharynx, NOS	NULL-C10.9	262

The above table was constructed by querying CONDITION_OCCURRENCE and filtering on the condition_concept_id where the concept is part of the 'ICDO3' vocabulary.

1.31 All MUMPS(Massachusetts General Hospital Utility Multi-Programming System) dates in the json strings in the Observation table in STARR-OMOP have now been converted to timestamp dates. In the confidential data, the dates have been jitters as per the patient's codebook. (EPIC Chronicles is written using the MUMPS language -This format returns the date as the number of days since 1st January 1841, and the time as the number of seconds since midnight.)

3.15 August Release (August 2023)

1.32 Location Table

1. Identified OMOP:
    a. The identified OMOP location table now contains detailed addresses for Care sites in the location source value column. These locations have been brought in using the Clarity effective_dept_id variable from the patient encounter table. These previously just denoted by a generic SHC and LPCH address.

   b. Identified OMOP location table now contains the latitude, longitude and Census block group for the patient data based on the 2020 Census and primarily using the API available from Census.gov (add link). If data is not available via the Census API, we use DeGAUSS or leave the fields as null.
2. Confidential(fka Deid) OMOP:
   a. Each caresite location has a distinct location_id in the location table compared to the generic location_id=1 (LPCH) and location_id=2 (SHC)
   b. No geolocation data is available in the confidential location table.

1.33 Care Site Table:
1. Identified OMOP:
   a. Every distinct care_site has a unique location_id which is reflected in the location table. In earlier versions, all SHC care_sites were assigned location_id=2, and LPCH caresites were location_id=1. Furthermore, the source of the address is now using the Clarity effective_dept_id variable from the patient encounter table which we believe to be more accurate.
2. Confidential(fka Deid) OMOP
   a. Every distinct care_site has a unique location_id which is reflected in the location table.

1.34 Person table:
1. Identified OMOP:
   a. care_site_id column and provider_id column has been populated using the patients current Primary Care Providers name and department if available in the clarity patient.cur_pcp_prov_id. Its left blank if not available
2. Confidential(fka Deid) OMOP
   a. care_site_id column and provider_id column has been populated using the provider_id and care_site_id from the corresponding provider and care_site table.
   b. person_source_value is now being populated with an anonymous MRN that can be used as a link to other datasets using the same cohort and IRB# (eg. STARR-Radio)

3.16 September 2023 release

1.35 The NetAgility instance which contained employee data is being phased out - and most of that employee data is being brought into EPIC in stages. As a result, we are now adding filters to remove encounter data that occurred in service areas marked as “Occupational Health Services” or “Workforce Health and Wellness”. This filtering started in mid September.

3.17 October 2023 Release

Feature Improvements

1.36 OMOP dataset name change
Beginning the week of Oct 9th, 2023 the OMOP dataset naming convention is modified as follows:
STARR-OMOP full

- starr_omop_cdm5_deid_latest → starr_omop_cdm5_**confidential**_latest
- starr_omop_cdm5_deid_yyyy_mm_dd → starr_omop_cdm5_**confidential**_yyyy_mm_dd

STARR-OMOP- lite

- starr_omop_cdm5_deid_lite_latest → starr_omop_cdm5_**confidential**_lite_latest
- starr_omop_cdm5_deid_yyyy_mm_dd → starr_omop_cdm5_**confidential**_yyyy_mm_dd

1% subsets will follow the same convention as above.
Note that for a period of time, users will continue to see some older datasets with the old naming convention (till those datasets get lifecycled out). This change is being made to underline the fact that while best efforts to de-identify these datasets have been made, they are still deemed High-risk by the Privacy Office

1.36 measurement_type_concept_id and visit_type_concept_id have been modified and now use standard concept ids from the domain Type Concept

1.37 Ophthalmology intraocular pressure from both eyes is now part of the measurement table. This data originally resides in Smartdata elements in EPIC.

1.38 We have added timing information for Outpatient and Office visits wherever available. This added timing information for about 10% of the visits-

1.39 In the OMOP measurement table, we are now prioritizing LOINC codes from clarity_component rather than lnc_db_main as we believe this table is actively maintained by SHC EPIC team. This also fixed the WBC LOINC codes that were being inaccurately mapped to the RBC base names.

Bug Fixes

1.40 We had noticed that there were images embedded in some of the notes text fields- These are now being actively removed.

1.41 There are duplicates in the Notes table. We are now joining on the src_flag to ensure that the correct provider (from LPCH or SHC) is mapped to the correct note. This has resulted in a reduction in duplicates and mislabelled notes.

1.42 Removed data coming in from order_proc_4- resulted in a reduction of rows in the measurement table where an order was repeated - once as an order and once with the result. Hasnt gotten rid of all of the “duplicates” -but reduced the measurement bloat. We estimate the reduction in rows to be ~20%.

1.43 Slight reduction in the drug_exposure table bloat as a result of ensuring that the correct src flags are used to get the values of the sig column.

3.18 November 2023 Release

Feature Improvements

1.44 We have added additional patient ethnicity data to the OMOP observation table whenever available. This data comes from the Clarity ethnic background table which is now getting populated at the hospitals and is in addition to the ethnic background information that was available in the person table (which primarily comes from the Clarity patient table). Note that as a result, in the recent OMOP has additional ethnic background information for ~570K patients. Note: there are patients for whom we have greater than 15 ethnicities/races that are self reported.

1.44 The OMOP vocabulary file has been updated to reflect these new ethnicity concepts

1.45 We have added individual measurement concept IDs for inspired and expired sevoflurane to the measurement table. This data previously resided in the observation table (as part of flowsheets). The adult hospital recently added separate fields for inspired sevoflurane/expired sevoflurane in the flowsheets and this is now reflected in OMOP. Please note that this level of granularity is not yet available for the Children’s hospital, and we continue to only report expired sevoflurane. Please note that the corresponding flowsheet rows have now been removed from the observation table, as they are already present in the measurement table.

Bugs Addressed

1.46 We had noticed a drop in the mapping rates in the admitting_source_concept_id, discharge_source_concept_id and visit_detail_concept_id. This was because we are doing an exact string matching to a concatenated string of values including department name, dept abbreviation, dept external name, and dept specialty; and this breaks if SHC adds any new

specialities etc. We have modified the ETL to fix that but the visit_detail source value continues to comprise of department name, dept abbreviation, dept external name, and dept specialty.

1.47 It was reported that drugs were missing from the OMOP drug_exposure table. We realized that this was because these drugs did not have a start date in the clarity order_med table (our primary source of getting the start dates); and start_dates are required in OMOP. We are now bringing in a reliable proxy for the start_date from a different column in the same table - “order_inst”, to the drug exposure ETL resulting in a 4% increase in the drug_exposure table. Note: we compared the dates brought in from order_inst with the dates in mar admin taken_date to ensure that they are consistent. The following table in BQ lists the drugs that are now part of the OMOP dataset as a result of this ETL change: `som-rit-phi-starr-prod.STARR_OMOP_Changes.23_11_19_drug_exposure_increase_in_pat_counts`

1.48 We are now protecting against joining providers & care sites from different Clarity systems that share the same ID by adding the source_flag in the join from provider to care_site. Please note that only a single department ID is duplicated across SHC and LPCH Clarity instances; this department was not indicated as the primary care site for any individual provider, so this change had no impact on the current OMOP.

3.19 December 2023 Release

Feature Improvements

1.49 Added more information on departments with non-specific names - there are some departments that have non-descriptive names (like B6,C8 etc) in the department name column in Clarity- however, often there is additional speciality/level of care information available in other Clarity fields that we can use to provide more specific department information that can be useful to researchers, and we are now adding that information to the care_site.name field by adding a pipe ‘ | ’ after the department name and then the specialty/level of care info. For example:

E3 → E3 | ONCOLOGY
D1 → D1 | ICU

1.50 ETL change to Observation _Period end date for the patients that have a non null death date: We have updated our Observation Period end date to reflect the OHDSI recommendation which is to use the earliest date from the following:

- Date of death + 60 days

    - This is a CDM convention to allow events after death (autopsy, final notes, etc).
- Last clinical event + 60 days
    - The assumption is that a patient will return to the same health provider if an adverse reaction/complication/unresolved condition occurs.
- Date of the data pulled from the system

Note that this logic is applied in the identified OMOP calculation of the observation_end_date, but gets jittered by +/-30 days in the confidential (fka de-id) dataset.

1.51 Increasing drug exposures brought into the table - we filter out all drugs without a valid start date since this is a key field. We found another source for start_date which allowed us to bring in more drug exposures (an increase of 4%) and improve the accuracy of the drug exposure table.

Bugs Addressed

1.52 Death_type_concept_id updated to a standard concept

1.53 Provider assignment in OMOP drug_exposure table had some issues- We had been incorrectly assuming that provider_id and user_id could be used interchangeably and were using them to join tables. We fixed the issue, and now have an additional join to go between provider_id and user_id correctly. As a result, 429 new providers have been brought in. We are also ensuring that the provider source tables from the corresponding hospital is being joined with the appropriate drug source tables (i.e. SHC to SHC and LPCH to LPCH)

3.20 Jan 2024 Release

Bugs Addressed

1.54 - Halved the number of historic RxNorm codes used in the drug_exposure table by adding code to select non-historic codes before historic ones where both are available in mappings from source
1.55 - Removed extraneous/incorrect mappings through ICD10 (international) vocabulary in condition_occurrence table; our source data only uses ICD10CM (U.S. version), but source codes were previously being treated as if they were in both ICD10 and ICD10CM vocabularies. This has been fixed so only ICD10CM is used. ICD9CM codes were unaffected.
1.56 - Corrected release date in cdm_source table.

1.57 - Use all available information to associate condition source codes with the appropriate vocabulary to avoid cases of ICD-9/10CM codes being classified as unmappable due to unknown vocabulary.
1.58 - Bring in ICD10CM source codes from transactions data table (previously only ICD9CM codes were brought in from the transactions table; ICD10CM codes were still being brought in from other source tables)
1.59 - Reassign the procedure_type_concept_ids in the procedure_occurrence to use standard type concepts

3.21 Feb 2024 Release

Feature Improvements

1.60 - In the SHC clinical workflow, the pharmacists' consults for 'Transition of Care(TOC)' are not marked as completed even though the consult was completed. Furthermore, when the patient is discharged, the system automatically cancels the remaining procedure orders. Since canceled orders are typically not brought into STARR-OMOP, we were missing a large number of TOC consults in OMOP. Specifically, code was added to the PROCEDURE_OCCURRENCE to bring in all 'Sent' and 'Cancelled' (only when the reason for cancellation is 'Patient Discharge') TOC consults.

The consults can be queried using the following filter:
where procedure_source_value in ('PHARMACIST MEDICATION EDUCATION','TRANSITIONS OF CARE (TOC) PHARMACIST CONSULT')
1.61 - The pathology reports (from the Beaker AP module) that had been missing from the notes table for the period after Oct 2022 are now in and will be updated regularly going forward. The following filter:
Where note_title = 'procedures' and load_table_id like '%impression_narrative'
will pull out these Beaker notes -however, please note that this filter will pull out the radiology report *as well.*

3.22 March 2024 Release

Feature Improvements

1.62—We have introduced a STARR-OMOP Feature Improvement/Bug Tracker Portal to provide transparency to our users and foster a proactive feedback loop between them and the development team. This portal should provide the status of bugs/issues reported, feature improvements planned, and an opportunity to report any other issues.

1.63 - Death Data from Social Security Administration Limited Access Death Master File (SSA LADMF) is now part of the identified OMOP dataset (_death_data_external column) and is available for Stanford researchers with an IRB. *Please note that we recommend you use this data with extreme caution.
1.64 - Self-pay data is now available for SHC and LPCH patients in the payer_plan_period table. This information is included in the payer_source_value column.

Bugs Addressed

1.65 - The ETL was fixed to re-instate the “source_code_description" column back to the “source_to_concept_map" table. This column was accidentally being dropped in the de-id process and although that table and column are empty; it was causing ATLAS to hang when the user tried to view the “Included Source Codes” of a concept set.

3.23 April 2024 Release

Feature Improvements

1.66 - Flowsheet data for inspiratory and expiratory sevoflurane is now brought in from the children’s hospital (previously, only inspiratory sevoflurane was available). This corresponds to the same addition as was done for the adult hospital in Nov ‘23.
1.67—The visit_source_value column in the visit_occurrence table now contains a text string that concatenates the fields used to determine the visit_type. The following fields are concatenated in this order, with ‘ | ‘ separators: encounter type (enc_type_c), admission/discharge/transfer patient class (adt_pat_class_c), hospital admission type (hosp_admsn_type_c), and visit type (appt_prc_id).

1.68 - Mother-baby linkage information is now available in the fact_relationship table. All relationships are directional, and each relationship is represented twice symmetrically within the FACT_RELATIONSHIP table. For example, if person_id = 1 is the mother of person_id = 2, there are two records are in the FACT_RELATIONSHIP table:

- If the relationship_concept_id=581437, which corresponds to Child to Parent measurement, then, then fact_id_1= person_id of the child and fact_id_2=person_id of the mother.
- If the relationship_concept_id=581436 which corresponds to Parent to Child measurement, then, then fact_id_1=person_id of the mother and fact_id_2=person_id of the child.

For more details, see here.

Bugs Addressed

1.69 - The fields visit_detail.discharge_to_source_value and visit_detail.visit_detail_parent_id have been added to the OMOP confidential dataset; they were previously being (accidentally) dropped in the process of de-identifying the OMOP dataset.
1.70 - The trace_id field on the fact_relationship table is now nullified in the confidential dataset as part of the de-identification process.
1.71- fix regex to avoid casting non-numbers to the value_as_number field.

3.24 July 2024 Release

Feature Improvements

1.72 - Note ids are now stable in identified, confidential and custom datasets.
1.73 - Procedure note types such as pathology, imaging, etc. are now specified explicitly in the note_title field of the NOTE table. We are still working on mapping these note types to concept ids.
1.74: Drug Exposure Overhaul in Production

- Limited source data to just three clarity tables - mar_admin_info, order_disp_meds, order_med
- Removed tables that were just subsets of mar or order_med (f_ip_hsp_sum_med_admin, pat_enc_ip_meds)
- Removed tables that should not be in there (mar_admin_info_edt which contained old / corrected MAR information)
- Only one row per drug exposure, not one row per drug exposure from each source
- Prioritized administration records, then dispensed records, then ordered records
- Removed PRN medications that were never administered (inpatient-mode medications from order_med that do not have a record in MAR)
- Included DISPENSED medication_ids for MAR records where possible (over ordered medication ids)
- Improved IV medication reporting by including only one row per continuous administration (defined as MAR actions within 3 hours of one another)
- Improved provenance labels (drug_type_concept_id) to make it clear whether medications are: EHR administration records, EHR order records, EHR dispense records or patient reported medications
- Improved order status inclusion to allow canceled, and suspended medications (these often just represent prescriptions that have been completed and are no longer needed)
- Improved MAR inclusion actions, including differentiating between SHC and LPCH mar actions, to include actions that mean 'the patient is currently receiving the medication'. The exception to this is 'rate check / verify' and 'rate change' which should be covered by IV medication section and would cause significant data bloat.

Changes to overall dataset:

- 63% reduction in drug exposure table
- ~40 providers removed from provider table
- ~1700 patients removed

3.25 October 2024 Release

Bugs Addressed

1.75 The dx_ids used in the Clarity tables are specific to each hospital; and hence the dx_id→ ICD9/10 mapping is *not the same* between hospitals. The ETL has been corrected to use the hospital-specific dx_id to ICD mappings for SHC and LPCH and this change is reflected in the condition_occurrence table. Also, we are utilizing multiple sources for mapping dx_ids to ICD codes, resulting in greater percentages of conditions having a standard concept ID assigned.

3.26 November 2024 Release

Feature Improvements

1.76 We plan to gradually bring in relevant labor and delivery data from the Clarity Stork module into OMOP as prioritized by the OHDSI Perinatal and Reproductive Health Working Group. In this phase we have brought in the following fields into the OMOP Observation table.

i) Blood loss during labor → mapped to observation_concept_id=4270157
ii) Parity (number of live births delivered) → mapped to observation_concept_id = 4264419
iii) Gravidity(number of pregnancies) → mapped to observation_concept_id=4060186

3.27 December 2024 Release

Feature Improvements

1.77 As part of the initiative to bring in relevant labor and delivery data from the Clarity Stork module, this month, we have brought in the following data from Clarity into the Observation and Measurement tables:

i) birthweight → mapped to measurement concept_id= 4264825
ii) Gestational age → mapped to observation concept_id = 44789950

iii) Apgar at 1 min → mapped to measurement concept_id= 3016704
iv) Apgar at 5 min → mapped to measurement concept_id= 3004221

3.29 February 2025 Release

Feature Improvements

1.79: We are in the process of bringing in medical imaging extension tables proposed by the OHDSI community. In the first phase, we plan to **only** populate the image_occurrence table, which will **initially be available** only in the identified OMOP. It will be made more broadly available in the confidential OMOP in the near future.

This table describes imaging events and links to DICOM images stored in PACS or VNA systems, representing collections of images acquired through imaging techniques. It aims to serve three purposes: linking DICOM images at the study or series level, capturing series-level attributes such as modality and anatomic site, and providing data provenance for the Image_feature table. The table includes study_UID and series_UID for retrieving DICOM data using the DICOMweb WADO standard. Additionally, standardized attributes like modality and anatomic site are included, with options to access data via local paths or URIs.

3.30 April 2025 Release

Feature Improvements

1.80: The *image_occurrence* table is now populated in the confidential dataset. It's still in progress, with ongoing updates to refine linkages to the VNA data.

Bugs Addressed

1.81: The gestational age values are now being translated by the de-identified pipeline and brought into confidential OMOP.

3.31 May 2025 Release

Feature Improvements

1.82: The following new columns in the *image_occurrence* table will be stable between releases:

- Image_study_uid
- Image_series_uid
- accession_number

Bug Fixes

1.83: Some IV medications were getting dropped in the ETL because the inclusion criteria for which medications 'IV' was based on a COALESCE(disp_med_id, dispensable_med_id). However, the IV section uses dispensable_med_id, and the non-IV section uses disp_med_id. As a result, some medications were getting skipped. This resulted in a 6.5% increase in the medications.

3.32 June 2025 Release

Bug Fixes

1.84: The stable note_ids in the OMOP note table changed between the May 2025 and June 2025 releases. This issue was caused by a bug that used outdated note_ids; and the issue was fixed in June. Going forward, note_ids will remain stable between releases.
Please note: if you are comparing note_ids before June 2025 vs. after June 2025, they will not align. Please contact starr-omop-support@stanford.edu with any questions or concerns.

3.33 July 2025 Release

Feature Improvements

1.85: The following columns were added to the *image_occurrence* table:

- _study_description
- _series_description

1.86: Allergy data was added to the observation table, resulting in ~2M row increase in the table

3.34 August 2025 Release

Feature Improvements

1.87: Expand the OMOP notes table to include unsigned notes, which we had previously excluded in favor of only signed notes. We decided to include them as they were important for certain projects. As a result, we see a 0.1 % increase in the notes table.

1.88: In the STARR-OMOP Common Data Model (CDM), encounters without an associated clinical event (condition, procedure, etc) are typically excluded from the visit_occurrence table, as they are considered non-events. However, to meet the requirements of specific ongoing projects using STARR OMOP, tumor board encounters will now be included in the visit_occurrence table to ensure completeness, regardless of whether any associated clinical events are present in OMOP.

3.35 November 2025 Release

Feature Improvements

1.89: Added _accession_number to the identified Notes table.

Glossary

- **BQ**: Google BigQuery
- **CDM**: Common Data Model
- **DOB**: Date of Birth
- **ETL**: Extraction, Transformation, and Load
- **LPCH**: Lucile Packard Children's Hospital Stanford (aka Stanford Children's Health)
- **MRN**: Medical Record Number
- **OHDSI**: Observational Health Data Sciences and Informatics
- **OMOP**: Observational Medical Outcomes Partnership
- **SHC**: Stanford Health Care

- **VNA:** Vendor Neutral Archive
- **PACS:** Picture Archive and Communication System