OHDSI Workgroup Objectives and Key Results (OKR)

May 2024 Update

Clinical Trials Workgroup leads: Mike Hamidi, Zhen Lin, Tom Walpole

CTWG Purpose

Objective: Our objective is to facilitate the integration of clinical trial data, formatted according to the CDISC SDTM standard, into the OMOP framework.

Approach: Our approach prioritizes minimal alterations to the OMOP CDM and Standardized Vocabularies. This ensures that OHDSI tools such as Atlas remain unaffected while enhancing the conversion process from SDTM to OMOP with minimal data loss. We propose introducing new concepts and modifiers through established conventions, without necessitating the creation of new CDM tables. Additionally, we offer guidance for ETL developers as needed. Originally based on OMOP CDM v6 and its Oncology extension, our proposals maintain backward compatibility with v5.3. With the integration of Oncology extension features into the standard v5.4, our modifications remain minimal, ensuring full compatibility with this latest version.

https://www.ohdsi.org/web/wiki/doku.php?id=projects:workgroups:clinicalstudy CTWG Overview

CTWG Accomplishments To Date

2024: Mapping specifications and ETL for the initial SDTM study are completed. Subsequent studies will adhere to the established conventions, with OMOP CDM guidelines being defined for implementation.

2023: Transitioned from Vivli real-world SDTM data due to environmental constraints, opting instead for tuberculosis study data sourced from C-Path, spanning over three clinical studies. Additionally, potential data sources from NIDA Data Share, including several real-world studies in SDTM format, and synthetic data options are identified. Mapping efforts have commenced for one of the C-Path TB clinical studies.

2022: The CTWG gained access to 20 Vivli clinical study packages in SDTM format. An inventory of these packages is underway to prioritize SDTM-to-OMOP mappings. Existing CTWG guidance topics are under review, with new topics being identified as needed.

2021: CTWG conducted an evaluation of clinical trial data providers offering SDTM data access, leading to engagements with Vivli for general data usage agreements and platform feasibility assessments.

2020: Utilized synthetic representations of CDISC SDTM data via PHUSE Test Data. Initial guidance topics were established but require validation with diverse real-world SDTM data. CTWG proposals were submitted to the OHDSI community in July 2020.

CTWG Challenges

Challenges arise from the Working Group's expertise limitations in mapping and reviewing SDTM-to-OMOP CDM tables. Additionally, there is a scarcity of technical SMEs available to develop ETL scripts and validate outcomes effectively. Despite these constraints, efforts are focused on maintaining a consistent set of rules applicable to forthcoming studies, balancing the conceptual and physical data application levels.

CTWG OKR

Objective #1:

To define the conceptual mappings and guidance to support CDISC SDTM-to-OMOP conversion

Key Result #1: ✓

Identify >=3 real-world SDTM clinical studies

– Sept. 2023: This has been achieved.

Key Result #2:

Develop SDTM-to-OMOP mapping specifications using a prioritized set of common SDTM domains (adverse events, vital signs, demographics, concomitant medications, laboratory test results, medical history, and procedures)

– May 2024: The first study has been mapped and currently undergoing further reviews. The next step is to develop the ETL scripts.

Key Result #3:

Publish draft SDTM-to-OMOP guidance by Q1 2024

- Conceptual mappings on key domains of interest

- Define OMOP CDM conventions

- Identified gaps, issues, and challenges

– May 2024: The mapping specifications and conventions are being document. The current activity is developing ETL scripts to test the mappings already defined. After which, the team will work on the guidance document.

CTWG Ask

- Additional sources of real-world clinical studies in SDTM format
- Any volunteers to support SDTM-to-OMOP ETL mappings
- Any organization actively working on SDTM-to-OMOP conversions that have lessons learned, supporting documentation, or ETL code

CTWG Mapping Overview

Tracker

CDM Table	SDTM Datasets	Mapping Owner(s)	Reviewer(s)	Status
PERSON	DM	Mike Hamidi		Ready for Review
OBSERVATION_PERIOD	TV, SV	Qi Yang		
VISIT_OCCURRENCE	SV	Qi Yang		Ready for Review
VISIT_DETAIL	TV, SV?	Do not populate		
CONDITION_OCCURRENCE	AE, MH, CE, PR	Mike Hamidi		Ready for Review
DRUG_EXPOSURE	EX, CM	Cynthia Sung		Ready for Review
PROCEDURE_OCCURRENCE	PR	Katy		Ready for Review
DEVICE_EXPOSURE		Do not populate		Not Applicable
MEASUREMENT	LB, MI, VS, MB, MS, RE, RP	Katy, Tom		In-Process
OBSERVATION		Qi Yang		In-Process
DEATH	DM, AE, DD, DS, SS	Mike Hamidi		In-Process
NOTE		Do not populate		Not Applicable
NOTE_NLP		Do not populate		Not Applicable
SPECIMEN	MB?	Do not populate		
FACT_RELATIONSHIP				
LOCATION				
CARE_SITE				
PROVIDER				
PAYER_PLAN_PERIOD				
COST				

Common Rules

Internal ID	Destination Field	Source Table/Field	Applied Rule	Comment
			RFSTDTC If --STDY is a negative number: Add --STDY days to RFSTDTC	**only** for de-identification purposes. The --STDY variables represent the study day of the start of an event/intervention, relative to RFSTDTC (subject reference start date / study day 1). However, RFSTDTC is nulled out in this dataset, also for de-id purposes. **We will use the study start date (23-Oct-14) as RFSTDTC for all patients.**
	ı.<table name>.<element>_end_date	src.<SDTM dataset>.<SDTM dataset root variable, e.g. --ENDY>	For each subject, assign an arbitrary RFSTDTC. Then perform the following calculation to derive a x_start_date value: If --ENDY is a positive number: Add (--ENDY - 1) days to RFSTDTC If --ENDY is a negative number: Add --ENDY days to RFSTDTC	In the general observation class, specifically interventions and events, the SDTM domains will have similar applications (e.g., AEENDTC, CMENDTC). In those use cases, the end date is equivalent to the end of an event or intervention. This dataset supplies **relative dates only** for de-identification purposes. The --ENDY variables represent the study day of the end of an event/intervention, relative to RFSTDTC (subject reference start date / study day 1). However, RFSTDTC is nulled out in this dataset, also for de-id purposes. **We will use the study start date (23-Oct-14) as RFSTDTC for all patients.**
	ı.<table name>.<element>_occurrence_id	N/A (refer to comment)	primary key, auto-generated by ETL	The src.<SDTM domain>.--seq can be used in combination with another identification methodology to ensure record uniqueness. However, it may be preferred to generate this programmatically.
	ı.<table ie>.<element>_source_concept_id	cdm.<table name>.<element>_source_value	c1.concept_id LEFT JOIN cdm.concept c1 ON c1.concept_code = cdm.<table name>.<element>_source_value AND c1.vocabulary_id = < source vocabulary_id >	<source vocabulary id> is a vocabulary of the source entity, i.e., MedDRA or custom vocabulary If custom mappings are in the source_to_concept_map form, then populate with stcm.source_concept_id LEFT JOIN cdm.source_to_concept_map stcm ON stcm.source_code = cdm.<table name>.<element>_source_value AND stcm.source_vocabulary_id = < source vocabulary_id >
		cdm.<table name>.<element>_source_value	Coalesce(c2.concept_id, 0*)	If custom mappings are in the source_to_concept_map form, then search for the target concept as follows:

CDM Mapping Specification

SDTM Source(s): ae, mh, ce, pr

Destination Field	User Guide	ETL Conventions	Datatype	Required	Primary Key	Foreign Key	FK Table	FK Domain	Source Table/Field	Applied Rule	Comment
condition_occurrence_id	The unique key given to a condition record for a person. Refer to the ETL for how duplicate conditions during the same visit were handled.	Each instance of a condition present in the source data should be assigned this unique key. In some cases, a person can have multiple records of the same condition within the same visit. It is valid to keep these duplicates and assign them individual, unique, CONDITION_OCCURRENCE_IDs, though it is up to the ETL how they should be handled.	integer	Yes	Yes	No			Not Applicable	CR.004	
person_id	The PERSON_ID of the PERSON for whom the condition is recorded.		integer	Yes	No	Yes	PERSON		src.ae.usubjid; src.mh.usubjid; src.ce.usubjid; src.pr.usubjid	CR.001	For PR, only map when src.pr.prindc is populated.
condition_concept_id	The CONDITION_CONCEPT_ID field is recommended for primary use in analyses, and must be used for network studies. This is the standard concept mapped from the source value which represents a condition	The CONCEPT_ID that the CONDITION_SOURCE_VALUE maps to. Only records whose source values map to concepts with a domain of "Condition" should go in this table. Accepted Concepts.	integer	Yes	No	Yes	CONCEPT	Condition	cdm.condition_source_value	cdm.condition_concept_id will require mapping in association to MedDRA (in most cases) dictionary equivalent standards terms via cdm.condition_source_value as part of Usagi processing	In src.mh.mhterm there are values pertaining to "HIV STATUS", where the src.mh.mhoccur="Y" means HIV positive. This needs to be logically treated as there are other in src.mh.mhterm that have an src.mh.mhoccur correlation (e.g., TB diagnosis)
condition_start_date	Use this date to determine the start date of the condition	Most often data sources do not have the idea of a start date for a condition. Rather, if a source only has one date associated with a condition record it is acceptable to use that date for both the CONDITION_START_DATE and the CONDITION_END_DATE.	date	Yes	No	No			src.ae.aestdy; src.mh.mstdy	CR.002	In SDTM, this is represented in ISO 8601 format (e.g., YYYY-MM-DD). Unfotunately, this study does not have src.ae.aestdtc, rather the calculation of time from the src.dm.rfstdtc reference anchor (e.g., via AESTDY, MHSTDY) Note: CE is N/A and PR is not available.
condition_start_datetime		If a source does not specify datetime the convention is to set the time to midnight (00:00:0000)	datetime	No	No	No			Not Available		Time is generally not collected
condition_end_date	0	Most often data sources do not have the idea of a start date for a condition. Rather, if a source only has one date associated with a condition record it is acceptable to use that date for both the CONDITION_START_DATE and the CONDITION_END_DATE.	date	No	No	No			src.ae.aeendy; src.mh.mhendy	CR.003	In SDTM, this is represented in ISO 8601 format (e.g., YYYY-MM-DD). Unfotunately, this study does not have src.ae.aeendtc (and src.mh.mhendtc), rather the calculation of time from the src.dm.rfstdtc reference anchor (e.g., via AEENDY, MHENDY) Note: CE is N/A and PR is not available.