Proposal: Clinical trial data conventions for the OMOP Common Data Model

v1.0, July 2020
OHDSI Clinical Trials Working Group

Table of contents

Executive summary

The Clinical Trials Working Group proposes conventions for the OMOP CDM and Standardized Vocabularies to capture clinical trial specific data.

Our use case is the conversion of clinical trial data in CDISC SDTM format to OMOP, with a view to allowing trial planning optimization. SDTM was chosen as it is a clinical trials' submission standard that is "required" by the FDA and PMDA, "preferred" by the China NMPA, and "accepted" by the EMA. All our proposals assume the source data is in SDTM format and represent the final set of data from a clinical trial.

We advocate minimum changes to the OMOP CDM and Standardized Vocabularies to minimize impact on OHDSI tools like Atlas, whilst providing a value-add SDTM-to-OMOP conversion with minimum data loss.

Our proposals cover eight main topics for which there is currently insufficient support in the OMOP CDM and Standardized Vocabularies. They include introducing new concepts and modifiers, but no new CDM tables. Furthermore, we provide guidance for ETL developers when dealing with some data that is more complicated in nature, or certain scenarios that may be present in clinical trial submitted datasets (e.g., non-unique subject ids).

This document details our proposals for each of those eight topics, built on OMOP CDM v6 and the Oncology extension, with v5.3.1 backward compatibility.

Background

The OMOP common data model (OMOP CDM) maintained within the OHDSI community is used for storing and analyzing observational health data from various sources (e.g. EHR records or administrative claims data). This comprises a patient-centric relational database model in which each type of clinical data (e.g. diagnoses, treatments) has its own distinct table for storage. In addition to the actual patient data, it contains an extensive standardized vocabulary schema, to which all the clinical terms are mapped.

The data types that are generally collected during clinical trials have a large degree of overlap with those of observational data. For example, the same kind of lab measurements and condition occurrence reporting takes place for both data sources. There are however some distinct features inherent to clinical trial data collection that do not have an obvious storage location within the OMOP CDM, like adverse event severity and causality, and information on trial arms.

In this document, we describe a proposal for storing the main characteristics of clinical trial specific data types within the OMOP CDM.

Approach

We took a **"minimal changes"** approach that refrains from making structural changes in the form of new tables. Instead, it captures most of the clinical trial source data using the existing OMOP CDM and existing vocabularies. In some cases we propose new standard concepts to capture trial specific events.

CDM version compatibility

This proposal is built on both OMOP CDM v6 and the Oncology extension.

We are aware that OHDSI tools do not support OMOP CDM v6 currently, and hence provide backward compatibility with CDM v5.3.1.

The following four integer fields should be added to OMOP CDM v5.3.1 in order to have a compatible data model to consume the proposals in this document.

- Observation (v6 attributes)
    - observation_event_id
    - obs_event_field_concept_id
- Measurement (oncology extension)
    - modifier_of_event_id
    - modifier_of_field_concept_id

Future scope

We have identified three future areas of interest. They are listed here as illustrations, without firm plans or priorities established as yet.

- This proposal document focuses solely on clinical trial data. We anticipate extending this approach in the future to other types of data such as registries.

- This proposal document assumes data comes from a single clinical trial. In future, we anticipate catering for scenarios where multiple trials are linked, and a person's data can be a combination of different clinical trials.

- This proposal document addresses SDTM-to-OMOP conversion. In future, we may want to cover OMOP-to-SDTM conversion too, with a view to aiding submission to regulatory bodies.

Topic 1: Trial enrollment and trial outcome

The first thing one needs to be able to capture is the time period for which a person was enrolled into a clinical trial and what this particular trial was. This is similar to what is stored in the existing OMOP observation period table, which records time spans for which a person is at-risk to have clinical events recorded within the source systems (concept 44814723, 'Period while enrolled in study'). However, such an observation period is not trial-specific and does not offer the granularity needed for trial enrollment. Patient events that cause trial granularity are for example informed consent, patient eligibility, trial arm randomisation and trial withdrawal. Therefore we will have to look at different options to capture the different trial enrollment events or statuses a person will have during the trial enrollment, especially during trial start and trial end.

Proposed convention

We propose to store trial enrollment and outcomes as an observation for each event related to a person's trial status. This does not require making any changes to the existing CDM definition. The start and end dates of a person's enrollment in a trial can be captured in two separate observations. Examples of these observations are shown in Table 1 and Table 2. The observation concepts used are children of the SNOMED concept "Research administrative status" (4204933), which subsumes several child concepts indicating the start of a trial enrollment, trial status updates or the reason for ending it.

Figure 1 shows a diagram of patient events and their corresponding concepts during a clinical trial. For three events associated with patient eligibility and trial arm randomization, there are currently no appropriate concepts available:

- For the two eligibility concepts, there are only general research study counterparts available: "Not eligible for participation in research study" (concept id 44811374) and "Eligible for participation in research study" (concept id 44811245). Ideally these would be trial specific.
- Clinical trial arm randomization could potentially be captured by the observation "Clinical trial arm" (concept id 37208111).

Novel concepts are discussed in Topic 5.

In figure 2, outcome events happening during trial end are captured. There are two possible outcome events: "Completion of clinical trial" (concept id 4042840) and "Patient withdrawn from trial" (concept id 4163733). These are stored as observation records. Reasons for withdrawal for the observation "Patient withdrawn from trial" are stored as value_as_concept_id of this observation. Table 2 and 3 show examples of these types of observations.

Concepts on three patient withdrawal reasons (patient's decision to withdraw, withdrawal because of a serious adverse event, and withdrawal decided by the investigator) are missing. Novel concepts need to be created in order to capture all trial outcome events.

Although this approach allows the capture of a person's start and end of a trial enrollment, as well as other trial status updates, it does not provide a straightforward way to further specify the person's treatment protocol. For a convention for arm assignment, see Topic 4.

Upon re-screening, oftentimes the patient will get a new study ID number. In OMOP, we want a person row to represent a single patient and have a way to link them to each of their study ID numbers and associated visits/data. We leave it to the discretion of the ETL implementer as to what extent to go to identify the same persons with different subject_ids.

Observation period

When building cohorts or conducting characterization analysis in Atlas, only the current observation period is considered. Hence, instead of capturing clinical trial epochs as separate records in the OMOP CDM observation period table, we propose to store one observation period record only per clinical trial subject, so analysis can be completed across the whole clinical trial duration.

The single observation period record per person covers the timespan for which a person is "at-risk" to have clinical events recorded within the source systems. Concept 44814723, 'Period while enrolled in study' suits the period type of this observation period.

- Store only one observation period record per clinical trial per trial subject.
- The observation period start date will be the date when a person gave informed consent.
- Observation period end date can be set to the last recorded date among all events for a person (concept 44814723, 'Period while enrolled in study').

The clinical trial epoch information is not lost. We propose to store it as part of the visit_occurrence.visit_source_value field, as described in Topic 2.

Examples

See also: https://forums.ohdsi.org/t/omop-cdm-and-clinical-trials/2109/7

Trial enrollment

Consent given

person_id	1
observation_concept_id	4163733 (Patient consented to clinical trial)
observation_date	2015-10-01

Table 1. Example of observation record excerpts indicating a person's trial enrollment.

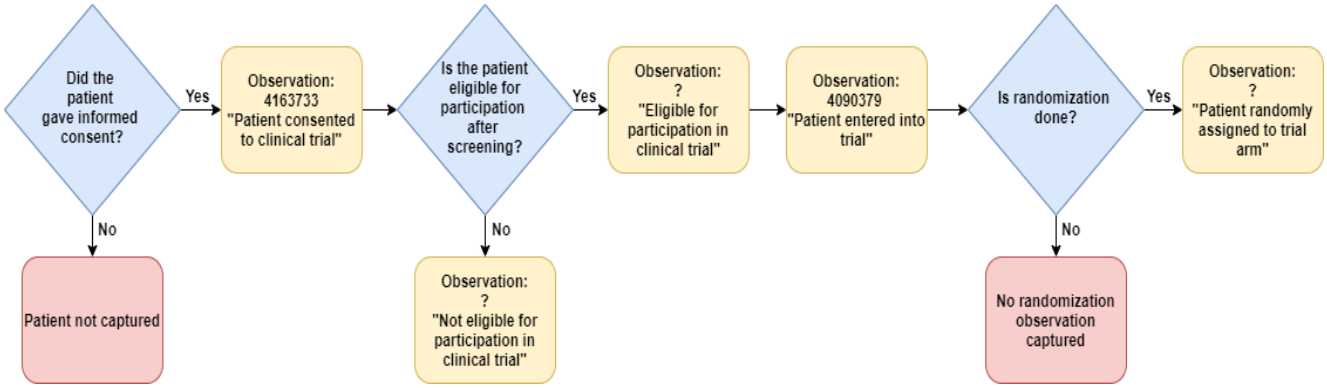

Figure 1. Patient events and corresponding mappings during trial start.

Trial outcome

Full completion of clinical trial

person_id	1
observation_concept_id	4042840 ('Completion of clinical trial')
observation_date	2016-05-30

Table 2. Example of observation record excerpts indicating a person's full trial completion.

Patient withdrawn from trial

person_id	1
observation_concept_id	4163733 ('Patient withdrawn from trial')
observation_date	2016-05-30
value_as_concept_id	44811247 ('Lost to clinical trial follow-up')

Table 3. Example of observation record excerpts indicating a person's trial withdrawal.

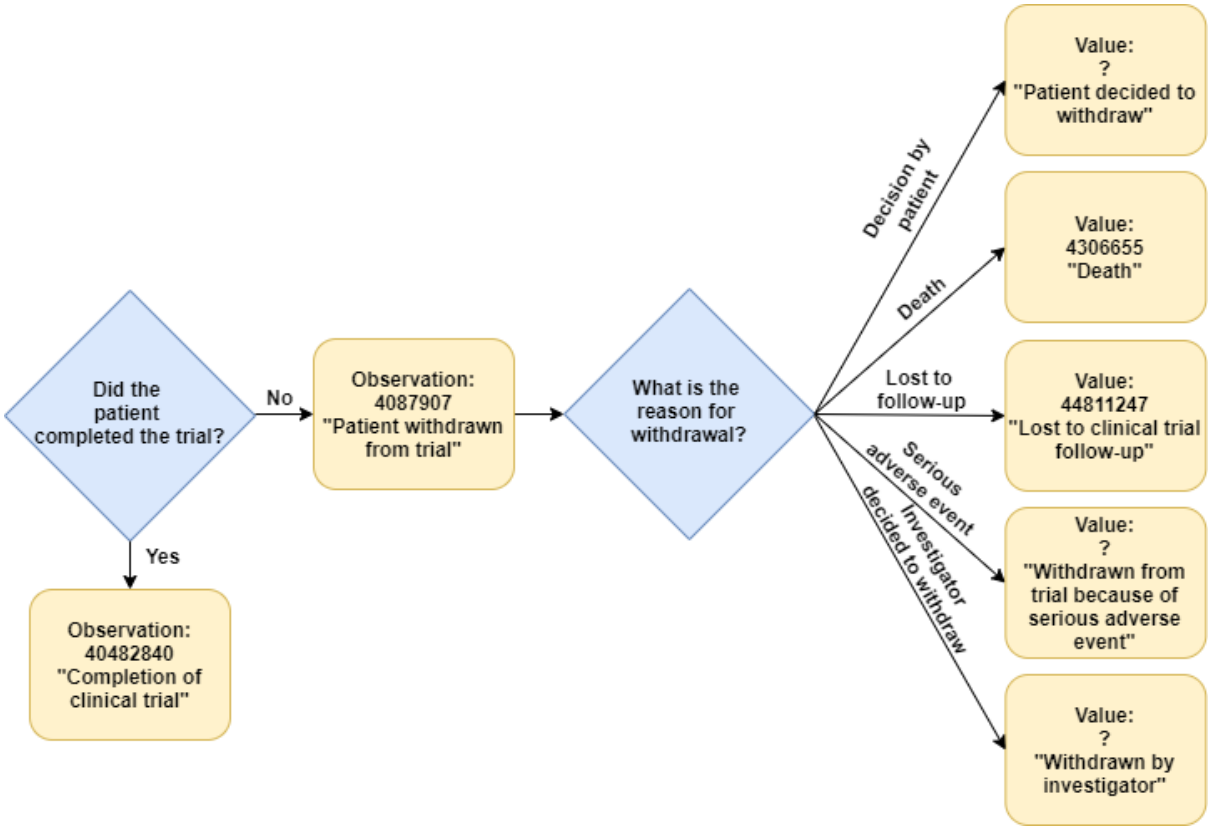

Figure 2. Trial outcomes and corresponding mappings.

Topic 2: Trial visits

The key events at which data is gathered from clinical trial participants is during planned visits (traditionally visits at trial sites, but increasingly also as virtual visits / telehealth). Unlike patient visits in observational data, most clinical trial visits are scheduled according to what the protocol of the clinical trial arm prescribes.

A clinical trial typically consists of several epochs, e.g. screening and treatment. A clinical trial subject can have multiple visits within an epoch. These visits are characterized by combinations of various time indicators (e.g. Cycle2Week1Day4, WEEK 7, UNSCHEDULED 1A, FOLLOW-UP 3). Visits can occur "on time", or they can be "delayed" or "missed".

The current OMOP visit occurrence table is already largely able to capture the relevant information regarding trial visits. To distinguish a clinical trial visit from the observational kind, the visit concept is used. Currently, the visit vocabulary contains a small number of standard concepts to use as visit_concept_id (inpatient visit, outpatient visit, home visit, etc), most of which don't apply to the typically observed visits that would occur as part of clinical trials.

For capturing the provenance of a visit concept using the OMOP type concepts, see Topic 6.

Proposed convention

We propose:

- to extend OMOP CDM vocabularies to capture the different trial visit concepts across clinical trial epochs
- to have composite source names to capture time indicators within an epoch.

Each convention is described in detail in the sections below.

In a future version, we will also propose how best to capture and use prior scheduled visits to determine the timeliness of trial visits. Many protocols define what is an acceptable delay to a scheduled visit, and what constitutes a missed visit; in those cases, that data could be stored in the CDM too.

Extension to OMOP CDM vocabularies

From the extensive list of visit concepts within the controlled SDTM terminology, we selected four that ensure coverage of the possible epochs and visits - screening, follow-up, treatment and unscheduled.

We propose the following five new visit concepts, where the 'Clinical Trial Visit' is the parent concept of the other four. This hierarchy will facilitate analysis in Atlas.

SDTM Study visit (epoch)	**Suggested concept_name**
Screening	Screening visit
Follow-up	Follow-up visit
Treatment	Scheduled visit
Unscheduled	Unscheduled visit
-	Clinical Trial visit

Composite source values

In order to capture *when/in what sequence* during an epoch a visit occurred, we propose to have the visit_occurrence.visit_source_value field be the concatenation of the epoch and visit values separated by a colon.

Example

Epoch	**Visit**	**visit_source_value**	**visit_concept_id**
TREATMENT	WEEK 7	TREATMENT:WEEK 7	Scheduled visit
FOLLOW-UP	FOLLOW-UP 3	FOLLOW-UP:FOLLOW-UP 3	Follow-up visit
TREATMENT	UNSCHEDULED 1A	TREATMENT:UNSCHEDULED 1A	Unscheduled visit

Topic 3: Seriousness, severity and causality

A typical use case of observational data in OMOP CDM is comparing treatment groups or treatment versus non-treatment on the incidence rates of certain outcomes. In clinical trials, it is of great interest to record any adverse events (AEs) that may occur after a treatment has been started, along with their severity and causality. AEs may be classed as serious (SAEs), requiring timely reporting to regulators.

For example:

- an AE may be 'possibly related' to the treatment drug
- an AE may be of 'mild' severity
- an AE may be an 'SAE' if the subject requires hospitalization.

Additionally, as a condition is diagnosed in a clinical trial, extra measurements can be taken to qualify that condition. For example, in a cancer diagnosis, the grade may be captured.

Seriousness, severity and causality are not existing attributes in any of the OMOP tables. Tumor size and grades are available in the Oncology Extension only.

Proposed convention

We propose to use both oncology extensions and attributes from OMOP CDM v6.

Specifically, to use the following fields to link an observation or condition to another record:

- Observation (v6 attributes)
    - observation_event_id
    - obs_event_field_concept_id
- Measurement (oncology extension)
    - modifier_of_event_id
    - modifier_of_field_concept_id

The _event_id contains the primary key of the record that is to be modified (e.g. a condition_occurrence_id).

The _field_concept_id is a concept specifying to which table the given _event_id points (e.g. concept_id 1147663 refers to the condition_occurrence_id).

These columns make it explicit that the information in the record does not represent an independent measurement or observation, but only serves the purpose of modifying another record. A record can be modified by multiple observations and measurements.

Examples

An adverse event with two additional attributes, the severity and relatedness to study drug.

Subject_id	Diagnosis	Date	Severity	Related
1	Cough	2018-02-21	Mild	Possible

This results in one condition occurrence record and two modifiers records.

condition_occurrence_id	4321
person_id	1
condition_concept_id	254761 ('Cough')
condition_start_date	2018-02-21

Table 4. Excerpt of the condition_occurrence table containing a record that is modified by referencing in an observation (see Tables 5,6)

person_id	1
observation_concept_id	4077563 ('Severity')
observation_date	2018-02-21
value_as_concept_id	4116992 ('Mild')
observation_event_id	4321
obs_event_field_concept_id	1147663 (Condition_Occurrence.condition_occurrence_id)

Table 5. Excerpt of the observation table (in OMOP CDM v6) containing one record that modifies a record from the condition occurrence table by indicating a mild severity.

person_id	1
observation_concept_id	45912709 ('Relationship to study drug') [non-standard, CIEL]
observation_date	2018-02-21
value_as_concept_id	4162850 ('Possible')

observation_event_id	4321
obs_event_field_concept_id	1147663 (Condition_Occurrence.condition_occurrence_id)

Table 6. Excerpt of the observation table (in OMOP CDM v6) containing one record that modifies a record from the condition occurrence table by indicating the possible relationship to the study drug.

Topic 4: Study information and arm assignment

Clinical trial protocols are usually well defined including expected patient inclusion/exclusion criteria, patient disposition, expected visits, treatment and procedure plans, etc. Unlike retrospective observational databases, many events are only recorded if listed as part of the study protocol outcomes or recognized as an adverse event.

Therefore, contextual metadata about the study protocol needs to be stored for subsequent evaluation and comparison. This extends to trial identifiers (e.g., from clinicaltrials.gov or EudraCT), whether the trial is registered with one entity or multiple entities.

Proposed Convention

We propose to use the COHORT and COHORT_DEFINITION tables in OMOP CDM v5 and v6 to store clinical trial inclusion/exclusion requirements, treatment arm definitions (e.g., patient dispositions), outcomes to be measured, and trial visit plan information.

The idea of the COHORT and COHORT_DEFINITION tables as part of the OMOP CDM schema is to capture retrospective information about cohorts as distributed with data. We propose storing information about which trial arm the individual participants are in using the COHORT table, along with any trial identifiers (e.g., NCT number). We propose storing information about the trial design and trial arms in the COHORT_DEFINITION table (examples below).

Example

cohort_definition_name	'Inclusion/Exclusion Criteria'
cohort_definition_description	<NCT Number> 'Males and postmenopausal females at least 50 years of age. Diagnosis of probable AD as defined by NINCDS and the ADRDA guidelines. MMSE score of 10 to 23.'
definition_type_concept_id	44819246 'Cohort'

Table 7. Excerpt of the cohort_definition table (in OMOP CDM v6) describing Inclusion/Exclusion criteria from the trial, with reference back to CDISC TI (Trial Inclusion) table.

COHORT_DEFINITION field	Example Value	CDISC Origin
cohort_definition_id	1	Table: TE Data Element: < ItemGroupData data : ItemGroupDataSeq = ''> From XML file format
cohort_definition_name	'Trial Arm - Placebo'	Table.Field: TE.ELEMENT or TE.ETCD
cohort_definition_description	'First dose of treatment to 1 month after'	Table.Field: TE.TESTRL and TE.TEENRL
definition_type_concept_id	44819246 'Cohort'	

Table 8. Excerpt of the cohort_definition table (in OMOP CDM v6) describing the Placebo arm of a trial, with reference back to CDISC TE (Trial Element) table.

COHORT_DEFINITION field	Example Value	CDISC Origin
cohort_definition_id	2	Table: TE Data Element: < ItemGroupData data : ItemGroupDataSeq = ''> From XML file format
cohort_definition_name	'Trial Arm - High Dose'	Table.Field: TE.ELEMENT or TE.ETCD
cohort_definition_description	'High Dose - First dose of treatment to 1 month after'	Table.Field: TE.TESTRL and TE.TEENRL
definition_type_concept_id	44819246 'Cohort'	

Table 9. Excerpt of the cohort_definition table (in OMOP CDM v6) describing the High Dose arm of a trial, with reference back to CDISC TE (Trial Element) table.

COHORT field	Example Value	CDISC Origin
cohort_definition_id	1 ('Trial Arm - Placebo')	Table.Field: TE.ELEMENT or TE.ETCD
subject_id	123 (based on cohort_definition.subject_concept_id - this refers to person.person_id = 123)	
cohort_start_date	1/1/2010	Table.Field: DM.RFSTDTC or DM.RFXSTDTC
cohort_end_date	4/1/2011	Table.Field: DM.RFENDTC or DM.RFXENDTC

Table 10. Excerpt of the cohort table (in OMOP CDM v6) describing a patient in the Placebo arm of a trial, with reference back to CDISC TE (Trial Element) and DM tables.

Topic 5: Novel concepts

Vocabulary extensions can be used for events that are part of trials but are not supported by existing coding vocabularies. For example, drugs not yet on the market or new interventions:

- drugs
- bio-assays
- interventions
- procedures
- devices

Proposed convention

For drug concepts, only on the ingredient level, single new concepts can be added without substantial effort. We propose an improved and simplified process to add semantic clinical drug level drug concepts as RxNorm extensions.

Some drugs cannot be standardised as they haven't been 'seen' before. In that case a 0 should be used; or a custom concept used within the CDM (2B+).

- Possible solution: get 'trial drug names' ('AZxyz', 'BMSabc') as published on clinicaltrials.gov from e.g. FDA or other central instance.

For other concepts, OMOP extensions can be introduced, presumably without the need for building novel concept classes. For consistency reasons, the new concept should be integrated with the existing hierarchy along with the Standard concepts. A dedicated "Clinical Trial" classifier concept could be linked to the newly created concepts to identify all Clinical Trial related concepts.

Topic 6: Type concept ids

Type concepts in OMOP give the provenance of a record. The existing type concept ids are mostly claims and EHR specific, e.g. 'Primary Condition' or 'Derived from EHR'. The only trial specific type concepts are the following visit type and observation period type:

- Visit occurrence: 44818519 - Clinical Study Visit
- Observation period: 44814723 - Period while enrolled in study

For other domains, we need similar new type concepts to indicate the provenance from a clinical trial.

The supplemental QORIG (data origin) column in SDTM provides information about the origin of the data. Possible values are "Assigned" (for medical dictionary coding terms), "Derived", "CRF", "eDT" (electronic data transfer, e.g. data from external labs) and "Protocol". Other examples of data origin in studies are data collected by doing an oral interview, data collected through an app/wearable, or data collected by email.

A consolidation of type concepts is currently in progress at OHDSI: https://forums.ohdsi.org/t/concept-type-consolidation-please-take-a-look/8306.

Most important conventions:

- All Concepts are now called in such a way that they can finish the sentence "This record was obtained from a …".
- Only generic type concepts are to be used, eliminating the need for domain-specific concepts.
- Type concepts have a simple hierarchy. The hierarchy is single-parent.

Proposed convention

We define new type concept ids for the missing trial provenances, see the table below. Relevant existing type concepts are given in italic at the bottom.

The parent term for the newly suggested type concepts will be 'Clinical study'.

Suggested type concept_name	Condition Type	Death Type	Device Type	Observation Type	Drug Type	Mease Type	Procedure type	Note Type	Specimen Type	Episode Type*
Case Report Form - medically captured	x	x	x	x	x	x	x	x	x	x
Case Report Form - self-reported (app, email, telephone, online questionnaire/ePRO)	x	x	x	x	x	x	x	x	x	x
Case Report Form - Derived (e.g. from protocol)	x	x	x	x	x	x	x	x	x	x
Existing type concepts										
Randomized drug					x					
Derived from EHR	x	x	x	x	x	x	x	x	x	
Case Report Form	x	x	x	x	x	x	x	x	x	x
Lab result				x		x				

Topic 7: Planned drug dose

During studies, the actual administered and planned drug doses can be different. Deviations from the protocol can be relevant, but this does require recording the planned dose.

Proposed convention

To keep administered as well as planned (ordered) drug doses in a way that makes comparing them possible, we propose to use a type concept id in the drug exposure table that allows to distinguish planned vs. administered. The type concept "prescription written" is currently available and could serve the purpose. There are however plans to become slightly more granular by having two concept types to replace this one: "EHR prescription issue" record for the time of issuing the prescription and "EHR planned dispensing record" for scheduled time of dispensing or administration. See also this discussion on prescribed vs dispensed on the OHDSI forums.

A link between those records can be built during ETL by creating a fact_relationship record, but this link cannot be used with standard tools like Atlas.

Example

From SDTM to drug exposure records we potentially could use the following model using the representation of scheduled and performed administrations (blinded) in EC:

USUBJID	1	1
ECTRT	hydroxychloroquine	hydroxychloroquine
ECMOOD	SCHEDULED	PERFORMED
ECDOSE	50	40
ECSTDTC	2019-04-22	2019-04-22
ECENDTC	2019-04-22	2019-04-22

Table 11. Example record for scheduled and performed doses in EC.

drug_exposure_id	254
person_id	1

drug_concept_id	1777087 ('hydroxychloroquine')
drug_exposure_start_date	2019-04-22
drug_exposure_end_date	2019-04-22
quantity	50
drug_type_concept_id	38000177 - Prescription written

Table 12. Example record for a planned (scheduled) dose.

drug_exposure_id	256
person_id	1
drug_concept_id	1777087 ('hydroxychloroquine')
drug_exposure_start_date	2019-04-22
drug_exposure_end_date	2019-04-22
quantity	40
drug_type_concept_id	38000175 Prescription dispensed in pharmacy OR 581452 Dispensed in Outpatient office OR 38000180 Inpatient administration

Table 13. Example record for a dispensed or administered (performed) dose.

Topic 8: Relative dates

In some clinical trials, for example when a trial is anonymized, the timepoints of events are given as days offset from a patient's Informed Consent or randomization date. The OMOP CDM always requires a date to be associated with an event. In general, even for anonymized trials, we can assume that at least the year of randomization or consent is given. This can be used to derive the dates of the events.

Proposed convention

If days since a reference is given, calculate the date using the given patient's reference date. The reference date is typically the Informed Consent date or the randomisation date. If a reference date is not specified, then use the patient's date of Informed Consent.

If the reference date is given in month or year precision, use the first day of the month and/or first month of the year as the reference date.

A record can be stored in the metadata table to indicate that the dates within the data set are derived, such that it is clear that e.g. seasonal effects might be obfuscated.

Example

SDTM source data records:

a. SDTM Demographics record

SUBJID	RFSTDTC
1	2019-03

b. SDTM Drug Exposure record

SUBJID	EXTRT	EXSTDY
1	Hydroxychloroquine	52

Table 14. a. Example record in the SDTM Demographics domain, containing a subject identifier (SUBJID) and a subject reference start date (RFSTDTC). b. Example record in the SDTM Drug Exposure domain, containing a subject identifier (SUBJID), treatment name (EXTRT) and study day of start of observation expressed in integer days relative to the reference date (EXSTDY).

Mapping these SDTM source data records to the OMOP CDM will lead to the following records in the drug exposure, observation and metadata tables:

drug_exposure_id	255
person_id	1
drug_concept_id	1777087 ('hydroxychloroquine')

drug_exposure_start_date	2019-04-22

Table 15. Excerpt of the drug_exposure table containing a record with drug_exposure_start_date calculated from the reference date (2019-03-01 + 52 days).

observation_id	123
person_id	1
observation_concept_id	4163733 ('Patient consent to clinical trial')
observation_date	2019-03-01

Table 16. Excerpt from the observation table containing a record with a reference date for a particular patient, which in this case is a date of a consent given. The original date of consent (2019-03) has been imputed according to the recommendations above (with the first date of month).

metadata_concept_id	0
metadata_type_concept_id	0
name	'Subject reference start date'
value_as_string	'Observation | Patient consent'
value_as_concept_id	4163733 ('Patient consent to clinical trial')
metadata_date	CURRENT_DATE

Table 17. Excerpt from the metadata table containing a record that indicates that a reference date is used. The concept used as a reference is stored in value_as_concept_id.