CRC Cohort mapping and conversion to the OMOP CDM tables

Revision history

Revisions	Authors	Reviewers	Main changes
V0 - 04th Aug 2022	Cecilia Mascia, Francesca Frexia, Vittorio Meloni (CRS4)	Alessandro Sulis, Giovanni Delussu, Mauro Del Rio (CRS4)	First version: extended description of the mapping and the implementation choices made for the openEHR to OMOP CDM transformation of the dataset

Scope and Objectives of this document

In the context of the EOSC-Life WP1 Demonstrator “Cloudification of BBMRI-ERIC CRC-Cohort and its Digital Pathology Imaging” (APPID 1228), it has been developed a demonstrator to improve the access and re-use to the ColoRectal Cancer-Cohort (CRC-Cohort)[1] dataset, providing clinical and imaging data, with related annotations, in documented formats using common and open approaches and tools.
To improve the integrability and re-usability of the dataset, the demonstrator enables the conversion of the CRC-Cohort clinical data to the openEHR[2] format, in order to save them in an openEHR-based repository, from which they can be exported in OMOP[3] and HL7 FHIR[4], in order to support the future integration into the BBMRI-ERIC Federated Platform, as shown in Figure 1.

[1] https://www.bbmri-eric.eu/scientific-collaboration/colorectal-cancer-cohort/

[2] https://www.openehr.org/

[3] https://www.ohdsi.org/data-standardization/the-common-data-model/

[4] https://hl7.org/FHIR/

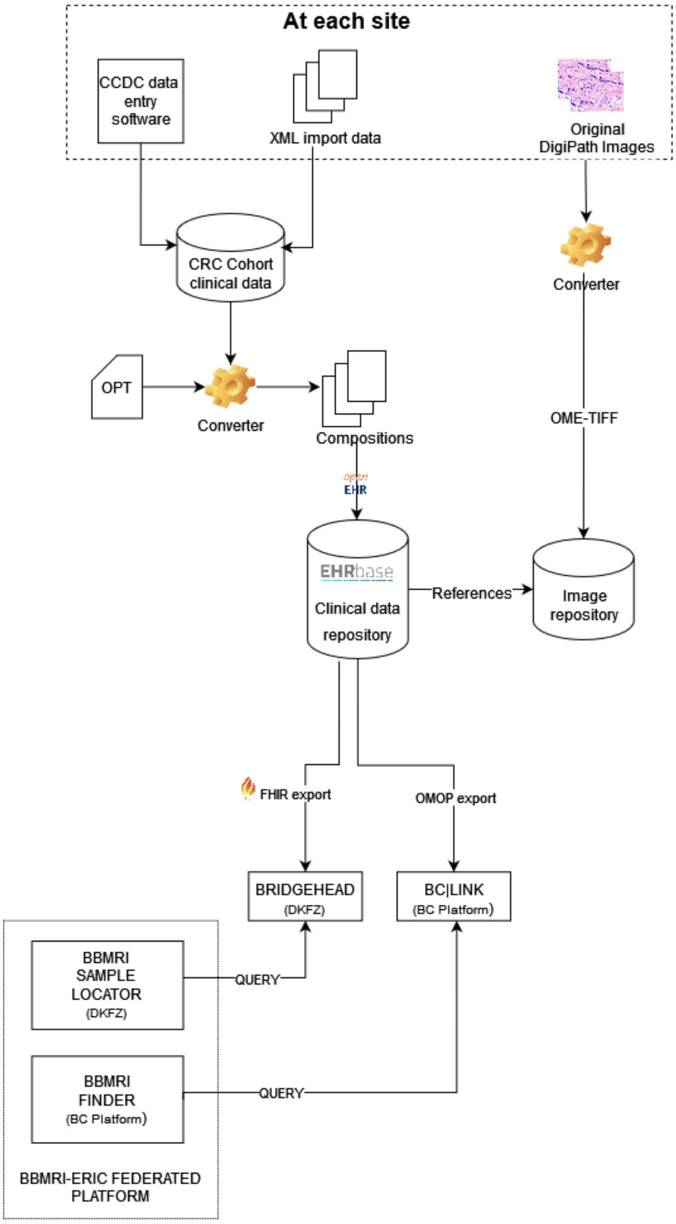

Figure 1 - Architectural diagram of the demonstrator

At present, all the fields in the CRC-Cohort Data Model have been mapped onto an openEHR template[5], and part of them have been mapped to FHIR Resources compliant with the Sample Locator[6] FHIR Profiles and to a fraction of the OMOP tables required by the BC Platform search components. The set of developed tools enables the extraction and the transformation of the data needed to fill in the following tables, according to the proposed mapping:

- Person
- Observation period

[5] https://github.com/crs4/crc_cohort_modelling

[6] https://www.bbmri-eric.eu/scientific-collaboration/colorectal-cancer-cohort/

- Condition occurrence
- Specimen

The mapping of other tables (e.g., Procedure occurrence and Observation) of the OMOP CDM is being finalised and the documentation of the mapping will be integrated into future versions of this document.

This document details how the openEHR fields are expressed according to the OMOP CDM, in order to **support both the sharing of the mapping/conversion approach and the validation of the mapping/conversion itself with domain experts**, which is **required to verify the quality of the content**.

OMOP CDM mapping

This mapping relates the current openEHR data model[7], expressed in the form of a template, to the elements of the OMOP Common Data Model, v5.4[8].

The mapping has the CDM OMOP tables as a starting point, and in particular, all the elements marked as mandatory. For each of them, an analogous element was sought in the openEHR model that was similar in terms of semantics, data type and constraints (e.g., allowed choices for a coded textual element). We define the mapping of each element as "valid" when these aspects are considered to be sufficiently overlapping or when it is possible to adapt the source element (openEHR) to the intended form of the target element (CDM table elements). However, given the difference in the content structure of the two approaches, there are many special cases for which a direct mapping is not so obvious or even not possible. These cases, marked as "uncertain", "not applicable" or "not done", are dependent on specific choices to be made in the ETL phase.

In the following paragraphs, we will give an **overview of the current mapping** and the **implementation/conversion choices** made so far in the tools to extract data from the openEHR database and convert them to the OMOP CDM tables. In particular, for each CDM table, we present a mapping table structured as follows:

- the *OMOP/openEHR* columns and subcolumns contain, respectively, the details of the destination and the source of the data transformation for each data element;
- the *Notes/implementation choices* column contains any remaining critical point, the choices made (*Current version*) and the actions to be taken for the continuation of the work (*Action required*);
- the *Mapping confidentiality* column contains the level we assign to the mapping of each element, with the marks: **V** - valid; **U** - uncertain; **NA** - not applicable; **ND** - not done.

When relevant, a specific table containing the mapping of the values to the standard OMOP vocabularies is also included. When designing an openEHR template, it is possible to associate each data element and its permitted values with a codification, via code and terminology id. In our case,

[7] CRC model: https://github.com/crs4/crc_cohort_modelling/blob/main/templates/opt/crc_cohort_rev.opt

[8]OMOP Documentation:https://ohdsi.github.io/CommonDataModel/cdm54.html#Clinical_Data_Tables

then, the template is already aligned to the standard OMOP terminology by codifying the values using the OMOP concept IDs.

General assumptions

Many tables, apart from the “Person” one, have elements to express the provenance of the reported data (e.g., period_type_concept_id in the Observation Period table). Considering that the data have been collected within a research project, we chose the *Case Report Form* option (Concept ID: 32809) among all the OMOP accepted concepts. This assumption is maintained in every table that requires this information.

PERSON

For each patient in the data source, a row of the PERSON table should be filled according to the following mapping table. Most of the required information is present in the data source and only a few points need to be checked.

PERSON mapping table

OMOP			openEHR			Notes/implementation choices	Mapping confidentiality[9]
cdmFieldName	isRequired	Datatype	node name(XSD label)	occurrences	datatype		
person_id	**Yes**	integer				Two possible choices: - keep the original id if integer (patient pseudonym) - generate a progressive number. **Current version:** as the IDs were not integers in the source data, we opted for a progressive number that is automatically generated during the ETL	V
gender_concept_id	**Yes**	integer	Biological sex (XSD label: Dataelement_85_1)	Mandatory	TEXT	Possible choices in the openEHR template are already expressed using OMOP codes (see the mapping of the values in Value mapping table A). **Current version:** "OTHER" is used even if is not among the accepted concepts in the CDM. **Action required:** check if it causes validation issues on the OMOP CDM side	V

[9] Mapping confidentiality levels: **V** - valid; **U** - uncertain; **NA** - not applicable; **ND** - not done

year_of_birth	**Yes**	integer	– – –	– – –	– – –	This information is missing in the data source. **Current version:** during the ETL, this value is calculated as: year of diagnosis (XSD label: Dataelement_51_3) - age at diagnosis (XSD label: Dataelement_3_1)	U
month_of_birth	No	datetime	– – –	– – –	– – –	Not required and no info from the source	NA
day_of_birth	No	datetime	– – –	– – –	– – –	Not required and no info from the source	NA
birth_datetime	No	datetime	– – –	– – –	– – –	Not required and no info from the source	NA
race_concept_id	**Yes**	integer	– – –	– – –	– – –	**Current version**: as a clear race or ethnic background cannot be established, we set the value to 0 as indicated in the OMOP specifications for cases where this information is not known.	V
ethnicity_concept_id	**Yes**	integer	– – –	– – –	– – –	**Current version:** as we do not have a source of this information, we set the value to NULL **Action required:** Check if non-standard concepts are acceptable, like 44814650 = No information from the PCORNet vocabulary.	U
location_id	No	datetime	– – –	– – –	– – –	Not required and no info from the source	NA
provider_id	No	integer	– – –	– – –	– – –	Not required and no info from the source	NA
care_site_id	No	integer	– – –	– – –	– – –	Not required and no info from the source	NA
person_source_value	No	varchar(50)	Patient pseudonym (XSD label: Identifier)	Mandatory	TEXT	This field holds the original pseudonym.	V

gender_source_value	No	varchar(50)	Biological sex (XSD label: Dataelement_85_1)	Mandatory	TEXT	This field holds the value of the Biological sex of the source. **Current version:** directly mapped to Biological sex (Dataelement_85_1), considering as the original source the openEHR repository (first colon of the Value mapping table A)	V
gender_source_concept_id	No	Integer	– – –	– – –	– – –	**Current version**: set to 0	V
race_source_value	No	varchar(50)	– – –	– – –	– – –	Not required and no info from the source	**NA**
race_source_concept_id	No	Integer	– – –	– – –	– – –	**Current version**: set to 0	V
ethnicity_source_value	No	varchar(50)	– – –	– – –	– – –	Not required and no info from the source	**NA**
ethnicity_source_concept_id	No	Integer	– – –	– – –	– – –	**Current version**: set to 0	V

Mapping for “Biological sex” values - Value mapping table A

Source value	**OMOP Gender Concept ID**
FEMALE	8532
MALE	8507
OTHER	8521 **(Non-standard)**

OBSERVATION PERIOD

We consider a single observation period for each patient. The main doubts in this mapping concern the boundaries of the time period, as these are not explicitly defined in the data source.

OBSERVATION PERIOD mapping table

OMOP			openEHR			Notes/implementation choices	Mapping confidentiality
cdmFieldName	isRequired	Datatype	node name(XSD label)	occurrences	datatype		
observation_period_id	**Yes**	integer	– – –	– – –	– – –	**Current version:** a progressive number is automatically generated during the ETL	V
person_id	**Yes**	integer	– – –	– – –	– – –	**Current version:** automatically set equal to PERSON.person_id	V
observation_period_start_date	**Yes**	date	Year of sample collection (XSD Label: Dataelement_89_3)	Mandatory	DATE	No such information in the data source, so the limits of the period should be inferred by the data. **Current version:** we selected the YEAR_OF_SAMPLE_COLLECTION as the start and the VITAL_STATUS_TIMESTAMP as the end of the observation period. **Action required:** Check if it is acceptable to consider this time period and how the date should be expressed, since the present version of the script considers the format (YYYY)	U
observation_period_end_date	**Yes**	date	Timestamp of last update of vital status (XSD Label: Dataelement_6_3)	Optional	DATE		U

period_type_concept_id	**Yes**	Integer	– – –	– – –	– – –	The value of this element should be determined between the accepted concepts. **Current version:** we used the Case Report Form option, OMOP Concept ID: 32809 **Action required**: Check if this assumption is OK	U

CONDITION OCCURRENCE

We consider a single condition period for each patient, corresponding to the primary tumor specified in the Histopathology section of the data source.

Questions to be answered:

- Is it correct to consider the primary tumor as the only condition to be recorded or should we consider other data?
- there are patients with multiple samples and more than one Histopathology section, in this case, more Condition records have to be created for each Person.

CONDITION OCCURRENCE mapping table

OMOP			openEHR			Notes/implementation choices	Mapping confidentiality
cdmFieldName	isRequired	Datatype	node name(XSD label)	occurrences	datatype		
condition_occurrence_id	**Yes**	bigint	– – –	– – –	– – –	**Current version:** a progressive number is automatically generated during the ETL	V
person_id	**Yes**	bigint	– – –	– – –	– – –	**Current version:** automatically set equal to PERSON.person_id	V
condition_concept_id	**Yes**	integer	Localization of primary tumor (XSD Label: Dataelement_92_1)	Mandatory	CODED TEXT	**Current version**: we assigned value to the element considering the primary tumor diagnosis code mapped to the OMOP terminology as showed in the value mapping table B (second column).	V
condition_start_date	**Yes**	date	Date of diagnosis (XSD Label: Dataelement_51_3)	Optional	DATETIM E		V
condition_start_datetime	No	datetime	– – –	– – –	– – –	Not required and no info from the source	NA
condition_end_date	No	date	– – –	– – –	– – –	Not required and no info from the source	NA
condition_end_datetime	No	datetime	– – –	– – –	– – –	Not required and no info from the source	NA

condition_type_concept_id	Yes	integer	– – –	– – –	– – –	**Current version:** we assumed that the provenance of the Condition record is a Case Report Form, OMOP Concept ID: 32809. **Action required**: Check if this assumption is OK	U
condition_status_concept_id	No	integer	– – –	– – –	– – –	Not required and no info from the source	NA
stop_reason	No	varchar(20)	– – –	– – –	– – –	Not required and no info from the source	NA
provider_id	No	integer	– – –	– – –	– – –	Not required and no info from the source	NA
visit_occurrence_id	No	integer	– – –	– – –	– – –	Not required and no info from the source	NA
visit_detail_id	No	integer	– – –	– – –	– – –	Not required and no info from the source	NA
condition_source_value	No	varchar(50)	Localization of primary tumor (XSD Label: Dataelement_92_1)	Mandatory	CODED TEXT	**Current version:** This element has been assigned a value with the verbatim value from the source data (first column of the next value mapping table B)	V
condition_source_concept_id	No	integer	– – –	– – –	– – –	Not required, potentially mappable in future versions	NA
condition_status_source_value	No	varchar(50)	– – –	– – –	– – –	Not required and no info from the source	NA

Mapping for “Localization of primary tumor” values - Value mapping table B

Source value	OMOP Condition Concept ID	OMOP Name
C 18.0 - Caecum	432837	Primary malignant neoplasm of cecum
C 18.1 - Appendix	433143	Primary malignant neoplasm of appendix
C 18.2 - Ascending colon	4247719	Primary malignant neoplasm of ascending colon
C 18.3 - Hepatic flexure	438979	Primary malignant neoplasm of hepatic flexure of colon
C 18.4 - Transverse colon	432257	Primary malignant neoplasm of transverse colon
C 18.5 - Splenic flexure	437798	Primary malignant neoplasm of splenic flexure of colon
C 18.6 - Descending colon	441800	Primary malignant neoplasm of descending colon
C 18.7 - Sigmoid colon	436635	Primary malignant neoplasm of sigmoid colon
C 19 - Rectosigmoid junction	438699	Primary malignant neoplasm of rectosigmoid junction
C 20 - Rectum	74582	Primary malignant neoplasm of rectum

SPECIMEN

This table of the CDM holds one or more records for every person, each of which related to a different sample.

SPECIMEN mapping table

OMOP			openEHR			Notes/implementation choices	Mapping confidentiality
cdmFieldName	**isRequired**	**Datatype**	**node name(XSD label)**	**occurrences**	**datatype**		
specimen_id	**Yes**	integer	– – –	– – –	– – –	**Current version:** a progressive number is automatically generated during the ETL	V
person_id	**Yes**	integer	– – –	– – –	– – –	**Current version:** automatically set equal to PERSON.person_id	V
specimen_concept_id	**Yes**	integer	Material type (XSD Label: Dataelement_54_2)	Mandatory	CODED TEXT	Possible choices in the openEHR template are already expressed using OMOP codes (see the mapping of the values in Value mapping table C). **Current version:** "Other" is used even if is not among the accepted concepts in the CDM. **Action required:** check the mapping of the values	U
specimen_type_concept_id	**Yes**	integer	– – –	– – –	– – –	**Current version:** we assumed that the provenance of the Specimen record is a Case Report Form, OMOP Concept ID: 32809. **Action required**: Check if this assumption is OK	U

specimen_date	**Yes**	date	Year of sample collection (XSD Label: Dataelement_89_3)	Mandatory	DATE	In the data source there is only the year and not the full date. **Current version**: we only set the year of acquisition (YYYY) **Action required**: Check if the format of the date is ok or it should be set differently	U
specimen_datetime	No	datetime	– – –	– – –	– – –	Not required and no info from the source	**NA**
quantity	No	float	– – –	– – –	– – –	Not required and no info from the source	**NA**
unit_concept_id	No	integer	– – –	– – –	– – –	Not required and no info from the source	**NA**
anatomic_site_concept_id	No	integer	– – –	– – –	– – –	Not required and no info from the source	**NA**
disease_status_concept_id	No	integer	– – –	– – –	– – –	Not required and no info from the source	**NA**
specimen_source_id	No	varchar(50)	Sample ID	Mandatory	IDENTIFIER	This field holds the original Sample ID	V
specimen_source_value	No	varchar(50)	Material type (XSD Label: Dataelement_54_2)	Mandatory	CODED TEXT	**Current version:** Directly mapped to Material type (Dataelement_54_2), corresponding to the labels in the first colon of the Value mapping table C)	V
unit_source_value	No	varchar(50)	– – –	– – –	– – –	Not required and no info from the source	**NA**
anatomic_site_source_value	No	varchar(50)	– – –	– – –	– – –	Not required and no info from the source	**NA**
disease_status_source_value	No	varchar(50)	– – –	– – –	– – –	Not required and no info from the source	**NA**

Mapping for “Material type” values - Value mapping table C

Source value	OMOP Specimen Concept ID	OMOP Name
Healthy colon tissue	4134449	Tissue specimen from colon
Tumor tissue	4122248	Tumor tissue sample
Other	4163599 **(Non-standard)**	Other specimen type

Next steps

The next actions we will carry out are the finalisation of the mapping of the missing data elements and the extension of the tools for their extraction and conversion from the openEHR DB into OMOP CDM tables, starting with the Procedure occurrence and Observation tables.

In parallel, general feedback on the whole mapping and any input on the following specific task would be useful for the continuation of the work:

- Check by domain experts of the mapping from the CRC-Cohort Data Model to the openEHR data model, available at:

https://github.com/crs4/crc_cohort_modelling/blob/main/documentation/CRC-Cohort_openEHR_mapping_documention_v0_Aug_2022.pdf

- Check each specific question highlighted in the text.
- Each mapping of the values (i.e., tables A, B and C) should be checked and validated by a domain expert.
- Check if it is possible, from an OMOP CDM perspective, to use non-standard concepts when a proper mapping is not feasible (e.g., for specimen type and biological sex).
- Implementation choices made to overcome the issue of missing or incomplete dates must be checked and corrected if necessary.
- Decide how to set the limits of the observation period.
- Check if the assumption made about using the *Case Report Form* concept for the provenance of the data is appropriate.