Introduction

The MMCI convector processes data from the CRC-cohort from MMCI and converts them into a chosen standard, FHIR or OMOP. After successful conversion, a set of quality checks is run. The tool also provides Wizzard, which questions the user and, based on provided information, recommends a better standard.

System Requirements

Operating system: Windows 10

Programming Language: Python 3.7.9.+

Python libraries are listed in setup.py.

Installation

Clone Repository:

git clone https://github.com/linartova/data-quality.git

cd data-quality

Install Dependencies:

pip install -r requirements.txt

Run the tool:

cd mmci-convector

python gui.py

Using the tool

This section describes running FHIR or OMOP conversions on data storage and input data. Then, there is a description of the step-by-step wizard, including questions for the user. Last, a table compares the converted fields and input files and more information about quality checks.

Prerequisites

The user must have the FHIR server and the OMOP CDM database to run the conversion and QC on both standards.

FHIR server

The recommended FHIR server is Blaze, which runs as a Java jar and is downloaded here.

OMOP CDM database

To use the OMOP conversion and QC tool, the user must download and set the OMOP CDM database. The tutorial is here. Recommended RDBMS is PostgreSQL.

Input file with data

The input file must be in XML. The tool uses the script "mmci_convertor/input_validation.py" to validate the XML input. The correct example of the input is in test_data.xml. In case of incorrect input, a short hint will appear.

Input data are described further in this documentation. Additional data will not be converted.

Step-by-step wizard

In the MMCI convector, the user can answer the set of questions. In the end, the answers will be evaluated and summarized, and the user will see the recommendation of which standard will be the best. The questionnaire will have two parts: the general questions and questions about which quality checks the user demands. There will also be a table with a comparison of quality checks.

Questions about general usage

What is the use case of the data? If it includes Data exchange, then the FHIR is the best option. If it is a Longitudinal analysis, then use OMOP.

Does the user have a license for SNOMED CT? The vast majority of OMOP use SNOMED CT, so if the user needs a license, it is better to go for FHIR, where the license is unnecessary.

What technical solution is better for data storage? FHIR uses a server with RESTfull API, and data are stored in Resources and files in JSON and XML formats. On the other hand, the OMOP CDM uses an SQL database for data storage with all the concepts and vocabularies that must be downloaded. The OHDSI also provides many open-source tools in R.

Does the user have additional data that needs conversion as well? If the user needs to convert more data not converted in the MMCI convector, then it is usually faster and more developer-friendly in FHIR. On the other hand, if the user knows OMOP, the conversion of the OHDSI open-source tool can be accelerated [].

What is the size of the data set? In FHIR, the large amount of data can cause performance issues. The OMOP using SQL database is a better choice for extensive data.

Does the user need a storage model or an interoperability model? For the final storage model, where data will not be changed again, the OMOP is the best solution. For repeatable data exchange between many facilities, the FHIR is the best.

Does the user need to customize the input data and minimize data loss? The best option is FHIR. The FHIR implements only a minimal model and is designed to provide many extensions and specifications. This goal is thanks to "profiling". Each facility has its profile, which can be found in Simplifier.

Does the user need the global model with standardized vocabularies and codes to simplify data? Then, the best option is OMOP.

Is it beneficial to use the OHDSI open-source tools? On the OMOP CDM works, OHDSI provides tools such as White Rabit for ETL, Achilles for database visualization, and the Data quality dashboard with many data quality checks.

Is it necessary to add custom constraints? The FHIR is the way to go.

Questions about QC

Tick all the needed QC from the list:

.

Final evaluation

The evaluation contains:

- Summarization of user answer
- Recommendation of standard based on the answer
- Table with needed QC and in which standard they are implemented

Conversion

In the tables below, the user can see how the data are mapped in the tool.

Tables list all elements from the input XML in the first row. Then, the blue rows are mapped into OMOP or FHIR, depending on the specific table.

BasicData FHIR				
Tag	Attribute	FHIR resource	FHIR object	Correction
2_2	Participation in clinical study			
22_4	KRAS exon 4 (codons 117 or 146) mutation status			

4_3	Time of recurrence (metastasis diagnosis)			
20_3	KRAS exon 2 (codons 12 or 13)			
6_3	Timestamp of last update of vital status			
87_1	BRAF, PIC3CA, HER2 mutation status			
24_4	NRAS exon 3 (codons 59 or 61)			
61_5	Liver imaging			
31_3	CT			
14_3	Microsatellite instability			
16_3	Risk situation (only HNPCC)			
85_1	Biological sex	Patient	gender	
3_1	Age at diagnosis (rounded to years)	Patient	birthDate	Dataelement_51_3 - Dataelement_3_1
21_5	KRAS exon 3 (codons 59 or 61)			
5_2	Vital status			
7_2	Overall survival status			
25_3	NRAS exon 4 (codons 117 or 146)			
88_1	Colonoscopy			
23_5	NRAS exon 2 (codons 12 or 13)			
63_4	Lung imaging			
30_3	MRI			
51_3	Date of diagnosis	Condition	recordedDate, onsetDateTime	
15_2	Mismatch repair gene expression			
Identifier		Patient	Identifier	

BasicData OMOP CDM				
Tag	Attribute	OHDSI table	OHDSI attribute	Correction
2_2	Participation in clinical study			
22_4	KRAS exon 4 (codons 117 or 146) mutation status			
4_3	Time of recurrence (metastasis diagnosis)			
20_3	KRAS exon 2 (codons 12 or 13)			

6_3	Timestamp of last update of vital status	OBSERVATION PERIOD	observation_period_end_date	
87_1	BRAF, PIC3CA, HER2 mutation status			
24_4	NRAS exon 3 (codons 59 or 61)			
61_5	Liver imaging	PROCEDURE OCCURRENCE	procedure_concept_id, procedure_source_value	mapping on right description
31_3	CT	PROCEDURE OCCURRENCE	procedure_concept_id, procedure_source_value	mapping on right description
14_3	Microsatellite instability			
16_3	Risk situation (only HNPCC)			
85_1	Biological sex	PERSON	gender_concept_id, gender_source_value	
3_1	Age at diagnosis (rounded to years)	PERSON	year_of_birth	
21_5	KRAS exon 3 (codons 59 or 61)			
5_2	Vital status			
7_2	Overall survival status			
25_3	NRAS exon 4 (codons 117 or 146)			
88_1	Colonoscopy	PROCEDURE OCCURRENCE	procedure_concept_id, procedure_source_value	mapping on right description
23_5	NRAS exon 2 (codons 12 or 13)			
63_4	Lung imaging	PROCEDURE OCCURRENCE	procedure_concept_id, procedure_source_value	mapping on right description
30_3	MRI	PROCEDURE OCCURRENCE	procedure_concept_id, procedure_source_value	mapping on right description
51_3	Date of diagnosis	PERSON	year_of_birth	calculation
		CONDITION OCCURRENCE	condition_start_date	
		PROCEDURE OCCURRENCE	procedure_date	only for diagnostic

				procedures
15_2	Mismatch repair gene expression			
Identifier		PERSON	person_source_value	

Pharmacotherapy OMOP CDM				
Tag	Attribute	OHDSI table	OHDSI attribute	Correction
10_2	Date of start of pharamacotherapy	DRUG EXPOSURE	drug_exposure_start_date, drug_exposure_start_datetime	
11_2	Date of end of pharamcotherapy	DRUG EXPOSURE	drug_exposure_end_date, drug_exposure_end_datetime	
59_5	Scheme of pharmacotherapy	DRUG EXPOSURE	drug_concept_id, drug_source_value	mapping
81_3	Other pharmacotherapy scheme	DRUG EXPOSURE	drug_source_value	only if scheme is other

Surgery OMOP CDM				
Tag	Attribute	OHDSI table	OHDSI attribute	Correction
8_3	Time difference between initial diagnosis and surgery	PROCEDURE OCCURRENCE	procedure_date	
49_1	Surgery type	PROCEDURE OCCURRENCE	procedure_concept_id, procedure_source_value	mapping into right values
67_1	Other surgery type	PROCEDURE OCCURRENCE	procedure_concept_id, procedure_source_value	only if Dataelement_49_1 is "Other"
9_2	Surgery radicality			
93_1	Location of the tumor			

Sample FHIR

Tag	Attribute	FHIR resource	FHIR object
56_2	Sample ID		
54_2	Material type	Specimen	type
55_2	Preservation mode		
89_3	Year of sample collection	Specimen	collection.collectedDateTime

Sample OMOP CDM				
Tag	Attribute	OHDSI table	OHDSI attribute	Correction
56_2	Sample ID	SPECIMEN	specimen_source_id	
54_2	Material type	SPECIMEN	specimen_concept_id, specimen_source_value	mapping on right description
55_2	Preservation mode			
89_3	Year of sample collection	OBSERVATION PERIOD	observation_start_date	choose the oldest date
		SPECIMEN	specimen_date	

Histopathology FHIR				
Tag	Attribute	FHIR resource	FHIR object	Correction
75_1	Distant metastasis			
83_1	Grade			
70_2	Stage			
92_1	Localization of primary tumor	Condition	code	mapping on right description
73_3	UICC version			
58_2	Availability invasion front digital imaging			
71_1	Primary Tumor			
77_1	Regional lymph nodes			
68_2	Localization of metastasis			
91_1	Morphology			
53_3	WHO version			
57_3	Availability digital imaging			
82_1	Biological material from recurrence			

	available			

Histopathology OMOP CDM				
Tag	Attribute	OHDSI table	OHDSI attribute	Correction
75_1	Distant metastasis			
83_1	Grade			
70_2	Stage			
92_1	Localization of primary tumor	CONDITION OCCURRENCE	condition_concept_id, condition_source_value	mapping on right description
73_3	UICC version			
58_2	Availability invasion front digital imaging			
71_1	Primary Tumor			
77_1	Regional lymph nodes			
68_2	Localization of metastasis			
91_1	Morphology			
53_3	WHO version			
57_3	Availability digital imaging			
82_1	Biological material from recurrence available			

Radiation therapy OMOP CDM				
Tag	Attribute	OHDSI table	OHDSI attribute	Correction
12_4	Date of start of radiation therapy	PROCEDURE OCCURRENCE	procedure_concept_id, procedure_source_value	mapping into right values
13_2	Date of end of radiation therapy			

Quality checks

Here is the list of quality checks provided by the tool. At the end, there is a comparison with the original R script running on the original input data.

Tables list all warnings and reports from the original R script working with input XML data in the first row. Then, the blue rows are mapped into OMOP or FHIR, depending on the specific table.

OMOP Warnings - part 1				
#	Original warning	Arguments	OMOP warning	OMOP CDM Tables
1	Vital check date precedes initial diagnosis date	(6_3 - 51_3)/7 < 0	observation_end_precedes_condition_start	cdf, odf
2	Vital check date is equal to initial diagnosis date	(6_3 - 51_3)/7 ==0	observation_end_equals_condition_start	cdf, odf
3	Suspicious survival information	7_2 and more		
4	Suspiciously young patient	3_1 < 15	too_young_person	pdf, cdf
5	Suspiciously long survival	7_2 and more		
6	Vital status timestamp missing	5_2 and more		
7	Vital status timestamp is in the future	6_3	observation_end_in_the_future	odf
8	Initial diagnosis date is in the future	51_3	condition_start_in_the_future	cdf
9	Pharmacotherapy scheme description is missing while pharmacotherapy scheme is Other	59_5, 81_3	missing_drug_exposure_info	ddf
10	Suspicious description of pharmacotherapy	59_5, 81_3	sus_pharma	ddf
11	Missing specification of used substances in pharmacotherapy description	59_5, 81_3		
12	Suspicious characters or words in description of pharmacotherapy	59_5, 81_3	sus_pharma_other	ddf
13	Surgery and histological location do not match	93_1, 92_1		
14	Surgery and histological location do not match (but multiple surgeries per patient)	93_1, 92_1		

15	Mismatch between surgery location and surgery type	93_1, 49_1		
16	Negative event (treatment/response) duration: end time is before start time	13_2, 36_1, 11_2	drug_end_before_start	ddf

OMOP Warnings - part 2				
#	Original warning	Arguments	OMOP warning	OMOP CDM Tables
17	Event (treatment/response) starts or ends after survival of patient	7_2 and more		
18	Start of response to therapy is before diagnosis	34_1		
19	Suspect incomplete followup: patient died of colon cancer while last response to therapy is Complete response	response, overall_survival		
20	Start of response to therapy is in the future	response		
21	Start of therapy is before diagnosis	12_4, 35_3, 8_3, 10_2	therapy_start_before_diagnosis	cdf, ddf, prdf
22	Start of treatment is in the future	12_4, 35_3, 8_3, 10_2, 51_3	treatment_start_in_the_future	ddf, prdf
23	End of treatment is in the future	13_2, 36_1, 11_2, 51_3	drug_exposure_end_in_the_future	ddf
24	Non-surgery therapy starts and ends in week 0 since initial diagnosis (maybe false positive)	12_4, 35_3, 8_3, 10_2, 13_2, 36_1, 11_2, treatment_type, pharma_scheme	sus_early_pharma	cdf, ddf
25	Suspiciously short pharma therapy - less than 1 week (may be false positive)	12_4, 35_3, 8_3, 10_2, 13_2, 36_1, 11_2, pharma_scheme	sus_short_pharma	cdf, ddf

26	Mismatch between provided and computed stage value	71_1 and more		
27	Suspicious TNM value combination for given UICC version (e.g., N2a for UICC version 6) or uncomputable UICC stage	71_1 and more		
28	pNX provided in TNM values, while UICC stage is determined (how?)	71_1 and more		

FHIR Warnings				
#	Original warning	Arguments	FHIR warning	FHIR Resources
4	Suspiciously young patient	3_1 < 15	too_young_person	pdf, cdf
8	Initial diagnosis date is in the future	51_3	condition_start_in_the_future	cdf

OMOP Reports - part 1					
#	Original report	Arguments	Comment	OMOP report	OMOP CDM Tables
1	createPlotWithSampleYears	89_3		missing_specimen_date	pdf, sdf
2	createPlotWithoutSampleYears	89_3			
3	createPlotWithoutSampleID	56_2		patients_without_specimen_source_id	pdf, sdf
4	createPlotWithoutPreservationMode	55_2			

5	createPlotWithoutMaterialType	54_2		patients_without_specimen_source_value_concept_id	pdf, sdf
6	createPlotsWithoutHistoValues	92_1	partially, other values are missing in OMOP	patients_without_condition_values	pdf, cdf
7	createPlotsWithoutSurgeryValues	8_3, 49_1	partially, other values are missing in OMOP	patients_without_surgery_values	pdf, prdf
8	createPlotsWithoutPatientValues	3_1, 85_1 and diagnostic values	partially, other values are missing in OMOP	missing_patient_and_diagnostic_values	pdf, prdf
9	createPlotsWithoutTargetedTherapy	35_3	partially, other values are missing in OMOP	missing_targeted_therapy_values	pdf, prdf
10	createPlotsWithoutPharmacotherapy	11_2, 10_2, 59_5		missing_pharmacotherapy_value	pdf, ddf
11	createPlotsWithoutRadiationTherapy	12_4	partially, other values are missing in OMOP	missing_radiation_therapy_values	pdf, prdf
1	createPlotsWithoutResponseTo	respon			

2	Therapy	se values			
13	createPlotForAllMissedValues	all values		completeness	
14	getMissingSampleWithoutPreserverationMode	55_2			

OMOP Reports - part 2					
#	Original report	Arguments	Comment	OMOP report	OMOP CDM Tables
15	getMissingSampleRecordSet	89_3, 56_2, 54_2	partially, other values are missing in OMOP	completeness	
16	getMissingHistoRecordSet	92_1	partially, other values are missing in OMOP		
17	getMissingSurgeryRecordSet	8_3, 49_1	partially, other values are missing in OMOP		
18	getMissingPatientRecordSet	3_1, 85_1 and diagnostic values	partially, other values are missing in OMOP		
19	getMissingTargetedTherapyRecordSet	35_3	partially, other values are missing in OMOP		
20	getMissingPharmacotherapyRecordSet	11_2, 10_2, 59_5			

21	getMissingRadiationTherapyRecordSet	12_4	partially, other values are missing in OMOP		
22	getMissingResponseToTherapyRecordSet	response values			
23	getAllSampleRecordSet				
24	getAllHistoRecordSet				
25	getAllSurgeryRecordSet				
26	getAllPatientRecordSet				
27	getAllTargetedTherapyRecordSet				
28	getAllPharmacotherapyRecordSet				
29	getAllRadiationTherapyRecordSet				

OMOP Reports - part 3					
#	Original report	Arguments	Comment	OMOP report	OMOP CDM Tables
30	getAllResponseToTherapyRecordSet				
31	getSampleRecordSetWithPreserverationModeFFPE				
32	getAllPatientsWithLocations				
33	getAllPatientsWithtTNMStageConspicuousness				

34	getAllTherapysandReponsesTogether				
35	getListsoFDataFramesWithCountsOfAllValues	all values	getAll* functions	counts_of_records	all
36	getListsoFDataFramesWithCountsOfAllMissingValues	all missing values	getAll* functions	completeness	all
37	getListsoFDataFramesWithCountsOfAllMissingValues				all
38	getCountFormsPerBiobank				all
39	getCountFormsWithMissingValuesPerBiobank				all
40	getPatientsWithPreservationModeBUTWithoutFFPE	FFPE values			
41	getPatientsWithoutSurgery	patient, surgrery		get_patients_without_surgery	pdf, prdf
42	getPatientsWhereNewTreatmentAfterCompleteResponseButNoProgressiveDiseaseOrTimeofRecurrenceAfterIt	response values			

FHIR Reports					
#	Original report	Arguments	Comment	FHIR report	FHIR Resources
1	createPlotWithSampleYears	89_3		missing_specimen_date	pdf, sdf
2	createPlotWithoutSampleYears	89_3			

3	createPlotWithoutSampleID	56_2		patients_without_specimen_source_id	pdf, sdf
6	createPlotsWithoutHistoValues	92_1	partially, other values are missing in OMOP	patients_without_condition_values	pdf, cdf

Dashboard and export

In the dashboard, the user can see the visualization of quality checks. Then, there are tables with failed rows. Users can download the graphs and failures in the zip files if needed later.

There is also a report with a table of failed rows, visualization, and a short description of the concrete quality check. At the beginning of the report, there is a brief message with a summary of the conversion and QC.

Conclusion and further resources

I hope you find this tool helpful and valuable, and there are resources for more information:

https://www.ohdsi.org/software-tools/

https://hakkoda.io/resources/fhir-to-omop/

https://medblocks.com/blog/which-health-it-standard-to-pick-fhir-openehr-or-omop

https://confluence.hl7.org/download/attachments/81018297/FHIR%20to%20OMOP%20Cookbook_v04.pdf?version=1&modificationDate=1707852008416&api=v2