Fully Automated Generation of COVID-19 SDTM Datasets from Electronic Health Records

Jozef Aerts, XML4Pharma

ABSTRACT

Starting from the COVID-19 Interim User Guide, and the newly published LOINC codes for COVID-19 tests, we developed a mapping between these LOINC codes and the SDTM-MB domain. A RESTful web service, extending the existing one for the LOINC-SDTM-LB mapping, was established and the API made public.
Together with the existing LOINC-LB mapping and a new, in development, mapping for VS (and its RESTful web service), this enables fully automated generation of DM, SV, VS, MB and LB SDTM datasets and the corresponding CM and MH datasets (and a RELREC dataset connecting these) from any Electronic Health Record (EHR) system that supports the HL7-FHIR API.
We also developed a mechanism to embed the FHIR source record into the SDTM dataset record, providing traceability to the source data, which may be of interest to regulatory reviewers.
The Java source code of the software has been made available on Sourceforge.

INTRODUCTION

Mapping of collected clinical research data to CDISC SDTM [1] is an enormous task, and usually takes a lot of time. Reason is that in many cases, the mappers do not know in advance what they will receive, especially regarding executed lab and lab-related tests. Background of this is that clinical research protocols do not provide exact information about what needs to be collected. For example, "measure glucose in urine" will lead to a multitude of different tests and test results, which all need to be mapped to SDTM-LB controlled terminology. This is easily avoidable by providing the set of allowed LOINC codes [2] for such tests already in the protocol, but such practice is unfortunately still very rare. Also, the CDISC Therapeutic Area User Guides (TAUGs) [3] very seldom contain LOINC codes for suggested tests.

Even now, CDISC has only started embracing LOINC codes for lab tests because of the new FDA requirement to provide them in SDTM-LB datasets. Although that would be extremely useful, there is no promotion for use of LOINC in other domains such as vital signs (VS), microbiology (LB) and questionnaires (QS). After all, we need to realize that healthcare uses LOINC for tests in general, and does not use CDISC controlled terminology at all.
Therefore, CDISC has started developing a mapping between the most used LOINC lab tests and the SDTM-LB controlled terminology. Such a mapping should essentially be unnecessary as will be argued later, as the LOINC code of a test is the worldwide unique identifier for the test.

During the course of the COVID-19 pandemic, LOINC published a set of new codes [4] related to the detection, quantification and worldwide spreading of the Corona virus. These codes are now implemented by all major manufacturers of test kits, and are even mandated for reporting purposes to the health authorities [5] such as the US Centers for Disease Control and Prevention CDC, who also published vendor lists and mappings to SNOMED-CT [6]. Also the US Department of Health and Human Services HHS requires COVID-19 lab tests to be reported using LOINC coding through the CARES act [7].
Unfortunately, CDISC does not use these at all, but is developing its own controlled terminology. So, in order to use test results from healthcare, for example from electronic health records (EHRs), a mapping between the new LOINC codes and CDISC controlled terminology is necessary.

As will be shown in this article, such a mapping enables the automated generation, usually taking a few minutes only, of submission-ready SDTM datasets from EHRs. In the classic case (collection from Case Report Forms – CRFs), this is a process that usually takes weeks or even months.

MAPPING OF LOINC CODES FOR COVID-19 to SDTM

During the COVID-19, CDISC published the "Interim User Guide – COVID-19" [8], explaining how COVID-19 related collected data needs to be mapped to CDISC-SDTM in the classic way. From the publication, it is clear that lab tests that are related to the detection of Corona virus (usually through its RNA) need to go into the MB (Microbiology) dataset, and not into LB. Little guidance however is provided how this should be done. The newly developed LOINC codes for COVID-19 are even not mentioned.

In order to enable (automated) generation of submission-ready SDTM datasets directly from EHRs, we developed such a mapping between the new COVID-19 LOINC codes and CDISC controlled terminology. The methodology used was very similar to the one used by the CDISC-Lab team to develop the LOINC-to-LB mapping [9] for the 2000+ most popular LOINC "classic" lab test codes [10]. Our COVID-19 mapping is currently not 100% stable yet, as LOINC is regularly adding new COVID-19 related codes, and as the new CDISC controlled terminology for COVID-19 for microbiology tests and methods is not final (status August 2020).
So we decided to currently not make the mapping available as an Excel file or any other tabular form, but implement it as a RESTful web service, so that it can be used in COVID-19 related clinical research, in applications that generate MB datasets from data that has the LOINC code for the test included. This means that currently, the mapping as in the underlying database is always a snapshot.

THE RESTFUL WEB SERVICE FOR LOINC-MB MAPPING

Publishing mappings between coding systems as worksheets is sub-optimal. It makes version control very difficult, and forces every individual implementer of the mapping to generate software systems and databases to implement the mapping. A much better method is to set up a publicly available RESTful web service and to publish the API for it. This is also the approach we followed.

The RESTful web service for the LOINC-CDISC-MB mapping is documented at http://xml4pharmaserver.com/WebServices/LOINC2CDISC_webservices.html#loinccorona2mb. It means that any application that has an internet connection, can use the service to obtain the mapping to CDISC-MB from a given COVID-19 LOINC code, and use that directly, e.g. to fully automatically generate MB datasets from EHRs.

The HTTP construct for the RESTful web service is:

http://www.xml4pharmaserver.com:8080/CDISCCTService2/rest/LOINC2SDTMCorona/{loinc_code}

where {loinc_code} represents the LOINC code for the COVID-19 test.
For example, to obtain the mapping to CDISC-MB for the new LOINC code 94500-6 [11], "SARS-CoV-2 (COVID-19) RNA [Presence] in Respiratory specimen by NAA with probe detection", the HTTP call will be:

GET *http://www.xml4pharmaserver.com:8080/CDISCCTService2/rest/LOINC2SDTMCorona/94500-6*

The HTTP response can be obtained as either XML or JSON (depending on the MIME-type requested). For XML for example, the response is:

```
<XML4PharmaServerWebServiceResponse ServerDateTime="2020-05-03T06:35:52">
  <WebServiceRequest>
    http://www.xml4pharmaserver.com:8080/CDISCCTService2/rest/LOINC2SDTMCorona/94500-6
  </WebServiceRequest>
  <Remark>
    This is a preliminary mapping! Results with a * in front are placeholders for future CDISC Controlled Terminology.
  </Remark>
  <Response>
    <MBTESTCD>SAR2RNA</MBTESTCD>
    <MBTEST>SARS-CoV-2 RNA</MBTEST>
    <MBTSTDTL>DETECTION</MBTSTDTL>
    <MBCAT/>
    <MBPOS/>
    <MBLOINC>94500-6</MBLOINC>
    <MBSPEC/>
    <MBLOC>RESPIRATORY SYSTEM</MBLOC>
    <MBMETHOD>NUCLEIC ACID AMPLIFICATION TEST</MBMETHOD>
    <MBANMETH/>
    <MBFAST/>
    <MBTPT/>
    <MBEVLINT/>
    <MBEVINTX>PT</MBEVINTX>
    <SUPPMB.RSLTYP>PrThr</SUPPMB.RSLTYP>
    <SUPPMB.RSLSCL>Ord</SUPPMB.RSLSCL>
    <SUPPMB.SYM/>
  </Response>
</XML4PharmaServerWebServiceResponse>
```


Stating that for LOINC code 94500-6, LBTESTCD="SAR2RNA", MBTEST="SARS-Cov-2 RNA", MBTSTDTL="DETECTION", MBLOC="RESPIRATORY SYSTEM", MBMETHOD="NUCLEIC ACID AMPLIFICATION", and MBEVINTX="PT" (point in time). Also mappings to the supplemental qualifiers RSLTYP="PrThr" ("result type = presence") and RSLSCL=Ord ("result scale=ordinal") are provided. In case of tests on specific genes of the RNA, the variable -SYM ("genetic symbol"), foreseen for the next version of the PGx standard, will be populated.
As one can see, such information could be used to automatically populate MB records directly from EHRs, as most EHR systems use LOINC coding for every type of test, and not only for classic lab tests as is often, but incorrectly, believed. This is necessary, as there are no EHR system that uses CDISC controlled terminology.

IMPLEMENTING THE MAPPING AND RESTFUL WEB SERVICE TO AUTOMATE GENERATION OF SDTM-MB-DATASETS

EHR records are more and more used as a source of data in clinical research. It is expected that in 5-10 years from now, they will be the primary source of data for clinical research [12]. Especially EHR systems that have an HL7-FHIR interface [13] are extremely promising to be used as a primary source for clinical research, due the highly standardized and very well documented FHIR-API [14] making it extremely easy to retrieve EHR information for patients that are also enrolled in a clinical study, even when the EHRs are not all in the same system or at the same hospital, when the patient has given consent to use the EHRs for the clinical study. Although this is not a topic of the current paper, we can state that FHIR has a very modern security and authentication mechanism [15] making this possible.

For our pilot, we used different publicly available "synthetic data" EHRs that provide an HL7-FHIR interface. The results reported here are from the "COVID19-under-FHIR" EHR system from SmileCDR [16]. The "base" of the FHIR RESTful web service is: *https://covid19-under-fhir.smilecdr.com/baseR4*.

We implemented the FHIR API, and our new mapping RESTful service in a simple Java program, using the Jersey libraries [17] for using the different RESTful web service. For the SDTM output, we selected not to use the outdated "SAS Transport 5" (XPT) format, but to use the modern CDISC Dataset-XML standard format [18]. The reasons for this will be explained further. For visualization of the datasets, we used the open source "Smart Submission Dataset Viewer" [19] which has "smart" features that go far beyond what is used by the regulatory authorities. A screenshot of a typical MB dataset thus generated is shown in Fig.1:

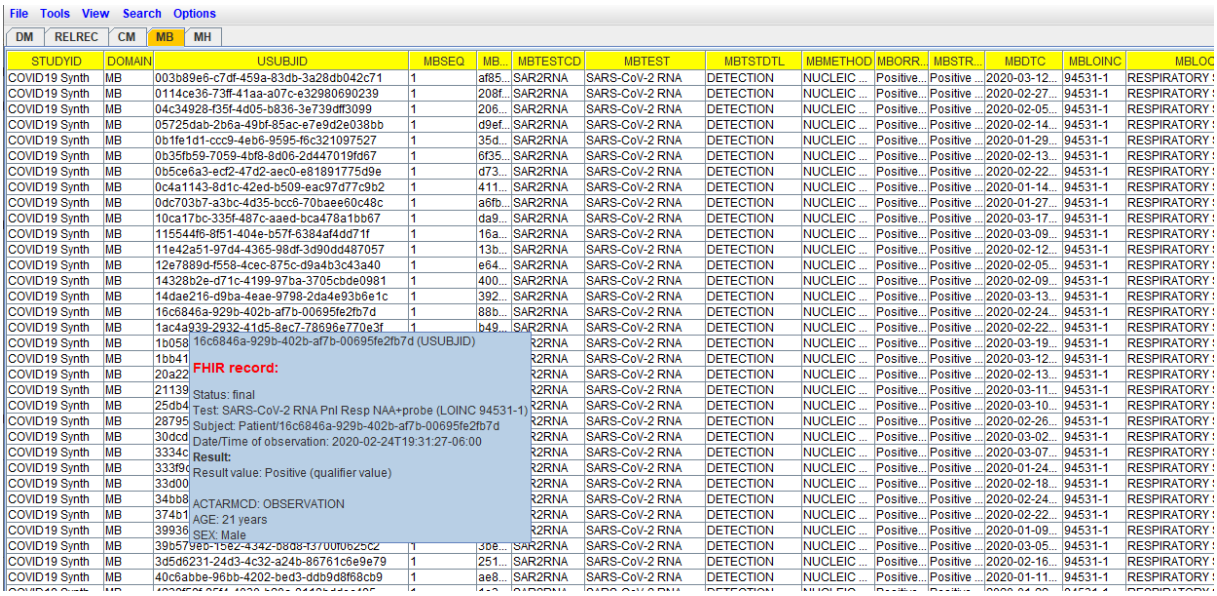

Fig.1: Screenshot of a typical MB dataset automatically generated from the EHR system, as displayed in the "Smart Submission Dataset Viewer". For each record, a human-readable representation of the source FHIR record is displayed as a tooltip.

For the value of USUBJID, we initially took the patient ID from the EHR. Later, we changed this by using the FHIR resources "ResearchSubject" [20] in combination with "ResearchStudy" [21**]**, where "ResearchSubject" acts as the liaison between patient information and study subject information, i.e. as the de-identifier.
The STUDYID was hard-coded. Everything else was retrieved from the EHR FHIR-resource instances itself, or derived on the fly.

EXTENDING THE SYSTEM: GENERATING ADDITIONAL SDTM DATASETS FROM THE EHR SYSTEM

After we were successful in generating SDTM-MB datasets directly and automatically from the EHR system through the use of the FHIR API and our mapping and their RESTful web services, we decided to extend the software to also generate other SDTM datasets.
For some of these, this is pretty easy, as the FHIR resource instances contain all information to populate the SDTM variables needed, where we used SDTMIG-3.2 as the reference [22]. So, we easily could generate the following datasets fully automatically: DM, CM, MH, LB, MB and RELREC (for relations between CM and MH). We will not provide details about the technical implementation here - these can be found in the source code on Sourceforge [23], but the approach is very similar to that of Zopf, Abolafia and Reddy [24] and its successor [25] used in a Phuse pilot project. The main difference is however that CDISC CDASH and ODM standards were not used as intermediates. We also generated a SUPPMB dataset, as we have mappings for RSLTYP ("result type") and RSLSCL ("result scale"), which are regarded as "Non-Standard Variables" (NSVs) and thus must be "banned" to a SUPPQUAL dataset.
We could also have generated a VS dataset, but we choose not to do so, as our LOINC-VS mapping is still incomplete (see further).

When the prototype software is started, it first asks for the EHR system to connect to:

```
Which FHIR-EHR System would you like to use?
0 - SyntheaStudy - server base: https://syntheticmass.mitre.org/v1/fhir
1 - HAPIFHIR - server base: http://hapi.fhir.org/baseR4
2 - Vonk - server base: https://vonk.fire.ly
3 - SPARK Firely - server base: http://spark.furore.com/fhir
4 - Azure - server base: http://sqlonfhir-stu3.azurewebsites.net/fhir
5 - Pyro - server base: https://stu3.test.pyrohealth.net/fhir
6 - COVID19 Synth - server base: https://covid19-under-fhir.smilecdr.com/baseR4
6
```

We select the "COVID19-under-FHIR" system. As this is a public system with synthetic data, no authentication is required.

The system then asks for whether subjects need to be selected based on a specific condition, or from the FHIR resources "ResearchSubject" and "ResearchStudy". We created such resources on a different FHIR server, as none of the systems we tested has implemented these resources yet. Also this is a nice presentation of using FHIR resources in a distributed way.

In this article, we choose to select subjects based on a specific condition. The system than asks for the SNOMED-CT code for the disease/condition for which patients need to be selected and information retrieved. For COVID-19, we start from the COVID-19 SNOMED-CT code "840539006" [26]:

```
Would you like to select Subjects based on a FHIR ResearchStudy (0) or on a Condition SNOMED-CT code (1)?
1
Please provide the SNOMED-CT code for the given condition.
Examples are:
COVID-19: 840539006
Diabetes Mellitus (general): 73211009
Diabetes Mellitus Type 2: 44054006
Diabetes Mellitus Type 1: 46635009
Parkinson's Disease: 49049000
Hypertension: 38341003
Dyspnea: 267036007
840539006
```

The system then submits a RESTful web service that searches for all patients which have a diagnose (FHIR resource Condition [27]) of COVID-19, and returns a "Bundle" [28] with "Condition" resource instances, from which the patient IDs and references to the "Patient" resources [29] are retrieved. The latter will then be used to generate the DM dataset fully automatically:

```
840539006
2020-08-22 11:44:23,609 INFO  Looking for Patient records with SNOMED-CT Condition = 840539006 in EHR System = https://covid19
2020-08-22 11:44:28,211 DEBUG FHIR Server HTTP Response Status = 200
2020-08-22 11:44:29,389 INFO  Creating new patient-subject = 1b0580b9-1ee3-4353-b555-64c797d57564 - for condition = 840539006
2020-08-22 11:44:29,572 INFO  Creating new patient-subject = 0c4a1143-8d1c-42ed-b509-eac97d77c9b2 - for condition = 840539006
2020-08-22 11:44:29,572 INFO  Creating new patient-subject = 40c6abbe-96bb-4202-bed3-ddb9d8f68cb9 - for condition = 840539006
2020-08-22 11:44:29,590 INFO  Creating new patient-subject = e5c6bf5f-772f-4fee-8d72-4d05bca3027d - for condition = 840539006
2020-08-22 11:44:29,592 INFO  Creating new patient-subject = 75575e6f-1b07-44af-a8f3-3223b1b9baf5 - for condition = 840539006
2020-08-22 11:44:29,592 INFO  Creating new patient-subject = af29ede4-c160-4b72-9239-e98926830b69 - for condition = 840539006
2020-08-22 11:44:29,600 INFO  Creating new patient-subject = c8b7b590-7dd1-4819-bc0a-0ec247487a67 - for condition = 840539006
2020-08-22 11:44:29,602 INFO  Creating new patient-subject = 3334cb34-7789-41e4-bae8-8f9af507972b - for condition = 840539006
2020-08-22 11:44:29,602 INFO  Creating new patient-subject = f5c3b3fe-8823-42ad-9488-abb2ff1e18ab - for condition = 840539006
2020-08-22 11:44:29,610 INFO  Creating new patient-subject = c29133bc-97c9-488e-92a5-3571f1ca2833 - for condition = 840539006
2020-08-22 11:44:29,613 INFO  Creating new patient-subject = 69821e92-aff2-4f80-8ea8-365e7c1d6804 - for condition = 840539006
2020-08-22 11:44:29,613 INFO  Creating new patient-subject = 89317528-2555-49bc-a92e-99fae6ad7ca8 - for condition = 840539006
2020-08-22 11:44:29,623 INFO  Creating new patient-subject = 6f811a24-99ee-481d-9f80-e629958f85d9 - for condition = 840539006
```

In our case, this results in 250 patients/subjects:

```
2020-08-22 11:44:32,587 INFO  Creating new patient-subject = 80717e5a-
2020-08-22 11:44:32,597 INFO  Creating new patient-subject = 08c974c2-
2020-08-22 11:44:32,617 INFO  Creating new patient-subject = 8c99dca7-
2020-08-22 11:44:32,627 INFO  Creating new patient-subject = d8702f9b-
2020-08-22 11:44:32,645 INFO  Creating new patient-subject = 2b03e4f1-
2020-08-22 11:44:32,657 INFO  Creating new patient-subject = 7af273fc-
2020-08-22 11:44:32,657 INFO  # of subjects = 250
```

In the next step, the system asks for a set of LOINC codes of all the tests that need to go into the CDISC "Findings" datasets. Alternatively, one can request the system to deliver all "Observations" that are available for each subject. The system will then try to generate LB, MB and/or VS datasets, depending on the mappings to CDISC that are available. In our case however, we start with a subset of the LOINC microbiology COVID-19 test codes, as has been published by LOINC [4]:

```
2020-08-22 11:44:32,657 INFO  # of subjects = 250
Do you want to obtain ALL observations for the selected subjects
 or use a selected set of LOINC codes for observations for the selected subjects?
0 = all observations
1 = observations for set of selected LOINC codes
1
Which set of LOINC codes would you like to use?
0 - Urinalysis : 17 codes
1 - Corona Virus Tests : 5 codes
2 - Hepatitic function 2000 panel : 7 codes
3 - Blood Type test : 13 codes
4 - CDISC Mapping Lab codes Chemistry : 1502 codes
5 - CDISC Mapping Lab codes Hematology : 292 codes
6 - CDISC Mapping Lab codes Coagulation : 107 codes
7 - CDISC Mapping Lab codes Toxicology : 241 codes
8 - CDISC Mapping Lab codes Urinalysis : 115 codes
9 - CDISC Mapping Lab codes Serology : 8 codes
10 - CDISC Mapping Lab codes Serology : 42 codes
11 - Synthea Lab codes : 41 codes
12 - Mini-mental Score Examination (MMSE) : 2 codes
1
```

For the demo, the subset currently consists of the LOINC codes 94531-1, 94499-1, 94503-0, 94504-8, 94306-8, as we know that these are the used codes in this EHR system for microbiology tests. In real life, one could of course select all new published COVID-19 LOINC codes, or even ask the system to retrieve all observations, and then generate and populate all SDTM Findings datasets automatically.

The system now starts retrieving the "Observation" FHIR resource instances for the given subjects and microbiology LOINC codes, resulting in 250 Microbiology (MB) records for our 250 subjects. In a real system, there will probably much more records, as there may be test repetitions over the course of the disease, and also different types of tests (RNA detection, RNS quantification, antibody test, …)

```
2020-08-22 12:05:42,519 INFO  HTTP Response Status = 200
2020-08-22 12:05:42,533 INFO  # of FHIR Observations = 0
2020-08-22 12:05:42,533 INFO  Looking for Observations with LOINC code = 94306-8 - for Subject = 7af273fc-2eb0-40f8-aac2-8354c40981bd
2020-08-22 12:05:42,713 INFO  HTTP Response Status = 200
2020-08-22 12:05:42,713 INFO  # of FHIR Observations = 0
2020-08-22 12:05:42,713 INFO  # of subjects so far = 250
2020-08-22 12:05:42,713 INFO  # records in LB dataset = 0
2020-08-22 12:05:42,713 INFO  # records in MB dataset = 250
2020-08-22 12:05:42,713 INFO  Some afterwork ... - domain = MB
2020-08-22 12:05:42,713 INFO  # records in LB dataset = 0
2020-08-22 12:05:42,713 INFO  # records in MB dataset = 250
2020-08-22 12:05:42,713 INFO  Now sorting by USUBJID and xxDTC
2020-08-22 12:05:42,713 INFO  Now adding xxSEQ to records
```

This process takes about 5 minutes, the time limiting step being the response from the FHIR server.

In the next step, "standardization" to SDTM –RESC and -RESN can be performed, either standardization to SI units or to "US conventional" units. Also for this, we developed a RESTful web service that was donated to NLM, and has recently been deployed on an NLM server [30] that does unit conversions between these two types of unit, using the LOINC code of the test (providing the molecular weight of the analyte, needed for the conversion). Remark here that the service uses UCUM notation for the units, as such unit conversions are impossible when using CDISC units. Fortunately, there is a good overlap.
In our case, we only have "detection" tests, so there is no quantitative result, and no need to do any conversions:

```
Standardization for --STRESC, --STRESN, --STRESU ...
Following options are available:
0 : no standardization
1 : standardization to SI units
2 : standardization to US Conventional units
Please select your choice (0, 1, or 2)
0
```

In the next step, the DM dataset is generated from the FHIR "Patient" resource instances. The information will however be incomplete, as the "Patient" resource of course does not have information from assigned or actual arm, and also none of the start- and end- datetimes of study participation, exposure to study drug, etc.. These can however be added automatically in a later stage, either from medication information from the EHR (resource "MedicationAdministration" [31] and or from the resource "ResearchSubject" which is containing all the information about envisaged and actual arm, informed consent etc. As "ResearchSubject" was not implemented on the test server, we have set up a separate FHIR server for it, containing mock data for these.

```
2020-08-22 12:17:43,549 INFO  Now looking up patient information for subject ID = 2b03e4f1-228c-4a96-9f10-d7582b5bb9f3
2020-08-22 12:17:43,680 INFO  HTTP Response Status = 200
2020-08-22 12:17:43,695 INFO  Demographics: local patientId = 2b03e4f1-228c-4a96-9f10-d7582b5bb9f3 - sex = F birthDate = 2005-02-18
2020-08-22 12:17:43,695 INFO  Demographics: local patientId = 2b03e4f1-228c-4a96-9f10-d7582b5bb9f3 - race =  - ethnicity =  - country = USA
2020-08-22 12:17:43,695 INFO  Now looking up patient information for subject ID = 7af273fc-2eb0-40f8-aac2-8354c40981bd
2020-08-22 12:17:43,849 INFO  HTTP Response Status = 200
2020-08-22 12:17:43,849 INFO  Demographics: local patientId = 7af273fc-2eb0-40f8-aac2-8354c40981bd - sex = M birthDate = 1999-10-20
2020-08-22 12:17:43,865 INFO  Demographics: local patientId = 7af273fc-2eb0-40f8-aac2-8354c40981bd - race =  - ethnicity =  - country = USA
2020-08-22 12:17:43,865 INFO  ***** 0 SDTM-LB records generated
2020-08-22 12:17:43,865 INFO  ***** 250 SDTM-DM records generated
2020-08-22 12:17:43,865 INFO  Writing SDTM DM dataset to file = FHIR2SDTMResults\DM_COVID19 Synth.xml
2020-08-22 12:17:43,997 INFO  # of subjects = 250
```

The system then asks whether also a "Medical History" dataset must be generated. If so, FHIR "Condition" resource instances are retrieved from the server, and MH records generated from them. This is very straightforward. The exact implementation can be looked up in the source code.

```
Do you also want to generate an MH (Medical History) dataset?
Y
```

Resulting in e.g.:

```
2020-08-22 12:30:08,770 INFO  MHTERM = COVID-19 - subject = 7af273fc-2eb0-40f8-aac2-8354c40981bd
2020-08-22 12:30:08,770 INFO  MHCAT =  - subject = 7af273fc-2eb0-40f8-aac2-8354c40981bd
2020-08-22 12:30:08,771 INFO  MHDTC = 2020-06-13T04:42:24.175+00:00 - subject = 7af273fc-2eb0-40f8-aac2-8354c40981bd
2020-08-22 12:30:08,771 INFO  MHDECOD = 840539006 - subject = 7af273fc-2eb0-40f8-aac2-8354c40981bd
2020-08-22 12:30:08,771 INFO  MHSTDTC = 2020-03-28T02:32:45-05:00
2020-08-22 12:30:08,771 INFO  MHENDTC = 2020-04-11T02:32:45-05:00
2020-08-22 12:30:08,771 INFO  MHTERM = Fever (finding) - subject = 7af273fc-2eb0-40f8-aac2-8354c40981bd
2020-08-22 12:30:08,771 INFO  MHCAT =  - subject = 7af273fc-2eb0-40f8-aac2-8354c40981bd
2020-08-22 12:30:08,771 INFO  MHDTC = 2020-06-13T04:42:24.175+00:00 - subject = 7af273fc-2eb0-40f8-aac2-8354c40981bd
2020-08-22 12:30:08,771 INFO  MHDECOD = 386661006 - subject = 7af273fc-2eb0-40f8-aac2-8354c40981bd
2020-08-22 12:30:08,771 INFO  MHSTDTC = 2020-03-28T01:11:45-05:00
2020-08-22 12:30:08,771 INFO  MHENDTC = 2020-04-11T02:32:45-05:00
2020-08-22 12:30:08,851 INFO  Starting writing MH dataset to = FHIR2SDTMResults\MH.xml
2020-08-22 12:30:08,938 INFO  # of records written = 500
2020-08-22 12:30:08,968 INFO  # of records written = 1000
2020-08-22 12:30:08,999 INFO  # of records written = 1500
Do you also want to generate an CM (Concomitant Medications) dataset?
Y
```

Very similarly, one can also generate an CM (Concomitant Medication) dataset.
When doing so, also a RELREC dataset is generated to document the relations between CM and MH, as the FHIR records also contain the "reason for medication", referencing the corresponding "Condition".

WHY USE CDISC DATASET-XML FORMAT FOR the GENERATED DATASETS?

In our prototype application, we used CDISC Dataset-XML format instead of XPT format for the generated datasets. There are several reasons for this:

- CDISC Dataset-XML is a modern format and does not have any of the limitations of the outdated XPT format
- CDISC Dataset-XML allows to incorporate the source record into the SDTM record itself.

The latter is extremely important. Regulatory authorities are always stating that they would like to have access to the source record, but the currently mandated SAS Transport 5 (XPT) format has no possibility to do so. With CDISC Dataset-XML, this is however "piece of cake". There are two possibilities to do so, either to have a copy of the source record (either as XML or JSON) directly into the XML of the Dataset-XML, or by adding an API reference to the source record. The latter would then indeed require that the regulatory authority obtains access to the source system through the API and using an authentication mechanism. It would probably also require that some of the information is filtered out for patient privacy reasons.

In our demo case, we choose to incorporate the source record into the SDTM record (Fig.2)

```
<ItemGroupData ItemGroupOID="IG.MB" data:ItemGroupDataSeq="1">
    <ItemData ItemOID="IT.STUDYID" Value="COVID19 Synth"/>
    <ItemData ItemOID="IT.DOMAIN" Value="MB"/>
    <ItemData ItemOID="IT.USUBJID" Value="003b89e6-c7df-459a-83db-3a28db042c71"/>
    <ItemData ItemOID="IT.MBSEQ" Value="1"/>
    <ItemData ItemOID="IT.MBREFID" Value="af85d327-78ea-4170-8e38-b0495b110baf"/>
    <ItemData ItemOID="IT.MBTESTCD" Value="SAR2RNA"/>
    <ItemData ItemOID="IT.MBTEST" Value="SARS-CoV-2 RNA"/>
    <ItemData ItemOID="IT.MBTSTDTL" Value="DETECTION"/>
    <ItemData ItemOID="IT.MBORRES" Value="Positive (qualifier value)"/>
    <ItemData ItemOID="IT.MBSTRESC" Value="Positive (qualifier value)"/>
    <ItemData ItemOID="IT.MBLOINC" Value="94531-1"/>
    <ItemData ItemOID="IT.MBLOC" Value="RESPIRATORY SYSTEM"/>
    <ItemData ItemOID="IT.MBMETHOD" Value="NUCLEIC ACID AMPLIFICATION TEST"/>
    <ItemData ItemOID="IT.MBDTC" Value="2020-03-12T07:10:20-05:00"/>
    <ItemData ItemOID="IT.MBDY" Value="1"/>
    <ItemData ItemOID="IT.MBEVLINT" Value="PT"/>
    <Observation xmlns="http://hl7.org/fhir">
       <id value="af85d327-78ea-4170-8e38-b0495b110baf"/>
       <meta>
                   <versionId value="1"/>
                   <lastUpdated value="2020-03-26T23:26:26.719+00:00"/>
                   <source value="#ochzJoHpusa8Ljo7"/>
                   <profile value="http://hl7.org/fhir/us/core/StructureDefinition/us-core-observation-lab"/>
                   <tag>
                      <system value="https://smarthealthit.org/tags"/>
                      <code value="Covid19 synthetic population from Synthea"/>
                   </tag>
                </meta>
       <fhir:text xmlns:fhir="http://hl7.org/fhir" xmlns:xhtml="http://www.w3.org/1999/xhtml">
          <fhir:status value="generated"/>
          <xhtml:div>
             <xhtml:p>Status: final</xhtml:p>
             <xhtml:p>Test: SARS-CoV-2 RNA Pnl Resp NAA+probe (LOINC 94531-1)</xhtml:p>
```

Fig.2: Incorporation of a FHIR source records into an SDTM-MB record, using the CDISC-Dataset-XML standard format. Such incorporations are impossible when using the by the FDA mandated SAS-XPT format.

In the "Smart Submission Dataset Viewer", this information can then easily be visualized as shown in Fig.1.

CONCLUSIONS AND RECOMMENDATIONS

Automated generation of SDTM datasets from EHR systems, especially from those implementing the FHIR API, is pretty easy, at least when mappings between LOINC codes for tests and SNOMED-CT (for coded results) to CDISC-CT are available.
In fact however, such mappings should essentially not be necessary, as when the LOINC code is provided, everything is clear and set, and values for -TESTCD, -TEST, -SPEC, -METHOD in SDTM are superfluous. Review tools can easily use RESTful web services to obtain the meaning of each code, as is already implemented in the Smart Submission Dataset Viewer [32]. So, we strongly recommend that current identifying variables such as -TESTCD, -TEST, -SPEC, -METHOD etc. in SDTM are set to "conditionally required": if the LOINC code is provided, there should be no necessity to populate them [33]. Doing so would further enable to fully automate the generation of SDTM datasets for almost all Findings/Observations in just tens of minutes, whereas the usual process (mapping by SAS programming) often takes weeks or even months. Especially for the case of COVID-19, this could save ten thousands of lives.

REFERENCES

1. CDISC SDTM Implementation Guide: https://www.cdisc.org/standards/foundational/sdtmig
2. LOINC: Logical Observation Identifier Names and Codes: https://loinc.org
3. CDISC Therapeutic Area Guides: https://www.cdisc.org/standards/therapeutic-areas/published-user-guides
4. SARS-CoV-2 and COVID-19 related LOINC terms: https://loinc.org/sars-cov-2-and-covid-19/
5. Centers for Disease Control and Prevention: How to report COVID-19 Laboratory Data: https://www.cdc.gov/coronavirus/2019-ncov/lab/reporting-lab-data.html
6. LOINC In Vitro Diagnostic (LIVD) Test Code Mapping for SARS-CoV-2 Tests: https://www.cdc.gov/csels/dls/sars-cov-2-livd-codes.html
7. COVID-19 Pandemic Response, Laboratory Data Reporting: CARES Act Section 18115: https://www.hhs.gov/sites/default/files/covid-19-laboratory-data-reporting-guidance.pdf
8. CDISC COVID-19 Interim User Guide: https://www.cdisc.org/standards/therapeutic-areas/covid-19
9. LOINC to LB Mapping Home: https://wiki.cdisc.org/display/LOINC2LB/LOINC+to+LB+Mapping+Home
10. LOINC Mapping Spreadsheet: https://www.cdisc.org/system/files/members/standard/terminology/LOINC_to_LB_Mapping_File.zip
11. LOINC code 94500-6: SARS-CoV-2 (COVID-19) RNA [Presence] in Respiratory specimen by NAA with probe detection: https://loinc.org/94500-6/
12. Use of EHRs data for clinical research: Historical progress and current applications: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6508843/
13. Fast HealthCare Interoperability Resources: https://www.hl7.org/fhir
14. HL7 FHIR REST API: https://www.hl7.org/fhir/http.html
15. HL7 FHIR Security: https://www.hl7.org/fhir/security.html
16. FHIRBall COVID-19 Efforts: https://www.fhirball.org/fhirball-covid-19-efforts/
17. Eclipse Jersey: https://eclipse-ee4j.github.io/jersey/
18. The CDISC Dataset-XML Standard: https://www.cdisc.org/standards/data-exchange/dataset-xml
19. The open source Smart Submission Dataset Viewer: https://sourceforge.net/projects/smart-submission-dataset-viewe/
20. HL7 FHIR Resource ResearchSubject: https://www.hl7.org/fhir/researchsubject.html
21. HL7 FHIR Resource ResearchStudy: https://www.hl7.org/fhir/researchstudy.html
22. CDISC Study Data Tabulation Model Implementation Guide v.3.2: https://www.cdisc.org/standards/foundational/sdtmig/sdtmig-v3-2
23. FHIRLOINC2SDTM Open Source project: https://sourceforge.net/projects/fhirloinc2sdtm/
24. R. Zopf, J. Abolafia and B. Reddy, Use of Fast Healthcare Interoperability Resources (FHIR) in the Generation of Real World Evidence (RWE): https://www.phusewiki.org/docs/Conference%202017%20RW%20Papers/RW04.pdf
25. S. Hume, J. Abolafia, and G. Low, Use of HL7 FHIR as eSource to Pre-populate CDASH Case Report Forms using a CDISC ODM API: https://www.lexjansen.com/phuse-us/2018/tt/TT16.pdf
26. March 2020 SNOMED CT International Edition Interim Release: Up to Date COVID-19 content available: https://www.snomed.org/news-and-events/articles/march-2020-interim-snomedct-release-COVID-19
27. HL7 FHIR Resource Condition: https://www.hl7.org/fhir/condition.html
28. HL7 FHIR Resource Bundle: https://www.hl7.org/fhir/bundle.html
29. HL7 FHIR Resource Patient: https://www.hl7.org/fhir/patient.html
30. NLM UCUM Unit Web Services: https://ucum.nlm.nih.gov/ucum-service.htmll
31. HL7 FHIR Resource MedicationAdministration: https://www.hl7.org/fhir/medicationadministration.html
32. Blog CDISC End-to-End: Why LBLOINC is so important - web services: http://cdisc-end-to-end.blogspot.com/2014/08/why-lbloinc-is-so-important-web-services.html
33. Blog CDISC End-to-End: Post-coordinated versus Pre-coordinated Controlled Terminology: http://cdisc-end-to-end.blogspot.com/2020/09/post-coordinated-versus-pre-coordinated.html

ACKNOWLEDGEMENTS

Special thanks are due to the members of the "CDISC COVID-19 Interim User Guide" development team, especially to Jon Neville, and to the members of the "CDISC Controlled Terminology" development team, in particular Erin Muhlbradt, for interesting discussions and information exchange.