PharmaSUG 2021 - Paper RW-063

# Generating FDA-ready Submission Datasets directly from EHRs

Jozef Aerts, XML4Pharma

## ABSTRACT

The use of Real-World-Data (RWD) in clinical research is growing rapidly, especially due to the rise of the HL7-FHIR standard [1] for electronic health record (EHR) exchange, with its standardized API. FDA however requires study data to be submitted as tables using the SDTM standard and in the outdated SAS Transport 5 format [2]. A new way to automatically generate SDTM data sets with embedded source FHIR records directly from the EHR system is presented. It mostly uses existing and by us developed LOINC-CDISC mappings [3] and a by us developed RESTful web service[4].
We also discuss aspects of the new FHIR-to-CDISC project [5], and how it can further boost the use of RWD in clinical research. Furthermore, the by the FDA so much desired "real rolling data review" is discussed in the scope of APIs, RESTful web services and modern exchanges formats such as XML, JSON and RDF.

## INTRODUCTION

The use of Real-World-Data (RWD) in clinical research, e.g. for use in synthetic placebo arms, is growing rapidly, especially due to the rise of the HL7-FHIR standard [1]. FDA however requires study data to be submitted as tables using the SDTM standard and in the outdated SAS Transport 5 format [2]. The traditional way is then to use the RWD for populating electronic case report forms (eCRFs), and then develop mappings between the operational data in the Electronic Data Capture (EDC) system and SDTM in the traditional way. Too often, this mapping process is only started around or after database closure, and requires a lot of programming, thus taking considerable time. Although high quality, SDTM-savvy software is available on the market allowing to already start developing the mappings once the study design is final (so even before the start of the trial), the traditional way is still to have SAS programmers develop and execute the mappings late in the process.
In the case of RWD however, when the EHR system has a standardized API, such as the HL7-FHIR standard [1], SDTM data sets could be generated "on the fly", directly from the EHR data itself. This however requires that mappings between LOINC codes, used in the EHR system to uniquely define tests, and CDISC controlled terminology [6] is available, as CDISC-SDTM uses post-coordination for describing tests, whereas the more precise LOINC coding system [7] is based on pre-coordination. For standardized results, CDISC developed its own controlled terminology, which is not used by the much larger medical world, who mostly uses SNOMED-CT [8]. So, in order to use data from EHRs in clinical research that needs to be transformed to SDTM data sets, in some cases, SNOMED-CT to CDISC-CT mapping will be necessary. Such a mapping is not available from CDISC, as there is little awareness and understanding within the organization and community about coding systems used in the medical world. The Unified Medical Language System (UMLS) [9] however partially provides such mappings, which can be generated or retrieved using the UMLS API and RESTful web services [10].

## METHODS

As most EHR systems, and especially through the use of the HL7-FHIR standard, use LOINC coding [7] for specifying tests, direct generation of SDTM data sets for observations require that a mapping between LOINC codes and SDTM variables (test codes and names, specimen- and body location codes, etc.) is available. Triggered by the mandated use of LOINC coding in laboratory SDTM data sets, CDISC developed and published such mapping, containing 2304 mappings for 1402 LOINC codes [3]. We soon found out that this is far from sufficient for real world usage, and therefore extended the mapping considerably, then containing 9514 mappings for 6171 LOINC codes. These were than implemented as a RESTful web service [4] which is much more suitable for automation than the by CDISC published Excel worksheet.

As this mapping is only suitable for laboratory observations that need to be mapped to the SDTM domain (remark that there are also laboratory observations that map to other domains, such as microbiology and

genetics findings), we also developed LOINC-CDISC mappings for observations to other domains.
As the FDA did not (yet) mandate the use of LOINC coding for COVID-19 related tests, which are categorized in the MB (microbiology) or into the PF (Pharmacogenomics Findings), or, as of SDTM-IG 3.4, in the GF (Genetic Findings) domain, we first developed the mappings for the MB domain for COVID-19 LOINC tests [4]. Remark that CDISC is currently not interested in such mappings, with the argument that FDA does not require LOINC coding for MB [11]. At the moment of writing (as new COVID-19-related LOINC codes are regularly published and our mappings are updated), we made 146 mappings for 97 LOINC codes available.
Another important set of LOINC codes is for vital signs. Also here, it is expected that CDISC will not develop mappings, as the FDA does not mandate the use of LOINC coding for the VS domain. For use with RWD however, such a set of mappings is of utmost importance, so we developed them ourselves. Currently, this comprises 593 mappings for 534 unique LOINC codes [4]. Also for these, a RESTful web service was established [4].

A Java program was then developed that implements the HL7-FHIR API [12] and RESTful web services. It also implements the new FHIR "ResearchStudy" [16] and "ResearchSubject" [34] resources. For the population of post-coordinated SDTM Findings identifier variables such as the test code (-TESTCD), test name (-TEST), specimen (-SPEC), and timing variables, the program uses our LOINC-CDISC RESTful web services for laboratory, COVID-19 related microbiology, and vital signs test codes [4].
For the generation of Interventions and Events data sets, the situation is less complicated, as these require less coding, except for MedDRA coding. A SNOMED-CT to MedDRA mapping has been developed as part of the WEB-RADR project [13], but at the time of writing, the results have not been published yet. They are expected to be published by May 2021 [14].

## SOFTWARE

The software program was developed using the Java programming language. As the software is meant as a "proof of concept" and pilot software, no graphical user interface has been developed yet. So, for the moment, the software runs from the line command.

When started, the program asks the user which EHR system needs to be queried. For the pilot, the program provides a list of demo EHR systems that implement the FHIR API (Fig.1).

```
C:\FHIR2SDTM_Software>java -jar FHIR2SDTM.jar
# Log Appenders = 0
logging goes to: logs\FHIRLOINC2CDISC_LOG_2021_4_8_8-54-39.txt
changing logging level to "ALL"
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
```

**Fig.1 List of demo EHR systems used by the demo software**

In a real life application, the next step would be an authentication step, which can again be required (or inherited) at the level of the study and/or the subject/patient. As the software is meant for demonstration and "proof of concept" purposes only, we did not implement any authentication, except for the Synthea system [15], which requires an API key.

The program then asks the user whether selection of patients/subjects should be either based on an existing study definition (as an ID of a FHIR "ResearchStudy" resource [16]) or on a SNOMED-CT code for a condition on which the patients will be selected. The latter can e.g. be used in the case that SDTM data need to be generated for a placebo or alternative treatment "synthetic" arm.

In the next step, patients or subjects are selected. In case the subject selection is based on a SNOMED-CT code for an existing medical condition, the selection is based on matching the SNOMED-CT code with the provided code in any of the "Condition" resource instances.

In the following step, the user is asked whether either all observations (instances of the "Observation" resource) should be used, or that observations should be selected on basis of a list of LOINC test codes. Just for the demo, a list of sets of pre-selected LOINC codes is then presented to the user. For example, for a COVID-19 study, this list contains the most used LOINC codes in Corona virus testing.

The program then queries the selected EHR system using the FHIR API, for all, or the by the user selected LOINC codes, observations for the selected subjects/patients. For each observation, it is checked whether a mapping to CDISC is available, and to which domain. If no mapping is found, the system currently skips the observation and reports this in the log file. If a mapping is found, an in-memory SDTM record is generated and the information retrieved from the EHR observation is added.

The user is than asked whether for numeric findings, an attempt should be performed to either standardize the results (in STDM variable -ORRES) to either SI units or to US Conventional units, which then go into the -STRESN/-STRESC/-STRESU (standardized values) variables. The latter is made possible through a by us developed RESTful web service for unit conversion that has been deployed by the National Library of Medicine NLM [17]. It uses the LOINC code to retrieve the molecular weight of the molecule under investigation to then do the unit conversion. Remark that this only works when the UCUM notation for units is used, which is usual in EHR systems. It would not work with CDISC units (although there is some overlap, especially for concentrations), as CDISC units is just a list, and not a system. Also, the CDISC unit list does not contain any conversion factors in a machine-readable form at all. Therefore, we also decided to not implement any UCUM-to-CDISC transformation for units, although some mappings have been published [18]. Reason is that we consider the UCUM system superior to the CDISC unit list: implementing the by CDISC published UCUM-to-CDISC mapping would lead to considerable quality loss, also making it very hard or impossible to perform unit conversions (e.g. when comparing studies) by regulatory reviewers. We also feel that CDISC should deprecate its unit list in favor of UCUM notation, and the regulatory bodies such as the FDA should encourage the use of UCUM notation for units.
In the case of vital signs data, the user is also asked whether results should be standardized to metric units or not. If so, this e.g. takes care that -STRESN/-STRESN/-STRESU is populated with metric units such as "cm" for body length or to "kg" for body weight. Also here, the NLM RESTful web service is used [17]. Blood pressures however remain untouched of this, they still are reported using "mm[Hg]" (millimeter mercury column) units.

In the next step, the system orders the generated SDTM records by subject and also assigns the SDTM sequence numbers -SEQ. The DM (Demographics) data set is then generated. As the contents in the EHR are very similar to those in the DM data set, this step is pretty straightforward.

Many of the mappings contain by the LOINC-CDISC [3] mapping team develop non-standard variables, such as "result type" (-RESTYP), "result scale" (-RESSCL), or "planned duration" (-PDUR). These non-standard-variables have been added to the mappings by CDISC, as it was (finally) recognized that the "classic" identifier variables in SDTM Findings domains cannot uniquely define tests. In our software, such non-standard variables result in the generation of "Supplemental Qualifier" data sets, such as SUPPLB or SUPPVS.

When the DM and Findings data sets have been generated, the user is asked whether also MH (Medical History) data sets need to be generated. When so, information for the selected subjects is retrieved from the "Condition" resource [19] instances for the selected subjects/patients. Also this is a pretty straightforward step. As stated before, coding to MedDRA (variable MHDECOD) has not been included yet, but this would easily be possible once a RESTful web services to transform SNOMED-CT codes to MedDRA is available. Also here, all in-memory records are then sorted by subject and the sequence number -SEQ is assigned.

In the next step, the user is asked whether also a CM (Concomitant Medications) needs to be generated. If so, the system is again queried for all "MedicationAdministration" [20] and "MedicationStatement" [21]

FHIR resource instances for the selected subjects/patients. Also here, no attempt is yet made to code the verbatim reported name of the drug or therapy, this time to WHO-Drug in CMDECOD.

Last but not least, the user is asked whether "related records" as a RELREC data set should be generated for the relation between CM and MH records. This is possible by the virtue of the "reasonReference" attribute in the FHIR "MedicationAdministration" and "MedicationStatement" resources, that usually references a "Condition" directly.

Finally, the system reports on which data sets have been generated and with how many records for each of them, and writes them to file using the CDISC Dataset-XML format. We selected not to generate SAS Transport 5 files, but to solely use the modern CDISC Dataset-XML format [22]:
- although often claimed so, SAS Transport 5 is not an open standard, and surely not "vendor neutral"
- generation of SAS Transport 5 would possibly additionally require splitting of records for values of more than 200 characters
- SAS Transport 5 only support US-ASCII characters. Especially in the case of verbatim items, there is no guarantee at all that the content is only using US-ASCII characters. If for example, the Spanish verbatim term "Acné", or "Cáncer" would be present in the EHR, this cannot be automatically be translated to ASCII-only characters, and the text would become unreadable in any software that uses SAS-Transport 5. The information would even become unusable in review by the regulatory authorities, as it is not clear what the original character encoding in the source EHR was (although nowadays, UTF-8/Unicode is used in most cases).
The major reason why we decided to generate the SDTM data sets in modern CDISC Dataset-XML format is however another: it allows to embed the source FHIR record into the SDTM record itself (Fig.2-3). This enormously enhances the traceability of the information, a feature so long hoped for by regulatory authorities. Furthermore, modern review tools such as the open source "Smart Submission Dataset Viewer" [23] already allow to visualize the embedded source record (Fig.4).

```
<ItemGroupData data:ItemGroupDataSeq="1" ItemGroupOID="VS">
   <ItemData ItemOID="VS.STUDYID" Value="CDISCPILOT01"/>
   <ItemData ItemOID="VS.DOMAIN" Value="VS"/>
   <ItemData ItemOID="VS.USUBJID" Value="01-701-1015"/>
   <ItemData ItemOID="VS.VSSEQ" Value="1"/>
   <ItemData ItemOID="VS.VSTESTCD" Value="DIABP"/>
   <ItemData ItemOID="VS.VSTEST" Value="Diastolic Blood Pressure"/>
   <ItemData ItemOID="VS.VSPOS" Value="SUPINE"/>
   <ItemData ItemOID="VS.VSORRES" Value="64"/>
   <ItemData ItemOID="VS.VSORRESU" Value="mmHg"/>
   <ItemData ItemOID="VS.VSSTRESC" Value="64"/>
   <ItemData ItemOID="VS.VSSTRESN" Value="64"/>
   <ItemData ItemOID="VS.VSSTRESU" Value="mmHg"/>
   <ItemData ItemOID="VS.VISITNUM" Value="1"/>
   <ItemData ItemOID="VS.VISIT" Value="SCREENING 1"/>
   <ItemData ItemOID="VS.VISITDY" Value="-7"/>
   <ItemData ItemOID="VS.VSDTC" Value="2013-12-26"/>
   <ItemData ItemOID="VS.VSDY" Value="-7"/>
   <ItemData ItemOID="VS.VSTPT" Value="AFTER LYING DOWN FOR 5 MINUTES"/>
   <ItemData ItemOID="VS.VSTPTNUM" Value="815"/>
   <ItemData ItemOID="VS.VSELTM" Value="PT5M"/>
   <ItemData ItemOID="VS.VSTPTREF" Value="PATIENT SUPINE"/>
   <Observation xmlns="http://hl7.org/fhir">
      <id value="blood-pressure"/>
      <meta>
         <profile value="http://hl7.org/fhir/StructureDefinition/vitalsigns"/>
      </meta>
      <text> <status value="generated"/> <div xmlns="http://www.w3.org/1999/xhtml"><p> <b> Generated Narrative with Details</b> </p> <p> <b> id</b> : blood-pres
         given as 'Vital Signs'})</span> </p> <p> <b> code</b> : Blood pressure diastolic supine<span> (Details : {LOINC code '8455-8' = 'Diastolic blood pressu
                  pressure supine'})</span> </p> <p> <b> value</b> : 64 mmHg<span>  (Details: UCUM code mm[Hg] = 'mmHg')</span> </p> <p> <b> interpretation</b>
               <system value="urn:ietf:rfc:3986"/>
               <value value="urn:uuid:187e0c12-8dd2-67e2-99b2-bf273c878281"/>
            </identifier>
      <!-- demonstrating the use of the baseOn element with a fictive identifier   -->
      <basedOn>
```

**Fig.2. Embedding the source HL-FHIR record within an SDTM record using the modern CDISC Dataset-XML format (upper part).**

```
<ItemData ItemOID="VS.VSTPTREF" Value="PATIENT SUPINE"/>
<Observation xmlns="http://hl7.org/fhir">
   <id value="blood-pressure"/>
   <meta>
      <profile value="http://hl7.org/fhir/StructureDefinition/vitalsigns"/>
   </meta>
   <text> <status value="generated"/> <div xmlns="http://www.w3.org/1999/xhtml"><p> <b> Generated Narrative with Details</b> </p> <p> <b> id</b> : blood-pressure</p>
      given as 'Vital Signs'})</span> </p> <p> <b> code</b> : Blood pressure diastolic supine<span> (Details : {LOINC code '8455-8' = 'Diastolic blood pressure--supin
               pressure supine'})</span> </p> <p> <b> value</b> : 64 mmHg<span>  (Details: UCUM code mm[Hg] = 'mmHg')</span> </p> <p> <b> interpretation</b> : Below
            <system value="urn:ietf:rfc:3986"/>
            <value value="urn:uuid:187e0c12-8dd2-67e2-99b2-bf273c878281"/>
         </identifier>
   <!-- demonstrating the use of the baseOn element with a fictive identifier   -->
   <basedOn>
      <identifier>
         <system value="https://acme.org/identifiers"/>
         <value value="1234"/>
      </identifier>
   </basedOn>
   <status value="final"/>
   <category>
      <coding>
         <system value="http://hl7.org/fhir/observation-category"/>
         <code value="vital-signs"/>
         <display value="Vital Signs"/>
      </coding>
   </category>
   <code>
      <coding>
         <system value="http://loinc.org"/>
         <code value="8455-8"/>
         <display value="Diastolic blood pressure--supine"/>
      </coding>
      <text value="Diastolic blood pressure--supine"/>
   </code>
```

**Fig.3. Embedding the source HL-FHIR record within an SDTM record using the modern CDISC Dataset-XML format (middle part).**

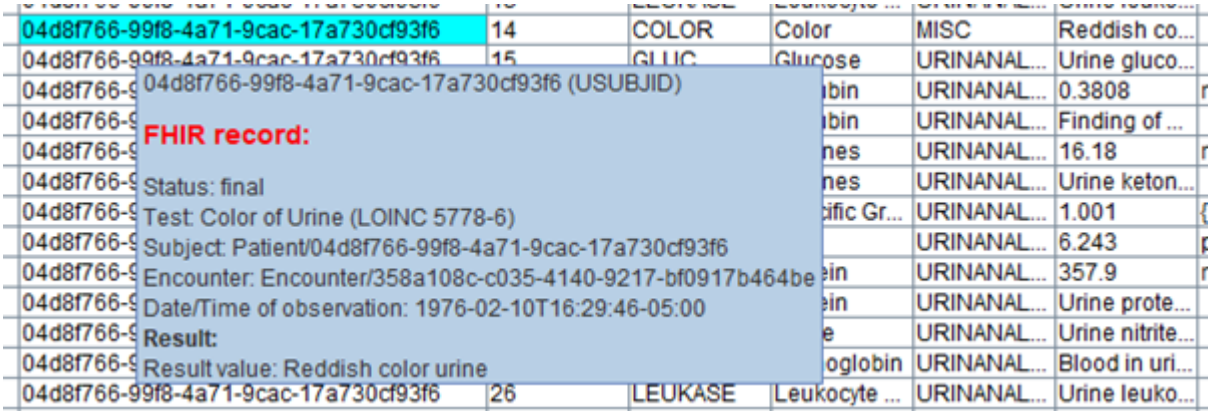


**Fig.4: Visualization of the source EHR FHIR record for an SDTM record in the LB data set, by the open source "Smart Submission Dataset Viewer"**

## RESULTS

In the pilot, and using our software, a good number of demo SDTM submission data sets have already been generated. Except for MedDRA and WHO-drug coding, the thus SDTM data sets were pretty complete, and SDTM compliant.
Generating such a set of SDTM submission data sets directly from an EHR system with FH7-FHIR interface is very fast compared to the usual procedures of retrieving the information from classic EDC systems with subsequent programming and execution of the mappings. In our case, generating a complete set of submission data sets is a matter of minutes or tens of minutes, mostly limited by the speed of the RESTful web service that is querying the EHR system.

The source code of the software is available from SourceForge at: https://sourceforge.net/projects/fhirloinc2sdtm/. Please do not expect to be running "out of the box". For example, the use of the Synthea demo EHR system [16] requires an API key, which need to be obtained by the end user himself. The source code however already comes with a set of pre-defined LOINC codes for use with the software in case not all observations for each subject should be treated.

Although we regularly update the source code and the sample result files on SourceForge, the software and all files are "as is", and there is no guarantee that everything is correct. So, the available source code should be regarded as a possible starting point for further development of real-life applications.

## LIMITATIONS

The application is currently limited to Findings domains for which there is a mapping available between LOINC codes and SDTM. This is currently the case for LB, VS and COVID-19-MB [4]. A mapping for EG (electrocardio) is currently in development. The latter would however require about 50 new CDISC test codes and names, which have however been refused by the EG-CT team, with the argument that such codes are only used in calculations or would mostly not be submitted to the FDA. Our argument that this would enhance completeness for the RWD case was rejected with the argument that more codes would confuse the (human) mappers. It thus looks as automation and use of RWD is not regarded as a use case for EG test codes. Another interesting domain is QS (Questionnaires): LOINC has a good number of codes for questionnaires and questionnaire questions, and it would be very interesting to see how these overlap with the questionnaires published by CDISC [24], and, if there is a reasonable amount of overlap, develop the mappings.
We must also take into account that the assignment of the correct domain is a challenge in itself. The classification in "domains" in SDTM is a rather arbitrary one, and not always logical. For example, if oxygen saturation is measured by a pulse oximeter, the data goes into vital signs (VS), but when it is measured from a blood sample, the data go into the laboratory domain (LB). For viruses, the information of the presence in a sample or quantification of the RNA does not go into LB, although the data is measured in the laboratory, but goes into MB. If the RNA of the virus is sequenced, the data goes into the old PF (Pharmacogenomics Findings) or in the new GF (Genetic Findings) domain [25]. FHIR and LOINC do not make such classifications: in FHIR, all observations/findings are represented by a single "Observation" resource, and LOINC only provides a "class" as a suggestion, which is however not used as an identifier at all. The GF domain mapping for COVID-19 tests that do not go into MB is also one of our first next goals.
So, if FDA keeps requiring SDTM data sets with CDISC-CT for submissions also of RWD, one would first need to decide for each LOINC code to which SDTM domain it maps to, and then first start to generate the appropriate mappings. Although we have already done so for almost 8,000 LOINC codes, the task is and remains a huge one, also as it would require a large number of "CDISC New Term Requests", as we have observed during this work that CDISC-CT is far from complete, and even does not aim for completeness.

Another limitation of the current work is that no standardization for non-numeric results in -STRESC or for other highly standardized variables in Events and Interventions domains is done. These are usually coded using SNOMED-CT [8], LOINC Answers [26], but also e.g. ICD-10 [27]. The development of mappings for such has just started, and can for a part already be obtained by using RESTful web services of the Unified Medical Language System (UMLS) [9] as has been used in one of our other projects, i.e. the automated generation of CDISC Biomedical Concepts directly from LOINC codes for panels and tests [28]. This is something that we want to add to the software in the future.

The current application does also not generate AE (Adverse Events) data sets. Reason is that the development of the FHIR "AdverseEvent" resource is still at "maturity level" 0, meaning that it is a very first draft, and that there currently is still very limited test data. It also remains open whether "AdverseEvent" [29] is the only resource to be queried for, as there also are other, related resources such as "AllergyIntolerance" [30].

The use of Dataset-XML is not considered to be a limitation of the current method. At the contrary. The objections of the FDA regarding file size are only apparent ones, as XML files can easily be compressed down to 2-3% of their original size, and modern computer applications do not even need to de-compress such files to read them. As such, binary XML files can be considerably smaller in size than SAS Transport 5 binary files.

## DISCUSSION

With the current regulatory submission mechanisms, the generation of submission data sets for RWD, mandated to be in CDISC-SDTM format, requires a massive amount of mapping between codes from coding systems used in healthcare (LOINC, SNOMED-CT, ICD, ...) and codes developed by CDISC. Unfortunately, there is no alignment at all, nor planned, between the two types of systems. It is only due to the FDA mandate for LOINC coding in SDTM-LB, that some awareness for LOINC has been created within the clinical research standard community and CDISC. Too often however, LOINC is still regarded as a burden instead of as an opportunity. This can also be seen by the fact that in the new CDISC COVID-19 Therapeutic Area User Guide [31] no mention of LOINC nor SNOMED-CT is made at all. SNOMED-CT is not used at all within the clinical research community. This also means that for the case of RWD, mappings would need to be developed between SNOMED-CT and CDISC-CT. For example, the SNOMED-CT code for "Severe acute respiratory syndrome coronavirus 2 RNA" is "1240411000000107" [32] but this code is not at all used in CDISC controlled terminology. Instead, CDISC developed its own test code "SAR2RNA" with CDISC-NCI code "C171531" and test name "SARS-CoV-2 RNA". As there are many thousands of bacteria and viruses and components of them, this would require an huge mapping effort, and this already only for microbiology. Of course, the better would be that CDISC adapts and implements SNOMED-CT. Even then, the current SDTM standard requires that CDISC test codes "may not be longer than 8 characters, and may not start with a number", which immediately excludes all LOINC and SNOMED-CT codes. The reason for this rule is the still by the FDA mandated use of the outdated SAS Transport 5 format [2] (a binary format stemming from the IBM mainframe time).
Another major issue for the use of RWD in clinical research and especially in electronic submissions to the regulatory authorities, is that all data information must be submitted in the form of tables. "If you only have SAS Transport 5, everything is a table". The only exception is the metadata of such submissions, which is done in the modern CDISC Define-XML format [33]. HL7-FHIR, which has become the de-facto standard for exchange of EHR data and of RWD, does not have such a restriction of having to be a "table". It is even independent of the actual exchange format being used: standardized representations exist in JSON, XML and RDF-Turtle, and others could easily be added.

As we expect that the use of RWD in clinical research will further grow, including the use of EHRs as a primary source for interventional clinical research data (reducing the need for "duplicate data entry"), this raises the question whether SDTM, in its current form, can further be the submission format of the future. Indications in this sense are the recent development of the "ResearchStudy" [16] and "ResearchSubject" [34] resources in HL7-FHIR. With these new resources, FHIR also becomes eligible as a submission format, with the advantage that all data does not be in the same physical place anymore, but is linked together trough the FHIR referencing mechanism and the use of modern APIs and RESTful web services. Such use of modern APIs and RESTful web services opens the door for "real rolling data" reviews, meaning that at least in theory, regulatory review can start as soon as a few percent of the data has been collected. As the amount of available data is then growing, review can be, even automatically be updated, and final submission decisions could be made very short time after the last data point is collected. This "real rolling data" idea and the use of APIs was discussed a lot during the June 2020 "Modernizing FDA's Data Strategy" meeting [35], but no effort was made to come to ideas about the "how". It is however clear that the use of SAS Transport 5 is not compatible with any such future implementation. FHIR however, with its modern transport formats, is.
On the other hand, it is clear that the use of categorization in domains and the use of tables, at least for visual inspection, is an important factor for "ease of review" for regulatory reviewers. It however also means that due to the categorization in SDTM domains, highly related data becomes disparate. For example, quantification or presence of Corona virus RNA must be looked up in one domain (MB), whereas information about a genetic modification of it is to be looked up in another domain (PF or GF). Of course, artificially generated RELREC tables can be generated to provide the relationship, but it is and remains cumbersome and error prone, and a non-ideal solution at all.
So, the question comes up whether a "best of both worlds" is possible, i.e. ease of review, combined with the use of multidimensional, even distributed, but highly explicitely linked data, exchanged using a modern format that is ideally suited for the use with APIs and RESTful web services.

We don't have the answer yet to this question, but it looks as the use of the CDISC Dataset-XML format, for which also a very compact JSON implementation is currently being developed, containing SDTM records with embedded source EHR records, for enhanced traceability, can be a first step. We hope to be able to present these ideas in a more elaborated form at the US Interchange in Autumn 2021.

## FUTURE WORK

As there are still a number of limitations to this approach, we want to further extend the mappings. Electrocardiography data (mapping to the EG domain) is the first on the list. We will also start investigating whether RWD from questionnaires can be mapped to CDISC-SDTM (QS domain). First tests show that this will not be an easy task.

The next interesting target domain is the GF (Genetics Findings) domain of the draft SDTM-IG 3.4 [36]. This is especially important in the context of the COVID-19 pandemic, as all sequencing and other genetic data of Corona virus mutations and variations have to be mapped to this domain.

The FHIR-to-CDISC project [5] will also provide us a large number of FHIRPath expressions that can easily be used in the software. This can replace many of the hard-coded parts of the software for linking FHIR resources with SDTM domains, especially for non-Findings domains. It doesn't however solve the domain categorization issues for FHIR observations to Findings domain, which still need to go over the LOINC code.

## CONCLUSIONS

We extended the existing LOINC-CDISC mapping, originally developed for only a relative small set of classic laboratory LOINC codes (with the aim of satisfying the LBLOINC FDA requirement), with several thousands of additional mappings, including mappings for COVID-19-related microbiology tests, and with mappings for vital signs tests.

We then developed a software for generating SDTM data sets directly from electronic health records for which an HL7-FHIR interface is available. This allows to fully automatically generate SDTM data sets for the DM, LB, VS, MB and for MH and CM domains, with a RELREC data set for describing the relation between CM and MH data sets. The data sets are generated in the modern CDISC Dataset-XML format, with the source FHIR resource data being embedded into the SDTM record itself, in order to demonstrate traceability back to the source EHR record.

The source code of the software has been made available as "open source" on the SourceForge website.

The growing use of RWD in clinical research and our results raise the question whether SDTM in its current form (2-dimensional tables and the use of SAS Transport 5) is still fit for the future. We don't think it is, as it looks to be a "showstopper" for the future use of APIs and "real rolling data" by the FDA. Some first ideas about a way to "combine the best of both worlds" (SDTM and FHIR) are being presented. We hope to be able to present these ideas in a more elaborated form at the US Interchange in Autumn 2021.

## REFERENCES


1. HL7 FHIR Release 4. https://www.hl7.org/fhir/
2. FDA Study Data Technical Conformance Guide, March 2021. https://www.fda.gov/media/147233/download
3. LOINC to SDTM-LB Mapping: https://www.cdisc.org/system/files/members/standard/terminology/LOINC_to_LB_Mapping_File.zip
4. XML4Pharma LOINC to CDISC Mapping Web Services. http://xml4pharmaserver.com/WebServices/LOINC2CDISC_webservices.html
5. D. Raths. FHIR-to-CDISC Project Seeks to Facilitate Use of EHR Data in Clinical Research. https://www.hcinnovationgroup.com/interoperability-hie/standards/news/21137358/fhirtocdisc-project-seeks-to-facilitate-use-of-ehr-data-in-clinical-research
6. CDISC Controlled Treminology. https://www.cdisc.org/standards/terminology/controlled-terminology

7. Logical Observation Identifiers Names and Codes. https://loinc.org
8. Systematized Nomenclature of Human and Veterinary Medicine. https://www.snomed.org/
9. Unified Medical Language System. https://www.nlm.nih.gov/research/umls/index.html
10. UMLS REST API. https://documentation.uts.nlm.nih.gov/rest/home.html
11. CDISC. Personal communication.
12. HL7-FHIR RESTful API. https://www.hl7.org/fhir/http.html
13. WEB-RADR: Recognising Adverse Drug Reactions. https://web-radr.eu/
14. P. Revelle, J. Millar. Personal communication
15. About SyntheticMass. https://synthea.mitre.org/about
16. HL7-FHIR resource ResearchStudy. https://www.hl7.org/fhir/researchstudy.html
17. NLM UCUM Web Services. https://ucum.nlm.nih.gov/ucum-service.html
18. UCUM-CDISC codetable. https://www.cdisc.org/system/files/members/standard/terminology/Unit-UCUM_Codetable_2021-03-26.xlsx
19. HL7-FHIR resource Condition. https://www.hl7.org/fhir/condition.html
20. HL7-FHIR resource MedicationAdminstration. https://www.hl7.org/fhir/medicationadministration.html
21. HL7-FHIR resource MedicationStatement. https://www.hl7.org/fhir/medicationstatement.html
22. The CDISC Dataset-XML Standard. https://www.cdisc.org/standards/data-exchange/dataset-xml
23. The Smart Submission Dataset Viewer (open source). https://sourceforge.net/projects/smart-submission-dataset-viewe/
24. CDISC Standard QRS. https://www.cdisc.org/standards/foundational/qrs
25. CDISC Standard PGx. https://www.cdisc.org/standards/foundational/pgx
26. LOINC Answers. https://loinc.org/answer-file/
27. International Statistical Classification of Diseases and Related Health Problems (ICD). https://www.who.int/standards/classifications/classification-of-diseases
28. J.Aerts. Automated Generation of CDISC Biomedical Concepts Starting from Healthcare Terminologies. CDISC European Interchange 2021.
29. HL7-FHIR resource AdverseEvent. https://www.hl7.org/fhir/adverseevent.html
30. HL7-FHIR resource AllergyIntolerance. https://www.hl7.org/fhir/allergyintolerance.html
31. CDISC Therapeutic Area User Guide COVID-19. https://www.cdisc.org/standards/therapeutic-areas/covid-19
32. SNOMED-CT code for "Severe acute respiratory syndrome coronavirus 2 RNA". https://confluence.ihtsdotools.org/display/snomed/SNOMED+CT+COVID-19+Related+Content
33. The CDISC Define-XML standard. https://www.cdisc.org/standards/data-exchange/define-xml
34. HL7-FHIR resource ResearchSubject. https://www.hl7.org/fhir/researchsubject.html
35. FDA Meeting Announcement. Modernizing FDA's Data Strategy. June 30, 2020. https://www.fda.gov/news-events/fda-meetings-conferences-and-workshops/modernizing-fdas-data-strategy-06302020-06302020
36. CDISC SDTMIG v.3.4. https://www.cdisc.org/standards/foundational/sdtmig/sdtmig-v3-4

## ACKNOWLEDGMENTS

We thank many of our colleagues at CDISC and in the CDISC community, at HL7, and in the Phuse community, for the many interesting discussions. Thanks also to Helena Sviglin (FDA) for suggestions on the mappings between some special FHIR resources and SDTM domains.

## CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Jozef Aerts
XML4Pharma
Jozef.Aerts@XML4Pharma.com
www.xml4pharma.com / www.xml4pharmaserver.com