What's Up with LOINC and UCUM? From EHR Records to LB Dataset in Just a Few Minutes

Jozef Aerts, XML4Pharma

ABSTRACT

The FDA has mandated the use of LOINC coding for lab tests in submissions, and recently endorsed UCUM for representing units. What does this mean for CDISC-standards based submissions?
The CDISC-Lab Team together with FDA, Regenstrief, NIH, and others, developed a CDISC-to-LOINC mapping, allowing to generate LBTESTCD/LBTEST, LBCAT, LBSPEC, LBMETHOD, LBFAST, LBLOC, LBPOS, LBTPT and LBEVLINT values starting from one of the >2400 most used LOINC codes.
We developed a RESTful web service to implement this mapping. It can be used from within any software in any modern software language (SAS, R, Java, Python, C#, …) and on any platform, to automate the mapping process. This RESTful web service will be an enormous help for those generating SDTM-LB datasets for lab results when the LOINC code is available.
We also recently solved the problem of SI-to-conventional units conversion and vice versa. Also here, we developed a RESTful web service, for which the code was donated to the National Library of Medicine (NLM), and at the moment of writing, is being implemented on an NLM server.
These RESTful web services, together with the RESTful API provided by the HL7-FHIR standard, enable to auto-generate complete SDTM LB and DM datasets. During the conference presentation, a life demo was performed where tentousands or LB records were generated starting from an Electronic Health Record (EHR) repository in a few minutes only, including automatic conversion from US conventional units to SI units for the population of LBSTRESC/LBSTRESN, and including the generation of the corresponding DM dataset.
This paper explains the underlying principles and mechanism for such automated generations in detail.

INTRODUCTION

Real World Data (RWD) in clinical research is mostly defined as the use of observational electronic health records for use in clinical studies. Also the FDA has recognized that this is an important topic and has published some definitions and guidance [1]. So far, electronic health records have mostly been used in clinical research to find and select suitable patients for specific clinical studies, like in the EHR4CR project [2]. Direct usage of EHRs to populate CRFs is also relative common nowadays [3], but requires programming specific for the peculiarities of the CRF and the study.

The HL7-FHIR standard has been a game changer in the healthcare world [4]. The recent addition of a number of Research resources "ResearchStudy" and "ResearchSubject" define an interface between the clinical research world, where CDISC standards are used, and the healthcare world, where HL7 standards are used. It is expected that HL7 will develop more research FHIR resources, thus extending this interface between the two worlds.

That these are still two worlds is demonstrated by the fact that CDISC submission standards are mostly based on post-coordinated controlled terminology [5] whereas FHIR resources use pre-coordinated controlled terminology. This is especially visible in the use of especially laboratory test codes, where CDISC uses a combination of different controlled terminology terms for the analyte (LBTESTCD), the specimen (LBSPEC), and the method used (LBMETHOD). HL7-FHIR however use pre-coordinated LOINC

controlled terminology, where the combination of analyte, propery (what is measured), time aspect, specimen, scale (quantitative, qualitative, ordinal, …) and method is represented by a single code.
In 2016, the announcement of the FDA that it will require the use of LOINC codes in submission laboratory data sets [6] caused a lot of concern in the clinical research and CDISC community. Although LBLOINC was already an SDTM variable, it was marked as "permissible" in the SDTM-IG, so almost nobody was using it. Also, SDTM required and still requires the use of postcoordinated terms for describing the laboratory tests, with the problem that this does not guarantee test uniqueness description [7,8]: for example, it does not allow to discriminate between quantitative and ordinal tests, expect from an interpretation of the results themselves.
The FDA announcement triggered the initiation of a working group, consisting of representatives from FDA, CDISC, Regenstrief Institute (the developers of LOINC) and the National Institute of Health (NIH) [9], to develop a mapping between the most used LOINC codes (pre-coordinated) and CDISC controlled terminology (post-coordinated) for the Laboratory (LB) SDTM-IG domain. This gigantic task was finalized in 2019, and the "LOINC to CDISC Mapping" was recently published for public review [10], and is expected to be published for "final" in the next months. For testing purposes, we developed a RESTful web service for using this mapping, as will be explained in the "Methods" section of this paper.

Another main difference between the CDISC clinical research world and the HL7 healthcare world in on the use of units. Healthcare (and also many other areas, like aviation, engineering etc) use UCUM notation for units [11], whereas CDISC developed its own list of units. A major difference between both is that UCUM notation is a system (not a list) and allows automated unit conversion, where this is not the case at all for CDISC units [12]. In the past, we already developed a RESTful web service for such conversions based on UCUM, which has been deployed for public use by the National Library of Medicine NLM [13]. This does however not solve the problem of "US conventional" to "SI" units. The latter has become important, due to the statements of the FDA that SI units are preferred, but that some reviewers will still require US conventional units [14]. The PMDA however requires the use of SI units [15].
Most US laboratories however use US conventional units uniquely, meaning a huge conversion effort, which, when using CDISC units, cannot be automated. Using a combination of UCUM notation and LOINC coding for the laboratory tests however, this can easily be accomplished. Therefore, we also developed a RESTful web service for conventional to SI unit (and vice versa) conversion, further described in the "Methods" section.

Using RESTful web services for unit conversions, in combination with existing LOINC RESTful web services and the FHIR RESTful web services API, make it possible to develop an application that retrieves FHIR "Observation" resources/records of lab results, and to automatically convert these into CDISC-SDTM-LB datasets, including standardization from conventional to SI units (LBORRES to LBSTRESC/LBSTRESN), as explained further on.

METHODS

RESTFUL WEB SERVICES

In the past, we gained a lot of experience with the development of RESTful web services for use in healthcare and clinical research. All these RESTful web services are described on our server at: http://xml4pharmaserver.com/WebServices/index.html. They are all written in Java (version 8) and deployed using a Tomcat application server (version 8) of a Linux host.

For the RESTful web service for the "CDISC to LOINC Mapping", we used the Excel worksheet provided by CDISC during the "internal review period", and converted that into a relational database (2402 rows). On the application server, when a request is obtained to return the CDISC variable values for a given LOINC code, the database is queried, and the results transformed into XML that is then returned as the HTTP response.

The RESTful web service for conversion of US conventional to SI units and vice versa is more complicated. It uses almost the same API methods as the conventional by the NLM deployed unit conversion RESTful web service [https://ucum.nlm.nih.gov/ucum-service.html], i.e.:

```
[base]/{source_quantity}/from/{source_unit}/to/{target_unit}
```

When a "conventional to SI" (or vice versa) conversion is requested, two additional parts in the request string are however required, either proving the LOINC code, or providing the molecular weight of th analyte, i.e.

```
[base]/{source_quantity}/from/{source_unit}/to/{target_unit}/LOINC/{loinc_code}
```

or

```
[base]/{source_quantity}/from/{source_unit}/to/{target_unit}/MOLWEIGHT/{molecular_weight}
```

First of all, the obtained request string is splitted into the different parts. If it is clear that it is a "classic UCUM conversion", so not a "conventional to SI" (or vice versa) conversion, the usual UCUM conversion using the "ucum-essence.xml" file [16] is executed [17]. In case of a "conventional to SI" (or vice versa) conversion however, the value of the parameter {loinc_code} or {molecular_weight} is read out.
If the LOINC code is provided, the "LOINC Component Part Number" [18] of the analyte is retrieved from a local implementation of the LOINC database, and the molecular weight retrieved from a simple CSV file that contains mappings between LOINC analyte part numbers, analyte name, and molecular weight of the analyte.
It is then detected whether the conversion is a "conventional to SI" or an "SI to conventional unit" conversion. The "source_unit" and "target_unit" are then decomposed into parts as in the classic method, and in the case of "SI to conventional" conversion, each occurrence of "mol" as the base compound is replaced by the molecular weight and "g" (gram). For example, "mmol/L" for "glucose" with molecular weight of 180.2, is converted as follows:

mmol/L = 10-3 * 180.2 g / L

In the case of "conventional to SI, the inverse of the molecular weight is used to convert "g" as a base unit to "mol". For example for glucose:

dg/mL = 10-2 * (180.2)-1 mol / 10-3 L

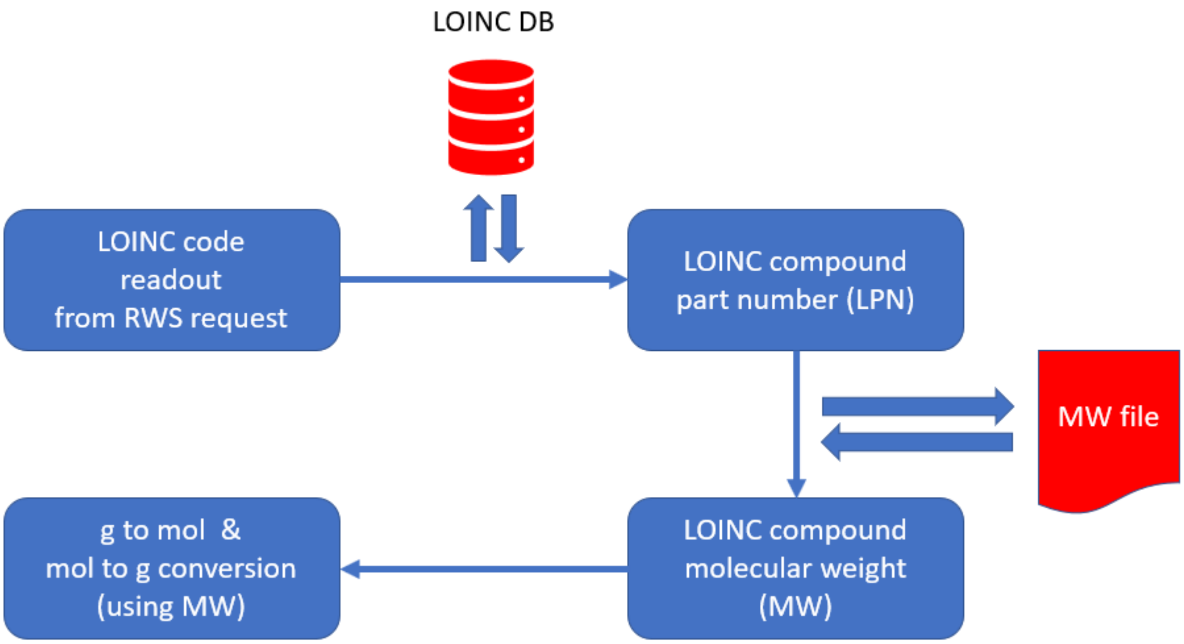
LOINC DB
LOINC code readout from RWS request
LOINC compound part number (LPN)
MW file
g to mol & mol to g conversion (using MW)
LOINC compound molecular weight (MW)

Figure 1. Workflow for conventional to SI (and vice versa) conversion using LOINC and UCUM.

Automated generation of SDTM-LB and DM records from EHR repositories

A large number of FHIR test servers, implementing the FHIR API are currently publicly available [19]. In our application, written in Java, we implemented six of these servers, ranging from the Synthea repository [20], a very large dataset containing realistic but fictional residents of the state of Massachusetts, to the Pyro server. An overview is provided in Table 1.

FHIR public test server	Description	API base
Synthea	https://synthea.mitre.org/about	https://syntheticmass.mitre.org/v1/fhir (*)
HAPIFHIR	http://hapi.fhir.org/	http://hapi.fhir.org/baseDstu3
Vonk	https://vonk.fire.ly/R4	https://vonk.fire.ly/R4
SPARK Furore	http://spark.furore.com/	http://spark.furore.com/fhir
Azure	http://nprogram.azurewebsites.net/	http://sqlonfhir-stu3.azurewebsites.net/fhir
Pyro	https://pyrohealth.net/	https://stu3.test.pyrohealth.net/fhir

Table 1. Public FHIR servers used for working with the FHIR to SDTM-LB application

(*) The Synthea server requires an access key (which can be obtained free of charge) and allows a maximum of 1000 hits per request.

The lists of LOINC codes to be used to retrieve (laboratory) test results from these servers contained two different sets of urinalysis tests, a set of test codes from the Synthea database, and a small set of Mini-mental Score Examination (MMSE) LOINC codes. The latter were used to test whether the approach may also be suitable to the SDTM QS (Questionnaires) domain, as LOINC also contains a good number of questionnaires codes.
The application was written as a Java simple application without graphical user interface, and uses the console for input.

After selection of the FHIR repository and the set of LOINC codes to be applied, the latter are submitted through the selected server, requesting "Observation" resources that have the selected LOINC code as identifier for the test. Essentially this reduces to the following GET request:

```
GET [BASE]/Observation?code=http://loinc.org|{loinc_code}
```

The obtained bundle of resources is then transformed to FHIR-XML when the response was obtained in JSON format (not all servers support FHIR-XML), and the XML is then parsed and split into "Observation" nodes. Each node is then transformed into SDTM using the following mapping table:

FHIR attribute	SDTM Variable
Patient reference ID	USUBJID
effectiveDateTime	LBDTC
Encounter reference ID	VISITNUM/VISIT
valueQuantity or valueCodeableConcept	LBORRES
unit	LBORRESU
coding	LBLOINC
referenceRange/low	LBORNRLO
referenceRange/high	LBORNRHI
interpretation/coding/display	LBNRIND
dataAbsentReason/coding/display	LBREASND

The value for STUDYID is assigned. In future, when the FHIR resources "ResearchStudy" and "ResearchSubject" are implemented on FHIR servers, the STUDYID will probably come from these, or just picked from a list of available studies on the FHIR server by the researcher.

The values for LBSTRESC, LBSTRESN, LBSTRESU, LBSTNRLO, and LBSTNRHI are first copied from the "original" equivalents, but can later be overwritten when standardization to another type of unit is requested.

The LOINC code was then used to generate values for LBTESTCD, LBTEST, LBCAT, LBSPEC, LBMETHOD, and LBEVLINT using the above described RESTful web service. In order to improve efficiency, values were cached in memory.

In the next step, the user is asked whether unit standardization needs to be performed. There are three choices: no standardization, standardization to SI units and standardization to US conventional units. In the latter two cases, the corresponding LOINC code in the "other" system is looked up from a local file retrieved from the LOINC database. For example, LOINC code 1751-7 represents "Albumin [Mass/volume] in Serum or Plasma", which is "conventional units". The "SI" counterpart has LOINC code 54347-0, representing "Albumin [Moles/volume] in Serum or Plasma". In order to find out to which unit in the counterpart system the original values need to be converted, a lookup for the "Example UCUM units" is performed in the counterpart LOINC code. For example, for LOINC 1751-7, the UCUM unit to be converted comes from the counterpart LOINC code 54347-0 which is umol/L.
Of course, for tests that are non-quantitative no such lookup is necessary, and the original values will be copied into the standardized values. For quantative tests for which a counterpart LOINC code and UCUM unit could be found, the unit conversions of the original values (LBORRES, LBORNRLO and LBORNRHI) to standardized units (LBSTRESC, LBSTRESN, LBSTNRLO and LBSTNRHI) are then performed using the RESTful web service. For efficiency, the conversion factors obtained from the RESTful web service are cached, and reused when possible.

Finally, the records are sorted by subject, and the values of LBSEQ calculated and assigned. They are then written to file in the CDISC Dataset-XML format (XPT files could easily be generated from them), with each FHIR resource as XML being embedded into each Dataset-XML record. This has the great advantage that it can be visualized as the "source record" by modern review tools such as the "Smart Submission Dataset Viewer" [21,22].

The unique patient IDs from the FHIR server are then reused to query the "Patient" FHIR resources, and convert the returned resources into a DM dataset. This is very easy, as the contents of the FHIR "Patient" resource and the SDTM-DM domain are very similar. Note however that "race" and "ethnicity" are non-core FHIR attributes, and need to be retrieved from country-specific extensions such as already exist for the USA [23].

RESULTS

The RESTful web services were successfully deployed on our application web service. They are not all publicly available yet, as they are intended to be moved to the NLM server, or be updated for the final LOINC to CDISC mapping.

During the demo, we generated SDTM datasets for urinalysis tests (18 tests), resulting in a total of 15,000 LB records for 33 unique subjects in the DM dataset, within 3 minutes.

LIMITATIONS

In the demo, which is meant to be a "proof of concept", we did not care about whether the patients are really enrolled into the study, nor about access restrictions (authorization, authentication) to the EHR repository. In a real world application, these will of course needed to be taken care off, so that researchers can only retrieve data of patients that are enrolled to the specific study and that have signed the informed consent. This can however be done using the FHIR resources "ResearchStudy" and "ResearchSubject" is explained in the section "Future implementations of similar applications".
We developed our demo application with version 3.2 of the SDTM-IG as target. We did not implement the by the CDISC Lab Team newly developed non-standard variables LBANMETH (Analysis method), LBRSLTYP ("Result type", corresponding to "SCALE" in LOINC), LBEVINTX (Evaluation Interval Text) that either have now become available through the SDTM model 1.7, or can be implemented as a non-standard variable. In future, we may decide on adapting the application for SDTM-IG v.3.3.
We did also do not any attempt to convert units in UCUM notation to CDISC units. The reason is that we strongly believe that UCUM notation should be allowed in SDTM submissions. After all, UCUM notation does not only have the advantage of being used worldwide (CDISC units are nowhere used outside CDISC), but also has the advantage of enabling unit conversions, also between US conventional and SI units, as shown in this paper. Such automated conversions are impossible using CDISC units.

FUTURE OF THE RESTFUL WEB SERVICES

At the moment of writing, the "LOINC to CDISC Mapping" RESTful web service is not publicly available yet. The reason is that it uses an early "for review" version of the mapping. As soon as the mapping is published by CDISC as "final", the RESTful web service will be updated, and the API will be made publicy available.
Once the mapping becomes available through the CDISC Library [24] however, we will retreat our own RESTful web service, and help the users to make the transition to the CDISC Library.

The RESTful web service for conversion of US conventional to SI units and vice versa has been deployed on our application server, and is currently being used in non-production mode by a number of pharma companies worldwide. People and organizations interested in using the RESTful web service can contact us and will then obtain a full description of the API and its methods. The source code of this RESTful web service has been donated to the NLM, and is currently in the process of being deployed on their server. Once this is completed, we will retreat our own web service.

FUTURE IMPLEMENTATION IN REAL-WORLD APPLICATIONS

The application that we developed is a prototype, a "proof of concept". Interesting is that it is very fast. In our Synthea case, the 15,000 SDTM records were generated in less than 3 minutes, including retrieval from the EHR system, generating post-coordinated SDTM identifier variable values, and standardization from US conventional to SI units.

In a real world application, there will of course many other factors to be taken into account, the most important being:

- Patient IDs need to be converted to Subject IDs
- Communication between the different parts where subject information is exchanged need to be secured
- Only records of patients enrolled in the study with a signed informed consent may be retrieved from the EHR system
- The investigator may only have access to records that are applicable to a study in which he/she is engaged.

This is where the new HL7 FHIR resources "ResearchStudy" and "ResearchSubject" come into play.
The ResearchStudy resource [25] describes essential information about the study, including the purpose, objective, sponsor, investigator, involved sites (references to "Location" resources), therapy, condition being studied, schedule of activities, and other key items.
The ResearchSubject has its own subject ID (which can be used for SUBJID) references a ResearchStudy, and further contains information about the assigned arm, the actual arm, and whether inform consent was signed. It also contains a reference to the corresponding Patient resource.
This means that by combining the information for the resources ResearchStudy, ResearchSubject and Patient, a filter can be generated to only retrieve information from subjects in a specific study for a specific set of tests (by the LOINC code). Additional filters can easily be applied, such as time period filters to when the test was performed. This information also allows to automatically populate ARMCD, ARM, ACTARMCD and ACTARM in SDTM-DM records.

As none of the FHIR-based test EHR systems has an implementation of ResearchStudy or ResearchSubject yet, we did not use such filtering. As FHIR however easily allows distribution of information, the server containing ResearchStudy and ResearchSubject resources does not be the same as the one for the EHRs. Therefore, we may set up a small FHIR server in future for extending the demo. In a real life implementation, the use of ResearchStudy and ResearchSubject will indeed be an important scenario for obtaining the right records for the right patients.

CONCLUSION

The use of electronic source records has become pretty custom in clinical research. The usual approach is to retrieve the information from the EHR into a CRF or eCRF, or in case of laboratory information, often into a database or file system. All these then need to be mapped to SDTM, an often tedious and time consuming exercise. This paper and the demo application that we developed demonstrate that, at least in the case of EHR systems with laboratory data and using FHIR, LOINC, and UCUM, SDTM datasets can be generated fully automatically, including conversions from conventional to SI units or vice versa, and this within minutes.

REFERENCES

1. Real World Evidence. FDA publication: https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence
2. Electronic Health Records for Clinical Research: http://www.ehr4cr.eu/
3. Source Data Capture from EHRs: Using Standardized Clinical Research Data. FDA publication: https://www.fda.gov/media/132130/download
4. Game changer: Creator of FHIR writes about approaching critical mass and a growing data sharing revolution, Healthcare IT news: https://www.healthcareit.com.au/opinion/game-changer-creator-fhir-writes-about-approaching-critical-mass-and-growing-data-sharing
5. CDISC-CT: we can do better. Working on and with CDISC Standards – blog: http://cdiscguru.blogspot.com/2018/09/
6. FDA Standards Catalog: http://www.fda.gov/downloads/ForIndustry/DataStandards/StudyDataStandards/UCM340684.xlsx
7. LOINC and the mapping to SDTM-LB. Working on and with CDISC Standards – blog: http://cdisc-end-to-end.blogspot.com/2017/10/loinc-and-mapping-to-sdtm-lb.html
8. From ACE to Zinc - Examples on the use of SDTM Controlled Terminology for lab tests. P. Vervuren : https://www.lexjansen.com/phuse/2010/cd/CD03.pdf
9. LOINC and the SDTM: https://www.cdisc.org/kb/article/loinc-and-sdtm
10. CDISC Controlled Terminology: https://www.cdisc.org/standards/terminology
11. The Unified Code for Units of Measure: https://unitsofmeasure.org
12. SDTM in non-submission Research (Part 2): Some Thoughts on Best Practices. Working on and with CDISC Standards – blog: https://cdiscguru.blogspot.com/2019/04/sdtm-in-non-submission-research-part-2.html
13. UCUM Web Service: https://ucum.nlm.nih.gov/ucum-service.html
14. Position on Use of SI Units for Lab Tests. FDA publication: https://www.fda.gov/media/109533/download
15. Overview of CDISC Implementation. Yuki Ando. PMDA publication: https://www.pmda.go.jp/files/000163676.pdf
16. ucum-essence.xml: http://unitsofmeasure.org/ucum-essence.xml
17. The Use of RESTful Web Services in Medical Informatics and Clinical Research and Its Implementation in Europe. J.Aerts: http://ebooks.iospress.nl/volumearticle/46463
18. LOINC Accessory Files: https://loinc.org/downloads/accessory-files/
19. Publicly Available FHIR Servers for testing: https://wiki.hl7.org/Publicly_Available_FHIR_Servers_for_testing
20. About Sythetic Mass: https://synthea.mitre.org/about
21. The open source Smart Submission Dataset Viewer: https://sourceforge.net/projects/smart-submission-dataset-viewe/
22. Source EHR records in SDTM. CDISC end-to-end – blog: http://cdisc-end-to-end.blogspot.com/2018/06/source-ehr-records-in-sdtm.html
23. HL7 FHIR US Core Implementation Guide: https://www.hl7.org/fhir/us/core/StructureDefinition-us-core-patient.html
24. The CDISC Library: https://www.cdisc.org/cdisc-library
25. HL7-FHIR Resource ResearchStudy: https://www.hl7.org/fhir/researchstudy.html

ACKNOWLEDGEMENTS

Special thanks are due to Erin Muhlbradt of NCI/NIH for many discussions about the use of LOINC and UCUM in clinical research, to many of the CDISC employees and volunteers who have an open ear for innovation and to many in the CDISC community for interesting discussions and for encourangement when it is about the use of LOINC and UCUM in clinical research.