Paper DS02

An open-source project to map lab terminology and standardize units for analysis

Sunil Kumar Pusarla, Omeros Corporation, Seattle, USA
Kevin Viel, Navitas Data Sciences
Dante Di Tommaso, Sanofi, Seattle, USA

ABSTRACT

FDA requires LOINC codes, an international standard that uniquely and unambiguously identifies laboratory measurements[(1)], for lab data supporting new drug and licensing applications, in studies initiated in 2020, and 2021[(2)]. The burden of preparing lab results for analysis drops whenever a laboratory provides LOINC codes. The FDA does recognize "inconsistencies in the specification and interpretation of LOINC codes submitted across tests, studies, and subjects."[(3)] Above and beyond analyte terminology, sponsors must convert results to standard analysis units. Moving forward, whether handling LOINC inconsistencies, or conducting studies that rely on lab results for which LOINC codes are not available, preparing laboratory results for robust and persuasive analyses will continue to present challenges of standardization and conversion. Sponsor variability is suboptimal in the context of public health policy.

The authors present and invite contributors to an open-source project to facilitate the standardization of lab test data, and overcome undesirable sponsor variability:

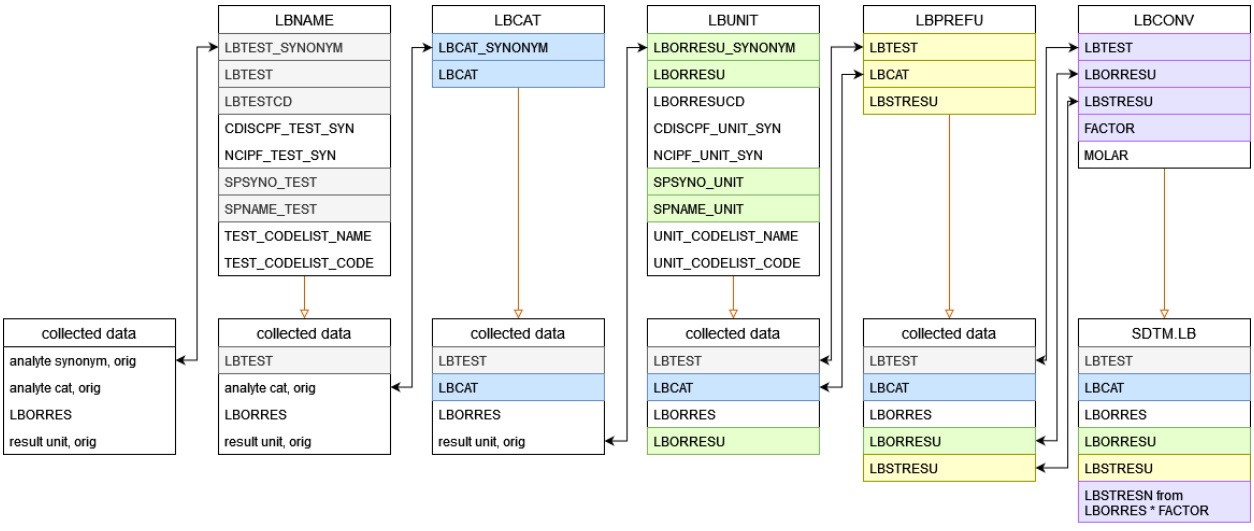
- map terms to industry and enterprise standards, and
- convert numeric results to preferred analysis and reporting units.

INTRODUCTION

Prior papers have addressed the importance of standardization of laboratory terminology and results and outlined key concepts and workflows. We will not rehash these fundamentals, and instead refer you to those thoughtful, comprehensive papers[(4)(5)]. However, they have not offered working solutions. While working for a mid-sized biotechnology company, the authors encountered circumstances that many have faced before: time and resource constraints, voluminous non-standard laboratory data essential to regulatory applications, varying international terminology and measurement conventions, study-specific requirements, and mandates for unimpeachable quality and transparency of data processing and analysis.

These circumstances took us on a daunting journey that resulted in the solutions which we now share. What we present today is basic and incomplete, and therefore a great opportunity to collaborate on an open-source project to ensure that basic capabilities are conducted consistently across the industry.

The workflow is simple, but accumulating, evaluating, and organizing the required information proved challenging:



Our current solutions, implemented first with The SAS System® and later with R software, do not cover some topics discussed in those prior papers. In particular, the Standardize Labs software

- focuses on *quantitative* results, and does not alter *qualitative* results,
- does not process test specimen and method terminology, and
- does not process normal ranges.

These are all potential extensions. Test specimen and method are *permissible* CDISC record qualifiers, and standardized terminology ensures accuracy of analyses and results. Note FDA data validation rule **FDAB030** (rule ID SD0007):

- "Standard Units (--STRESU) must be consistent for all records with the same Short Name of Measurement, Test or Examination (--TESTCD), Category (--CAT), Subcategory (--SCAT), Specimen Type (--SPEC) and Method of Test or Examination (--METHOD)."[13]

WHY OPEN SOURCE?

Our main challenges were to

1. map novel terminology to standard terminology, primarily for test names and original result units,
2. accumulate a database of conversion factors, and
3. deliver within a defined timeline and budget.

Extensive information is available in industry guidance documents and common references. Guides include CDISC controlled terminology[6], which specifies CDISC Submission Values for test names and results units, and further provides CDISC synonyms and National Cancer Institute (NCI) preferred terminology. Common references include publicly available websites and tools like http://unitslab.com/ and https://dwjay.tripod.com/conversion.html, as well as publications like the NEJM series based on the case records of Massachusetts General Hospital[7][8][9].

These guides and references, however, are fragmented and not available in a format that reflects FAIR data principles[10]. This necessitates redundant efforts across the industry to consolidate, qualify and maintain basic, non-competitive information. Such redundant effort is wasteful, subject to variability, and potentially error-prone.

Errors and variability in the standardization of terminology and conversion of test results can be difficult to detect and pernicious. Reasonable variability, alone, may influence decisions about health outcomes. This seems suboptimal, since choosing between a conversion factor of 40 or 40.1 or 40.078, should not compromise patient safety, or shift a public health decision. Collective engagement to develop, review, qualify, and maintain a central repository and tool can effectively minimize associated risks. A single, distributed, and shared effort can deliver a reliable, high-quality resource and tool.

In addition, the data processing logic of these tasks is relatively simple, which creates opportunities to incorporate more compelling benefits. Using SAS software, we implemented both data step hash and SQL workflows, in order to benchmark relative strengths*, and provide some measure of independent validation. These implementations produce identical results, so users can choose according to their preference or experience. By comparison, our R and shiny implementation is incomplete, and provides an easy-access learning experience for any other developers that share our interests in continually expanding and improving our technical capabilities. Delivery-based learning remains the fastest path to professional advancement, in our experience. Having a conceptually straightforward product on which to work is ideal for extending our technical skills.

*The authors' opinion? The SAS data step code using hash lookup may be less familiar and intuitive than SQL code, but does start to exhibit performance and debugging advantages for moderately large lab data sets.

As with any open-source software project, potential contributions are broad and varied: develop back-end and front-end features; improve robustness, reliability and transparency of existing features; write technical and user docs; identify, prioritize and design functionality; test and qualify features; share and learn technical skills and subject-matter expertise; build your reputation within your professional community. If you share our interests, please join us!

STANDARDIZATION AND CONVERSION WORKFLOW

As we progress through the standardization and conversion workflow, above, we will highlight key considerations, efforts and (your!) future contributions (i.e., current gaps in functionality 😊).

Our SAS implementation currently supports hierarchical standards:

- Enterprise standards, based on CDISC terminology,
    - Project configuration or conventions, and
        - Study configuration or conventions.

Project configuration extends and overwrites Enterprise standards, and Study configuration extends and overwrites Project conventions.

Based on our experience within organizations, a two-level hierarchy with Study configuration overwriting Enterprise standards is sufficient, while three levels are rarely necessary.

As an industry project, the LBPREFU deserves special consideration. A two-level hierarchy still seems sufficient, but these *may not be the same two levels* that seem appropriate for the other tables. Standard spellings for test names, categories and units are influenced by medical and standards organizations like NCI, CDISC and LOINC; unit conversions are of course based on immutable principles. For these tables, **"industry"** and "study" levels seem appropriate. The LBPREFU information, however, is situational and routinely requires customization. For LBPREFU, **"enterprise"** and "study" levels seem appropriate. We therefore consider cross-industry governance of LBPREFU less suitable. See Step 4: Preferred standardized units, LBPREFU table, below, for further discussion.

Preferred units should nonetheless:

1. adhere to NCI/CDISC synonyms, like "mmol/L" rather than a non-standard synonym like "mcmol/mL", and
2. be consistent within a clinical database and set of reports or deliverables.

The authors seeded test name synonyms and submission values from the CDISC SDTM Terminology codelists for:

- Laboratory,
- Immunogenicity Specimen Assessments,
- Microbiology, and
- Microorganism test names.

Note that SDTM Terminology does not yet include a "Microorganism Test Code" codelist corresponding to the "Microorganism" codelist of test names, unlike the other three codelists, above. To compensate for the lack of a specific "Microorganism Test Code" codelist, we simply used the code for each Microorganism test name as the corresponding test code.

To illustrate, compare SDTM Terminology (2021-09-24) and our LBNAME entries for:

- Micro**organism** test "ABIOTROPHIA", which includes its codelist code "C86077" as test code, versus
- Micro**biology** test "Acinetobacter", with corresponding test code "ACINETOB".

Also note that CDISC submission values for Microorganism test names are uniformly "UPPER CASE", while the NCI preferred terms are uniformly "Sentence case", but otherwise match ignoring case. Synonyms are not case sensitive in Standardize Labs, and therefore we do not have these case-specific NCI preferred terms for Microorganism tests; we record only the uppercase CDISC submission values.

STEP 1: TEST NAME, LBNAME TABLE

Each organization, in their development programs that do not yet fully incorporate LOINC laboratory standards, will encounter novel test name terminology from a variety of sources. We worked closely with our lab data expert to map novel and sometimes "creative" original terms to standard terms and categories. In some cases, we had to extend the CDISC codelists, like C67154 "Laboratory Test Name".

Industry agreement on all terminology is not feasible. To accommodate customization without sacrificing transparency, we include the following variables in the LBNAME table:

- SPSYNO_TEST: Sponsor-Collected Synonym for Test Name?
- SPNAME_TEST: Sponsor Extension Standard Test Name?

The objective of these NULL | Y=yes indicators is to identify and separate sponsor contribution and customization. The Standardize Labs project should include uncontentious, unambiguous information in the core, open database. If sponsor proposals receive broad support, we can of course incorporate these into the core database.

The current database contains mostly synonyms and standard terms from industry guides and references, like those noted above. In addition, we include examples of sponsor-collected synonyms and sponsor-extensions, for the purpose of illustration. For example, collected test names may include common CDISC codes in test names. We therefore include synonyms like "CALCIUM (CA)", to map to CDISC lab test name **Calcium**, with CDISC lab test code **CA**. We flag this as a sponsor-collected synonym, with SPSYNO_TEST assigned "Y". The National Institutes of Health provides the PubChem database, a comprehensive resource for chemical information[14].

We also include **ADAMTS13 Activity** as an example of a sponsor-extension standard test name, with test code **ADAMTS13**. This lab test is not yet covered by CDISC SDTM Terminology, and at this time these "sponsor" term and code do not conflict with LOINC values[11]. A robust application should be able to reconcile sponsor extensions as industry standards expand.

Davis (2009) highlights that collected test synonyms can be misleading, especially when results are presented as ratios, fractions or percentages. In his example, a Basophils result with original units of "%" most likely measures the

ratio Basophils/Leukocytes. While this is an important tip, our lab standardization does not consider units when standardizing lab name terminology. "10^9/L" is a common preferred unit for analyzing and reporting Basophils results. A misleading original result of Basophils as "%" will drop out the app's query to convert "%" into "10^9/L". This conversion does not exist, which forces a review of the problematic data points.

In fact, we encountered this exact scenario, and following review we updated the test name from Basophils to Basophils/Leukocytes based on the original unit of "%" and confirmation from the lab.

Test name synonyms are forced to upper-case in the Standardize Labs database, since capitalization is irrelevant for mapping to the appropriate mixed-case standard test name. A constraint on the SQL table ensures this consistency. Additional constraints ensure that the NULL | Y variables contain only those expected values.

Terminology overrides, for instance a study term that takes precedent over an enterprise term, should not fragment synonym mappings. The current database includes the following synonyms for **Hemoglobin**, code **HGB**:

- H**A**EMOGLOBIN
- HEM**A**GLOBIN
- HEM**O**GLOBIN

The application should catch Study extensions that fragment "Hemoglobin" synonyms. For example, it should block or at least discourage mapping the following synonyms to different test name / code **Haemoglobin** / **HAEMGB**:

- HAEMOGLOBIN
- HEMOGLOBIN MONOMER

This is an error of fragmentation, since the synonym HAEMOGLOBIN is included in the initial set, and diverting it alone to a second set of terms effectively fragments results that should be analyzed together.

STEP 2: TEST CATEGORY, LBCAT TABLE

We initialize the lab categories based on CDISC's LOINC to LB Mappings, which the LOINC Working Group recommended for mapping "the most commonly submitted LOINCs in clinical research to the associated CDISC terminology". These mappings include "more than 1,400 LOINCs across the Chemistry, Hematology, Coagulation, Toxicology, Urinalysis, Serology, and Miscellaneous categories."[(12)] Lab categories are typically not case-sensitive, so Standardize Labs deals exclusively with upper-case synonyms and preferred categories.

Similar to the Davis (2009) discussion, above, correct lab categorization of results is essential for correct conversion and analysis of both quantitative and qualitative results. Blood and urine samples are commonly tested for some common analytes, although the results and interpretations will differ substantially. Conversions that Standardize Labs fails to complete, such as converting erythrocytes results in "/HPF" to preferred units of "10^12/L", will again force review of problematic data, and adjustment of recorded result.

Lab category standardization functionality could be extended easily to standardize lab sub-category, as well. Note that while category is *expected* in terms of CDISC compliance, sub-category is *permissible*.

STEP 3: RESULT UNITS, LBUNIT TABLE

Lab result units are of course more complicated than test names and categories. Synonyms like "10^3/mm3", "10^3/uL" and "10^9/L" must be standardized to a single spelling like "10^9/L". Standardizing unit spelling is similar to standardizing test name spellings, so the structure and purpose of the LBUNIT data set is analogous to that of LBNAME, as indicated in the workflow diagram, above.

The authors seeded unit synonyms and submission values from the CDISC SDTM Terminology codelists for:

- PK UNITS OF MEASURE,
- PK UNITS OF MEASURE – DOSE MG,
- PK UNITS OF MEASURE – DOSE UG,
- PK UNITS OF MEASURE – WEIGHT G,
- PK UNITS OF MEASURE – WEIGHT KG, and
- UNIT.

Lab results of course involve further layers of complexity due to the different scales of measurement, which the remaining workflow steps address.

STEP 4: PREFERRED STANDARDIZED UNITS, LBPREFU TABLE

The table LBPREFU is a declaration of preference, rather than a statement of fact. It is a declaration of preferred spelling and scale. Test names and categories, and preferred units should only include standard spellings, and no other synonyms. These preferences may depend on the intended audience: reviewers and decision makers.

For example, a panel of blood chemistry tests typically includes multiple electrolytes, like Bicarbonate, Chloride, Potassium and Sodium. The US conventional unit for these electrolytes is "mEq/L", while the international standard unit (SI) is "mmol/L". Although these are synonyms for measurements on the same scale, for these electrolytes, presenting a medical reviewer with results in "mEq/L" if s/he is accustomed to "mmol/L", or *vice versa*, introduces an unnecessary hurdle for review and interpretation. Those hurdles are greater when the differing units measure different scales, as for Cholesterol: SI units of "mmol/L" versus US conventional units of "mg/dL", with a conversion factor of 38.665 from SI to US conventional.

Selecting the appropriate scale may depend, as well, on medical considerations for a study population and expected outcomes of planned interventions. For example, if a study population are known to develop decreased or elevated cell counts, then the preferred reporting units for these results may not be the familiar "10^9/L", but rather a smaller or larger unit of measure.

Study overrides of enterprise preferred units are common. When organizing enterprise and study configuration, co-location of information eases the burden of configuration management, and heightens transparency. If standard test names are recorded as study configuration, preferred units for those same tests should be recorded as study configuration, as well.

These preferences are unlikely industry standards. While Standardize Labs does declare some preferences in the LBPREFU table, we consider these a starting point, only – a template for users. The authors consider these preferences ill-suited for an industry standard, and do not anticipate maintaining them as such.

STEP 5: CONVERSION FACTORS, LBCONV TABLE

Once the preceding terms are standardized and preferences declared, a query to the LBCONV table provides appropriate conversion factors to normalize lab results to a single scale, as required in FDA rule FDAB030, mentioned above.

As noted in the Introduction, compiling a reliable table of conversions proved challenging. Many references are available, but most are suitable only for manual look-up. After considering multiple options, we decided to programmatically consolidate ("scrape") conversions from the UnitsLab.com website. Challenges included unwinding the structure of their web front-end, identifying their sources of relevant information, interpreting the information as they recorded it, and recognizing and correcting some inconsistencies in those details. Altogether, this presented exceptional problems to solve, and these provided uniquely rewarding experiences both together as we assessed options, and designed, coordinated and qualified solutions, as well as individually as we each tackled our own responsibilities.

Wu and Wales[5] suggest separate reference tables for generic and test-specific conversions. We took a comparable but modified approach. Instead of separate reference tables, Standardize Labs has a single LBCONV reference table with a missing LBTEST for "generic" or test-independent conversions. Any conversion that relies on unique molecular properties, such as molar mass, is valid only for that test. If that unique property is common to both the original (collected) and standard (target) units, then the property effectively cancels out, and the conversion is again generic.

To illustrate these points, consider the following conversions:

- For electrolytes like Bicarbonate and Calcium, the conversion from mmol/L to umol/L is generic, the molar aspect of both original and standard units is equivalent. The conversion is therefore uniformly a factor of 1000. LBTEST is blank for this conversion in LBCONV.

lbtest	lborresu	lbstresu	factor
All	mmol ⊗	umol ⊗	All
	mmol/L	umol/L	1000

- However, for these same electrolytes, conversion from mmol/L to mEq/L is not generic. Moles measure a specific number of molecules, regardless of molecular structure. Milliequivalents is instead related to the unique molecular structure of each electrolyte. The conversion for Potassium is 1, and for Calcium 2, hinting at their different molecular structures.

lbtest	lborresu	lbstresu	factor
All	mmol	mEq	All
Bicarbonate	mmol/L	mEq/L	1
Calcium	mmol/L	mEq/L	2
Chloride	mmol/L	mEq/L	1
Magnesium	mmol/L	mEq/L	2

- Conversions from mmol/L to mg/L are also test-specific, since the conversions from molecular-count (moles) to weight (grams) is again specific to each analyte.

lbtest	lborresu	lbstresu	factor
All	mmol	mg/L	All
Acetoacetic Acid	mmol/L	mg/L	102.088
Acetone	mmol/L	mg/L	58.08
Beta-Hydroxybutyrate	mmol/L	mg/L	104.11
Bicarbonate	mmol/L	mg/L	61.0168
Bilirubin	mmol/L	mg/L	584.67

The Standardize Labs application produces a report of unsuccessful steps, whether attributable to an unrecognized synonym for name, category or units term, or an unrecognized conversion.

For remaining lab results, it delivers standardized spellings and numeric results. Again, qualitative records pass through unaltered – as do records for quantitative results that the app cannot map or convert.

As noted in FDA Validation Rule FDAB031, any character result that represents a numeric (quantitative) result should be equal to and equally precise as the corresponding numeric result, and Standardize Labs should ensure this consistency, as well.

TODO LIST

We publish both our SAS and R software in the Standardize Labs Gitlab repository:

- https://gitlab.com/DanteDT/standardize-labs/-/tree/main/man

The SAS components are for reference. Constructs and capabilities of the R ecosystem, such as package innovation and shiny interfaces, make this path far more suitable for an open-source application. As we implement the R package and shiny application, we improve the design of the app, which will continue to diverge from the original SAS implementation. We do not intend to update the SAS implementation as the R application advances.

While contributors should drive future functionality and improvements, there are a few enhancements that we consider high priority.

1. SDTM Terminology is versioned and released quarterly. Standardize Labs should support versioned CDISC terminology, as well as versioned sponsor extensions. When combining sponsor extensions with SDTM Terminology, including when updating the version of core SDTM Terminology, the application should reconcile any conflicts.
2. Incorporate the LOINC mappings discussed above[12].
3. Standardize terminology for qualitative results, such as for urinalysis results.
4. Confirm and protect expected behavior of the application with robust, automated tests.

CONCLUSION

Basic capabilities that vary between sponsors are strong candidates for industry collaboration, especially when those variations are arbitrary and can impact healthcare decisions. The often cumbersome and iterative standardization of

lab terminology and normalization of analysis units seems perfect candidates, a "low-hanging fruit" in a world that adores low-hanging fruit which bring tangible value for negligible investment.

We organized the SAS components as suggested by Jablonski (2021)[15], and the R package as described by Zhu et al. (2020)[16]. Join us and contribute: learn, teach and grow, and help your peers and organizations do the same.

REFERENCES

The authors accessed web site information in February, 2022, unless otherwise noted.

(1) https://loinc.org/principles/, LOINC Documentation, *Mission, Vision, and Principles for Open Terminology Development*.
(2) FDA Data Standards Catalog v8.0 (16. February, 2022), accessed February, 2022.
(3) FDA Study Data Technical Conformance Guide v4.8.1 (October, 2021), accessed February, 2022.
(4) Davis, Mat, *Using CDISC Lab Terminology and Determining Standard Units*, DM 06, PharmaSUG 2009.
(5) Wu, Charley; Wales, Dona-Lyn, *Laboratory Data Standardization with SAS*, AD 10, PharmaSUG 2017.
(6) SDTM Terminology, CDISC Library Archives, https://www.cdisc.org/members-only/cdisc-library-archives. Available exclusively to CDISC member organizations.
(7) Laposata, Michael, SI Unit Conversion Guide, NEJM Books, Boston, MA, USA. 1992.
(8) Kratz, Alexander; Lewandrowski, Kent, Normal Reference Laboratory Values, New England Journal of Medicine, v. 339, n. 15, October 8, 1998, pp. 1063-1072.
(9) Kratz, Alexander; Ferraro, Maryjane; Sluss, Patrick; Lewandrowski, Kent, Laboratory Reference Values, New England Journal of Medicine, v. 351, n. 15, October 7, 2004, pp. 1548-1563.
(10) FAIR data, Wikipedia, https://en.wikipedia.org/wiki/FAIR_data.
(11) LOINC database is available for free download, https://loinc.org/downloads/.
(12) CDISC, *ReadMe for LOINC to LB Mapping*, https://www.cdisc.org/standards/terminology/controlled-terminology, August 08, 2020.
(13) FDA, *The Validator Rules v1.5*, March 2021, https://www.fda.gov/industry/fda-data-standards-advisory-board/study-data-standards-resources.
(14) National Institutes of Health, National Center for Biotechnology Information, National Library of Medicine, *PubChem*, https://pubchem.ncbi.nlm.nih.gov/.
(15) Jabłonski, Bartosz, *My First SAS Package - a How To*, Paper 1079-2021, SAS Global Forum 2021.
(16) Zhu, Yalin; Jajoo, Rinki; Bai, Clare; Nepal, Sarad; Woodie, Daniel; Anderson, Keaven; Zhang, Yilong, *R Package Oriented Software Development Life Cycle in Regulated Clinical Trial Environments*, TT12, PHUSE Connect 2020.

ACKNOWLEDGMENTS

Thanks to Omeros Corporation for supporting this work, and permitting the resulting open-source project.

CONTACT INFORMATION

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

Sunil Kumar Pusarla	https://www.linkedin.com/in/sunil-kumar-pusarla-6807a016/
Kevin Viel	https://www.linkedin.com/in/kevin-viel-abb58123/
Dante Di Tommaso	https://www.linkedin.com/in/dante-di-tommaso-2097891/