

The automation of unit verification, units conversion, and converted results verification through the UCUM, LOINC, and neural machine translation

Sam Tomioka
Data Science
Sunovion Pharmaceuticals

1. Motivation
2. What is UCUM?
3. Method
4. Use Cases
5. Summary

Background

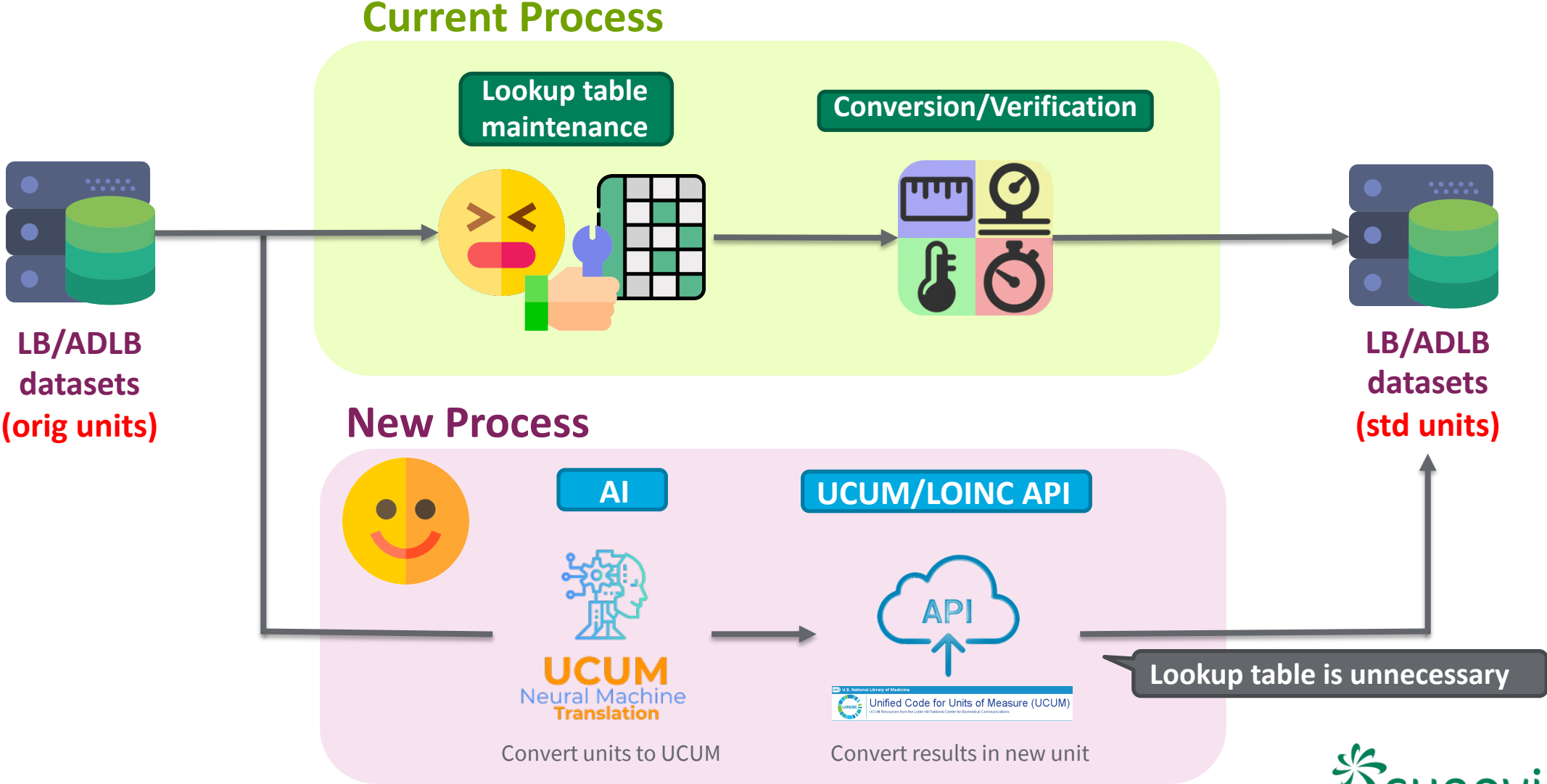
Results are reported in **original unit**, **SI unit**, and **US conventional** unit by the lab vendors

Units must be verified

LBTESTCD	LBORRES	LBORRESU	LBSTRES	LBSTRESU	LBCVRES	LBCVRESU	PARAM	AVAL
ALB	49	g/L	49	g/L	4.9	g/dL	Albumin (g/L)	49
BILI	6	umol/L	6	umol/L	0.4	mg/dL	Bilirubin (umol/L)	6

Results must be verified

Motivation



Benefits of UCUM approach

We can verify/convert even when the units are not found in the lookup table

We can leverage **LOINC** for conversion involving molar mass

Vendor provided result was incorrect! Molar mass of **Phosphorus** was used instead of **Phosphate**.

UCUM-LOINC verification



LBLOINC	LBTEST	LBTESTCD	LBORRESU → LBSTRESU
35221-1	Phosphate	PHOS	$\frac{3.0\text{mg}}{\text{dL}} = \frac{0.03\text{g}}{\text{L}} * \frac{\text{mmol}}{0.030974\text{g}} = 0.969 \text{ mmol/L}$
35221-1	Phosphate	PHOS	$\frac{3.0\text{mg}}{\text{dL}} = \frac{0.03\text{g}}{\text{L}} * \frac{\text{mmol}}{0.094971381\text{g}} = 0.316 \text{ mmol/L}$

Phosphorus $\text{MM}^{-1} * 1/1000$

Phosphate $\text{MM}^{-1} * 1/1000$



Ask GPT-2 1.5B Model about UCUM

UCUM can be used to express the same concept in CDISC data model

BI Markets Insider

OpenAI text-generating tool GPT2 won't be released for fear of misuse

From that headline alone, GPT2 was able to generate quotes from UK Labour leader Jeremy Corbyn as well as a fictional spokesman for Prime ...
Feb 15, 2019

The use of Unified Code for ...
in all data models, but is

The Verge

OpenAI unveils multitalented AI that writes, translates, and slanders

OpenAI's new algorithm, named GPT-2, is one of the most exciting examples yet. It excels at a task known as language modeling, which tests a ...

In-Depth · Feb 14, 2019

a conversion tool when converting
om one system to another.

S ScienceAlert

Scientists Developed an AI So Advanced They Say It's Too Dangerous to Release

OpenAI, a research non-profit based in San Francisco, says its "chameleon-like" language prediction system, called GPT-2, will only ever see a ...

Feb 19, 2019



What is UCUM?

Scope	All units of measure being contemporarily used in international science, engineering, and business.
Purpose	To facilitate unambiguous semantically defined and computable communication of physical quantities with units between computer systems.
History	Inspired by ISO 2955-1983, ANSI X3.50-1986, and HL7 ISO+.
Benefits	Allows automatic conversion No need for lookup tables to account for all possible unit combinations for each test
Adoption	Healthcare Data Standards --- DICOM, HL7, FHIR, OMOP, Regulatory Agencies --- FDA, Europe Commission, PMDA, ICH ICSR E2B (R3) submissions

Specifications: <http://unitsofmeasure.org/ucum.html>

UCUM Grammar – Character Set, Syntax

- Character Set

ASCII characters in the range of 33–126 (excluding “, (,), +, -, ., /, =, [,], {, })

		Examples
() in HL7 ISO+ unit replaced by [] or enclose one full unit	[]	mm(Hg) → mm[Hg] (ka_u) – if not enclosed kilo – a_u
annotation	{ }	{cells}/[HPF]

Terminal unit symbols

			Examples
algebraic	Multiplication	.	kg.m
	Division	/	m/s
	exponents	+, -,	m2, 1 cm3=10*-6m3
	nested	()	mg/(12.h)

UCUM Grammar - Terminal Symbols

			Examples
Prefixes (20)			Kilo – k , deci – d , milli – m , ...
Unit Atom	Base Units (7)		length – m , time – s , mass – g , ...
	Derived Units	Dimensionless	10ⁿ , % , ppm , ...
		SI	mol , Hz , Pa , °C , ...
		ISO 1000, ISO 2955 and ANSI X3.50	Liter - L , week - wk , ...
		⋮	⋮

Any derived units can be derived from the base – $L = \frac{1}{1000} m^3$, $Hz = s^{-1}$



Examples

CDISC Code	CDISC Submission Value	UCUM
C98759	10^6/hpf	10*6/[HPF]
C117967	BU/mL	[beth'U]/mL
C135518	cmH2O*s/mL	cm[H2O].s/mL
C67473	mEq/dL	meq/dL
C49670	mmHg	mm[Hg]

Proposed Method



APIs (RESTful, Python)

NIH U.S. National Library of Medicine



Unified Code for Units of Measure (UCUM)
UCUM Resources from the Lister Hill National Center for Biomedical Communications

<https://ucum.nlm.nih.gov/>

XML4Pharma

http://xml4pharmaserver.com/WebServices/UCUM_webservices.html

	US NLM	XML4Pharma
	Ver 1	Ver 2
Unit Conversion (no molar conversion)	✓	
Unit Validation	✓	✓
Base Unit Conversion	✓	✓
Unit Conversion with LOINC		✓
Unit Conversion with option to supply MM		✓
Derive UCUM from LBORRESU and LBSTRESU		
Consider valence of ion		✓

pyucum 0.1.3

<https://pypi.org/project/pyucum/>

pyucum
Ver 0.1.3
✓
✓
✓
✓
✓



Step 1: Convert LBORRESU and LBSTRESU into UCUM

```
from pyucum import *

patterns = [("%", "%25"),
            ("\A[xX]?10[^E]", "10*"),
            ("IU", "%5BIU%5D"),
            ("\Anan", ""),
            ("\ANONE", ""),
            ("\A[rR][Aa][Tt][Ii][Oo]", ""),
            ("\ApH", ""),
            ("Eq[l]?", "eq"),
            ("\ATI/L", "TR/L"), #update T to TR
            ("\AGI/L", "GA/L"), #update G to GA
            ("V/V", "L/L"),
            ("[a-z]{0,4}/HPF", "/%5BHPF%5D"),
            ("[a-z]{0,4}/LPF", "/%5BLPF%5D"),
            ("fraction of 1", "1"),
            ("sec", "s"),
            ("1.73m2", "%7B1.73_m2%7D"),
            ("\AG/L", "GA/L")
]

dfconverted, ucumlist=orresu2ucum(df1,patterns)
```

`orresu2ucum(df_,patterns)`

`df_` is a data frame which contains LBORRESU and LBSTRESU columns

`patterns` is a list of regular expression and replacing value.
[(‘regexp to identify LBORRESU or LBSTRESU to be updated’,
‘replacing value’),...]

Because the values are passed through HTML, percent encoding [1] should be used for the replacing value

"[" by "%5B"

"]" by "%5D"

"{" by "%7B"

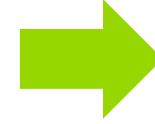
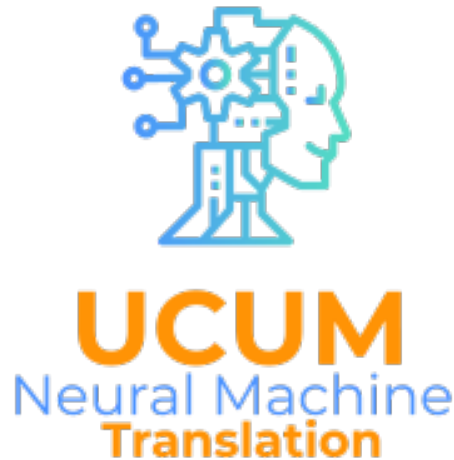
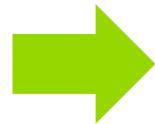
"}" by "%7D"

Improved Step 1: Neural Machine Translation

Who wants to write regular expressions to convert units to UCUM representation?

CDISC C112247

day*g/mL/kg



UCUM

d.g/mL/kg

Demo of UCUM NMT

Character-base neural translation for UCUM

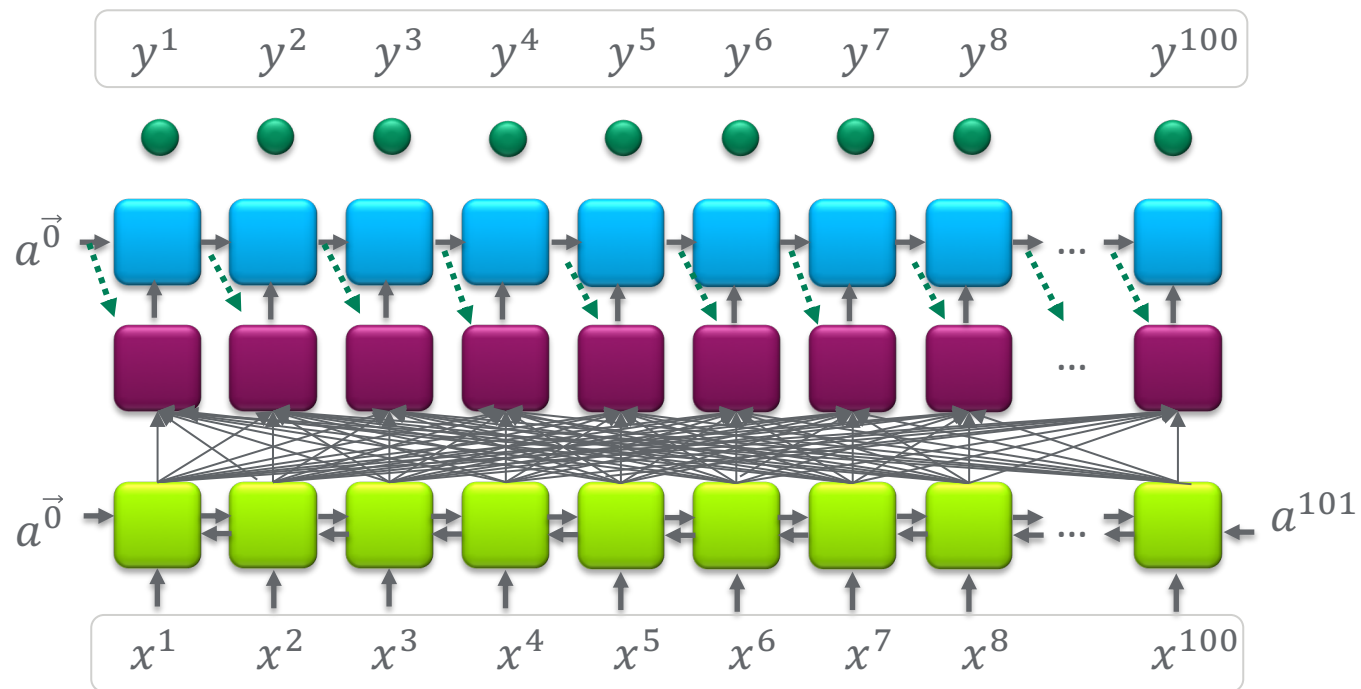
Output

d.g/mL/kg



Input

day*g/mL/kg



47,179 records of training data are used

■ bi-LSTM ■ Attention ■ LSTM ● Softmax



Step2: Validate UCUM Expression using UCUM API

```
from pyucum import *  
  
url='http://xml4pharmaserver.com:8080/UCUMService2/rest'  
#url=https://ucum.nlm.nih.gov/ucum-service/v1/  
ucumVerify(ucumlist, url)
```

Output

```
['g/dL = true',  
 'mg/dL = true',  
 'ng/mL = true',  
 '10*3/uL = true',  
 '%25 = true',  
 '10*6/uL = true',  
 '10*3/mm3 = true',  
 'meq/L = true',  
 'mL/min = true',  
 'ng/L = true',  
 'ng/dL = true',  
 'u%5BIU%5D/mL = true',
```

After updating the LBORRESU and LBSTRESU to UCUM using `orresu2ucum(df_,patterns)`, verify UCUM units using `ucumVerify(ucumlist,url)`

`ucumlist` is a list object which consists of units converted by `orresu2ucum`.

`url` is the base RESTful web service URL.

Demo: <https://ucum.nlm.nih.gov/ucum-lhc/demo.html>



Step 3: Convert results in LBORRESU to results in LBSTRESU

```
from pyucum import *  
  
findings,full,response=convert_unit(nodupdf, url, patterns,loinonly=0)  
  
findings[(findings['fromucum'].notnull())]
```

Outputs

1. Data frame with records where LBSTRESN not equal to converted results
2. Data frame with all results
3. List of API responses

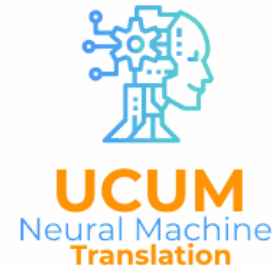
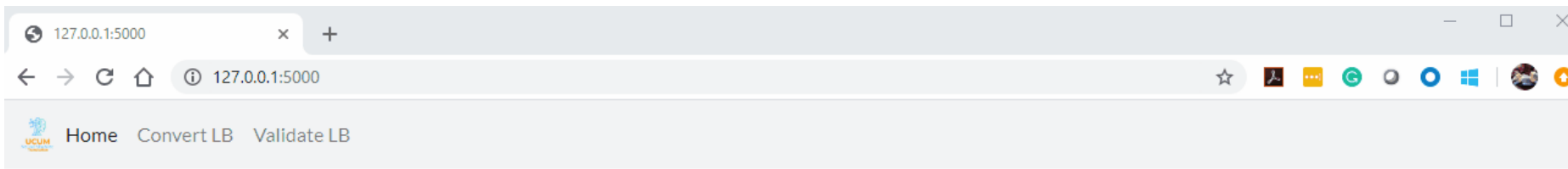
`convert_unit(df, url, patterns, loinonly=0)`

`df` is an input data frame containing LBORRES, LBSTRESN, LBORRESU, LBSTRESU, and LBLOINC

`url` is the base RESTful web service URL.

`patterns` is a list of regular expression.
[(‘regexp to identify LBORRESU or LBSTRESU to be updated’, ‘replacing value’),...]

`loinonly` (0=False, 1=True) if 1 (True) then conversion is done when LOINC is not missing.



1. Select Model

- ☒ Character Attention
- ☐ Unit Embedding Attention
- ☐ Character Attention w/ sub-c
- ☐ 3-1 Chars Attention

2. Enter Unit

Convert

3. Upload File

Choose File No file chosen

Convert

1. This small app can generate mapping file for units in SDTM to UCUM expression automatically
2. This app accepts SDTM and the mapping file from above to output converted results.

© 2019 S.Tomioka

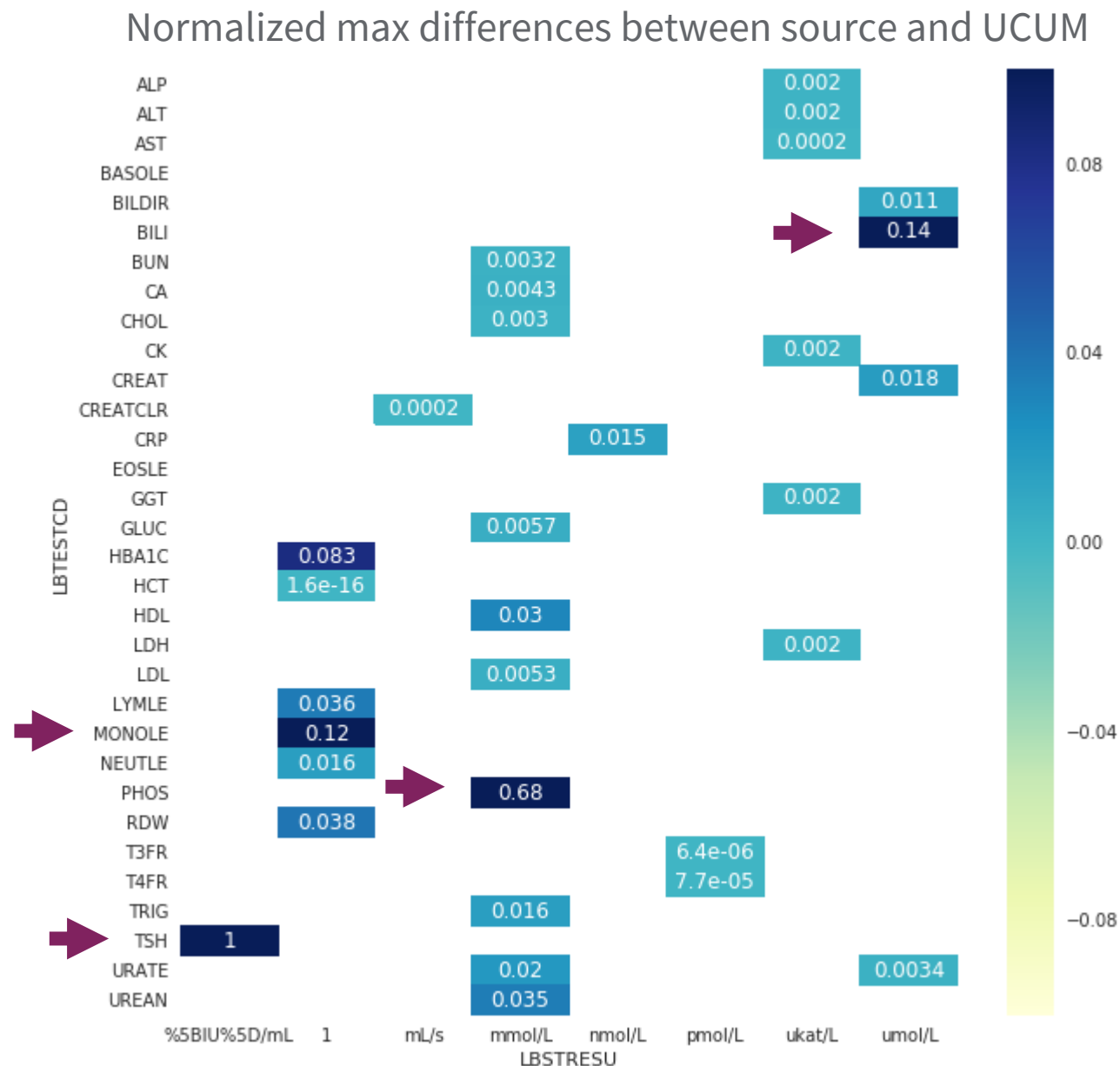
Use Case

17565 lab records from all studies populated in our Safety Dashboard were reviewed.

TSH and PHOS have issues

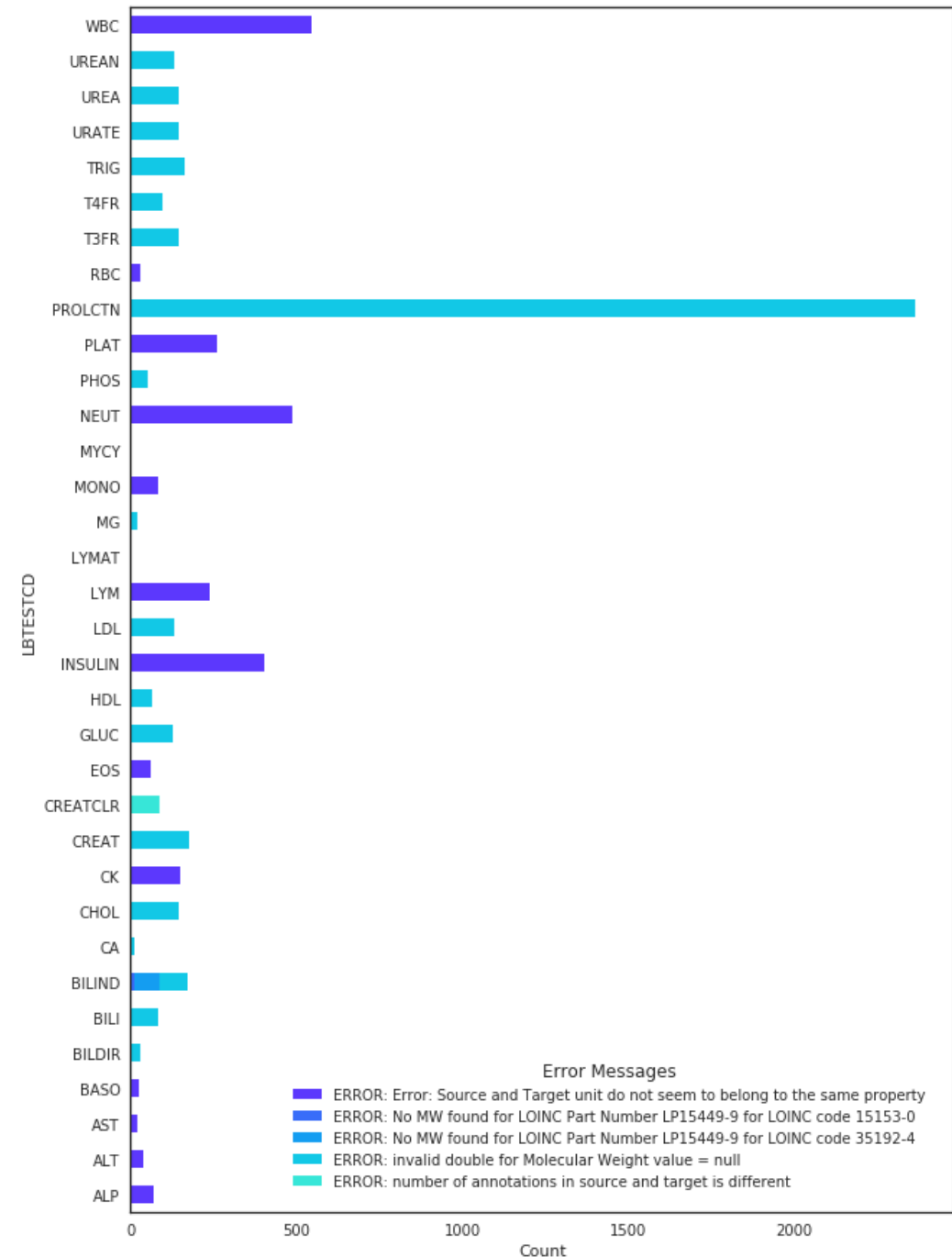
Maximum difference of BILI and MONOLE are high. MONOLE differences were caused by rounding. On the other hand BILI had conversion issue.

0.4mg/dL was converted as **6.0 µmol/L** by the vendor. Per LOINC 1975-2, BILI MM is 584.6621g/mol so $0.4\text{mg/dL} = 0.04\text{g/L} \times 0.005846621\text{g/}\mu\text{mol} = \textbf{6.84156 }\mu\text{mol/L}$.



Use Case (continued)

- The most frequent error was **ERROR: invalid double for Molecular Weight value = null** due to missing MM or LOINC when the conversion requires MM.
- There were 2 kinds of **No MW found for LOINC Part Number LPxxxx for LOINC code yyyy.**
- Creatinine Clearance (mL/min/1.73 m²) conversion failed with **ERROR: number of annotations in source and target is different.** The surface area 1.73 m² was added as annotation for the source but the target did not include the same annotation. This will be solved with <http://xml4pharmaserver.com:8080/UCUMService2/rest/ucumtransform/121/from/mL/min/%7B1.73m2%7D/to/mL/s/%7B1.73m2%7D>
- Several conversions failed with **ERROR: unexpected result: Error: Source and Target unit do not seem to belong to the same property.** (next page)



Use Case (continued)

ERROR: unexpected result: Error: Source and Target unit do not seem to belong to the same property

Index	LBTESTCD	LBORRESU	LBSTRESU
5652	INSULIN	uU/mL	pmol/L
6586	BASO	G/L	10*9/L
6588	EOS	G/L	10*9/L
6591	LYM	G/L	10*9/L
6594	MONO	G/L	10*9/L
6597	NEUT	G/L	10*9/L
6600	PLAT	G/L	10*9/L
6606	RBC	T/L	10*12/L
6609	WBC	G/L	10*9/L
7026	LYMAT	G/L	10*9/L
7828	MYCY	G/L	10*9/L
11395	ALP	%5BIU%5D/L	U/L
11400	ALT	%5BIU%5D/L	U/L
11403	AST	%5BIU%5D/L	U/L
11416	CK	%5BIU%5D/L	U/L

UCUM seems to have issue with G (giga) and T (Tera) even after the second letter was added (GA, TR), because these are prefixes and UCUM might be expecting that these are attached to either base unit or unit atom. G can be expressed as 10^9 and $T=10^{12}$ in UCUM.

The ucum-essence.xml seems to suggest that %5BIU%5D i.e. [IU] and U are equivalent but the conversion was not successful.

Use Case (continued)

PHOS and TSH conversion issues identified

Output from `convert_unit(dfin, url, patterns, loinonly=0)`

	LBTESTCD	LBTEST	LBORRES	LBORRESU	LBSTRESU	LBSTRESN	LBLOINC	checklist	fromucum	re:
15154	TSH	Thyrotropin	1.01	m%5BIU%5D/L	%5BIU%5D/mL	1.01	http://xml4pharmaserver.com:8080/UCUMService2/...		1.010000e-06	
15155	TSH	Thyrotropin	1.85	m%5BIU%5D/L	%5BIU%5D/mL	1.85	http://xml4pharmaserver.com:8080/UCUMService2/...		1.850000e-06	
15165	TSH	Thyrotropin	0.98	m%5BIU%5D/L	%5BIU%5D/mL	0.98	h...maserver.com:8080/UCUMService2/...		9.800000e-07	
15166	TSH	Thyrotropin	1.69	m%5BIU%5D/L	%5BIU%5D/mL	1.69	h...maserver.com:8080/UCUMService2/...		1.690000e-06	
15168	TSH	Thyrotropin	0.85	m%5BIU%5D/L	%5BIU%5D/mL	0.85	http://xml4pharmaserver.com:8080/UCUMService2/...		8.500000e-07	

m[IU]/L to [IU]/mL

$$\frac{\text{mIU}}{\text{L}} = \frac{\text{IU}}{1000\text{L}} = \frac{\text{IU}}{10^6\text{mL}}$$

1.01/10**6

1.01e-06



Summary

UCUM-based approach

- eliminates the need for lookup table
- accounts for charge of the molecule (Mg++ so 1eq=2 mol)
- accounts for the molecular weights of the analyte into the conversion between molar and mass concentrations to facilitate the conversion of the lab results, verification of the standardized lab results and LOINC code provided by the vendors
- may not work when prefixes such as G and T are used alone in the original units

Pyucum allows use these API without knowing the API syntax

Let's use UCUM based approach!

Thank you!



Sam Tomioka



GitHub

Source code for pyucum



Resources

A Unified Code for Units of Measure Software Library from the National Library of Medicine

<https://ucum.nlm.nih.gov/ucum-lhc/index.html>

The Unified Code for Units of Measure Specifications

<https://unitsofmeasure.org/ucum.html>

Jozef Aerts (2019) Instructions for testing the RESTful web service for converting "molar" concentrations to "mass" concentrations and vice versa using UCUM units and LOINC codes.

http://www.xml4pharma.com/UCUM/Instructions_for_testing_the_RESTful_web_service_for_molar_mass_unit_conversions.pdf

My code

<\\reese\\shared\\CDR Sam Team\\Python Hackathon\\docs\\ucum-master>

Here is the latest notebook in html.

\\reese\\shared\\CDR Sam Team\\Python Hackathon\\docs\\ucum-master\\docs\\ucum_20190526-test-large-sample.html

Exercise

1. Install pyucum
2. Convert units in LBORRESU and LBSTRESU to UCUM and Verify
3. Convert results in LBORRESU to the results in LBSTRESU using API