UCUM for CDISC: A Tutorial

Author: Jozef Aerts, Professor in Medical Informatics and CEO of XML4Pharma
Date last update: 2017-07-24

Introduction

This tutorial has been developed for the team members of the CDISC SDS (submission data standards) and CT (controlled terminology) teams. These teams have several times turned it down to allow UCUM notation in CDISC submissions (SDTM, SEND, ADaM). Arguments for this "prohibition" have often shown to be incorrect, mostly due to insufficient or lack of understanding of UCUM. So, within these groups, a lot of "myths" float around…
This tutorial can of course also be used by anyone else interested in the use of UCUM notation for units.

This tutorial explains what UCUM is, how it works, and how it can easily be applied to SDTM (human studies) and SEND (pre-clinical studies), and ADaM (analysis results). After doing this tutorial, users will understand why UCUM is important for CDISC, and why there is no need for very long codelists with units in CDISC controlled terminology.

What is UCUM?

UCUM stands for "Unified Code for Units of Measure" (www.unitsofmeasure.org). UCUM has been developed by the Regenstrief Institute to bring order in the chaos of units and especially their notation. For example, how do you write "foot" (a length unit). There are many different ways. In CDISC-CT, the abbreviation "ft" is used. However, the explanation that 1 foot is equal to 30.84 centimeters is only provided as narrative text in CDISC-CT, and cannot be used by machines (not machine-executable).
The UCUM notation is "[ft_i]" (we will explain this further in detail) allows machines and computer programs to ***automatically*** convert measurements in "foot" into "miles", "kilometer", "micrometer", "inches" – essentially in any other unit that represents a length.

UCUM notation is used a lot in healthcare: it is mandated to be used in the US "Meaningful Use", in CCD and CDA (electronic health records - EHRs), and is mandated to be used in any national EHR system I do know (I know quite a few…). The use of UCUM is also "highly recommended" in the new FHIR standard for exchange of medical information.
But UCUM notation is not only used in healthcare, it is used in many industries, like engineering, aviation, aerospace, and business. It is also used a lot in science, because it provides an ideal solution to resolve the "jungle of units". Essentially, the only organization that so far essentially has been "banning" the use of UCUM notation is … CDISC.

Principles of UCUM

UCUM is **NOT** a codelist! It is a **SYSTEM**!

CDISC only develops codelists – just "lists" of "things" without any connections to other "things". This may sound strange in the time of "semantic networks", but it has its history, as the CDISC codelists come from the paper world with checkboxes to be checked.
UCUM however is not at all a list, although it also has lists of "special" units.

UCUM base units and prefixes

UCUM defines seven base units: the "gram", the "meter", the "second", the "radian", the "Kelvin", the "Coulomb" and the "candela". Also the "mole"[1] is often used as a base unit (see later).

Then, UCUM defines prefixes, like "kilo" ("k"), "milli" ("m"), "micro" ("u"), "giga" ("G"), etc..
One can also call them "standardized multipliers"
There are 24 of them.
Any prefix can be combined with any base unit.
Simple examples are:

- "kg" ("kilogram")
- "Gm" ("Gigameter")
- "mC" ("milli-coulomb")

Remark that "liter" is not a base unit – also see later.

It is very important to understand how these combinations of "prefix" and "base unit" work. They are combinations, not lists. So, you will not find "kg" in any UCUM list, as UCUM is not a list, it is a system. This makes it extremely versatile, as it does not require to define lists with an almost infinite number of "units", which is what CDISC is doing.

UCUM units

Secondly, UCUM defines a list of units. These are called "unit atoms". There are 287 of them[2]. Important is that any of these units can be broken down into one or a combination of the seven base units. None of these units has a prefix. It is just the "naked" unit. Examples are:

- "W" ("Watt")
- "Ohm"
- "min" ("minute")
- "a" ("year")[3]

So, you won't find "kW" (kilowatt) in any "UCUM list", as most units can be combined with any of the prefixes. There are however some (well described) rules when prefixes may be used and when not. For example, one of the rules state that "only metric unit atoms may be combined with a prefix". So, it is allowed to write "kW" (kilowatt), but it is not allowed to have "kilo-feet", as "foot" is not a metric unit.

This brings us to the machine-readable specification. The Regenstrief Institute does not publish its standards as Excel files or so as CDISC does, it publishes them as machine-readable XML files. For UCUM, this is the "ucum-essence.xml" file.

For example, for "liter", it contains the entry:

[1] The number of moles is nothing else than the number of molecules divided by the Avogadro number (6.023 10^{23}). So essentially, "mole" is dimensionless.
[2] Remark that the CDISC codelist for „unit" (NCI code C71620) currently has almost 700 hundred entries, growing at each new release, as it is just a "list", without any system.
[3] „a" comes from „annum" (Latin)

```
<unit Code="l" CODE="L" isMetric="yes" class="iso1000">
   <name>liter</name>
   <printSymbol>l</printSymbol>
   <property>volume</property>
   <value Unit="dm3" UNIT="DM3" value="1">1</value>
</unit>
```

stating that a liter (notation "l") is a volume, and equals 1 dm3 (one cube decimeter). "Decimeter" ("dm") itself is a combination of "deci" and "meter":

```
<prefix Code="d" CODE="D">
   <name>deci</name>
   <printSymbol>d</printSymbol>
   <value value="1e-1">1 &#215; 10<sup>-1</sup>
   </value>
</prefix>
```

In the definition of "d" ("deci") it is stated (in a machine-readable way!) that "deci" means "1e-1" (or "0.1"). Using this information a simple computer program (there are many of them) can easily reduce "liter" to the base unit "m" and find out that 1 liter equals 0.001 m^3.

It is now time to bust one of the myths CDISC people have about UCUM. They say that UCUM cannot be used as "l" (liter) is not used in the USA. Instead, Americans use "L".
This has been recognized by UCUM, with the result that we find a second entry for "liter" in the "ucum-essence.xml" file:

```
<unit Code="L" isMetric="yes" class="iso1000">
   <name>liter</name>
   <printSymbol>L</printSymbol>
   <property>volume</property>
   <value Unit="l" value="1">1</value>
</unit>
```

It defines the unit "L" (American way of writing) and states (in a machine-readable way) that 1L = 1l. This can also be read by a computer program, and through "chaining" allows to find out that 1L = 0.001 m^3.

Let us now have a look at some other, less classic units. According to a formal statement of the CDISC SDS team lead, "*UCUM expressions, in order to support computability, represent familiar units in unfamiliar ways, with curly brackets and other symbols. This is off-putting to some users*".

Square brackets

So, let me explain how the "square brackets" work for the "off-putted users". I promise: it is easy.

The whole world (and especially science) is using SI units as much as possible. But we do still have some units that are "local", so not universally used. For example, American physicians will probably measure a patient's height in "inches" and the weight in "pounds".
How do you write "inches" and "pounds"? Each physician might use a different notation, as there is no standard for it. UCUM has standardized the notation for "inches" and "pounds". The standardized UCUM notations are:

- "[in_i]" for "inch"
- [lb_av]" for "pound"

The square brackets are indicating that the unit is not a "universal" unit, but a "local" unit. Other examples are:

- "[oz_av]" for "ounce"
- "[stone_av]" for "British stone"
- "[sc_ap]" for "scruple" (used in pharmacy)

So, instead of "promoting" one of the many notations for "foot", "pound" and so on to the official one, UCUM selected to define a completely new notation (therefore the statement "unfamiliar" by some) which is machine-readable and which allows to break down such units to base units.

***Exercise**: Time for a first exercise! Look up "scruple" in the ucum-essence.xml file and look up what the UCUM notation is. Then try to break down "scruple" to the base units (probably to "gram" as it is a weight). This will require some chaining, i.e. using the results of one (but simple) calculation as an input for the next calculation.*
Beware! "[gr]" does not mean "gram"! It is in square brackets, so it also is "local" unit.

Difficult? My (undergraduate) students solve this in less than 10 minutes.

One of the arguments of CDISC people is that this kind of units requires programming to work with them. Well, ladies and gentleman, the programming has already been done: there is a RESTful web service for such calculations. Just try[4]:

http://www.xml4pharmaserver.com:8080/CDISCCTService/rest/ucumtransform/from/%5Bsc_ap%5D/to/g

Now, we don't expect researchers to start writing "[in_i]" on their (paper) forms. When using EDC (electronic data capture), the system can display "inch" (the display "name" in the ucum-essence.xml) and keep "[in_i]" in the database. When (still) using paper with a "free text" unit written (no preprinted unit or checkbox), someone (a person) will still need to translated what the researcher has written as "unit" to a "standardized" unit anyway, so why not immediately use UCUM for the result unit?

When using EHRs, the units will come in UCUM notation anyway, as the use of UCUM is mandated in almost every EHR system, but also mandated by "Meaningful Use" in the USA and in almost every national EHR system (such as the Austrian "ELGA").

***Exercise**: Look up the (weight) unit "scruple" in the CDISC controlled terminology. Can you find it? If you find it there, using the entry, can you calculate the conversion factor to "gram"? Could a software program do so using the entry in the CDISC controlled terminology?*
Now do the same for "grain" (also a weight unit). Can you calculate the conversion to "gram" in an automated way using the CDISC controlled terminology? Using UCUM with ucum-essence.xml?

Square brackets are also used in UCUM for "constants". For example, one finds in the ucum-essence.xml:

[4] Remark that in the HTTP request, the character "[" is replaced by its entity "%5B" and the character "]" by its entity "%5D".

```
<unit Code="[pi]" CODE="[PI]" isMetric="no" class="dimless">
   <name>the number pi</name>
   <printSymbol>&#960;</printSymbol>
   <property>number</property>
   <value Unit="1" UNIT="1"
          value="3.14159265358979323846264338327950288419716939937510582097494459 23">&#960;</value>
</unit>
```

Other examples are:

- [h] Planck constant
- [eps_0] permittivity of vacuum

Exercise: look up the value of "pi" allowing you to converts "degrees" into "radians" in the CDISC controlled terminology.

Further use of square brackets

CDISC typically messes up the unit with "what is measured". For example, it defines the unit "mmHg" (millimeter of mercury column) as a unit for "pressure" (usually a blood pressure". Because the CDISC "UNIT" list is a "list", the meaning of "m" (prefix), "m" (meter) and "Hg" (mercury) is completely lost. CDISC also defines the unit "cm H2O" (centimeter of water column). Remark that in "mmHg" there is no blank, whereas in "cm H2O" there is a blank! Again, the CDISC "UNIT" list is a "list" without any system.
In UCUM, as a system, this is done slightly different, but in a much more consistent way. The same units in UCUM notations are:

- mm[Hg]
- m[H2O]

for the former, there are essentially three parts: the prefix "m" (milli), the base unit "m" (meter) and "what is measured" (mercury column) – in square brackets.
For the latter, there is no prefix, but only a base unit, and "what is measured" (water column)

Exercise: look up "mmHg" and "cm H2O" in the CDISC controlled terminology and calculate the conversion factor between them solely using the information in the CDISC file.

In the ucum-essence.xml, you will not find "mm[Hg]", but you will find "m[Hg]". The reason is that the you can combine the latter with any prefix, including "m" of "milli".

Conversions

Now, let us reduce "mm[Hg] to the base units:

- we first split off the prefix "m" and find out that it means "milli" or "0.001".
- in the ucum-essence.xml file, for "m[Hg]" we find:

```
<unit Code="m[Hg]" CODE="M[HG]" isMetric="yes" class="clinical">
   <name>meter of mercury column</name>
   <printSymbol>m Hg</printSymbol>
   <property>pressure</property>
   <value Unit="kPa" UNIT="KPAL" value="133.3220">133.3220</value>
</unit>
```

stating that 1 “m[Hg]” is 133.3220 “kPa”. The latter is again a combination of a prefix “k” (kilo – 1000) and the unit “Pa”

- So let us look up “Pa”:

```
<unit Code="Pa" CODE="PAL" isMetric="yes" class="si">
    <name>Pascal</name>
    <printSymbol>Pa</printSymbol>
    <property>pressure</property>
    <value Unit="N/m2" UNIT="N/M2" value="1">1</value>
</unit>
```

stating that 1 Pa = 1 N/m2 (Newton per square meter). “Meter” is a base unit, so we just can store it and continue with “N”

- The lookup for “N” gives:

```
<unit Code="N" CODE="N" isMetric="yes" class="si">
    <name>Newton</name>
    <printSymbol>N</printSymbol>
    <property>force</property>
    <value Unit="kg.m/s2" UNIT="KG.M/S2" value="1">1</value>
</unit>
```

stating that 1 N = 1 kg.m/s2
“g”, “m” and “s” are base units, so we are essentially done. We only need to combine everything again:

1 mm[Hg] = 0.001 x m[Hg] = 0.001 x 133.3220 kPa = 0.001 x 133.3220 x 1000 x N/m^2
= 133.3220 x $(kg.m/s^2) / m^2$ = 133.3220 x 1000 x $(g.m/s^2) / m2$
= 133322.0 $g/(m.s^2)$

This is a typical exam assignment for my students (1st year, undergraduate).

We can now do the same thing for “pounds per square inch”. In the ucum-essence.xml we find:

```
<unit Code="[psi]" CODE="[PSI]" isMetric="no" class="misc">
    <name>pound per sqare inch</name>
    <printSymbol>psi</printSymbol>
    <property>pressure</property>
    <value Unit="[lbf_av]/[in_i]2" UNIT="[LBF_AV]/[IN_I]2" value="1">1</value>
</unit>
```

And this again allows us to bring “pounds per square inch” to the base units “g”, “m” and “s”.

Now, imagine one of our American physicians provided a pressure in “pounds per square inch” and we need to convert that the “millimeter mercury column”. Suppose the value is 2.5 in “pounds per square inch”.

Exercise: look up ‘‘millimeter mercury column’’ and ‘‘pounds per square inch’’ in the CDISC controlled terminology and retrieve the conversion factor from the controlled terminology file.

With UCUM, calculating the conversion is easy: we can both reduce “mm[Hg]” and “[psi]” to the base units and get the conversion factor by simple division (our undergraduate student in medical informatics do this as an exercise).
One also notices that using the ucum-essence.xml file, such conversions become programmable.

Don't worry, we don't ask you to program this, others have already done that for you and e.g. made a RESTful web service available. The instructions and API for it can be found here.

So, you can just try out:

http://www.xml4pharmaserver.com:8080/CDISCCTService/rest/ucumtransform/2.5/from/[psi]/to/mm[Hg]

Resulting in:

```
- <XML4PharmaServerWebServiceResponse ServerDateTime="2017-05-20T09:15:55">
  - <WebServiceRequest>
      http://www.xml4pharmaserver.com:8080/CDISCCTService/rest/ucumtransform/2.5/from/%5bpsi%5d/to/mm%5bHg%5D
    </WebServiceRequest>
  - <Response>
      <SourceQuantity>2.5</SourceQuantity>
      <SourceUnit>[psi]</SourceUnit>
      <TargetUnit>mm[Hg]</TargetUnit>
      <ResultQuantity>129.28769</ResultQuantity>
    </Response>
  </XML4PharmaServerWebServiceResponse>
```

stating that 2.5 [psi] = 129.3 mm[Hg]

The RESTful web service can also be used from within your own applications (written in SAS, Java, C#, C++, Python, …)

Can you do this using CDISC controlled terminology?

Annotations

The part about UCUM annotations is the most badly understood part of the UCUM specification by CDISC people. They consider UCUM annotations as "something bad". The contrary is true.

Consider the unit "number of cells per milliliter". The UCUM "unit" for this is:

{cells}/mL

The part "{cells}" is called an annotation. It means "what is measured", but is not standardized by UCUM itself. It is left to the industry to make agreements/consensus on these. For example, in microbiology "cells" means something different than in electrical battery science. In the case of microbiology, the annotation "{cells}" has been defined by LOINC (also from the Regenstrief Institute).
Typical "annotations" in medical science already used are:

- {cells} number of cells
- {tablets} number of tablets (pills)
- {CFU} colony forming units

Some of the "UCUM annotations" already defined by LOINC are:

- {RBCs} red blood cells
- {WBCs} white blood cells
- {Hb} hemoglobin
- {creat} creatinine
- {beats} hearth/pulse beats

- {protein} protein

A list of the most occurring annotations in LOINC, and used in the UCUM units is given in appendix 1. CDISC could easily pick these up as they are already in LOINC and standardize them for use in CDISC too.

Let's elaborate the annotation "{CFU}" a bit further.

If we (CDISC) decide on agreeing that within the scope of medical research, the annotation "{CFU}" means "colony forming units", then AUTOMATICALLY all of the following UCUM units become valid:

- 10^3.{CFU}
- 10^3.{CFU}/mL
- 10^3.{CFU}/g
- 10^6.{CFU}
- 10^6.{CFU}/mL
- 10^6.{CFU}/g

For which all 6 equivalents can be found in the CDISC controlled terminology.

But also the following units automatically become valid just by agreeing on "{CFU}":

- 10^4.{CFU}
- 10^5.{CFU}/g
- 10^2.{CFU}/mg
- {CFU}/mg

and hundreds of others, none of which can be found in the CDISC controlled terminology. So, if you have an instrument that provides results in "colony forming units per milligram", you will either need to transform your results to one from the CDISC list manually (error prone), or make a request to CDISC to add "CFU per milligram" to the CDISC controlled terminology, a process that typically takes 3-6 months. Can you wait that long with your submission?

If we, in CDISC, agree on using the annotation "{CFU}" for "colony forming units", we can use any combination with existing UCUM notations, so we don't need extra terms anymore.

Remark that in CDISC-CT, we essentially already agreed that "CFU" means "colony forming units" implicitly (not explicitly), by adding 6 terms to the "UNIT" codelist, however not in a machine-readable way, but just as narrative text.

This will become very important when talking about the prejudice that UCUM cannot be used for SEND (i.e. for pre-clinical studies). It is just the other way around. Using UCUM allows to structure the chaos in units used in pre-clinical research, even doing conversions between them.

Exercise: do you see any other units in the CDISC controlled terminology "UNIT" codelist where the concept of the unit has been mixed up with "what is measured". Find a few of them and make a proposal for an "annotation".

UCUM for SDTM

UCUM becomes very important for SDTM for the cases where the information comes from:

- Electronic Health Records (EHRs)
- Lab and other automated instruments using LOINC coding

Almost any EHR system and standard (including HL7-CDA, CCD, consolidated-CDA, HL7-FHIR) mandate the use of UCUM units. An example from HL7-CDA (using SNOMED-CT coding):

```
−<Observation>
    <code code="271649006" codeSystem="2.16.840.1.113883.6.96" codeSystemName="SNOMED
    CT" displayName="Systolic BP"/>
    <effectiveTime value="200004071530"/>
    <value xsi:type="PQ" value="132" unit="mm[Hg]"/>
  </Observation>
```

Or from HL7-FHIR:

```
<valueQuantity>
  <value value="107"/>
  <unit value="mmHg"/>
  <system value="http://unitsofmeasure.org"/>
  <code value="mm[Hg]"/>
</valueQuantity>
```

As the use of UCUM in SDTM is currently not allowed, all values and units coming from EHRs need to be evaluated and often transformed to CDISC units: in this case from "mm[Hg]" to "mmHg". In some cases, this will mean recalculations (always error-prone). In other (but more seldom) cases, there will be no corresponding CDISC unit, and one will need to do a "new term request", which may retard a submission by months[5].

The second case is that when lab (and other) instruments use LOINC coding for defining which test is exactly performed. Now that the FDA has mandated LOINC coding for studies that start after March 15th 2018, it becomes evident to also use UCUM notation for the units. The background is that for each quantitative LOINC code for which there are units, LOINC provides a "preferred unit" which is in UCUM notation. For example (extract from the LOINC database):

[5] One can of course always extend the unit codelist, but this is not a good thing for the FDA, as it makes the unit incomparable with other studies and submissions.

LOINC_NUM	LONG_COMMON_NAME	EXAMPLE_UCUM_UNITS	
10676-5	Hepatitis C virus RNA [Units/volume] (viral load) in Serum or Plasma by Probe with amplification	[IU]/mL	☹
10577-5	Glucosidase [Enzymatic activity/volume] in Seminal plasma	ukat/L	
10578-3	Glycerophosphocholine [Moles/volume] in Seminal plasma	mol/L	
10351-5	HIV 1 RNA [Units/volume] (viral load) in Serum or Plasma by Probe with amplification	[IU]/mL	☹
10334-1	Cancer Ag 125 [Units/volume] in Serum or Plasma	[arb'U]/mL	☹
10501-5	Lutropin [Units/volume] in Serum or Plasma	m[IU]/mL	☹
10548-6	Phenytoin Free/Phenytoin.total in Serum or Plasma	%	
10634-4	Complement C1 esterase inhibitor.functional/Complement C1 esterase inhibitor.total in Serum or Plasma	%	
10864-7	Immune complex [Units/volume] in Serum or Plasma by Raji cell assay	[arb'U]/mL	☹
10874-6	Bombesin [Mass/volume] in Plasma	pg/mL	
11011-4	Hepatitis C virus RNA [Units/volume] (viral load) in Serum or Plasma by Probe and target amplification method	k[IU]/mL	☹
11014-8	Somatotropin Ab [Titer] in Serum or Plasma	{titer}	☹

(the ones with a "sad" smiley are not in CDISC controlled terminology)

Many modern laboratory instruments already work with LOINC codes. Those who do, often also (electronically) provide the results with UCUM notation for the units. So, when these values flow into SDTM, why shouldn't it be allowed to copy the unit "as is" in –ORRES, as the units is a standard unit, from UCUM. If it remains forbidden to use UCUM notation in SDTM, every single result that was transmitted electronically, either from a (lab) instrument or from an EHR, must be evaluated, transformed, recalculated, certainly leading to a decrease data quality and often to errors.

UCUM for SEND

An argument often heard is that UCUM doesn't work for SEND (pre-clinical research).

The "UNIT" codelist for SEND currently contains about 660 terms. As there are so many possible units in preclinical research, this list will probably grow to almost infinity (as it is a list). What is needed, is a SYSTEM.
Most of these 660 "units" can easily be written using UCUM notation. For example:

- /100 WBC => /10^2{WBCs} using the LOINC annotation for white blood cells
- /2000 RBC => /2*10^3{RBCs}
- /2500 RBC => /2.5*10^3{RBCs}
- /LPF => /[LPF] "LPF" is already standardized in UCUM

But now let us take something more difficult!

In our preclinical research, we have chickens that drink water, and we measure how much water they consume as function of the amount of space they have, amount of food, and time.
We however have one group in the US that measures this using gallons, square feet, ounces of food, and hours. In Europe however, the same measurement is made using liters, square meters, grams of food, and days. What is the conversion factor?

Exercise: look up in the CDISC SEND controlled terminology, whether you can find the units for this situation. If you can, calculate the conversion factor.

In UCUM, the notation for the measurements done in the USA is:

[us_gal]{waterconsumption}/([ft_i]2.{chicken}.[oz_av]{food}.h)

For the measurements done in Europe (or anywhere else in the world), the UCUM notation for the measurements are:

l{waterconsumption}/(m2.{chicken}.g.{food}.d)

Finding out the conversion factor using CDISC controlled terminology is an impossible task. With UCUM it is easy as everything can be broken down to the seven base units. You can try it manually - it will probably talk you half an hour or so. However, as these kinds of conversion calculations are easily programmable, and someone has already done that for you and made a RESTful web service available for it, you (or better: your application) can use the webservice.
Just try:

http://www.xml4pharmaserver.com:8080/CDISCCTService/rest/ucumtransform/from/l{waterconsumption}/(m2.{chicken}.g{food}.d)/to/[gal_us]{waterconsumption}/([ft_i]2.{chicken}.[oz_av][food].h)

Which returns some XML (or JSON if you like JSON more):


```
− <XML4PharmaServerWebServiceResponse ServerDateTime="2017-05-20T11:15:04">
  − <WebServiceRequest>
      http://www.xml4pharmaserver.com:8080/CDISCCTService/rest/ucumtransform
      /from/l%7Bwaterconsumption%7D/%28m2.%7Bchicken%7D.g%7Bfood%7D.d%29
      /to/%5Bgal_us%5D%7Bwaterconsumption%7D
      /%28%5Bft_i%5D2.%7Bchicken%7D.%5Boz_av%5D%7Bfood%7D.h%29
    </WebServiceRequest>
  − <Response>
      <SourceQuantity>1.0</SourceQuantity>
      <SourceUnit>l{waterconsumption}/(m2.{chicken}.g{food}.d)</SourceUnit>
    − <TargetUnit>
        [gal_us]{waterconsumption}/([ft_i]2.{chicken}.[oz_av]{food}.h)
      </TargetUnit>
      <ResultQuantity>0.028990207</ResultQuantity>
    </Response>
</XML4PharmaServerWebServiceResponse>
```


Providing you the conversion factor of "0.02899902"

Again, we do not expect researchers in pre-clinical research to use UCUM notation when writing results on paper. Many of these results are however calculated and there, UCUM notation can easily be introduced.

The use of UCUM notation for units in pre-clinical research may seem somewhat limited. It however surely makes sense if such conversions as demonstrated above must be done. The use of CDISC units for pre-clinical research however does not make sense at all – no possibility at all to use them for unit conversions, and based on a list that threatens to grow to infinity.

UCUM for SEND: conclusions

I hope that I have convinced you that, unlike many within the SDS team think, UCUM is very well possible for SEND. Even better, UCUM in combination with agreed-on ‘‘annotations’’ bring order in chaos, whereas the CDISC units for SEND have no logic at all.
All that the SEND and Controlled Terminology people need to do is to agree on standardizing the annotations. This is far easier than each time adding new entries to a list.

Myths about UCUM and their usage in CDISC standards

I often hear objections from CDISC people about allowing UCUM notation in CDISC submission datasets and especially SDTM and SEND datasets. Most of them are due to a lack of understanding of UCUM or of the strength of the combination of LOINC and UCUM.

The following statements come from communications with CDISC people. Of course, I don't mention names (that would not be fair). In each case, you will find my (sometimes altered/extended) answer.

Statement 1: *UCUM has "l" as symbol for "liter", but Americans use "L". So, we can't use UCUM.*

Just have a look at the contents of the "ucum-essence.xml" file, which essentially is the machine-readable UCUM specification. You will find:

```
<unit xmlns="" Code="l" CODE="L" isMetric="yes" class="iso1000">
    <name>liter</name>
    <printSymbol>l</printSymbol>
    <property>volume</property>
    <value Unit="dm3" UNIT="DM3" value="1">1</value>
</unit>
<unit xmlns="" Code="L" isMetric="yes" class="iso1000">
    <name>liter</name>
    <printSymbol>L</printSymbol>
    <property>volume</property>
    <value Unit="l" value="1">1</value>
</unit>
```

The first entry defines the symbol "l" and defines that it is not a base unit, but that it is equal (in a machine-readable way) to 1 dm^3.
The second entry defines the symbol "L" (American way of writing) and states (in a machine-readable way) that 1 L equals 1 l. By chaining, one then gets that 1 L = 1 dm^3.

Statement 2: "*The inherent flexibility of UCUM notation makes it impractical for regulators, who require minimized variability in the way data are represented …*"

It is correct that there is a lot of flexibility in UCUM notation, but it is based on a system and on science. There is less flexibility in CDISC-UNIT codelist, but it is neither based on a system nor on science, it is just an unorganized pragmatic approach governed by tradition. That the flexibility of UCUM would lead to too high variability for reviewers is another myth. It is just the opposite. UCUM brings order in chaos. Let me explain with a simple example:

My brother lives in Belgium. When I ask him about his blood pressure he states: 12-8 (sitting). The reason is that in Belgium, blood pressure is measured in centimeter mercury. So, for the systolic blood pressure his EHR states 12.0 and for the unit it states cm[Hg] (UCUM notation as is mandatory in most EHRs) with the code being a LOINC code: 8459-0 (systolic blood pressure sitting). When transferring this to SDTM VSORRES, a problem arises. The CDISC [UNIT] codelist does not know centimeter mercury, and we MUST use one from the CDISC [UNIT] codelist. Even if we extend the codelist in the define.xml with something like "cmHg", the validation tools of the FDA will throw an error. So, what we unfortunately need to do is to RECALCULATE the value into millimeter mercury and put that in VSORRES. Now in the SDTM-IG, I read the definition of VSORRES as "Result or Finding in Original Units", with the "CDISC note" being: "Result of the vital signs measurement as originally received or collected". In order to conform, I needed to recalculate, with the result that the value I need to put into VSORRES is not "original" at all anymore! Recalculation is always error prone. In this

case it is very simple as the conversion factor is 10. In other cases, it is more difficult however. When doing work for my customers, I have seen cases where –ORRES needed to be recalculated in order to obtain "minimized variability" and where the "new" values were simply wrong! So "original" in SDTM is not "original" anymore. Even worse is that the reviewer will never know whether the value is "as collected" or not. Traceability: zero! Coming back to my example, my brother's systolic blood pressure of 12.0 cm[Hg] (correct UCUM notation), was translated to VSORRES=140 with VSORRESU=mmHg (for the sake of "minimized variability"). How will the reviewer ever know that this is NOT the original value as collected? There is no way of finding out. Obviously, in this case a calculation error was made, but there is no way the reviewer can ever find out. He or she will simply think that 140 is really the value that was collected. Of course, we could put a copy of the EHR data point in the SDTM dataset, but this requires the use of (Dataset-)XML, which unfortunately is not accepted by the FDA yet.
When doing reviews, reviewers will probably look at both -ORRES (original result) and at -STRESN (standardized result) with the units (for which we must in principle use CDISC-CT) in -ORRESU and -STRESU. For the latter, it should be unique for each unique test. But what is a "unique test"? SDTM ----TESTCD does not describe the unique test. Even the combination of -TESTCD, -CAT, -METHOD, -SPEC and so on does not guarantee whether a test is unique or not. Only the LOINC code does. So, no wonder that the FDA recently decided to require LBLOINC for new studies (which CDISC blocked to have it "required" for so many years). LOINC and UCUM go closely together (both were developed by the Regenstrief Institute), but that still seems not to be understood within CDISC.
Let us go back to our example: the sitting systolic blood pressure (LOINC 8459-0) of my brother is 12.0 cm[Hg]. If I query the LOINC database (please try it) I find that the "preferred unit" is "mm[Hg]". So, we should take **that** for VSSTRESU (standardized result unit). It is worldwide valid and is understood by any other system outside CDISC that adheres to international standards (or do we really want to keep living on an island?). A simple automated RESTful web service query in my (SDTM-generating) system (I don't need to do anything) converts 12.0 cm[Hg] into 120 mm[Hg] automatically and adds that to VSSTRESN and VSSTRESU. This also works in case another researcher would have measured the blood pressure in pounds per square inch (UCUM: [psi]). So, my SDTM record would be something like:

VSLOINC	VSORRES	VSORRESU	VSSTRESN	VSSTRESU
8459-0	12.0	cm[Hg]	120	mm[Hg]
8459-0	2.5	[psi]	129.3	mm[Hg]

Remark that the values in VSSTRESN are automatically calculated through the use of a RESTful web service (well tested and validated) without any custom programming by a (SAS) programmer (error prone). Even better, the reviewer now really knows that the values in VSORRES are REALLY "original" and have not been recalculated (error prone) for the sake of "minimized variability".
Furthermore, the units are now in a notation that can also be understood outside the world of CDISC.

Statement 3: "*Part of the value of SDTM is having only a single controlled terminology value to represent a given concept. This means that if there are multiple ways to represent a unit, and each of those multiple ways is mathematically synonymous with the rest, the CDISC teams are required to choose one - and only one - value to represent that unit. An example is "International Unit per Milliliter per Gram" which has a CDISC Submission Value of "mIU/mL/mg". There are at least 12 different ways to express this unit using UCUM notation making it too flexible, and thus impractical for the SDTM dataset."*

For this example, "mIU/mL/mg" there is a test of course. If the test is worldwide standardized, there is a LOINC code for it, and there is a single "preferred UCUM unit". So, we take the "preferred UCUM unit" for LBSTRESU. This might be "m[IU]/mL/mg". Suppose however the data was captured (and

stored in the EHR) using “c[IU]/dL/g”. We use that “really original result unit” for LBORRESU, so that it is 100% sure for the reviewer that this really was the original collected value and unit, and standardize (can be automated using one of the RESTful web services) for this test (based on the LOINC code in LBLOINC, not on some combination of LBTESTCD with other things) to “m[IU]/mL/mg”, because that is the “preferred UCUM unit” for that test (by LOINC code). The results go into LBSTRESN and LBSTRESU. Traceability: very good.
If I would have followed the SDTM requirements, I would have had to RECALCULATE the value in order to satisfy the [UNIT] codelist, which is error prone, and the reviewer would even have never known that I did this recalculation (no traceability).

For lab tests, the combined use of LBLOINC (now finally required by the FDA) and UCUM notation for the units, makes everything “almost bomb proof”, whereas when using CDISC units, we are extremely error prone.

Statement 4: “*we do provide mappings to UCUM notation in the CDISC units codelists to help implementers who use UCUM understand the connections*”

Well, these “mappings” have been developed by Erin Muhlbradt (NCI) and myself. Most people within CDISC don’t seem to know or ignore that. We did NOT develop this list to provide a mapping from UCUM to CDISC units (as communicated by CDISC). We DID develop this list to learn people about UCUM, and to help them making the step to UCUM and away from the CDISC [UNIT] codelist, i.e. to allow them to replace the disadvantageous CDISC unit (no conversions possible, not a system, just a list) to the much more versatile UCUM notation (automated conversions, worldwide standard, even outside medical informatics). It is a petty to see that our list is now being abused as an argument against allowing UCUM notation in CDISC submission standards (--ORRESU and –STRESU variables).

Appendix 1: Annotations already assigned by LOINC

In the following table, the most used UCUM annotations already present in the LOINC are provided, together with their count (number of occurrences in the LOINC database, i.e. number of LOINC codes for which the unit contains this annotation), a short explanation, and an example.

Annotation	Count	Explanation	Example
{titer}	2597	Titer	102-4: Cefoperazone [Susceptibility] by Serum bactericidal titer Unit: **{titer}**
{creat}	1172	Creatinine	11141-9: Phosphate/Creatinine [Mass Ratio] in Urine Unit: mg/g**{creat}**
{protein}	144	Protein	1812-7: Alpha fucosidase [Enzymatic activity/mass] in Tissue Unit: nmol/h/mg**{protein}**
{copies}	177	Copies	21333-0: HIV 1 RNA [#/volume] in Serum Unit: {copies}/mL
{Hb}	62	Hemoglobin	16273-5: Calcium [Mass/mass] in Red Blood Cells Unit: mg/g**{Hb}**
{RBCs}	43	Red Blood Cells	2305-1: Galactokinase [Enzymatic activity/volume] in Red Blood Cells Unit: U/mL**{RBCs}**
{WBCs}	9	White Blood Cells	33990-3: Normoblasts/100 leukocytes [Ratio] in Blood Unit: /100**{WBCs}**
{platelets}	3	Platelets	42671-8: Serotonin [Entitic mass] in Platelets Unit: ng/10*9**{platelets}**
{Ehrlich'U}	4	Ehrlich Units	19161-9: Urobilinogen [Units/volume] in Urine by Test strip Unit: **{Ehrlich'U}/**dL

and many others …