Python Based Standardization Tools for ClinicalTrials.Gov

Jacob Barhak

Austin, Texas

http://sites.google.com/site/jacobbarhak/

COMBINE 2018

Boston University

October 8-12, 2018

Abstract

ClinicalTrials.Gov is a government database that stores clinical trials from around the world. This database is growing fast, partially because some requirements of reporting clinical trial results are now supported by U.S. law. Despite the growth of database, the data stored there is still not widely used for modeling and simulation, partially because the database is still a relatively new tool, and partially because the data there is not standardized. Since data is entered by multiple entities into this semi-structured text based database, and since clinical trials have a large variety, the data is not immediately suitable for modeling. Although the National Library of Medicine scrutinizes this data, the scrutiny level is for human comprehensible data, while modeling requires computer comprehension.

For this reason, there is a need to clean data towards tasks such as disease modeling. This work will discuss a set of Python tools that were used to process ClinicalTrails.Gov and aim towards standardization. The tools parse XML data, index the information for easier processing, and then uses Machine Learning and Natural Language Processing to cluster the data, The clustering algorithm is used to assist a human user look at similar information provided by a Graphical User Interface.

These Python tools were used to extract 21,094 units from 30,763 different clinical trails with results. These quantities show a clear need of standardization to be used in future computer modeling efforts.

Difficulties Making Medical Data Machine Comprehensible

Data Source	Difficulty
• Electronic Medical Records	• Restricted and splintered
• Public datasets and databases: – ClinicalTrials.Gov – Physionet – Many others ...	• New and needs standardization – Fast Growing – supported by law – Multiple datasets – Assorted
• Publications – Medical journals in electronic form – Web sites – Printed	• Free Text – not standardized – Made for humans + some cost – Unorganized non centralized data – Assorted and hard to access
• Knowledge held by physicians	• Not Machine Accessible

ClinicalTrials.Gov is a good source to start machine comprehension due to: accessibility, size, variety, and potential growth

Existing Medical Data Specifications/Standards

- Unified Medical Language System (UMLS)
- Snomed CT
- Clinical Data Interchange Standards Consortium (CDISC):
  - Foundational: PRM, SEND, CDASH, SDTM, SDTMIG, SDTMIG-Pgx, ADaM, QRS
  - Data Exchange: CTR-XML, ODM-XML, SDM-XML, Define-XML, DataSet-XML, LAB, RDF
  - Therapeutic Areas – 27 areas including Alzheimer's, Diabetes, Cardiovascular, Influenza, Cancers
  - CDISC Share
  - Semantics
  - Domain information Model
  - Has units classification
- Intentional Classification of Diseases (ICD)
- HL7 Fast Healthcare Interoperability Resources Specification (FHIR®)
- The Unified Code for Units of Measure - (UCUM)

So many specifications!
Yet, so much data is still not interchangeable
More work is needed!

About ClinicalTrials.Gov

- National Institutes of Health (NIH) Project
- Maintained by the National Library of Medicine (NLM)
- Accumulates Clinical Trial Data internationally
- Fast growing database

Date	Number of Trials	Studies with Results
20-Apr-2018	271,510	30,763
3-Nov-2017	258,046	28,785
12-Feb-2017	236,687	24,251
27-Sep-2016	226,460	22,614
7-Apr-2015	187,653	Not collected

- Now many clinical trials are required to register in this data base by law: PUBLIC LAW 110–85—SEPT. 27, 2007 - TITLE VIII—CLINICAL TRIAL DATABASES. Section 801 of the Food and Drug Administration Amendments Act of 2007. Online: https://www.gpo.gov/fdsys/pkg/PLAW-110publ85/pdf/PLAW-110publ85.pdf#page=82

ClinicalTrials.Gov Capabilities

- Has web entry system to collect data from multiple sources
- Entered data is scrutinized for accuracy **at the human level**
- Stores numeric trial results as well as clerical information
- Has a useful **human** web interface
  - Has a sophisticated search capability that handles medical terms, acronyms and filtering
  - Offers expert search for programmers
- Trial data can be exported to XML **for machine readability**
- XML scheme is published

ClinicalTrials.Gov is a great new tool !
Yet, it requires standardization for machine comprehension

Imported Data Can be Used for Population Generation

Data imported from ClinicalTrials.Gov can be used for generation of synthetic population to mimic statistics

- Heterogeneity = generate individuals
- Multiple characteristics per individual
- Allow correlations
- Allow restrictions

Correlated

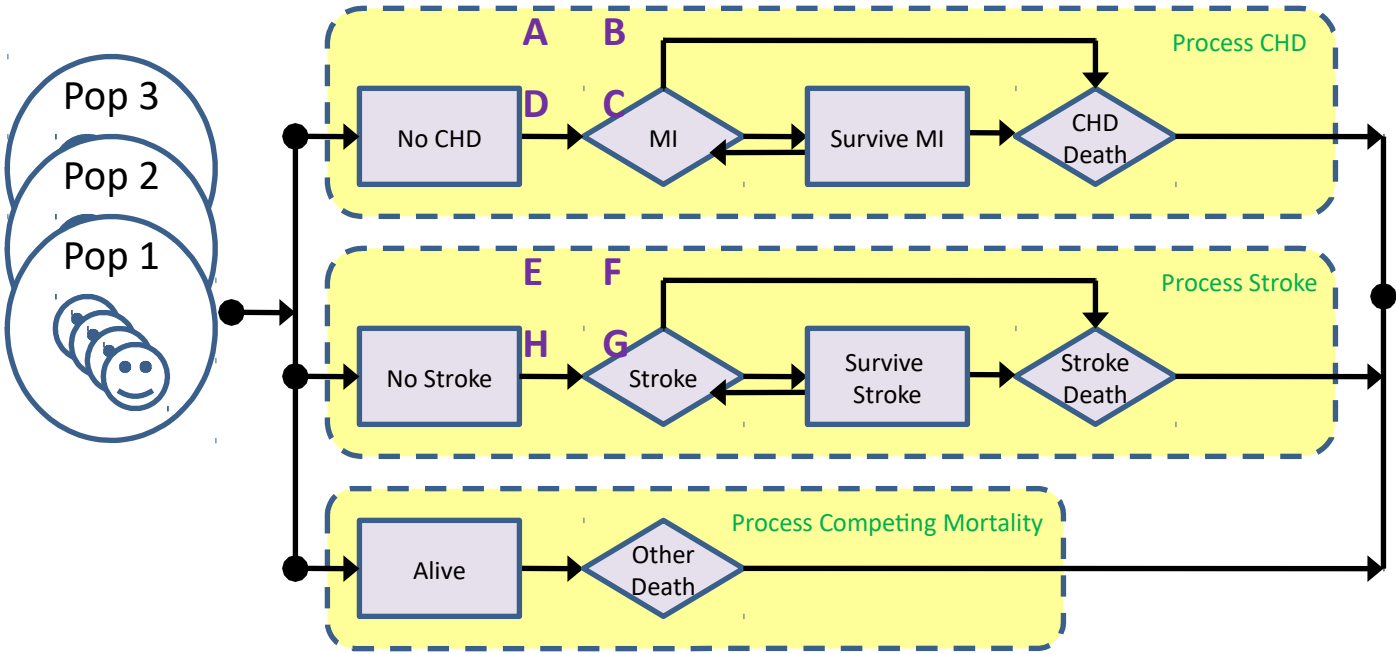
IndividualID	Male	Age	BP	…
0	0	50	140	…
1	1	45	135	…
2	0	22	120	…
3	1	85	145	…
4	1	14	125	…

Generation Code / Equations

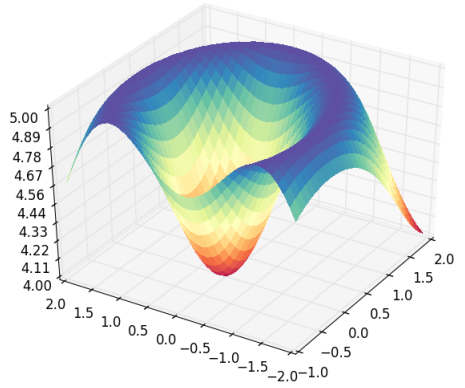
Restrict Age

Disease Model Validation with Simulation Using Outcomes

- The Reference Model for Disease Progression is an ensemble model
- The model accumulates knowledge: models= assumptions and facts = trial data



Eq EH	E	E	E	E	...
Eq AD	A	B	C	D	...
Pop 1	4	6	2	1	...
Pop 2	2	4	6	1	...
Pop 3	2	3	9	2	...
...	...	...	...	...	...

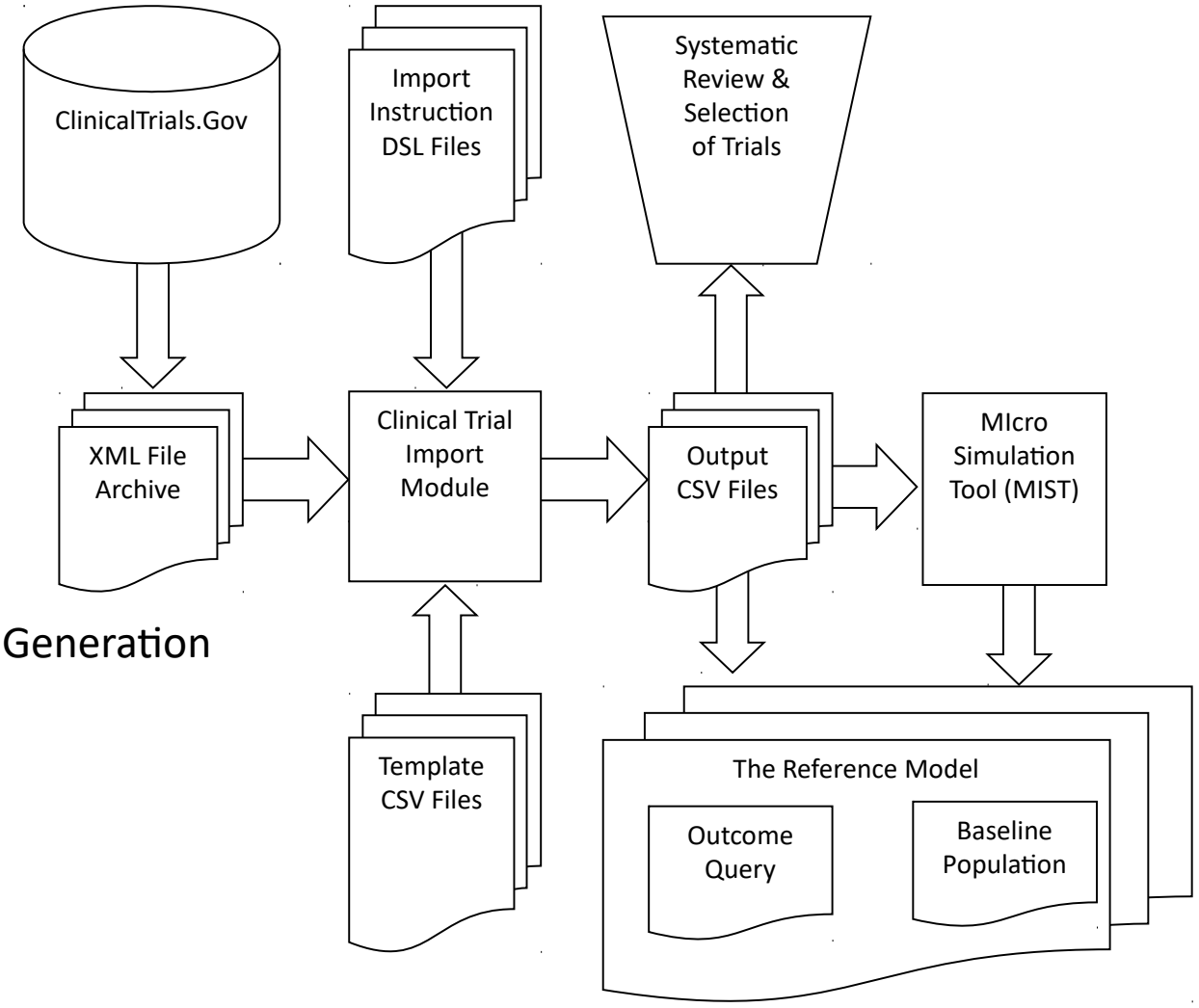


- Model combinations are validated against trial outcomes
- The Reference Model can calculate best model combination

Importing ClinicalTrials.Gov Data

Three Import stages:

- Systematic Review
  - Human Selection
- Import Populations
  - Synthetic Population Generation
- Import Outcomes
  - Model Validation



Issues Encountered During Import

- Systematic review
  - Organizing results in tabular human readable format
  - Extracting meaningful numbers from convoluted free text
- Population generation
  - Code generation from data
  - Name matching
  - Unit conversion – context sensitive
  - Time extraction from free text
- Outcome conversion
  - Scaling numbers to similar reference
  - Cohort matching
  - Calculation of missing outcomes for full trial

The human natural language used in ClinicalTrials.Gov is harder for machines to understand and process!

More importantly, machines cannot interpret numbers without scaling units!

Python Based Standardization of ClinicalTrials.Gov

- The entire database with results was downloaded on 20-Apr-2018
  - 30,763 XML files of trials were loaded
- A set of python scripts were written with the following purposes:
  - Import XML data and organize it
    - 61 batches were required to handle the big data
  - Index all XML tags and their values
    - 387 different tags = fields were processed
  - Index unit tags with association to title tags
    - 21,094 different units were detected
  - Find similarity between unit names
    - Include CDISC units in similarity matrix
  - Cluster units by similarity
    - 110 clusters were created
  - Create Graphical User Interface for users to manually map units

Several stages required work in smaller batches to fit into memory.

This issue should gain attention in the future due to database growth

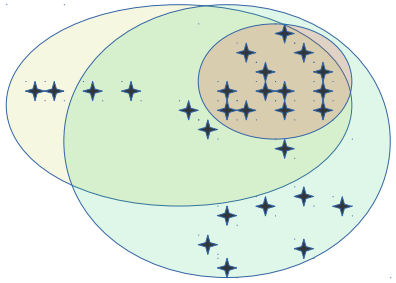
Natural Language Processing (NLP) Issues

- In addition to ClinicalTrials.Gov units, 6645 CDISC units and their synonyms, including UCUM mapping were added for auxiliary matches
- Unicode characters required special handling
- Synonyms of common words were replaced in pre-processing to help the system with matching, e.g. 'Percent' = '%'
- Distances between text of units was were using two different functions:
  - difflib - calculating charter changes to match words
  - scikit-learn - Term Frequency–Inverse Document Frequency (TFIDF) calculated using 3-6 characters combined with cosine similarity
- Similar units were further ranked:
  - CDISC units were considered more important
  - Common units are ranked first if similar to rare units

Those python libraries produced satisfactory results. An attempt was made with the spaCy library that parses text yet it did not produce better results than the tools used.

Clustering Units

- To help a user see similar units together, clustering methods were used
- The similarity score between units was used as a metric to cluster similar units
- The MiniBatchKMeans clustering function within scikit-learn was used to reduce memory footprint
- Three clustering passes were made with variations of the similarity score to handle both loose neighbors and very close neighbors



- Clusters in different passes were intersected to capture very close neighbors together
- After clustering, small clusters were merged to larger clusters to improve clustering results by re-merging close neighbors that were split

Graphic User Interface (GUI)

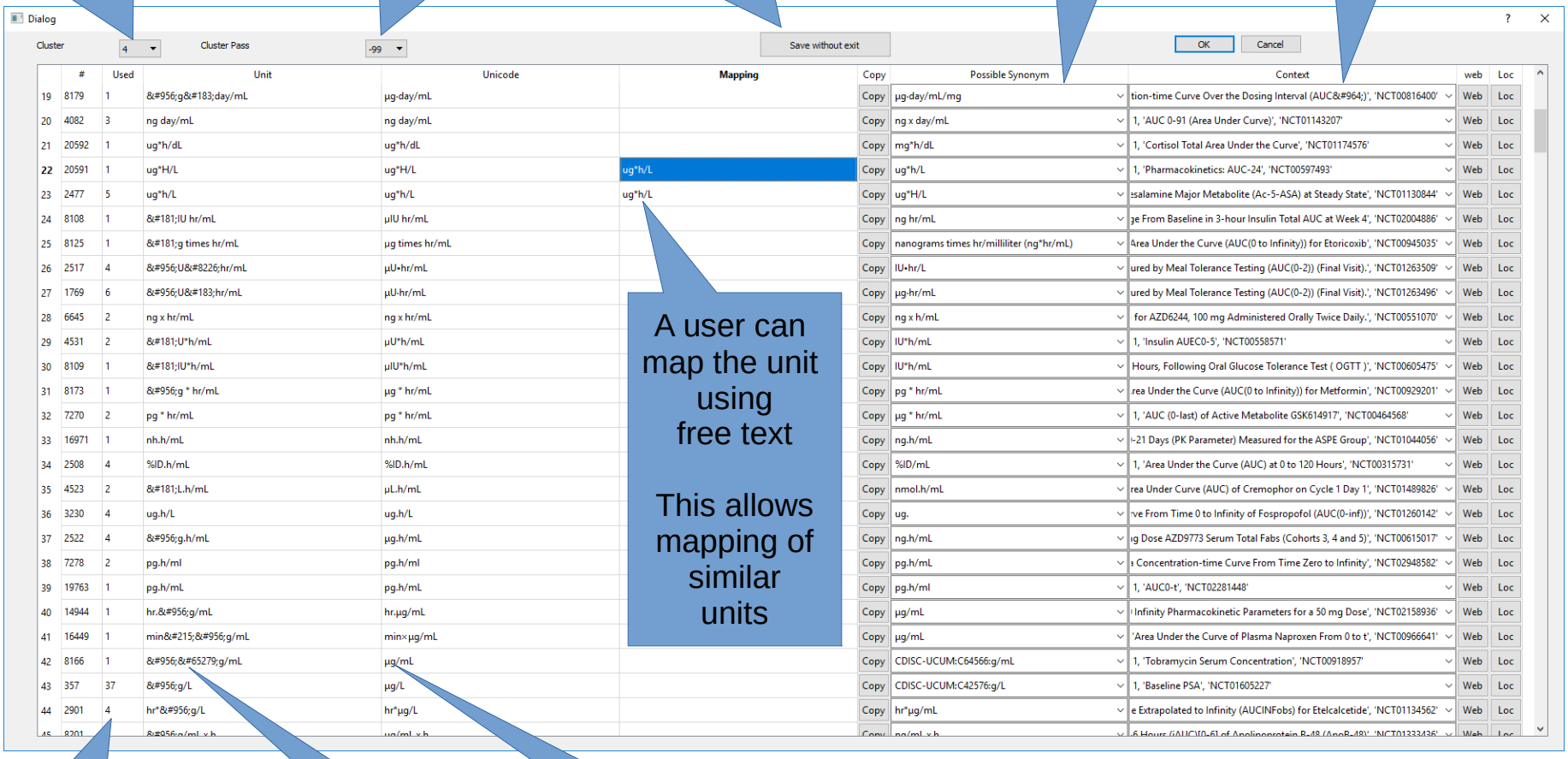
A user can summon specific cluster #

A user can explore clustering level

A user can save mapping with version #

Suggested synonym

Unit context in trial



A user can map the unit using free text

This allows mapping of similar units

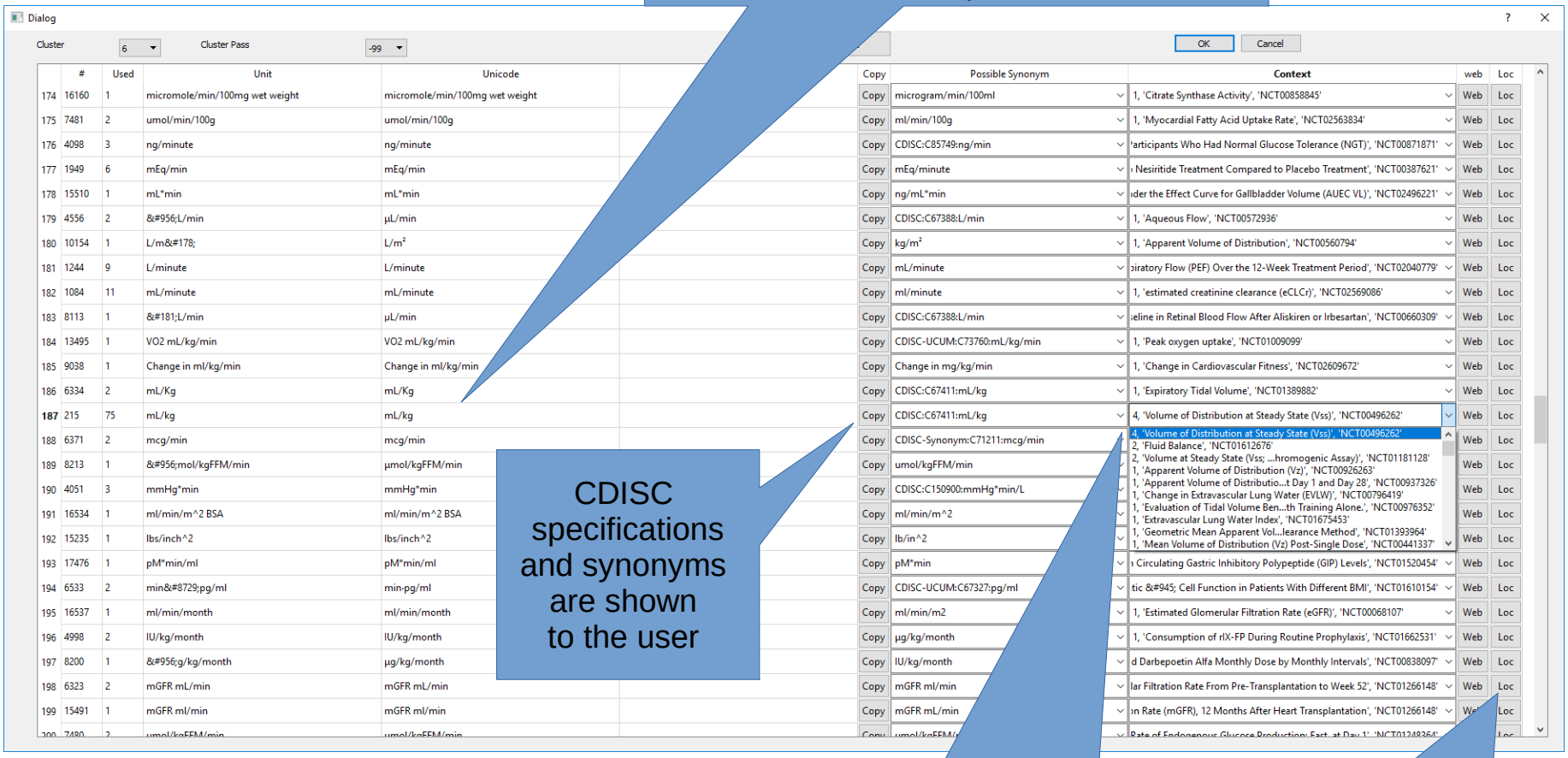
Shows number of unit occurrences

Show Unicode Escape

Shows the unit as user sees it

The GUI Assists Human User

For a human, those units are the same while a machine requires exact match

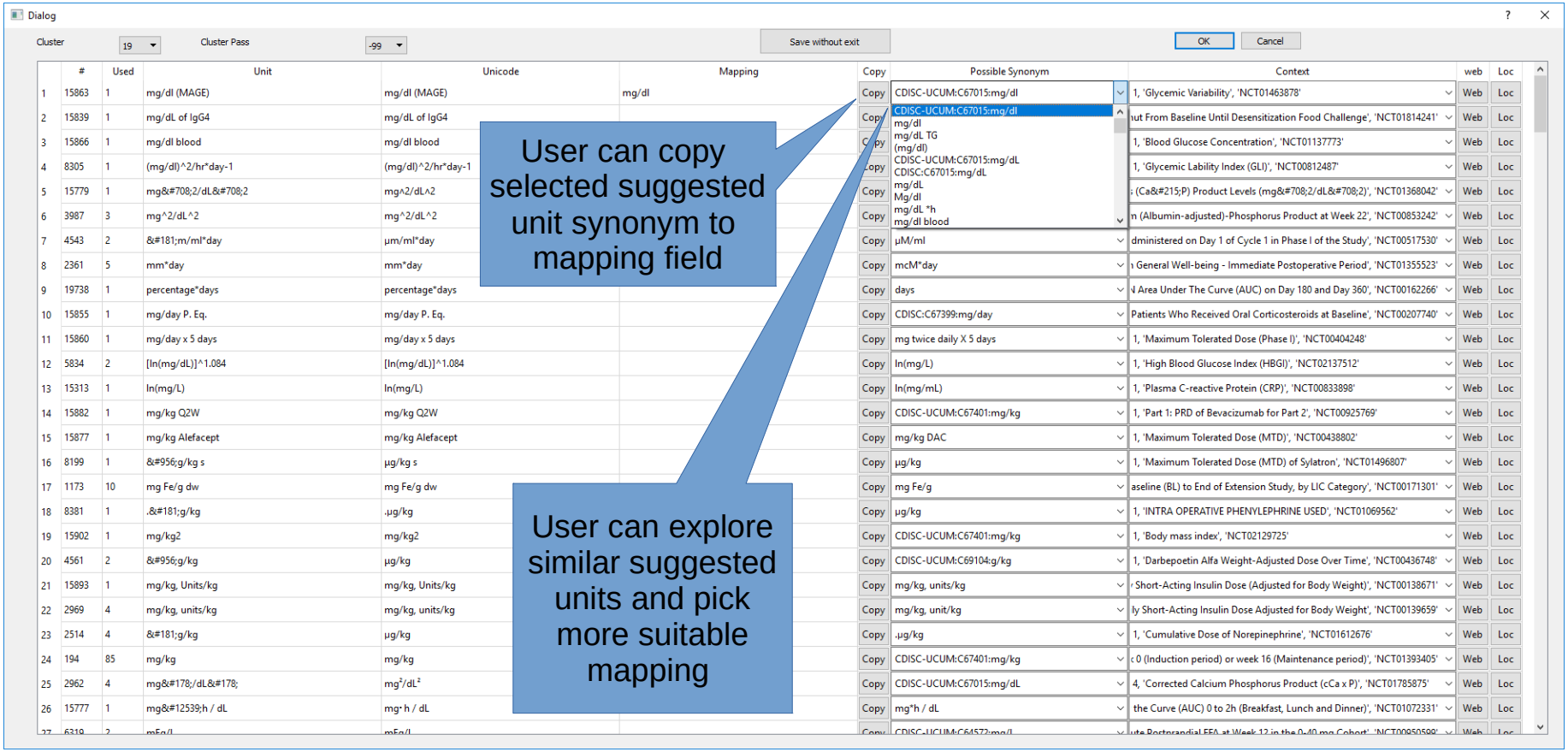


CDISC specifications and synonyms are shown to the user

User can view multiple contexts of the same unit and focus on those online

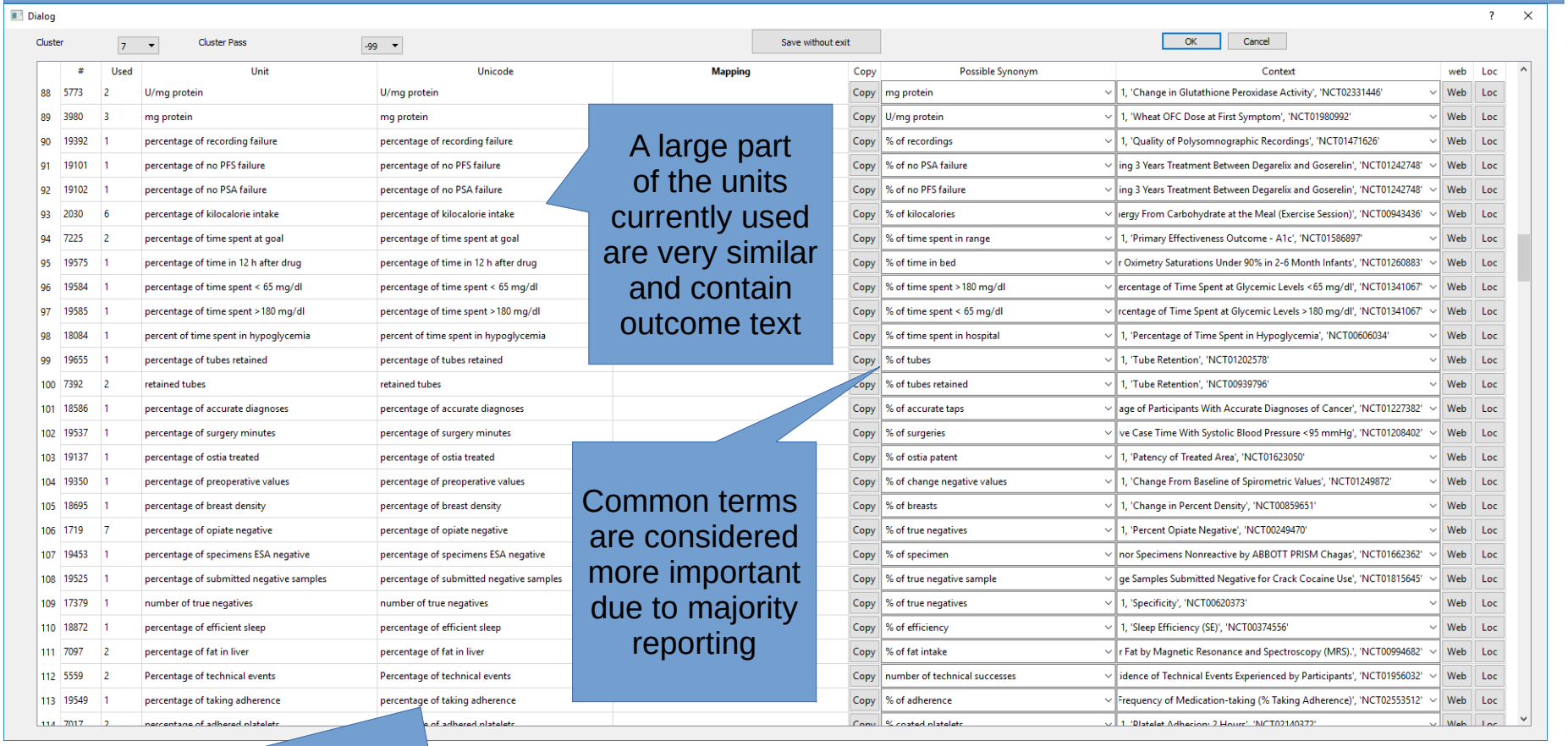
A user can go online focusing on the unit in the trial web page

The GUI Helps Save Time by Providing Suggestions



Very Common Units and Terms

Units appearing more than 1000 times: Participants , participants , years , units on a scale , percentage of participants , Percentage of participants , months , Units on a scale , Years , mg/dL , ng/mL , days , Subjects , mmHg , scores on a scale , Percentage of Participants , Scores on a scale , percent change , hours , Months , Days , mm , kg , mmol/L , subjects , minutes , pg/mL , score on a scale , ratio , kg/m^2



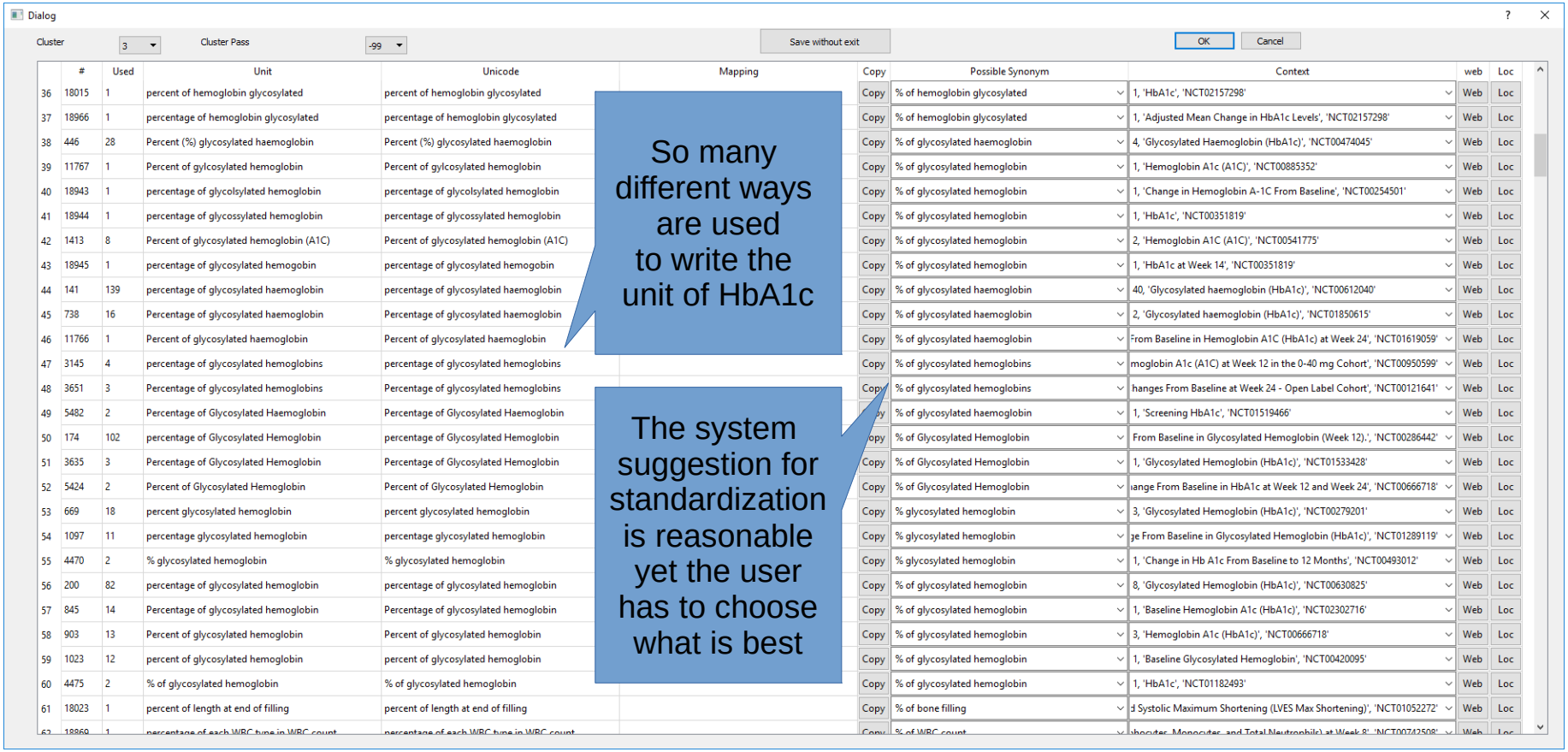
	#	Used	Unit	Unicode	Mapping	Copy	Possible Synonym	Context	web	Loc
88	5773	2	U/mg protein	U/mg protein		Copy	mg protein	1, 'Change in Glutathione Peroxidase Activity', 'NCT02331446'	Web	Loc
89	3980	3	mg protein	mg protein		Copy	U/mg protein	1, 'Wheat OFC Dose at First Symptom', 'NCT01980992'	Web	Loc
90	19392	1	percentage of recording failure	percentage of recording failure		Copy	% of recordings	1, 'Quality of Polysomnographic Recordings', 'NCT01471626'	Web	Loc
91	19101	1	percentage of no PFS failure	percentage of no PFS failure		Copy	% of no PSA failure	ing 3 Years Treatment Between Degarelix and Goserelin', 'NCT01242748'	Web	Loc
92	19102	1	percentage of no PSA failure	percentage of no PSA failure		Copy	% of no PFS failure	ing 3 Years Treatment Between Degarelix and Goserelin', 'NCT01242748'	Web	Loc
93	2030	6	percentage of kilocalorie intake	percentage of kilocalorie intake		Copy	% of kilocalories	ergy From Carbohydrate at the Meal (Exercise Session)', 'NCT00943436'	Web	Loc
94	7225	2	percentage of time spent at goal	percentage of time spent at goal		Copy	% of time spent in range	1, 'Primary Effectiveness Outcome - A1c', 'NCT01586897'	Web	Loc
95	19575	1	percentage of time in 12 h after drug	percentage of time in 12 h after drug		Copy	% of time in bed	r Oximetry Saturations Under 90% in 2-6 Month Infants', 'NCT01260883'	Web	Loc
96	19584	1	percentage of time spent < 65 mg/dl	percentage of time spent < 65 mg/dl		Copy	% of time spent >180 mg/dl	ercentage of Time Spent at Glycemic Levels <65 mg/dl', 'NCT01341067'	Web	Loc
97	19585	1	percentage of time spent >180 mg/dl	percentage of time spent >180 mg/dl		Copy	% of time spent < 65 mg/dl	rcentage of Time Spent at Glycemic Levels >180 mg/dl', 'NCT01341067'	Web	Loc
98	18084	1	percent of time spent in hypoglycemia	percent of time spent in hypoglycemia		Copy	% of time spent in hospital	1, 'Percentage of Time Spent in Hypoglycemia', 'NCT00606034'	Web	Loc
99	19655	1	percentage of tubes retained	percentage of tubes retained		Copy	% of tubes	1, 'Tube Retention', 'NCT01202578'	Web	Loc
100	7392	2	retained tubes	retained tubes		Copy	% of tubes retained	1, 'Tube Retention', 'NCT00939796'	Web	Loc
101	18586	1	percentage of accurate diagnoses	percentage of accurate diagnoses		Copy	% of accurate taps	age of Participants With Accurate Diagnoses of Cancer', 'NCT01227382'	Web	Loc
102	19537	1	percentage of surgery minutes	percentage of surgery minutes		Copy	% of surgeries	ve Case Time With Systolic Blood Pressure <95 mmHg', 'NCT01208402'	Web	Loc
103	19137	1	percentage of ostia treated	percentage of ostia treated		Copy	% of ostia patent	1, 'Patency of Treated Area', 'NCT01623050'	Web	Loc
104	19350	1	percentage of preoperative values	percentage of preoperative values		Copy	% of change negative values	1, 'Change From Baseline of Spirometric Values', 'NCT01249872'	Web	Loc
105	18695	1	percentage of breast density	percentage of breast density		Copy	% of breasts	1, 'Change in Percent Density', 'NCT00859651'	Web	Loc
106	1719	7	percentage of opiate negative	percentage of opiate negative		Copy	% of true negatives	1, 'Percent Opiate Negative', 'NCT00249470'	Web	Loc
107	19453	1	percentage of specimens ESA negative	percentage of specimens ESA negative		Copy	% of specimen	nor Specimens Nonreactive by ABBOTT PRISM Chagas', 'NCT01662362'	Web	Loc
108	19525	1	percentage of submitted negative samples	percentage of submitted negative samples		Copy	% of true negative sample	ge Samples Submitted Negative for Crack Cocaine Use', 'NCT01815645'	Web	Loc
109	17379	1	number of true negatives	number of true negatives		Copy	% of true negatives	1, 'Specificity', 'NCT00620373'	Web	Loc
110	18872	1	percentage of efficient sleep	percentage of efficient sleep		Copy	% of efficiency	1, 'Sleep Efficiency (SE)', 'NCT00374556'	Web	Loc
111	7097	2	percentage of fat in liver	percentage of fat in liver		Copy	% of fat intake	r Fat by Magnetic Resonance and Spectroscopy (MRS).', 'NCT00994682'	Web	Loc
112	5559	2	Percentage of technical events	Percentage of technical events		Copy	number of technical successes	idence of Technical Events Experienced by Participants', 'NCT01956032'	Web	Loc
113	19549	1	percentage of taking adherence	percentage of taking adherence		Copy	% of adherence	Frequency of Medication-taking (% Taking Adherence)', 'NCT02553512'	Web	Loc
114	7017	2	percentage of adhered platelets	of adhered platelets		Copy	% coated platelets	1, 'Platelet Adhesion: 2 Hours', 'NCT02140372'	Web	Loc

A large part of the units currently used are very similar and contain outcome text

Common terms are considered more important due to majority reporting

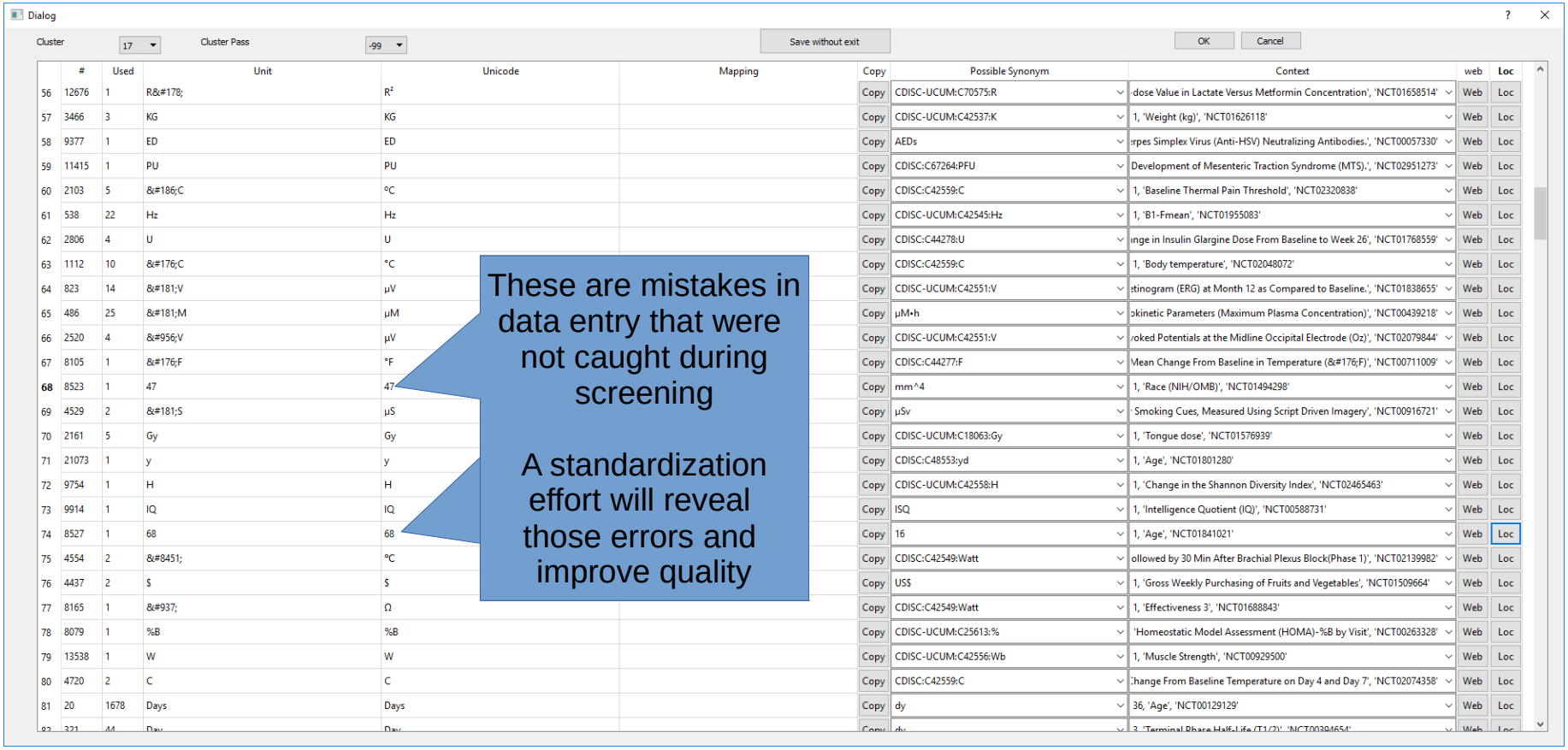
Many numbers are proportions and in that sense unit-less, yet represented as percentage So the unit text represents scaling

Standardization is Important



	#	Used	Unit	Unicode	Mapping	Copy	Possible Synonym	Context	web	Loc
36	18015	1	percent of hemoglobin glycosylated	percent of hemoglobin glycosylated		Copy	% of hemoglobin glycosylated	1, 'HbA1c', 'NCT02157298'	Web	Loc
37	18966	1	percentage of hemoglobin glycosylated	percentage of hemoglobin glycosylated		Copy	% of hemoglobin glycosylated	1, 'Adjusted Mean Change in HbA1c Levels', 'NCT02157298'	Web	Loc
38	446	28	Percent (%) glycosylated haemoglobin	Percent (%) glycosylated haemoglobin		Copy	% of glycosylated haemoglobin	4, 'Glycosylated Haemoglobin (HbA1c)', 'NCT00474045'	Web	Loc
39	11767	1	Percent of gylcosylated hemoglobin	Percent of gylcosylated hemoglobin		Copy	% of glycosylated hemoglobin	1, 'Hemoglobin A1c (A1C)', 'NCT00885352'	Web	Loc
40	18943	1	percentage of glycolsylated hemoglobin	percentage of glycolsylated hemoglobin		Copy	% of glycosylated hemoglobin	1, 'Change in Hemoglobin A-1C From Baseline', 'NCT00254501'	Web	Loc
41	18944	1	percentage of glycossylated hemoglobin	percentage of glycossylated hemoglobin		Copy	% of glycosylated hemoglobin	1, 'HbA1c', 'NCT00351819'	Web	Loc
42	1413	8	Percent of glycosylated hemoglobin (A1C)	Percent of glycosylated hemoglobin (A1C)		Copy	% of glycosylated hemoglobin	2, 'Hemoglobin A1C (A1C)', 'NCT00541775'	Web	Loc
43	18945	1	percentage of glycosylated hemogobin	percentage of glycosylated hemogobin		Copy	% of glycosylated hemoglobin	1, 'HbA1c at Week 14', 'NCT00351819'	Web	Loc
44	141	139	percentage of glycosylated haemoglobin	percentage of glycosylated haemoglobin		Copy	% of glycosylated haemoglobin	40, 'Glycosylated haemoglobin (HbA1c)', 'NCT00612040'	Web	Loc
45	738	16	Percentage of glycosylated haemoglobin	Percentage of glycosylated haemoglobin		Copy	% of glycosylated haemoglobin	2, 'Glycosylated haemoglobin (HbA1c)', 'NCT01850615'	Web	Loc
46	11766	1	Percent of glycosylated haemoglobin	Percent of glycosylated haemoglobin		Copy	% of glycosylated haemoglobin	From Baseline in Hemoglobin A1C (HbA1c) at Week 24', 'NCT01619059'	Web	Loc
47	3145	4	percentage of glycosylated hemoglobins	percentage of glycosylated hemoglobins		Copy	% of glycosylated hemoglobins	moglobin A1c (A1C) at Week 12 in the 0-40 mg Cohort', 'NCT00950599'	Web	Loc
48	3651	3	Percentage of glycosylated hemoglobins	Percentage of glycosylated hemoglobins		Copy	% of glycosylated hemoglobins	hanges From Baseline at Week 24 - Open Label Cohort', 'NCT00121641'	Web	Loc
49	5482	2	Percentage of Glycosylated Haemoglobin	Percentage of Glycosylated Haemoglobin		Copy	% of glycosylated haemoglobin	1, 'Screening HbA1c', 'NCT01519466'	Web	Loc
50	174	102	percentage of Glycosylated Hemoglobin	percentage of Glycosylated Hemoglobin		Copy	% of Glycosylated Hemoglobin	From Baseline in Glycosylated Hemoglobin (Week 12).', 'NCT00286442'	Web	Loc
51	3635	3	Percentage of Glycosylated Hemoglobin	Percentage of Glycosylated Hemoglobin		Copy	% of Glycosylated Hemoglobin	1, 'Glycosylated Hemoglobin (HbA1c)', 'NCT01533428'	Web	Loc
52	5424	2	Percent of Glycosylated Hemoglobin	Percent of Glycosylated Hemoglobin		Copy	% of Glycosylated Hemoglobin	ange From Baseline in HbA1c at Week 12 and Week 24', 'NCT00666718'	Web	Loc
53	669	18	percent glycosylated hemoglobin	percent glycosylated hemoglobin		Copy	% glycosylated hemoglobin	3, 'Glycosylated Hemoglobin (HbA1c)', 'NCT00279201'	Web	Loc
54	1097	11	percentage glycosylated hemoglobin	percentage glycosylated hemoglobin		Copy	% glycosylated hemoglobin	ge From Baseline in Glycosylated Hemoglobin (HbA1c)', 'NCT01289119'	Web	Loc
55	4470	2	% glycosylated hemoglobin	% glycosylated hemoglobin		Copy	% glycosylated hemoglobin	1, 'Change in Hb A1c From Baseline to 12 Months', 'NCT00493012'	Web	Loc
56	200	82	percentage of glycosylated hemoglobin	percentage of glycosylated hemoglobin		Copy	% of glycosylated hemoglobin	8, 'Glycosylated Hemoglobin (HbA1c)', 'NCT00630825'	Web	Loc
57	845	14	Percentage of glycosylated hemoglobin	Percentage of glycosylated hemoglobin		Copy	% of glycosylated hemoglobin	1, 'Baseline Hemoglobin A1c (HbA1c)', 'NCT02302716'	Web	Loc
58	903	13	Percent of glycosylated hemoglobin	Percent of glycosylated hemoglobin		Copy	% of glycosylated hemoglobin	3, 'Hemoglobin A1c (HbA1c)', 'NCT00666718'	Web	Loc
59	1023	12	percent of glycosylated hemoglobin	percent of glycosylated hemoglobin		Copy	% of glycosylated hemoglobin	1, 'Baseline Glycosylated Hemoglobin', 'NCT00420095'	Web	Loc
60	4475	2	% of glycosylated hemoglobin	% of glycosylated hemoglobin		Copy	% of glycosylated hemoglobin	1, 'HbA1c', 'NCT01182493'	Web	Loc
61	18023	1	percent of length at end of filling	percent of length at end of filling		Copy	% of bone filling	d Systolic Maximum Shortening (LVES Max Shortening)', 'NCT01052272'	Web	Loc
62	18869	1	percentage of each WBC type in WBC count	percentage of each WBC type in WBC count		Copy	% of WBC count	ibocytes, Monocytes, and Total Neutrophils) at Week 8', 'NCT00742508'	Web	Loc

Standardization Reveals Mistakes

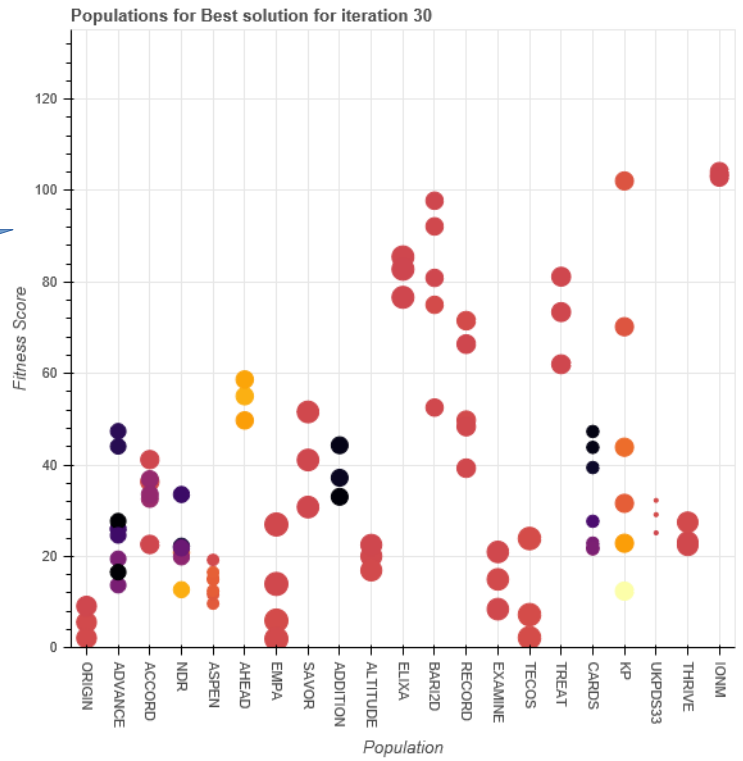


	#	Used	Unit	Unicode	Mapping	Copy	Possible Synonym	Context	web	Loc
56	12676	1	R²	R²		Copy	CDISC-UCUM:C70575:R	dose Value in Lactate Versus Metformin Concentration', 'NCT01658514'	Web	Loc
57	3466	3	KG	KG		Copy	CDISC-UCUM:C42537:K	1, 'Weight (kg)', 'NCT01626118'	Web	Loc
58	9377	1	ED	ED		Copy	AEDs	rpes Simplex Virus (Anti-HSV) Neutralizing Antibodies.', 'NCT00057330'	Web	Loc
59	11415	1	PU	PU		Copy	CDISC:C67264:PFU	Development of Mesenteric Traction Syndrome (MTS).', 'NCT02951273'	Web	Loc
60	2103	5	ºC	°C		Copy	CDISC:C42559:C	1, 'Baseline Thermal Pain Threshold', 'NCT02320838'	Web	Loc
61	538	22	Hz	Hz		Copy	CDISC-UCUM:C42545:Hz	1, 'B1-Fmean', 'NCT01955083'	Web	Loc
62	2806	4	U	U		Copy	CDISC:C44278:U	nge in Insulin Glargine Dose From Baseline to Week 26', 'NCT01768559'	Web	Loc
63	1112	10	°C	°C		Copy	CDISC:C42559:C	1, 'Body temperature', 'NCT02048072'	Web	Loc
64	823	14	µV	µV		Copy	CDISC-UCUM:C42551:V	tinogram (ERG) at Month 12 as Compared to Baseline.', 'NCT01838655'	Web	Loc
65	486	25	µM	µM		Copy	µM•h	okinetic Parameters (Maximum Plasma Concentration)', 'NCT00439218'	Web	Loc
66	2520	4	μV	µV		Copy	CDISC-UCUM:C42551:V	oked Potentials at the Midline Occipital Electrode (Oz)', 'NCT02079844'	Web	Loc
67	8105	1	°F	°F		Copy	CDISC:C44277:F	Mean Change From Baseline in Temperature (°F)', 'NCT00711009'	Web	Loc
68	8523	1	47	47		Copy	mm^4	1, 'Race (NIH/OMB)', 'NCT01494298'	Web	Loc
69	4529	2	µS	µS		Copy	µSv	Smoking Cues, Measured Using Script Driven Imagery', 'NCT00916721'	Web	Loc
70	2161	5	Gy	Gy		Copy	CDISC-UCUM:C18063:Gy	1, 'Tongue dose', 'NCT01576939'	Web	Loc
71	21073	1	y	y		Copy	CDISC:C48553:yd	1, 'Age', 'NCT01801280'	Web	Loc
72	9754	1	H	H		Copy	CDISC-UCUM:C42558:H	1, 'Change in the Shannon Diversity Index', 'NCT02465463'	Web	Loc
73	9914	1	IQ	IQ		Copy	ISQ	1, 'Intelligence Quotient (IQ)', 'NCT00588731'	Web	Loc
74	8527	1	68	68		Copy	16	1, 'Age', 'NCT01841021'	Web	Loc
75	4554	2	℃	°C		Copy	CDISC:C42549:Watt	ollowed by 30 Min After Brachial Plexus Block(Phase 1)', 'NCT02139982'	Web	Loc
76	4437	2	$	$		Copy	US$	1, 'Gross Weekly Purchasing of Fruits and Vegetables', 'NCT01509664'	Web	Loc
77	8165	1	Ω	Ω		Copy	CDISC:C42549:Watt	1, 'Effectiveness 3', 'NCT01688843'	Web	Loc
78	8079	1	%B	%B		Copy	CDISC-UCUM:C25613:%	'Homeostatic Model Assessment (HOMA)-%B by Visit', 'NCT00263328'	Web	Loc
79	13538	1	W	W		Copy	CDISC-UCUM:C42556:Wb	1, 'Muscle Strength', 'NCT00929500'	Web	Loc
80	4720	2	C	C		Copy	CDISC:C42559:C	hange From Baseline Temperature on Day 4 and Day 7', 'NCT02074358'	Web	Loc
81	20	1678	Days	Days		Copy	dy	36, 'Age', 'NCT00129129'	Web	Loc
82	321	44	Day	Day		Copy	dy	3, 'Terminal Phase Half-Life (T1/2)', 'NCT00394654'	Web	Loc

Future work

- **In the Foreseeable Future:**
  - Standardizing units from ClinicalTrials.Gov into into CDISC and UMLS and possible collaboration with the Simulation Interoperaility Standards Organization - SISO
  - Parallel efforts expanding SBML to represent disease models
  - The Reference Model is continuously absorbing new knowledge in the form of models and data to better map our cumulative computational understanding

A map that shows our computational gap of understanding of clinical trials is already possible! Much due to ClinicalTrials.Gov data availability



- **In the Much Farther Future:**
  - Computer comprehension of clinical data.
  - Allow applications such as:
    - Personal computerized medical adviser
    - Preventative medical predictor

Acknowledgments

- Thanks to CDISC consortium help
- Thanks to NIH persons who helped and specifically to:
  - Nick Ide from NLM ClinicalTrials.Gov team on advice
  - Erin E Muhlbradt from NCI for advice on CDISC unit data
- Thanks to the SBML community and DDMoRe consortium, specifically to:
  - Chris Myers, University of Utah, USA
  - Leandro Watanabe, University of Utah, USA
  - Lucian Smith, Caltech, USA
  - Jacek Swat, Simcyp Ltd. (Certara), UK

References

- **The Reference Model and ClinicalTrials.Gov**
  - J. Barhak, The Reference Model Models ClinicalTrials.Gov. SummerSim 2017 July 9-12, Bellevue, WA. Paper: https://doi.org/10.22360/SummerSim
  - J. Barhak, The Reference Model Visualizes Gaps in Computational Understanding of Clinical Trials, 2018 IMAG Futures Meeting March 21-22, 2018 @ NIH, Bethesda, MD. http://sites.google.com/site/jacobbarhak/home/Poster_IMAG_MSM2018_Map_Upload_2018_03_17.pdf
  - J. Barhak, The Reference Model: A Decade of Healthcare Predictive Analytics with Python, PyTexas 2017, Nov 18-19, 2017, Galvanize, Austin TX. Presentation: http://sites.google.com/site/jacobbarhak/home/PyTexas2017_Upload_2017_11_18.pptx Video: https://youtu.be/Pj_N4izLmsI
- **Related Future Standardization Efforts for Models and Data**
  - Jacob Barhak, Chris Myers, Leandro Watanabe, Lucian Smith, Maciek Jacek Swat , Healthcare Data and Models Need Standards. Simulation Interchangeability Standards Organization (SISO) 2018 Fall Innovation Workshop. 9-14 Sep 2018 Orlando, Florida. Presentation: http://sites.google.com/site/jacobbarhak/home/SISO_SIW_2018_08_14.pptx
  - L. Smith, M. J. Swat, J.Barhak . Sharing Formats for Disease Models. SummerSim 2016 24-27 July, Montreal, CA. Paper: https://doi.org/10.22360/SummerSim.2016.SCSC.010
  - L. Watanabe, J. Barhak, C. Myers, Towards Reproducible Disease Models using the Systems Biology Markup Language. Simulation 2018. http://dx.doi.org/10.1177/0037549718793214 Accompanying Source code: https://github.com/Jacob-Barhak/DiseaseModelsSBML