Title: Supervised Learning of Units of Measure

PI(s) Grant: N/A
Institution(s): N/A
Grant Number: N/A
Title of Grant: N/A

Abstract Authors:
Jacob Barhak
Joshua Schertz

Abstract Authors:

U.S. law requires registration of clinical trial data in ClinicalTrials.Gov. This NIH/NLM governed registry contributed much towards providing important modeling data information by accumulating over 300,000 clinical trials. However, despite the great effort by the government to centralize the data, the entities reporting data do not follow a predetermined standard. Therefore, numerical information entered is machine readable, yet not machine comprehensible, especially due to units being entered as free text. If a machine cannot comprehend the units, it cannot comprehend the numbers. This causes human intervention in the modeling process - slowing down modeling and the uses of this important registry.

The extent of the problem requires some machine learning, as of 12 Apr 2019, all 35,926 trials with results had 24,548 different units. The authors created solution infrastructure to address this problem. The solution includes:

1) Data extraction tools for ClinicalTrials.Gov that can index data and assemble clusters of data with unsupervised learning.

2) ClinicalUnitMaping.Com : a website for unit mapping that also demonstrates the extent of the problem.

3) A collection of existing unit standards used for medical purposes that currently holds data from CDISC, NIST / RTMMS / IEEE, Unit Ontology / Bio Portal, UCUM.

4) Supervised Machine Learning using neural networks that can predict the standardized unit given a non standard unit.

The supervised machine learning techniques are new, and their development involved many technical aspects and many attempts to solve the problem. This publication will discuss the difficulties and summarize multiple attempts, architectures, and solutions to resolve the problem.

The interavtive poster is accessible online through:
https://jacob-barhak.github.io/Poster_MSM_ML_IMAG_2019.html

Please use the 10 Simple Rules for Credible Models to describe your model:

1	Define context clearly	Discuss methods to use supervised machine learning so that machines can comprehend units of measure in a standardized manner
2	Use appropriate data	ClinicalTrials.Gov data collected by U.S. Law are used for training towards matching unit standard data
3	Evaluate within context	The products of this work are useful for clean and standardized data entry into models from clinical trials
4	List limitations explicitly	The current machine learning model is still under development. Unit mapping is still required and the final product accuracy still needs testing. When development is complete it will be useful to assist with data entry in models.
5	Use version control	Web site code is on Github and snapshots of data extraction and Machine Learning code is archived in files on a local development machine with backups.
6	Document adequately	Development is still under way - Multiple versions of this work have been published and made public here: https://clinicalunitmapping.com/about
7	Disseminate broadly	Publication had many details already: https://clinicalunitmapping.com provides a good view of the techniques and extent of the problem.
8	Get independent reviews	Multiple stakeholder that deal with unit standardization, including NIH/NIST/NATO and standardization organizations: SISO/CDISC and industry are aware of this project - an SBIR proposal was reviewed multiple times and feedbacks absorbed into the development process.
9	Test competing implementations	The approach in this project bundles several other unit standards - CDISC/RTMMS/UCUM/Unit Ontology - the machine learning approach does not replace them, it provides a unified solution that maps to those standards and therefore extends those rather than compete.
10	Conform to standards	CDISC/RTMMS/UCUM/Unit Ontology standards have been consulted. The Machine Learning techniques augment standardization.