Ensuring consistency of SEND Datasets with Study Reports using Machine Learning Algorithms

Suresh Madhavan, Raja Ramesh, Venkatesh Krishnan, Kurien Abraham, Mohit Mathew, Latha Prabakar
PointCross Life Sciences Inc.

ABSTRACT

Many factors in the SEND preparation process contribute to inconsistency with the authoritative and audited Study Report. But a persistent issue is the lack of standard terminology and consistent parsing of qualitative data such as in MI – Microscopic, MA-Macroscopic and CL-Clinical Observations that will improve quality and reduce costs.
This paper describes a continuously improving process using machine learning algorithms driven by a digital representation of the Study Report to provide recommendations automatically for parsing observations to STRESC, Modifiers and Severity.
The recommendation engine semantically recombines the SEND components to match the findings as reported in the Study Report allowing the automated comparator tool to check the consistency of the qualitative incidence counts and the quantitative data in SEND against the PDF Report

INTRODUCTION

Pathologist and veterinarians enter their original findings in Clinical Observations or Pathology data entry systems which is then mapped to ORRES in SEND data generation These ORRES are further split by rules or by Experts curating each unique terms, and converted to Base Pathological Process, Modifiers and Severity. Modifiers that are commonly used include organ-specific topography, distribution, character of the change and duration (Frame and Mann, 2008[3]).
These methods are not always reliable and consistent within or across studies. The FDA requires submission of Base Pathological Terms, Severity and modifiers, along with ORRES. To help reviewers understand the incidence count at different levels and attributes. The splitting of MI/MA lesions to the granularity as expect by FDA requires very scrupulous attention as it may easily compromise the ability to detect a test-article effect or may lead to the appearance of a test-article effect when none is actually present
Below is the illustration of FDA requirement of submitting Microscopic pathological Findings:

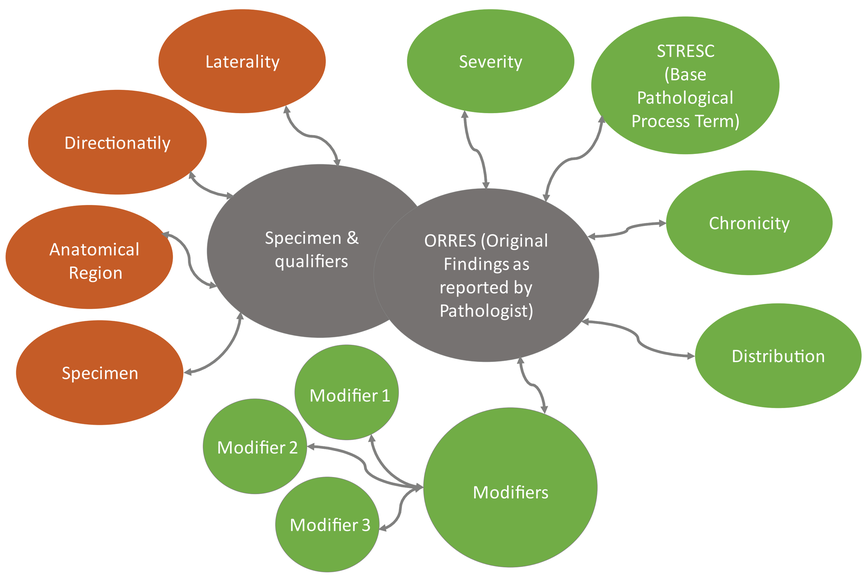
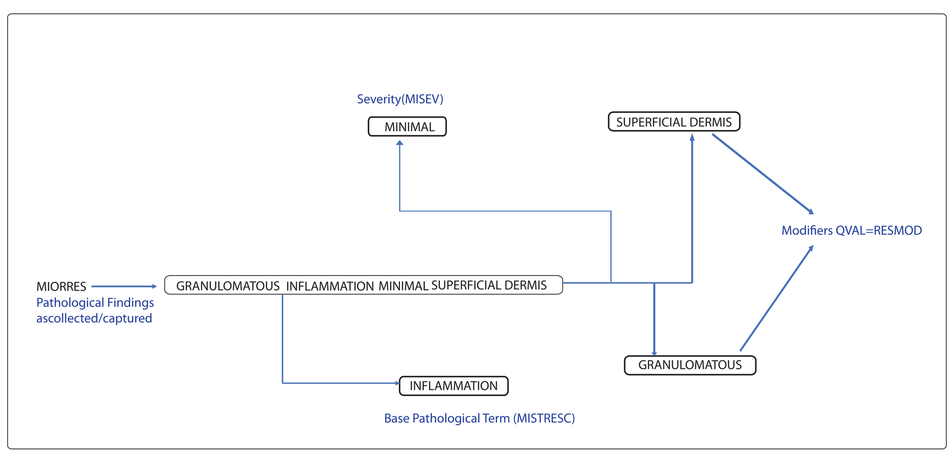


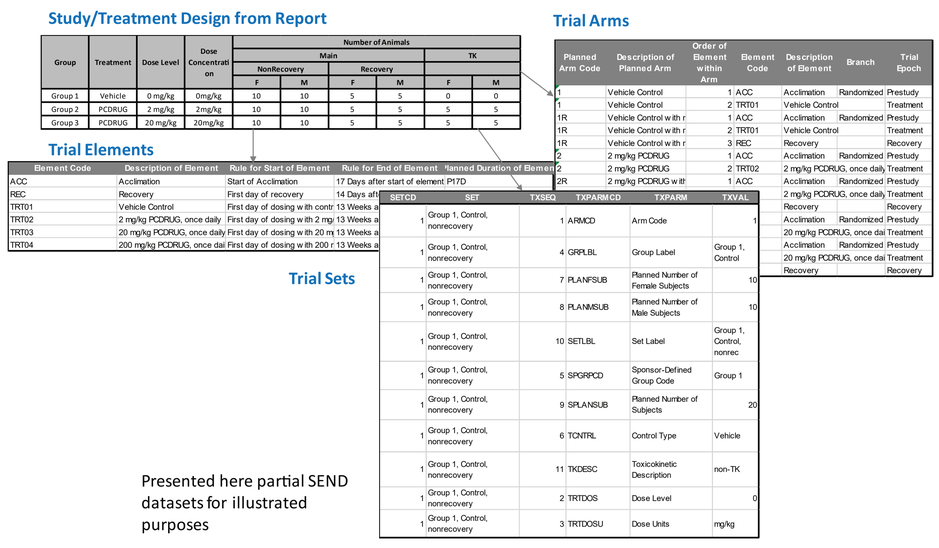
Illustration of MI/MA/CL Split Process and Mapping to SEND variables



A typical 1-month Toxicity study can easily comprise of approximately 25-75 unique MI Findings. And at times organ names can also be a part of ORRES that further compound the problem of segregating and assorting into different SEND variables

Disparate granularity of Study/Trial Design representation in Study Report versus SEND datasets

Below is the illustration of Study Design from PDF Reports and Trial Design domains as required by FDA:



Group	Treatment	Dose Level	Dose Concentration	Number of Animals					
				Main				TK	
				NonRecovery		Recovery			
				F	M	F	M	F	M
Group 1	Vehicle	0 mg/kg	0mg/kg	10	10	5	5	0	0
Group 2	PCDRUG	2 mg/kg	2mg/kg	10	10	5	5	5	5
Group 3	PCDRUG	20 mg/kg	20mg/kg	10	10	5	5	5	5

Planned Arm Code	Description of Planned Arm	Order of Element within Arm	Element Code	Description of Element	Branch	Trial Epoch
1	Vehicle Control	1	ACC	Acclimation	Randomized	Prestudy
1	Vehicle Control	2	TRT01	Vehicle Control		Treatment
1R	Vehicle Control w ith r	1	ACC	Acclimation	Randomized	Prestudy
1R	Vehicle Control w ith r	2	TRT01	Vehicle Control		Treatment
1R	Vehicle Control w ith r	3	REC	Recovery		Recovery
2	2 mg/kg PCDRUG	1	ACC	Acclimation	Randomized	Prestudy
2	2 mg/kg PCDRUG	2	TRT02	2 mg/kg PCDRUG, once daily		Treatment
2R	2 mg/kg PCDRUG w ith	1	ACC	Acclimation	Randomized	Prestudy
				2 mg/kg PCDRUG, once daily		Treatment
				Recovery		Recovery
				20 mg/kg PCDRUG, once dai		Treatment
				Acclimation	Randomized	Prestudy
				20 mg/kg PCDRUG, once dai		Treatment
				Recovery		Recovery

Element Code	Description of Element	Rule for Start of Element	Rule for End of Element	Planned Duration of Element
ACC	Acclimation	Start of Acclimation	17 Days after start of element P17D	
REC	Recovery	First day of recovery	14 Days af	
TRT01	Vehicle Control	First day of dosing w ith contr	13 Weeks a	
TRT02	2 mg/kg PCDRUG, once daily	First day of dosing w ith 2 mg	13 Weeks a	
TRT03	20 mg/kg PCDRUG, once daily	First day of dosing w ith 20 m	13 Weeks a	
TRT04	200 mg/kg PCDRUG, once dai	First day of dosing w ith 200 r	13 Weeks a	

SETCD	SET	TXSEQ	TXPARMCD	TXPARM	TXVAL
1	Group 1, Control, nonrecovery	1	ARMCD	Arm Code	1
1	Group 1, Control, nonrecovery	4	GRPLBL	Group Label	Group 1, Control
1	Group 1, Control, nonrecovery	7	PLANFSUB	Planned Number of Female Subjects	10
1	Group 1, Control, nonrecovery	8	PLANMSUB	Planned Number of Male Subjects	10
1	Group 1, Control, nonrecovery	10	SETLBL	Set Label	Group 1, Control, nonrec
1	Group 1, Control, nonrecovery	5	SPGRPCD	Sponsor-Defined Group Code	Group 1
1	Group 1, Control, nonrecovery	9	SPLANSUB	Planned Number of Subjects	20
1	Group 1, Control, nonrecovery	6	TCNTRL	Control Type	Vehicle
1	Group 1, Control, nonrecovery	11	TKDESC	Toxicokinetic Description	non-TK
1	Group 1, Control, nonrecovery	2	TRTDOS	Dose Level	0
1	Group 1, Control, nonrecovery	3	TRTDOSU	Dose Units	mg/kg

The increased granularity of Trial Design domains in SEND datasets pose the issue of mapping directly to the Sponsor Defined Dose Groupings in the Study Report and results in inconsistencies between the SEND datasets and Study Report such as differences in incidence counts and group mean data reported for sponsor dose groups.

Methodology

Parsing Original Pathological Findings to SEND Variables
Improving the quality and consistency in qualitative data with respect to SEND datasets by using a machine learning technique that can adopt to the internally developed training set.

Data Collection: A corpus was created taking unique observations/findings reported in the SEND datasets (PointCross synthesized and anonymized datasets)

Data Preprocessing : Sequential Based approach
Text is tokenized by removing the case differences and special characters and correcting the spacing. Knowledge from PointCross maintained global CT, Ontologies and CDISC CT is used to identify phrases (multi-word constructs) assingle entity for further processing, and identifying negation and double-negatives. The above output is used to create vector representation of the words using Word2Vec and the embedding layer is created by using Skip gram architecture and negative samples method. The output from this exercise is a dense vectors representations of words/phrases that retain the natural relationship between words in multi-dimensional space based on pathological descriptions and curated ontologies. A deep convolutional neural network is created and trained using the text from the training set and input is added using the neural embedding created above.

Pattern Based Approach
The data was processed to get the list of unique words as columns and list of words as rows in a matrix form with 0 or 1 as values, where 0 indicates the absence of that particular word in standard variable and 1 indicates its presence. MLP-Multilayer Perceptron belongs to a class of fully connected feed forward networks where each neuron is connected with other neurons at every next layer and it uses the supervised learning technique called back-propagation also known as backward propagation of error as a generic model with the following parameters for the training set:

- **A learning function** with a suitable learning rate between 0 and 0.2. If the function is taking time to converge, the learning rate is too small (may be close to 0). If the function fails to converge, the learning rate is too big (may be close to 1).
- **The maximum output difference** which measures how much error between output and target value. Basically, this parameter takes care of the model so that it is not over-fitting.
- **The initial function** with random weights between -0.3 and 0.3

Also, a network of associated words is built to support the supervised learning model in order to get reliable result

Results

We used a sample study PC201708 (http://info.pointcrosslifesciences.com/mysend) to test the above mentioned methods. The test study data is 13 –week repeat dose toxicity study conducted in rats, consists of a total of 4217 MI findings of which 92 are unique.

Pattern Based Approach
We used punctuation as a separator to create phrase by word matrix and kept a cut-off score of 0.6 for any term to appear in the output. The study has 154 unique terms out of which only 82 terms were selected based on the cutoff or due to the absence of those terms in the training set. The confusion matrix represents terms that are classified by MLP with the set-up as mentioned above and gives an accuracy of 89%

Confusion Matrix

Test Data – PC201708

Predicted Data	MISTRESC	MISPEC	MIANTREG	MISEV	QVAL	Total	Recall
MISTRESC	24	0	0	0	1	**25**	**0.96**
MISPEC	0	20	4	0	2	**26**	**0.77**
MIANTREG	0	0	0	0	0	**0**	**NaN**
MISEV	0	0	0	3	0	**3**	**0**
QVAL	1	0	1	0	26	**28**	**0.93**
Total	**25**	**20**	**5**	**3**	**29**		
Precision	**0.96**	**1**	**0**	**1**	**0.90**		

Accuracy	**0.89**

Sequential Based Approach
Using this approach we were able to get an accuracy of ~76% (using a cut off value of 0.6).
The learning capability will be enhanced in NN in future process which helps to increase its efficiency with good performance evaluation.

Conclusion

The industry has faced many challenges in being able to prepare SEND data sets with reliability, quality and at a reasonable cost. In this paper we have described how neural network techniques can be used in recommending and preparing SEND ready datasets. Based on our understanding, MLP gives better performance due to their ability to recognize patterns. Neural Networks is a tool which facilitates Sponsors and CROs in getting ready with SEND datasets. The critical evaluation of the MLP outputs are continuously improving and can contribute greatly to cost effective and responsive services for SEND Data.

References

- https://www.fda.gov/forindustry/datastandards/studydatastandards/default.htm
- http://www.phusewiki.org/docs/2016_Tokyo_SDE/SEND%20for%20Pathologists%20and%20Toxicologists.pdf
- Frame S.R. and Mann P.C. (2008). Principles of Pathology for Toxicology Studies. In: Principles and Methods of Toxicology, Fifth Edition (Hayes A.W, ed), pp 591-609,CRC Press. Boca Raton.
- https://code.google.com/archive/p/word2vec/
- Andrew Tomlinson. Medical applications for pattern classifiers and image processing. http://www.railwaybridge.co.uk, accessed April 27, 2005, 2000.