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

Suresh Madhavan, Raja Ramesh, Venkatesh Krishnan, Kurien Abraham, Mohit Mathew and Latha Prabhakar
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

Background

SEND is currently the preferred submission format for the US FDA and became a mandate on 18th December 2017[1]. In SEND, all findings (MI-Microscopic Findings, MA-Macroscopic Findings, CL-Clinical Observations) as reported by Pathologists/Toxicologist are needed, these original as collected observations are split appropriately and standardized to controlled terminology and then mapped to separate columns in SEND domains. These Result or Findings as Collected (ORRES) contain terms including base pathological process, severity, modifiers and at times specimen type.

The Global standard terminologies and ontologies for pathological findings are often limited. INHAND (International Harmonization of Nomenclature and Diagnostic Criteria) initiative provides a standardized nomenclature and differential diagnosis for classifying microscopic lesions observed in laboratory rats and mice in toxicity. It is often restricted to rats and mice and, as of August 2016, the total coverage was only 78% (1247 terms) of standard nomenclature for microscopic lesions, in addition to the fact that it is often restricted to only rats and mice and still needs to be developed for other species (currently it is not the scope of the project[2]). The published terms are still insufficient and INHAND itself still need to add and rectify. There is no global standard terminology for clinical observations and macroscopic findings and it is not even in the pipeline of development.

Similarly another challenge is SEND accepts a framework of study design in TE-Trial Elements, TA-Trial Arms and TX-Trial Sets domains. The trial design of a SEND dataset is designed to be machine-readable and it is very granular. It takes into account all variations in the trial arm of each dose group leading to more trial sets of fewer and more narrowly similar subjects having common set of experimental and sponsor defined parameters. Whereas study reports generally have study design defined at dose/treatment groups. SEND also requires standardized way of explicitly stating certain decision rules that may not be seamless in the study Report.

Hence increased granularity of SEND trial design domains compared to sponsor dose group assignment and the need of explicit rendition of base pathological process and modifiers introduce challenges when sponsors trying to create SEND datasets. This also leads consistency and quality issues between the SEND datasets and Study Report

Business Problem

In general Pathologist and veterinarians either 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 certain automation rules and patterns, 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]).

The Base Pathological terms and modifiers that are created either by the rule based approach or by Experts curating each unique terms. Often times these methods are not reliable and inconsistent within or across studies. The FDA requisite of submitting Base Pathological Terms, Severity and modifiers along with ORRES aids dynamic way of performing incidence count at different levels. As certain base pathological terms appear in more than an organ, the toxicity pattern of a drug in one or more targets organs will be exposed out easily when taking Base terms separately along with severity for group summary incidence tabulation. 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

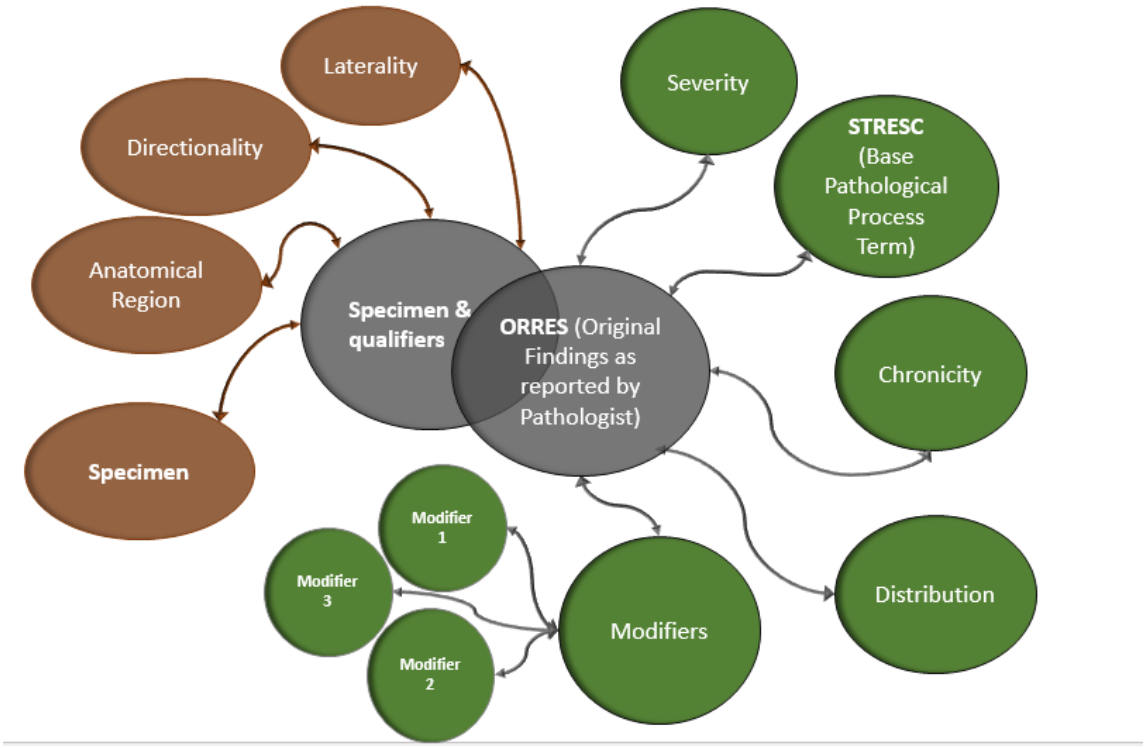
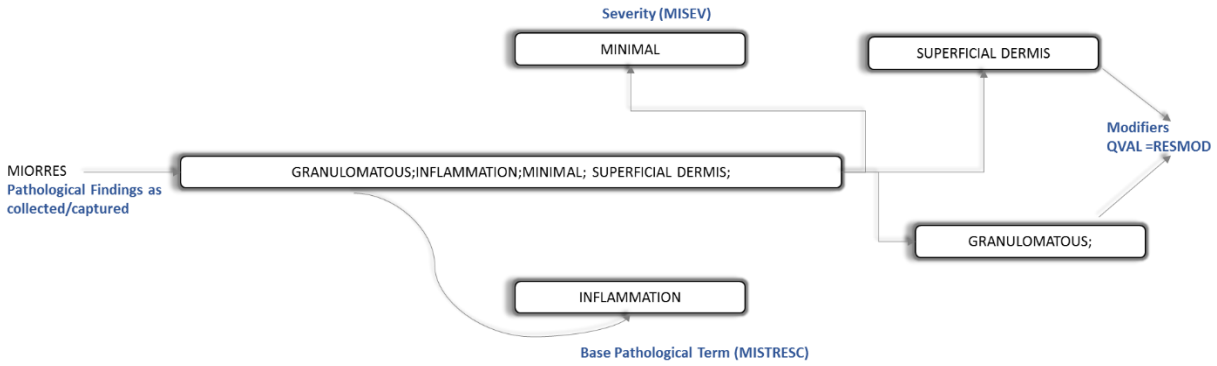


Illustration of MI/MA/CL Split Process and Mapping to SEND variables



A 1-month typical 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. The terminologies used and framework of representation of study design at the sponsor dose group level in PDF report is distinctive from Trial design representation in SEND datasets.

Study/Treatment Design from Report

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

Trial Elements

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 aft	
TRT01	Vehicle Control	First day of dosing with contr	13 Weeks a	
TRT02	2 mg/kg PCDRUG, once daily	First day of dosing with 2 mg/	13 Weeks a	
TRT03	20 mg/kg PCDRUG, once daily	First day of dosing with 20 m	13 Weeks a	
TRT04	200 mg/kg PCDRUG, once dai	First day of dosing with 200 r	13 Weeks a	

Trial Arms

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 with r	1	ACC	Acclimation	Randomized	Prestudy
1R	Vehicle Control with r	2	TRT01	Vehicle Control		Treatment
1R	Vehicle Control with 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 with	1	ACC	Acclimation	Randomized	Prestudy
				2 mg/kg PCDRUG, once daily		Treatment
				Recovery		Recovery
				Acclimation	Randomized	Prestudy
				20 mg/kg PCDRUG, once dai		Treatment
				Acclimation	Randomized	Prestudy
				20 mg/kg PCDRUG, once dai		Treatment
				Recovery		Recovery

Trial Sets

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

Presented here partial SEND datasets for illustrated purposes

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. This potentially introduce inconsistencies of SEND datasets with Study Reports. This results in differences in incidence counts and group mean data reported for sponsor dose groups. Hence industry needs an easy way of automated generation and suggestion of SEND trial design from Study Report and being able to compare with submitted SEND trial design that indirectly facilitates in comparing SEND data values against summarized data in study reports.

This paper reports a continuously improving process using machine learning algorithms (MLA) recommendations in classification of original observations to STRESC, Severity and Modifiers.

Methodology

In order to improve the quality and consistency in qualitative data with respect to SEND datasets, a generic model based technique would be a better suited solution. A machine learning technique that can adopt to the internally developed training set for improving the quality and consistency in qualitative data with respect to SEND datasets.

Parsing Original Pathological Findings to SEND Variables

Data Collection:

The PointCross synthesized and anonymized findings were used for developing the training set which includes --STRESC, --ANTREG, --SEV, --SPEC, --LAT, --DIR, --REASND and QUAL as part of SEND standard columns for each domains such as Macroscopic observations (MA), Microscopic observations (MI) and Clinical Observations (CL).

Data Pre-processing & applying available Semantics:

Sequential Based approach

- Text is tokenized by removing the case differences and special characters, correcting the spacing.
- Knowledge from existing PointCross maintained global CT, Ontologies and CDISC CT is used to identify phrases (multi-word constructs) as single entity for further processing, and identifying negation and double-negatives. (**Example 1.** "Gland, Adrenal", "Adrenal Gland", "Gland: adrenal" and other forms represented as "gland_adrenal", **Example 2.** "No visible lesions noted", "No significant pathologic

alterations", "No abnormalities detected", "NO NECROPSY OBSERVATIONS", "With Normal Limits" "NORMAL")

Creating Neural Word Embeddings and Convolutional Neural network for classifying ORRES to standard variables:

- We have used the resulting output from the above processing to create a vector representation of the words using Word2Vec ([4]).
- After a couple of trails we have decided to use the skip-gram architecture and negative samples methods to create the embedding layer.
- Output from this unsupervised learning 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 created by humans.
- 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

Feature Extraction: 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 that particular word is not present in standard variable and indicates that particular word is present in the standard variable.

Method: 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 & Discussions

The classification of ORRES to STRESC and other modifiers is generally a tedious effort and manual classification suffers from many inconsistencies which lead to leverage of machine learning algorithms. Neural network have proven to be useful in Pharmaceutical Research and in many other different clinical applications using pattern recognition; for example: diagnosis of breast cancer, interpreting electrocardiograms, diagnosing dementia, predicting prognosis and survival rates[5].

Here in this paper, to test our methodological approach we took PC201708 sample synthesized SEND study (Downloadable: http://info.pointcrosslifesciences.com/mysend).

- 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.
- We used punctuation as a separator to create phrase by word matrix.
- We kept 0.6 cut-off score for any term to appear in the output

Pattern Based Approach

After data preprocessing, we extracted 154 unique terms/phrases as the input for machine learning engine. With the current set-up, the MLP engine classified and predicted about 82 unique terms.

The reason for other remaining terms not appearing in the MLP output is, those terms do not have an appearance in the training dataset we used, and terms showed less score than the defined cut-off value of 0.6.

The confusion matrix represents terms that are classified by MLP with the set-up as mentioned above:

Classification Methods	Methodology	Data Processing	R Library	Accuracy (%)
Neural Network	Multilayer Perceptron	Punctuation (comma separator)	RSNNS	89

Confusion Matrix								
		Test Data – PC201708						
		MISTRESC	MISPEC	MIANTREG	MISEV	QVAL	Total	Recall
Predicted Data	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. Here in this paper we have described how computational techniques such as MLP 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. NN 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

1. https://www.fda.gov/forindustry/datastandards/studydatastandards/default.htm
2. http://www.phusewiki.org/docs/2016_Tokyo_SDE/SEND%20for%20Pathologists%20and%20Toxicologists.pdf

3. 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.
4. https://code.google.com/archive/p/word2vec/

5. Andrew Tomlinson. Medical applications for pattern classifiers and image processing. http://www.railwaybridge.co.uk, accessed April 27, 2005, 2000.

Appendix:

References of Corpus building

1. **NTP Atlas of Nonneoplastic Lesions in Rats and Mice:** Left side navigation panel

   https://ntp.niehs.nih.gov/nnl/index.htm

2. **NHAND nomenclature:**

   https://www.goreni.org/docs/INHAND_nomenclature.pdf

3. http://lab-ally.com/histopathology-resources/histopathology-glossary/
4. **International Harmonization of Rat Nomenclature:**

   https://reni.item.fraunhofer.de/reni/rat_nomenclature/index.htm

5. **Clinical Observations:**

http://onlinelibrary.wiley.com/doi/10.1002/9780470464151.app2/pdf

6. **Mouse Adult Gross Anatomy Ontology**

https://bioportal.bioontology.org/ontologies/MA?p=classes