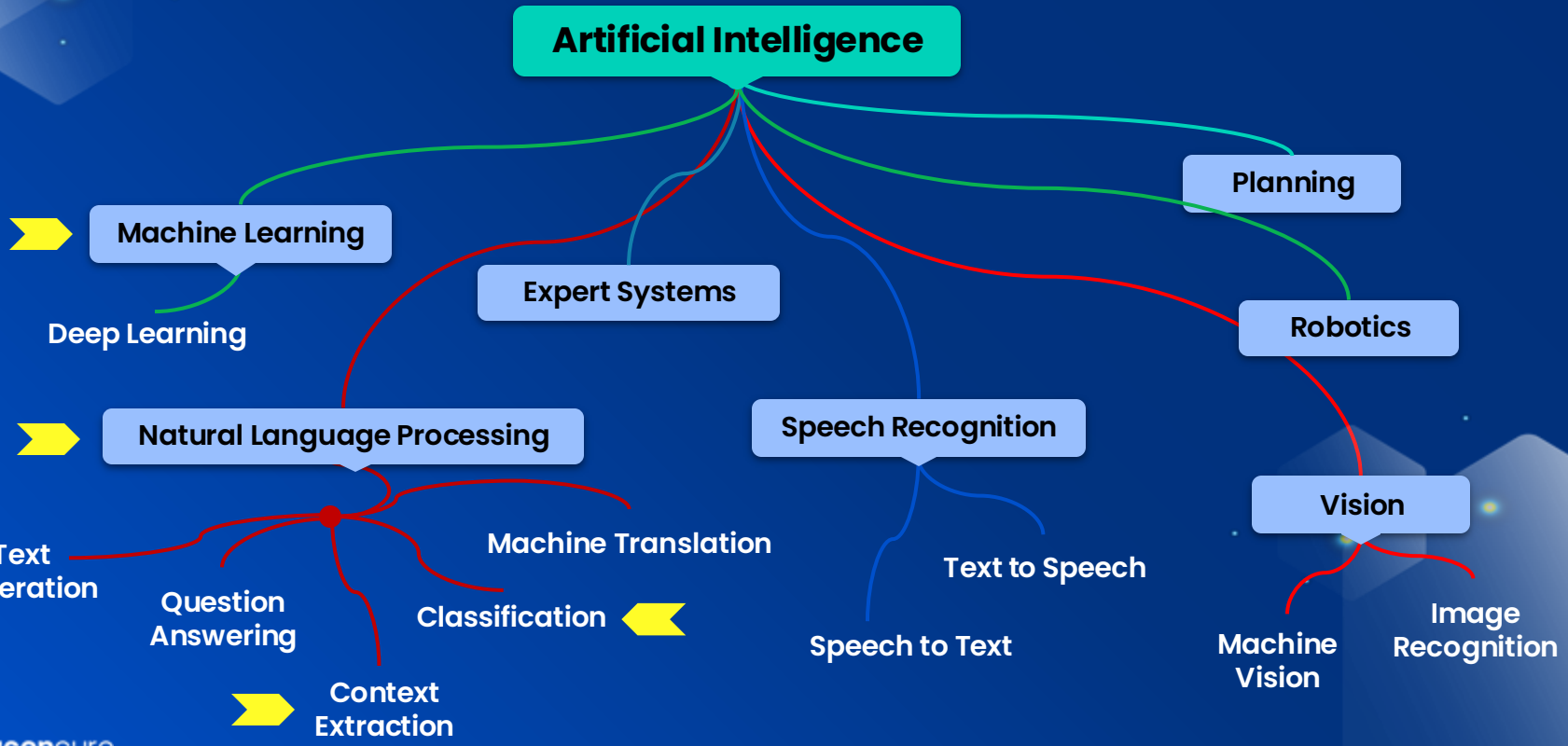
Beaconcure

Augmented Metadata and the Expanding Role of Machine Learning (ML) and Natural Language Processing (NLP) in TLF Generation and Validation

Steve Ross, Phuse SDE (Cary, US)

Nov 21, 2024

The Map of AI: Where do Machine Learning and Natural Language Processing fit in?



Beaconcure

Defining the problem: Challenges in Current TLF Generation and Review

- **Legacy Systems:** Outdated systems hinder integration and are often incompatible with modern automation technologies.
- **Resistance to Change (Legacy mindsets):** Cultural inertia within organizations can obstruct the adoption of innovative automation practices in clinical research.
- **Data Quality Concerns:** Inconsistent data quality across studies complicates analysis and limits the effectiveness of automated processes.
- **Increasing Complexity:** The growing intricacy of clinical trial data demands enhanced management and increases reliance on automation technologies.
- **Manual Efforts Persist:** Despite advancements, substantial manual efforts remain necessary for TLF generation, hindering efficiency and productivity.
- **Human Error Risks:** Inconsistent review processes and human involvement raise the likelihood of errors, compromising data integrity and quality.

Harnessing Machine Learning and Natural Language Processing

Machine Learning (ML) - using algorithms and models that "learn" from clinical data to automate processes that guarantee the accuracy, consistency, and quality of trial outputs. ML systems recognize patterns, detect anomalies, and streamline tasks when creating and validating tables, listings, and figures (TLFs).

ML Automation in TLF Creation: Machine learning automates repetitive tasks such as mapping variables, enhancing speed and accuracy of TLF generation.

Efficiency Boost: By reducing manual coding, ML enables programmers to focus on complex analyses, expediting the overall process.

Error Reduction: Utilizing ML algorithms minimizes human errors by providing consistent variable identification and data validation responses.

Harnessing Machine Learning and Natural Language Processing

Natural Language Processing (NLP) applies computational techniques to interpret, analyze, and extract meaningful information from unstructured text (clinical study protocols, regulatory guidelines, and annotations within tables, listings, and figures (TLFs). NLP enables automated review and quality checks by identifying and resolving inconsistencies, standardizing terminology.

NLP's Role in AI Interpretation: Natural Language Processing enables AI systems to decode unstructured clinical text, facilitating data extraction.

Enhancing Document Understanding: NLP algorithms analyze complex medical terminologies, improving comprehension of clinical documents across formats.

Streamlining Data Processing: Leveraging NLP allows rapid processing of vast textual datasets, enhancing automation workflows and efficiency.

The Role of Metadata in Clinical Trials

Metadata in Clinical Trials: Metadata serves as a crucial framework, enabling organized data management throughout the clinical trial lifecycle.

Standardization Enhancement: Leveraging metadata fosters standardization across various studies, ensuring consistent interpretations and reducing errors.

Streamlined Processes: Utilizing metadata facilitates automation in reporting, analysis, and compliance, significantly improving overall efficiency.

ML, NLP Application targets in the production, validation, and review of statistical outputs

01 Obtaining metadata through parsing

02 Comparing data and metadata at scale and at speed.

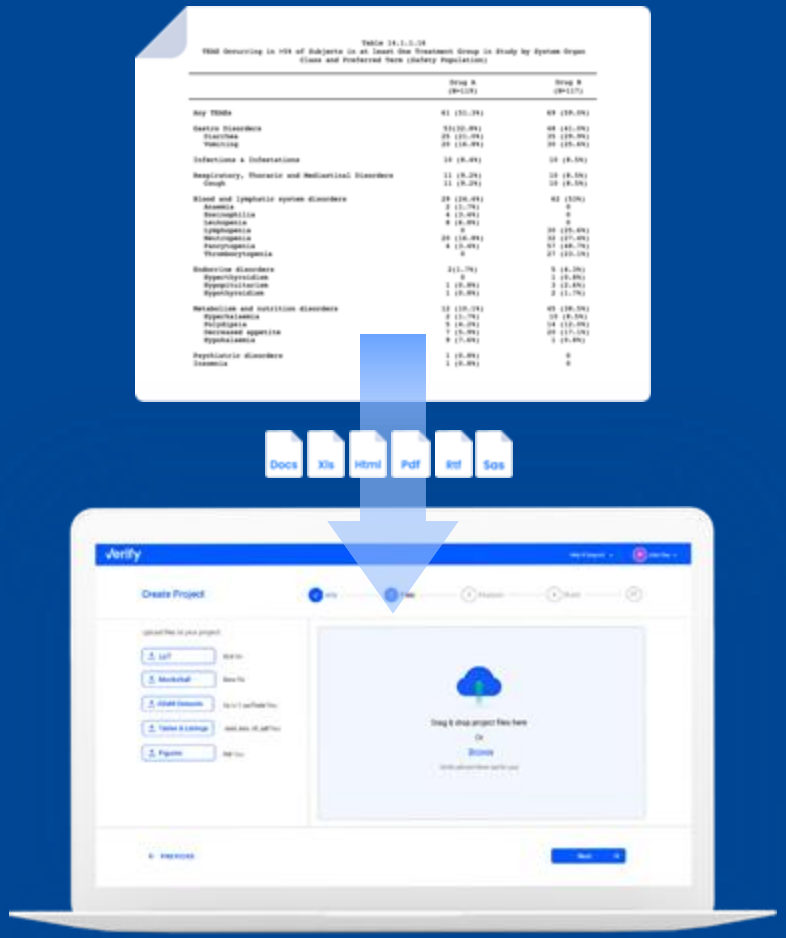
03 Understanding context

01 Obtaining metadata through parsing

Upload all your delivery files (TLFs, Table of Contents, TLF shell documents) to Verify application, or link directly through an API

Application parses your files, deconstructing down to the cell level, categorizing and linking objects

This enables us to ‘make connections’ in the table set. To a standard, to multiple places in a table or table set, even across studies in the case of ISS/ISE



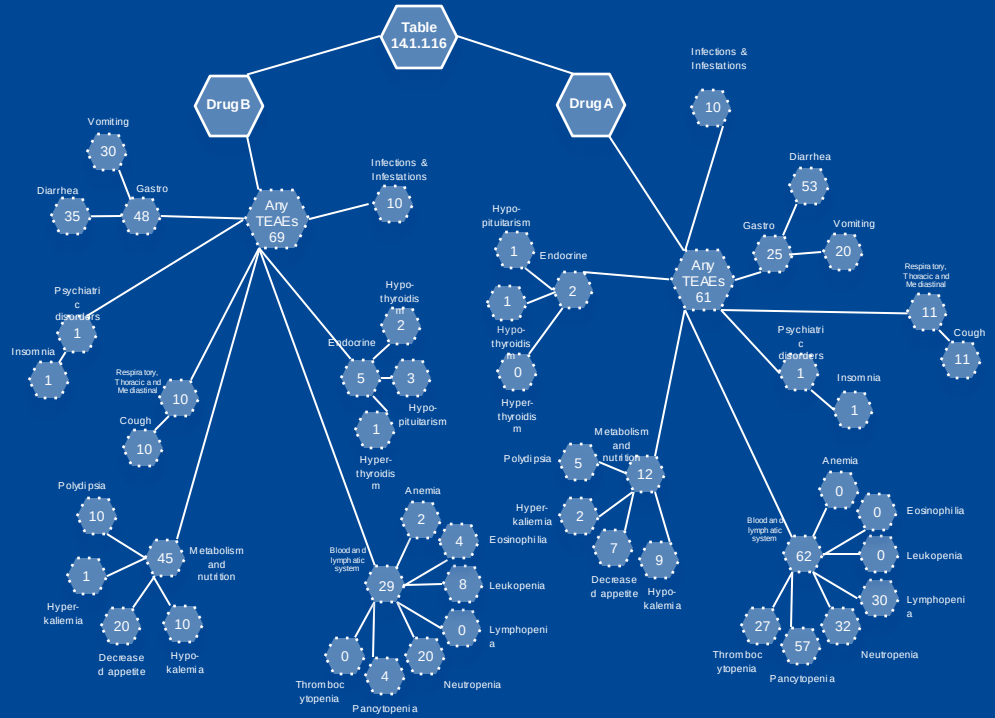
Beaconcure

AI-Enabled Parsing of TLFs into a Robust Structured Database

Table 14.1.1.16

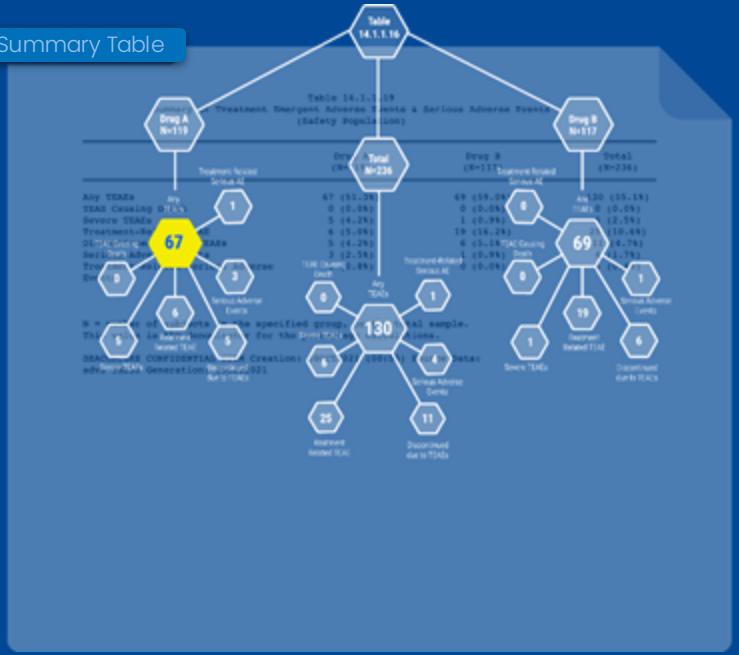
TEAE Occurring in >5% of Subjects in at least One Treatment Group in Study by System Organ Class and Preferred Term (Safety Population)

	Drug A (N=119)	Drug B (N=117)
Any TEAEs	61 (51.3%)	69 (59.0%)
Gastro Disorders	53(32.8%)	48 (41.0%)
Diarrhea	25 (21.0%)	35 (29.9%)
Vomiting	20 (16.8%)	30 (25.6%)
Infections & Infestations	10 (8.4%)	10 (8.5%)
Respiratory, Thoracic and Mediastinal Disorders	11 (9.2%)	10 (8.5%)
Cough	11 (9.2%)	10 (8.5%)
Blood and lymphatic system disorders	29 (24.4%)	62 (53%)
Anaemia	2 (1.7%)	0
Eosinophilia	4 (3.4%)	0
Leukopenia	8 (6.8%)	0
Lymphopenia	0	30 (25.6%)
Neutropenia	20 (16.8%)	32 (27.4%)
Pancytopenia	4 (3.4%)	57 (48.7%)
Thrombocytopenia	0	27 (23.1%)
Endocrine disorders	2(1.7%)	5 (4.3%)
Hyperthyroidism	0	1 (0.8%)
Hypopituitarism	1 (0.8%)	3 (2.6%)
Hypothyroidism	1 (0.8%)	2 (1.7%)
Metabolism and nutrition disorders	12 (10.1%)	45 (38.5%)
Hyperkalaemia	2 (1.7%)	10 (8.5%)
Polydipsia	5 (4.2%)	14 (12.0%)
Decreased appetite	7 (5.9%)	20 (17.1%)
Hypokalaemia	9 (7.6%)	1 (0.8%)
Psychiatric disorders	1 (0.8%)	0
Insomnia	1 (0.8%)	0

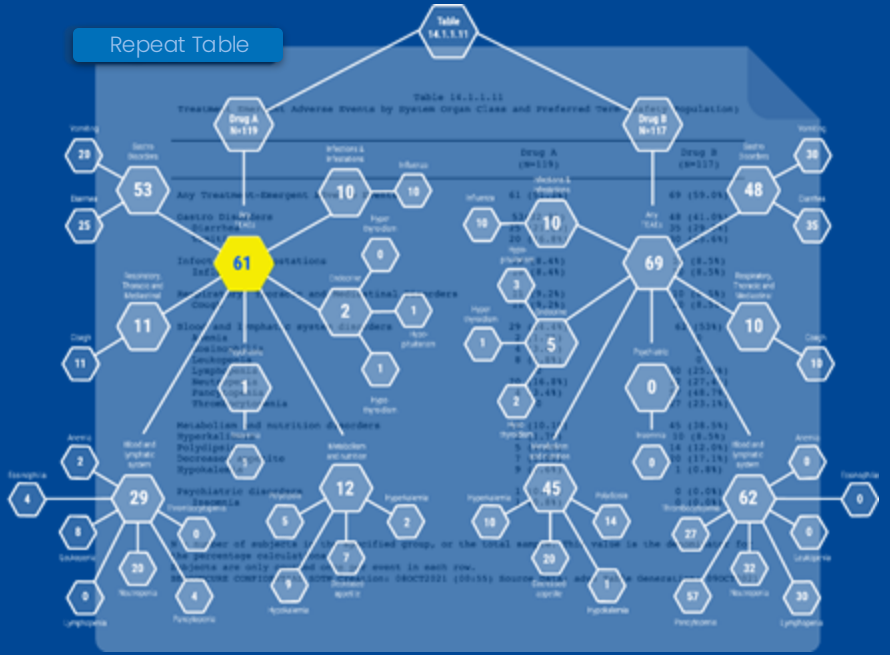


02 Fully parsed data facilitates semantic understanding of different tabular data

Summary Table



Repeat Table



Validation

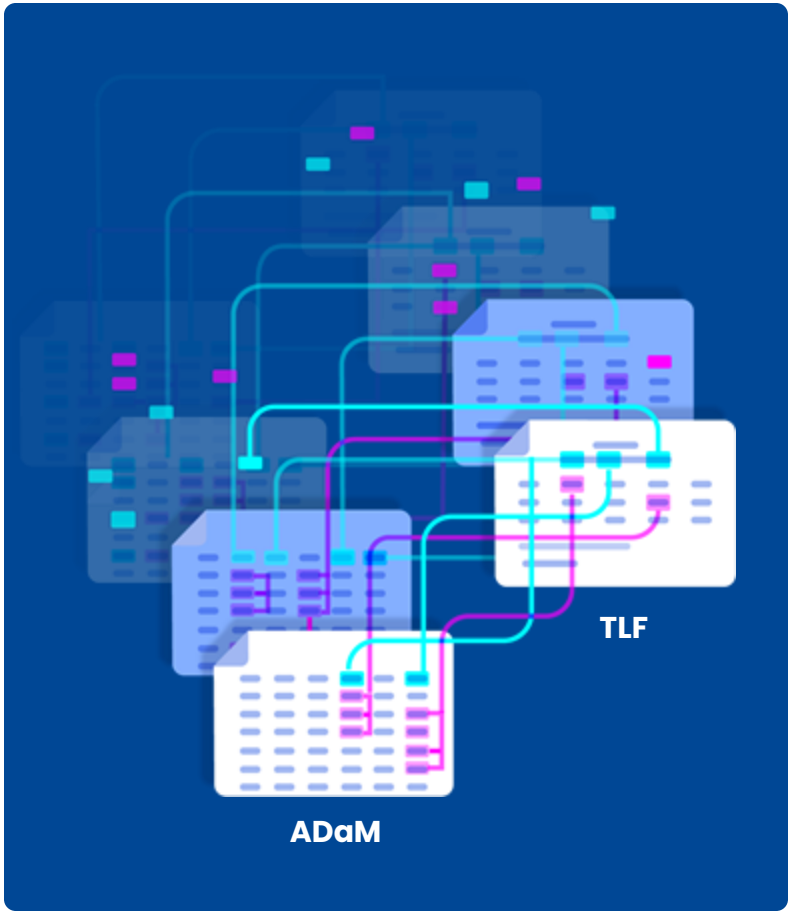
Collaboration

Architecture

03 Understanding Context

Extensive AI-Enabled Validation Check Library

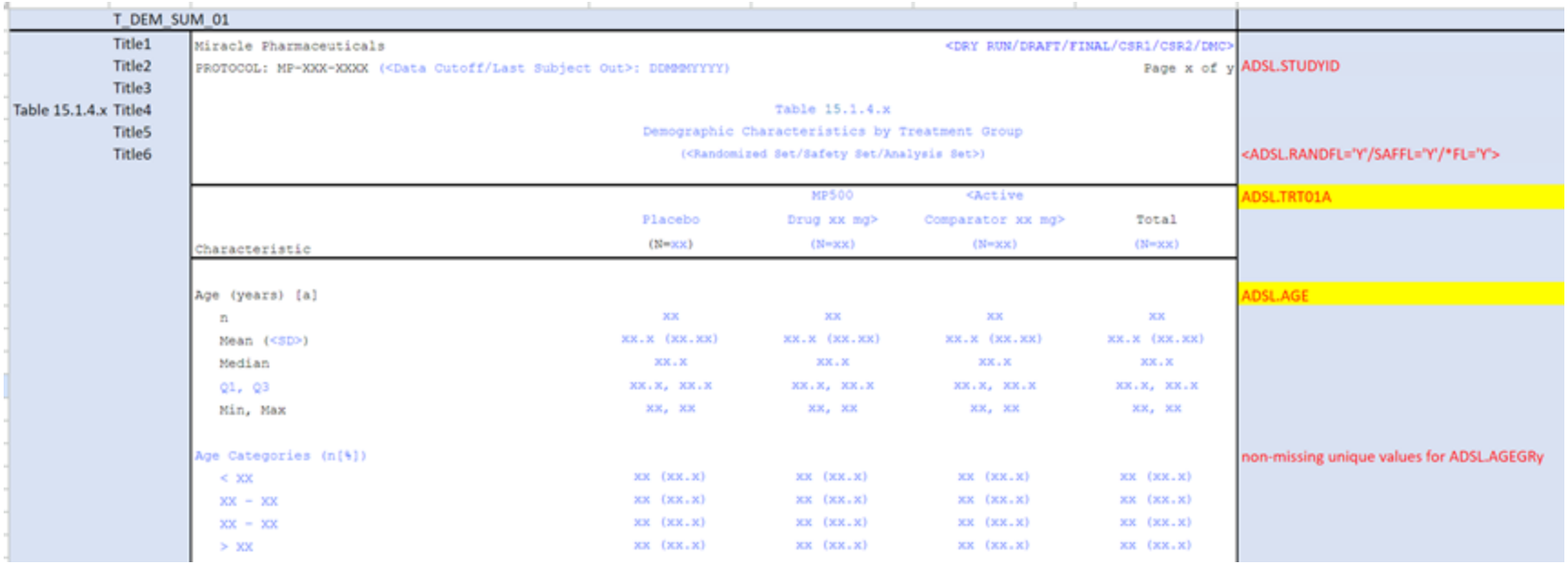
- ✓ **Format Checks**
  Validate format, titles, and footnotes
- ✓ **Reference Checks**
  Validate consistency of TLFs vs. reference documentation (mock shell & LoT)
- ✓ **Within-Table Checks**
  Validate arithmetic and hierarchies within TLFs
- ✓ **Cross-Table Checks**
  Validate cross-table consistency
- ✓ **Table vs. ADaM Checks**
  Validate consistency of TLFs vs. ADaM



Augmented Metadata in TLF development

- Now that we have deconstructed our data artifacts
  - TLF mock shells
  - ToC's
  - Output objects (.rtf, .html, .pdf)
  - ADaMs
- and have reassembled them with context, we have a metadata store than links results objects with method objects - "augmented metadata"

Standard Demographics Table Shell



	T_DEM_SUM_01					
Title1	Miracle Pharmaceuticals				<DRY RUN/DRAFT/FINAL/CSR1/CSR2/DMC>	
Title2	PROTOCOL: MP-XXX-XXXX (<Data Cutoff/Last Subject Out>: DDMMMYYYY)				Page x of y	ADSL.STUDYID
Title3						
Table 15.1.4.x Title4	Table 15.1.4.x					
Title5	Demographic Characteristics by Treatment Group					
Title6	(<Randomized Set/Safety Set/Analysis Set>)					<ADSL.RANDFL='Y'/SAFFL='Y'/*FL='Y'>
			MP500	<Active		ADSL.TRT01A
		Placebo	Drug xx mg>	Comparator xx mg>	Total	
	Characteristic	(N=xx)	(N=xx)	(N=xx)	(N=xx)	
	Age (years) [a]					ADSL.AGE
	n	xx	xx	xx	xx	
	Mean (<SD>)	xx.x (xx.xx)	xx.x (xx.xx)	xx.x (xx.xx)	xx.x (xx.xx)	
	Median	xx.x	xx.x	xx.x	xx.x	
	Q1, Q3	xx.x, xx.x	xx.x, xx.x	xx.x, xx.x	xx.x, xx.x	
	Min, Max	xx, xx	xx, xx	xx, xx	xx, xx	
	Age Categories (n[%])					non-missing unique values for ADSL.AGEGRy
	< XX	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)	
	XX - XX	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)	
	XX - XX	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)	
	> XX	xx (xx.x)	xx (xx.x)	xx (xx.x)	xx (xx.x)	

Standard Demographics Table

Table 15.1.4.1
Demographics
Safety Analysis Set

Characteristic	MP500 Single Agent Dose Escalation			
	MP500 8 ug/kg (N=3)	MP500 16 ug/kg (N=5)	MP500 16 ug/kg + 1 DEX (N=3)	MP500 16 ug/kg + 2 DEX (N=3)
Age (years) [a]				
n	3	5	3	3
Mean (SD)	66.7 (6.11)	70.6 (5.81)	57.3 (11.68)	57.0 (8.89)
Median	68.0	68.0	55.0	54.0
Min, Max	60, 72	65, 80	47, 70	50, 67
Age categories [a] [n (%)]				
18 to 64	1 (33.3)	0	2 (66.7)	2 (66.7)
65 to 84	2 (66.7)	5 (100)	1 (33.3)	1 (33.3)
85 or above	0	0	0	0
Sex [n (%)]				
Male	3 (100)	5 (100)	2 (66.7)	2 (66.7)
Female	0	0	1 (33.3)	1 (33.3)
Missing	0	0	0	0
Ethnicity [n (%)]				
Hispanic or Latino	0	1 (20.0)	1 (33.3)	0
Not Hispanic or Latino	3 (100)	4 (80.0)	2 (66.7)	3 (100)
Not Reported	0	0	0	0

DEX = Dexamethasone
[a] Age is calculated as the integer part of (Date of informed consent - Date of birth) / 365.25.
[b] Body Mass Index is defined as weight(kg) / [height (m)]2.

Analysis Results Data (+) Metadata

Identifiers		Group Vars			Result Vars		Statistics		
Number	Title	Dataset	Variable	Value	Variable	Label	Value	Name	Label
15.1.4.1	Demographics Safety Analysis Set	ADSL	TRT01a	MP500 8 ug/kg (n=3)	AAGE	Age (years) [a]	3	Count	N
15.1.4.1	Demographics Safety Analysis Set	ADSL	TRT01a	MP500 8 ug/kg (n=3)	AAGE	Age (years) [a]	66.7 (6.11)	Mean (standard deviation)	Mean (SD)
15.1.4.1	Demographics Safety Analysis Set	ADSL	TRT01a	MP500 8 ug/kg (n=3)	AAGE	Age (years) [a]	68	Median	Median
15.1.4.1	Demographics Safety Analysis Set	ADSL	TRT01a	MP500 8 ug/kg (n=3)	AAGE	Age (years) [a]	60, 72	Minimum, Maximum	Min, Max

AI-Enabled Validation Techniques

AI-Driven Validation Speed: Accelerates the validation process for output tables against ADaM datasets compared to traditional methods.

Accuracy Enhancement: Utilizing AI reduces human errors through automated checks, ensuring higher accuracy in data validation outcomes.

Resource Efficiency: AI-driven tools streamline validation tasks, freeing up resources for more strategic analytical activities within teams.

Through the looking glass - What does a future-state development and review cycle look like with the integration of ML and NLP into an organization's process map?

Expanding Role for ML Applications: Emerging ML applications aim to shorten deliverable timelines by automating more and more complex decision-making in clinical trials. Smarter models facilitate this.

NLP for Real-Time Insights: Natural Language Processing enables linking of analysis methods to study data, enabling faster design and analysis of clinical outputs.. Better parsing leads to better decision-making.

Integration of Adaptive Algorithms: Adaptive algorithms promote comprehensive learning, not just faster results checking. Semantic meaning!

Beaconcure

Key Takeaways from Automation Approaches

Transformative Automation Benefits: ML, NLP, and automation tools enhance TLF generation, improving efficiency, accuracy, and collaboration within clinical trials.

Streamlined Validation Process: Using AI-driven validation with tools like Verify increases consistency and significantly reduces time spent on manual checks.

Resource Optimization: Rebalance the workload of man and machine by allocating resources to strategic tasks by minimizing repetitive manual data handling.

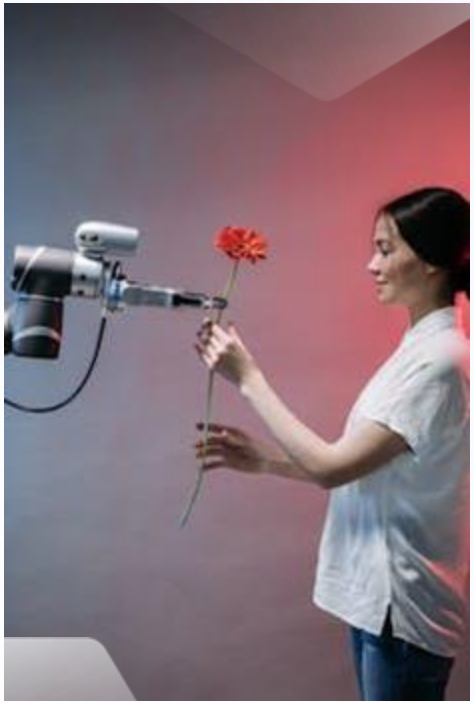


Photo by Pavel Danilyuk on Pexels

Thank you!

Contact:

Steve Ross

Director of Statistics and Life Sciences Consultancy

Beaconcure

Steve@Beaconcure.com