Paper ET-19 (Emerging Trends)

Evaluating Safety Compliance in Clinical Trials by Leveraging Patient Narratives and Deep Learning

Sherrine Eid, SAS Institute, Philadelphia, U.S.
Sundaresh Sankaran, SAS Institute, Cary, U.S.

ABSTRACT

Adverse Events (AEs) during clinical trials require identification, classification, risk quantification and treatment. Information embedded within clinical trial narratives can enhance understanding and insights. Leveraging best-in-class Artificial Intelligence (AI) methods helps us accurately and precisely assess Adverse Events, safety signals and physician compliance from patient narratives. Our analysis ensures valuable insights and productivity gains from a derived structured table with high accuracy and precision, reflected in time savings of hours versus weeks.

INTRODUCTION

Clinical Trials are important mechanisms to evaluate effects of proposed drug therapies and provide direction to product discovery and innovation. As this process is governed and subject to oversight, it's crucial to ensure accurate, interpretable, and quantified relationships between any Adverse Events (AEs) with and multiple activities over the course of a trial. Important considerations are the **effects of the study drug** and **actions of the physician.**

The study drug would have incurred significant costs and investment in prior phases, which are at risk because of Adverse Events and any indications that patient safety had gotten compromised. Actions of the physician**,** on the other hand, indicate deviation and an error-prone operation. In addition to the effect on health outcomes, these also risk compromising the study, delaying drug development and time to market. The potential financial impact to the developer (pharmaceutical), and social, economic, and psychological costs on patient lives can be significant.

Therefore, we need to understand and attribute factors driving adverse events to implement proper mitigation. Data that helps us in our analysis includes clinical trial narratives which record actions and observations per patient over the trial. While useful and valuable, leveraging unstructured patient narratives is challenging; it's labor intensive, subjective and complex to extract. This is especially so due to the need to contextualize, reconcile patient activities, and establish accurate relationships.

In this paper, we outline a Natural Language Processing (NLP)-based approach to identify entities and events of interest within the scope of a given context. This has been distilled from experiences with diverse organisations and different data formats and objectives. NLP is a branch of Artificial Intelligence (AI) which includes extraction of patterns from subjective text matter. Entity extraction examples include medication, treatments, and critical interventions over a trial.

After describing entity extraction approaches, we also suggest the use of other AI applications such as machine learning and deep learning to scale these up across a larger corpus. To preserve meaning and provide interpretability, we also describe how rule discovery mechanisms can provide us with a governable NLP model for inference on future datasets. Finally, we present a temporal view through a suggested dashboard, which may help stakeholders explore and drill down deeper on events of interest.

DATA PREPARATION

Clinical Trial narratives tend to be longer than the average narrative, due to rigorous description of all events over the course of a trial (which might last even longer than 3 years in some cases). Typical narratives, based on our observation, ranged around 10 to 15 paragraphs. While not exhaustive, here are some common challenges we've encountered:

1. **Critical information tends to mix with supporting details**: Not only do narratives contain a direct description of events, but they might also present broad details of the trial, regarding cohorts, groups, and

selection criteria. While this is valuable, such information is already available from other structured data requires an analyst to distinguish essential information of interest from supporting / tangential details.

2. **Text may not be directly extractable.** Disparate and decentralized system may result in data available only in final document formats such as PDFs or scans. In such cases, we need to make use of Optical Character Recognition (OCR) tools provided by cloud service providers (Azure, AWS etc.) or other third-party providers.

3. **Data not in chronological order.** An ideal narrative depicts all events and actions in the same order as they occurred. However, this cannot be assumed due to text's non-standard nature. To mitigate this challenge, we need to develop methods that accurately identify dates of occurrences (as opposed to just any date), time and other positional and temporal markers.

4. **References to dates and times are not explicit.** Sometimes, explicit dates (in a date format such as 01-JAN-2023) are not provided. Analysts need to develop NLP rules to identify temporal markers. An example of a temporal marker would be the phrase "on the next day", which would need to be extracted and then paired with an explicit date marker (study start date, or another prior explicit date) to arrive at a date / timestamp.

5. **Link narrative data to other structured information.** As is the case with many other processes with human involvement, it's likely that errors of commission and omission might have occurred while recording data. For example, while the narrative might have mentioned that the physician had administered a given, a corresponding medication table may contain no such detail. NLP helps us identify such inconsistencies and ensure that data is recorded and sanitized.

Upon conclusion of Data Preparation activities, the target dataset for analysis is expected to be a table with the following elements (illustrative):

Sl No	**Column**	**Description**	**Sample Data (**fictional**)**
1	Subject ID	Identifier pointing to the subject (patient id) in the clinical trial.	S0001
2	Narrative	A text observation containing the entire description of events which have occurred over the clinical trial. The observation may contain multiple paragraphs and may conclude either upon the completion of the trial, or a forced completion (such as the patient's death or drop out)	*Pt enrolled in trial ABC, location XYZ as part of cohort C001 for PQROL. Pt administered 0.25 mg of study drug on 1st Jan 2021…. On 2nd Jan, pt complained of severe nausea, following which ___ was prescribed. Pt's condition did not stabilise overnight. …. On Day 3, the dosage was tapered down to 0.1 mg, … patient was taken off drug…. Study concluded on 30 June 2021.*
3	Location	Location of the clinical trial in case this is across multiple locations	Site 1
4	Other details	Other subject-level / site-level details as applicable for the analysis. This comes from structured data.	

ENTITY EXTRACTION

Entity extraction, also known as concepts, fact extraction (a variety) or Named Entity Recognition (in machine learning contexts) is an NLP capability to help identify patterns of interest. The pattern extracted may conform to an abstract definition known as an entity. A layman's understanding of an entity may run along the lines of people, organisations, and locations. For example, all names or references to people (identified by their names) within a document can be thought of as an entity.

For the methods described in this paper, we made use of SAS Viya, a proprietary analytics and AI platform, with some applications specific for Natural Language Processing, namely, SAS Visual Text Analytics. We also used Python and the Bidirectional Encoding Representation through Transformers (BERT) architecture for some of the machine learning tasks. While a detailed description of the methods used are beyond the scope of this paper, we provide some references which may help readers obtain such information.

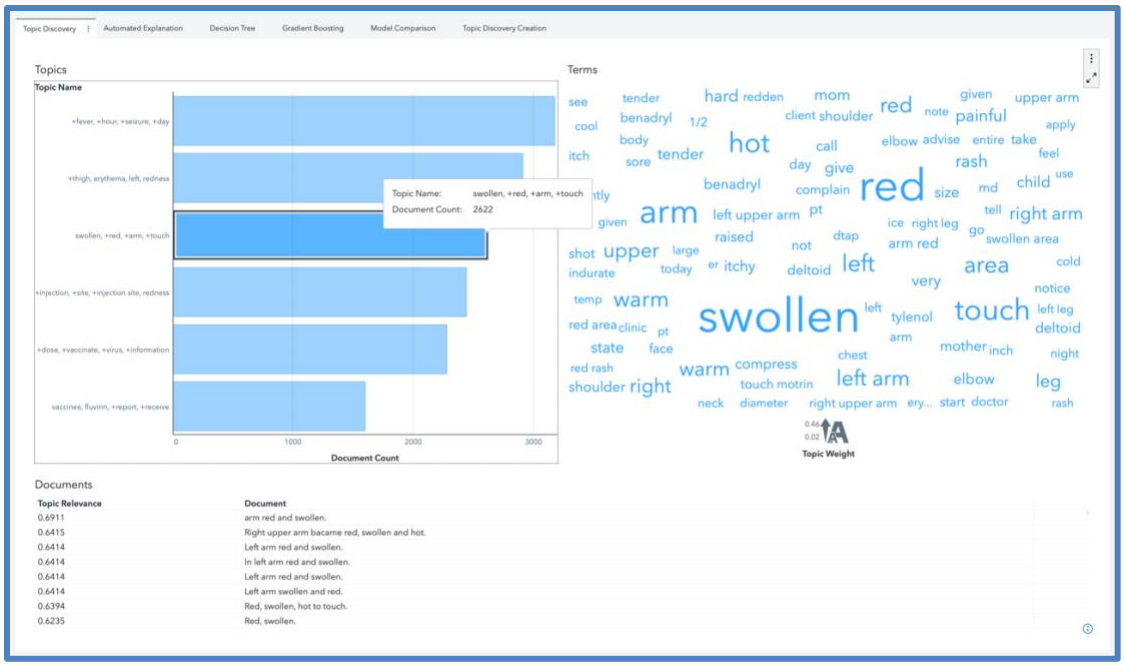


Figure 1: An example of text analytics (NLP) within the SAS Viya platform.

We defined entities that prove useful for downstream analysis and interpretation of incidents that occurred during a trial. Some definitions are simple, for example, the relevant study drug, and are used primarily as markers. Some others, however, happen to be quite complex and define an entity within a given context.

A correct understanding of context goes beyond mere extraction of patterns. Consider something seemingly trivial such as a date. The extraction of date as a pattern is simple enough and can be accomplished through patterns conforming to a structure such as 'dd-mon-yyyy' or 'mm/dd/yyyy' etc. But what if our objective is to identify the date of onset of a particular symptom, as illustrated in the following example:

"… the patient was injected with the study drug on 1st Jan 2021. The next day, they started exhibiting symptoms of …., prompting the physician to prescribe XXX on 4th Jan …."

As you may note, dates within the document include the date when the study drug was administered (1st Jan 2021), and a partial reference to when a prescription was made out to tackle an event (4th Jan, which can be assumed to be 4th Jan 2021). The date when a particular symptom manifest is not specifically called out but implied through the phrase "the next day", which leads us to the following approach.

1. Identify all date references: e.g. date_concept_1.
2. Identify all implied date reference: e.g. date_concept_2.
3. Identify patterns which indicate symptoms.

For the above, we use a rule authoring interface within SAS Visual Text Analytics which also provides us predefined capabilities for extraction of common entities such as names, dates etc.

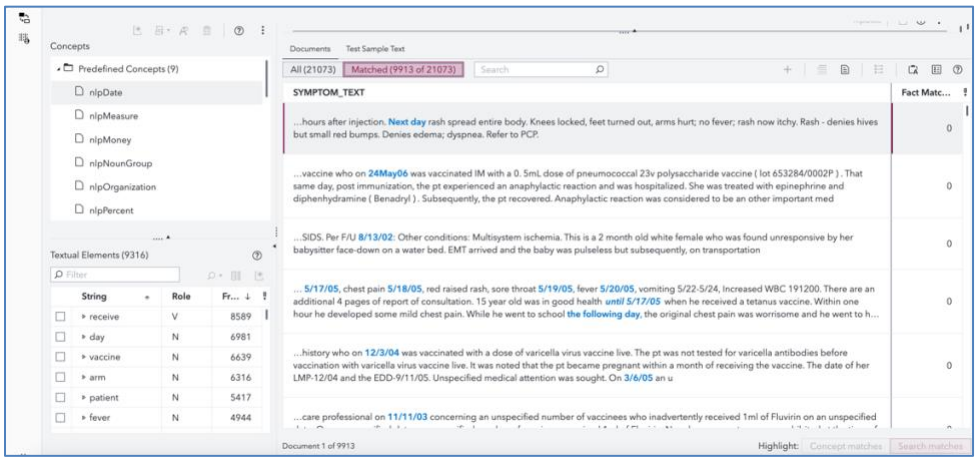


Figure 2: An example of concept extraction for dates (a predefined concept that can also be extended)

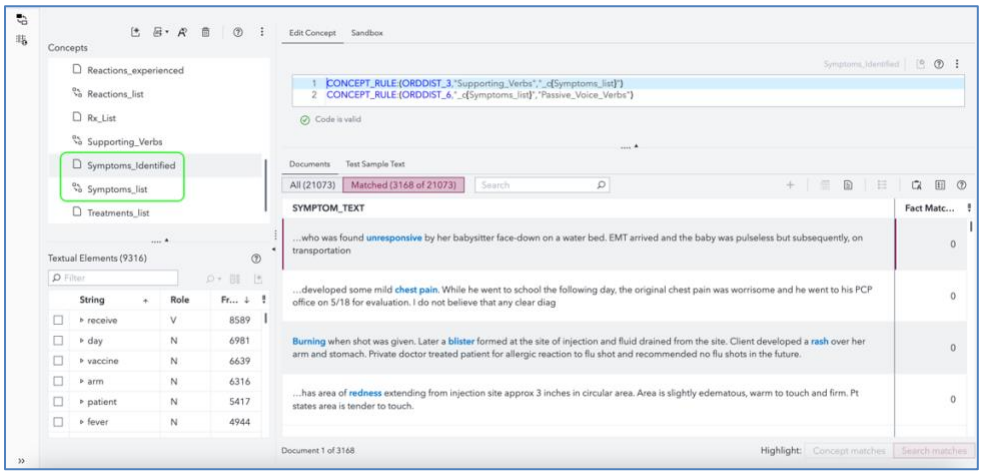


Figure 3: An example of defining a concept for symptoms and extracting given symptoms.

4. Identify specific references where a date reference was associated with the onset of a symptom.
5. If an implied date reference, associate with a prior explicit date reference (prior_date_reference). We tackle this through the special definition of an entity, known as a fact.

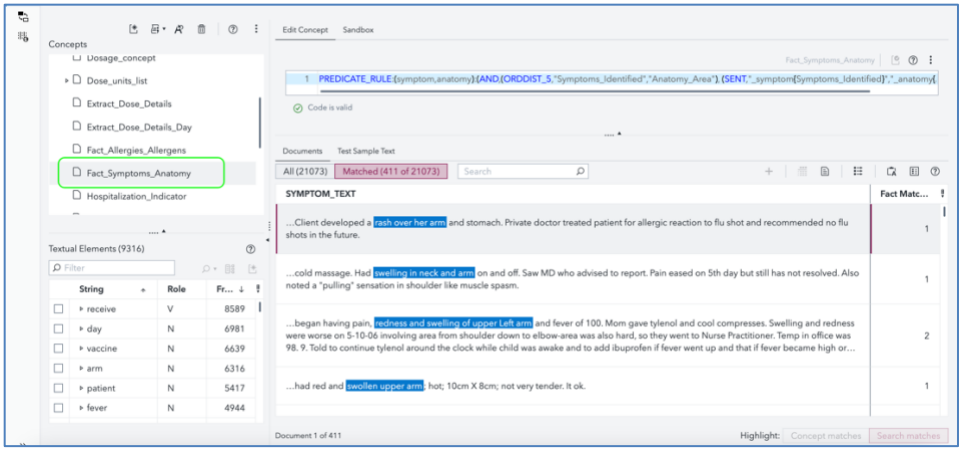


Figure 4: An example of fact extraction - in the example here, extracting a list of symptoms along with the anatomy affected.

6. Calculate offset as indicated by the implied date reference. Note that the more ambiguous the implied date reference (e.g. a *couple of days later*, or, heaven forbid, *some days later*), the more variable the offset might be. Simplified code:

```
data RESULT_TABLE_IMPLIED_DATES;
set EXTRACT_TABLE (where _concept_ = "date_concept_2");
if _match_text_ = "next day" then do;
   offset = 1;
   implied_date = intnx("day",prior_date_reference,-1);
end;
run;
```

With this crash course in concept definition out of the way, let's observe pertinent concepts that we've encountered when tackling the specific question of adverse events during clinical trials. We've generalised some of the concept definitions for what we encounter in discussions around most clinical trials.

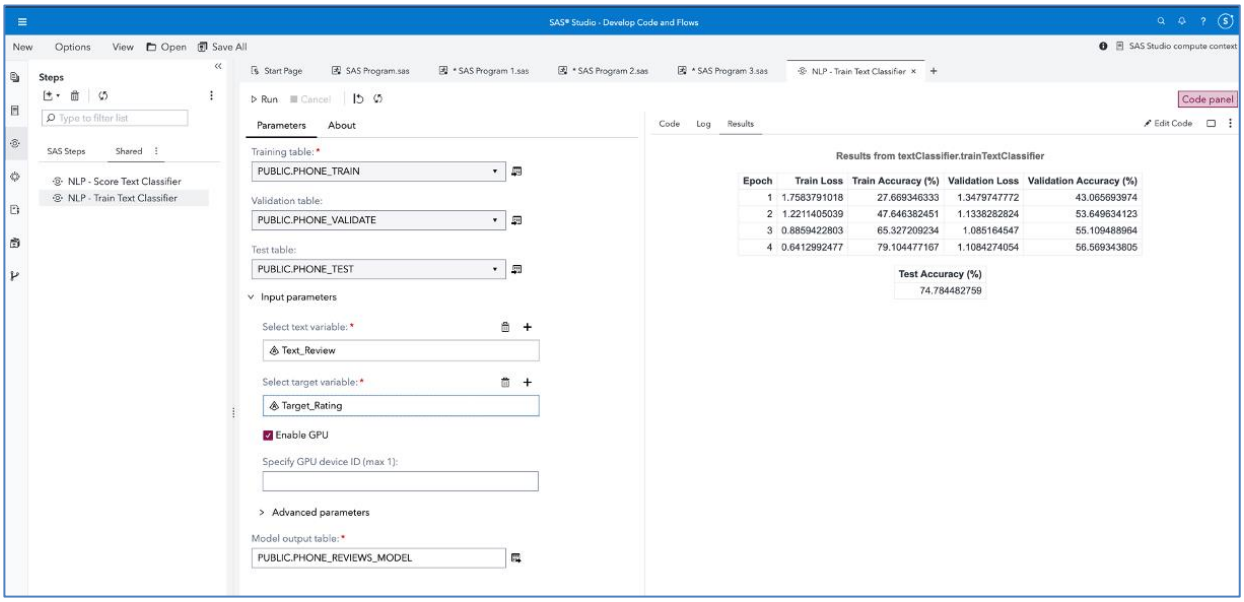
Sl No	Concept name	Definition
1	Study Drug	The study drug for the trial
2	Date of administration	Date when the patient was administered the study drug
3	Symptom Name	Symptom that arose in the patient over the course of the trial. Note that there might be multiple symptoms that might together constitute an adverse event, and these may run in parallel. Extracts relating to symptoms need to be coupled with the proper date / time markers for proper interpretability.
4	Date of onset	Date when the symptoms are recorded to manifest in a patient. Note #3 and the fact that these concepts are coupled together for proper interpretation
5	Date of resolution	Date of the symptom getting resolved
6	Dosage increase	Date & new dosage amount indicating when the physician administered increase in an associated drug. Note that this drug can refer to the study drug, concomitant medication or a drug intended to treat a symptom, therefore this must be coupled with a drug reference.
7	Dosage taper-down	Date and new dosage amount indicating when the physician decided to reduce the dosage of a drug. Note #6. This needs to be coupled with the drug reference since this could refer to either the study drug, or others.
8	De-escalations	Indicators of the risk characteristics of a symptom reducing in nature due to the patient stabilising.
9	Re-escalations	Indicators of a previously resolved symptom recurring or increasing in intensity.
10	Dose Characteristics: - Dose amount - Dose frequency	A more granular breakdown of dose indications to help analysts understand the characteristics in more detail. For example, a dose indication of "oral ABC 20 mg BID" indicates that the dose amount is 20, the dosage unit is mg, and the dose frequency is BID (twice a day or bis in die)
11	Study Day marker	Indicators for when an event occurred during the study period. The indicator could be an explicit date (01-JAN-2021) or an implicit marker (on Day 20).

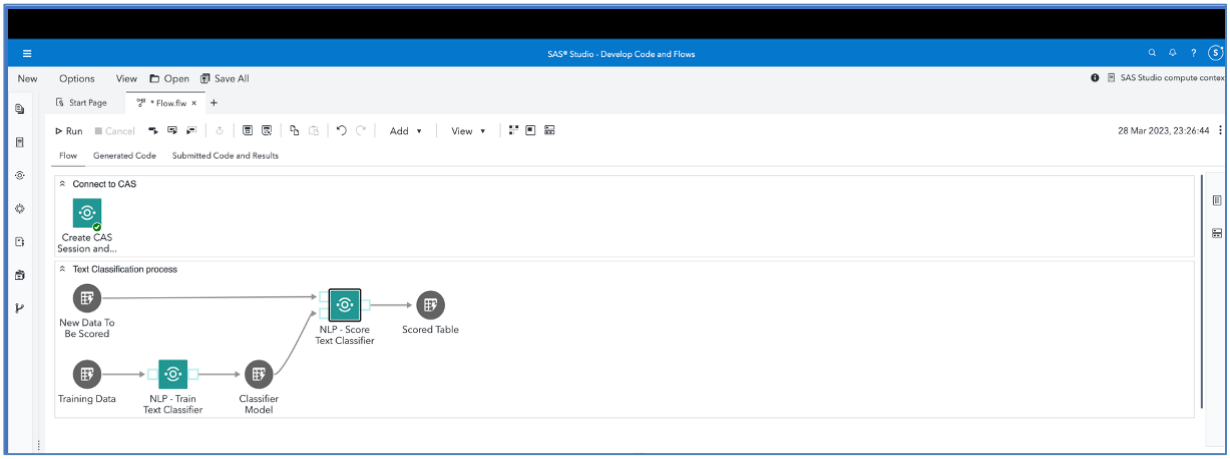
12	Patient Expiry Flag	Indicators regarding whether the patient expired during (or, if captured, after) the study period. Narrative with such indicators is subject to higher risk scrutiny.
13	Lab Test Flag	Indications and capture of any lab referrals ordered by the physician. Based on the amount of detail involved, this can be expanded to include lab test results and interpretation noted within the narrative (to be associated with treatment)
14	Concomitant Medications	Any concomitant medications taken by the patient
15	Steroid / Special Drug indicator	Indicators of any special or specific class of drugs administered.

DEEP LEARNING AND MACHINE LEARNING

A concept extraction approach based on rules alone may not scale well due to the effort involved in covering multiple patterns. The SAS Viya platform offers us supplemental methods through a set of AI techniques known as Deep Learning which is a subset of Machine Learning and approaches the problem from a different angle. Machine Learning & Deep Learning identify patterns through converting text extracts into numerical proxies known as embeddings and fitting models based on these embeddings. To harness Deep Learning, our approach has been multipronged: - we first use (SAS-) native methods to extract text features and construct an embeddings table, using Singular Vector Decomposition. We then fit deep learning architectures, primarily Recurrent Neural Networks and LSTMs on our extracted datasets (where concepts had already been extracted with rules) to train a model that identifies patterns close and like current embeddings.

In parallel, we also harness a language model trained using Bi-directional Encoding through Representation in Transformers (BERT). BERT has proven popular due to its making use of transformer architecture, which incorporates inputs both to the left and right of a word token while predicting the likelihood of a token within a sentence. In this case, our analysis on the SAS Viya platform is aided by close integration with other open-source platforms such as Python, and we can reach out to BERT implementations in Python to train a Named Entity Recognition (NER) model.





Figures 5 & 6: Typical training patterns for deep learning and machine learning exercises, which tend to require large amounts of labelled data and many iterations.

RULE DISCOVERY

The advent of complex and sophisticated entity extraction methods (encompassing rules, machine learning and deep learning) has led to a corresponding amount of uneasiness on the part of consumers of such information. Analysts require better understanding of analytics solutions and how AI solutions work. For this purpose, we follow a practice of running predictions from machine learning models through Boolean rule discovery. The resulting set of rules are easily understandable Boolean statements (AND/OR etc.) which identify the main keywords and surrounding "context" that the model is geared to look out for whenever it tries to extract certain entities. This practice of combining multiple analytical techniques is known as "Composite AI" and helps promote the message that good analytics practices should not constrain the developer to a limited set of methods.

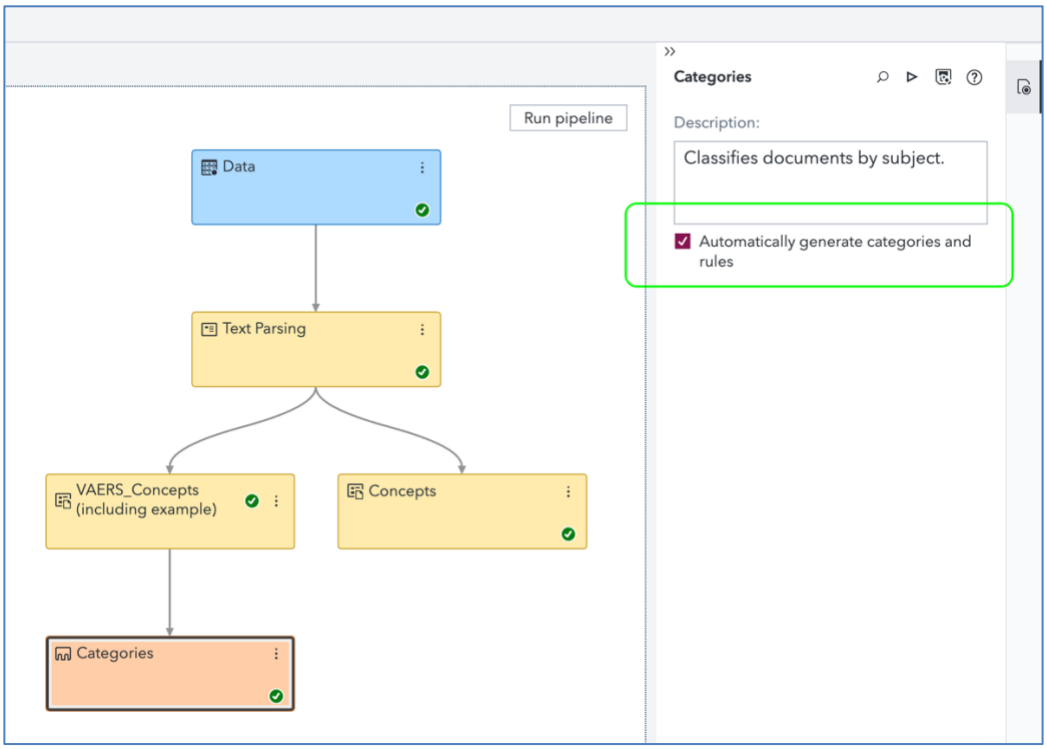


Figure 7: Automated rule generation facilitates new rule discovery.

REPORTING AND DASHBOARDS

During the final phase of our approach, we make use of SAS Visual Analytics, a reporting and visualization application which forms part of the SAS Viya platform to present results in a temporal framework. We have found that simple, easy to understand visualizations help stakeholders quickly attribute key factors and root causes with more confidence and conviction. The reports and dashboards that are created provide strong documentation to drive process redesign, a relook at the drug development process or any other change which addresses risks and inefficiencies that have occurred over the trial. An example dashboard with simplified, fictional details is presented. Note that these dashboards also allow for drilldowns and facilitate comparison with other structured data elements.

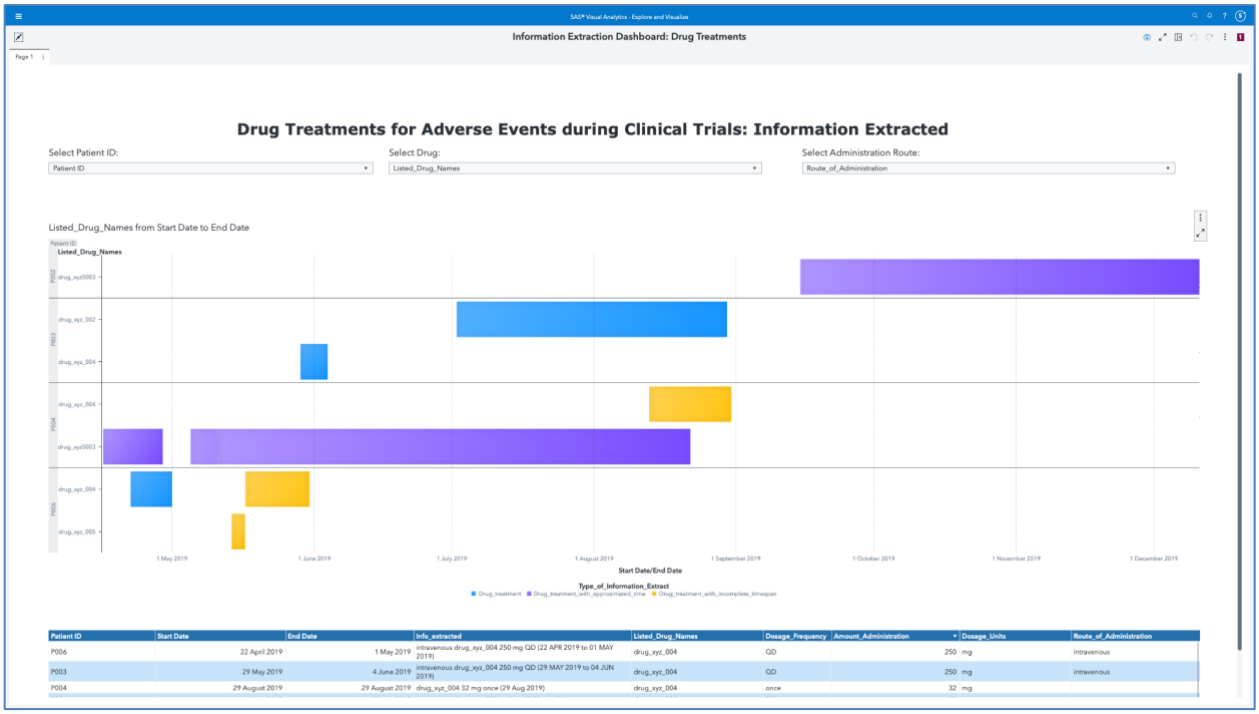


Figure 8: Example Dashboard showing temporal view of Adverse Events

CONCLUSION

An essential part of our analysis is that we have succeeded in **extracting key elements from unstructured data forming part of a clinical trial narrative into a structured data format, which aids analysis**. Wider ranging implications of this are that this approach enables key stakeholders – drug developers, research organisations, clinical trial operators and additional stakeholders who have an interest – to form an **objective, data-driven interpretation of outcomes** that have occurred over the course of a trial. This is especially **valuable when inspected through the lens of safety and patient centricity**.

With increasing advances in Natural Language Processing and its neighbouring broad umbrella of Generative Artificial Intelligence (Generative AI), our research into this area is set to continue. Technological advancements in Large Language Models (LLMs, of which BERT is a part) enable us to better investigate the problem of deriving complex and accurate entities which describe clinical trial outcomes. Another area of future research is to use text generation capabilities and summarize narratives into more concise forms while retaining the essence of what they convey.

REFERENCES

1. Sankaran, Sundaresh (2022), The Analytics Life Cycle for Natural Language Processing: A Healthcare and Life Sciences Example, https://www.linkedin.com/pulse/analytics-life-cycle-natural-language-processing-example-sankaran/
2. Sankaran, Sundaresh & Sabo, Tom (2023), Bringing it All Together: Applying the Analytics Life Cycle for Natural Language Processing to Life Sciences, https://www.lexjansen.com/pharmasug/2023/SD/PharmaSUG-2023-SD-167.pdf
3. Hatim, Qais; Sankaran, Sundaresh; Sabo, Tom; McRae, Emily; Laufe, Lauren; Blanchard, Robert (2023), Deep Learning to classify Adverse Events from Patient Narratives, https://www.pharmasug.org/proceedings/2023/SA/PharmaSUG-2023-SA-110.pdf
4. SAS Visual Text Analytics, forming part of the SAS Viya platform, https://www.sas.com/en_us/software/visual-text-analytics.html
5. SAS Visual Analytics, part of the SAS Viya platform, https://www.sas.com/en_us/software/visual-analytics.html
6. Albright, Russ; Cox, James; Jin, Ning (2016), Getting More from the Singular Vector Decomposition (SVD): Enhance Your Models with Document, Sentence and Term Representations, https://support.sas.com/resources/papers/proceedings16/SAS6241-2016.pdf

ACKNOWLEDGMENTS

Authors thank their prior collaborators and partners for continued discussion around use cases related to unstructured data.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Author Name: Sherrine Eid
Company: SAS Institute
Address: 100 SAS Campus Dr, Cary, NC 27560, USA
Work Phone: +1 919 531 3991
Email: sherrine.eid@sas.com
Website: www.sas.com

Author Name: Sundaresh Sankaran
Company: SAS Institute
Address: 100, SAS Campus Dr, Ste C5219, Cary, NC 27560, USA
Work Phone: +1 919 531 4921
Email: Sundaresh.sankaran@sas.com
Website: www.sas.com; www.linkedin.com/in/SundareshSankaran