PharmaSUG China 2024 - Paper DV-10023

Enhancing Patient Profile Visualization through Clustering Algorithms: An Exploration of Smart Data Presentation Techniques

Zhengyuan (Roy) Wen and Kang Zhao, AstraZeneca.

ABSTRACT

Visualization of patient profiles is a communication efficiency booster in the field of healthcare data analysis, particularly for early drug development and clinical decision-making. This article explores the use of clustering algorithms to enhance the smart visualization of patient data. By segmenting checking conditions within the Analysis Data Model (ADaM) specifications into coherent groups with similar characteristics, we aim to provide a clearer understanding of patient populations embedded in Clinical Study Protocol and Statistical Analysis Protocol for clinical physicians, revealing patterns that traditional visualization methods may obscure while easing the usages.

With traits of structured ADaM specifications at the derivation rule level, we are able to introduce a Proof of Concept (PoC) for a visualization platform underpinned by various clustering algorithms designed to pinpoint areas of interest for physicians to investigate further. This includes both subjects of interest and metrics of interest. The methodology employs exploratory analysis with natural language processing and expounds on the logic behind the algorithm selection. The challenges of integrating these algorithms into a user-friendly visualization tool tailored for healthcare professionals are also discussed.

Our results indicate that advanced visualization, supplemented with smart condition combinations and recommendations, holds promise for future research and development, as evidenced by feedback from clinical physicians. However, further questions and issues, such as regulatory concerns and the risk of unblinding, remain to be resolved. We advocate for continued exploration on devising more intelligent visualization techniques and to transition these advances from concept to clinical application.

Keywords: Patient Profile Analysis, Clustering Algorithm, Natural Language Processing, CDISC Analysis Data Model, Data Visualization

INTRODUCTION

BACKGROUND

In the healthcare industry, data visualization acts as a critical bridge from raw clinical data to actionable insights, enabling healthcare professionals to rapidly comprehend extensive information arrays, which is crucial for providing timely and effective patient care. Visualization of patient data supports clinicians' day-to-day decision-making and assists in identifying trends and outliers that may be obscured in textual or tabulated data. In essence, effective and timely data visualization is indispensable for enhancing patient care and clinician decision-making.

Effective visualization of healthcare data is pivotal in drug development and clinical decision-making. Despite its critical role, traditional visualization tools often struggle to capture the intricate details of patient profiles, particularly in the early-phase clinical trials. On the other hand, late phase clinical trials are more standardized and better structured.

To address the disconnected dots between data points of interest in early phase activities, and rich set of variations for variable systems from late phase studies, we propose a novel workflow for improving patient profiling. Also, we proposed a method to incorporate the derivation rule in ADaM Specification expand the power of language via Natural Language Processing and Clustering Analysis. Finally, we build up a PoC to illustrate our results.

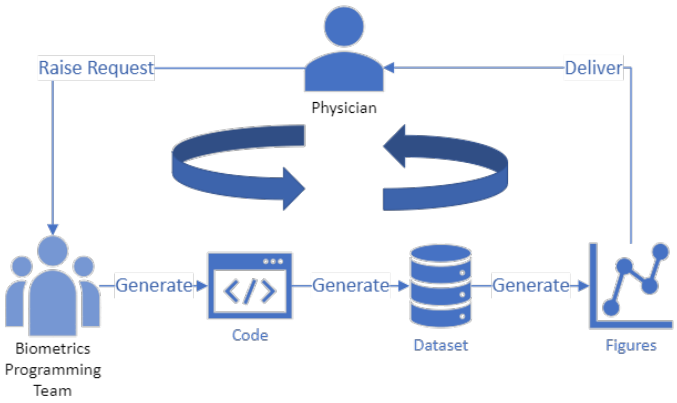


Figure 1. Traditional Visualization Workflow

Challenges with Traditional Visualization Methods

One of the foundational steps in the process of data analysis in Clinical Trial is the visualization of patient data, which traditionally follows a specific workflow. The conventional data flow (See Figure 1) typically involves,

1. Physicians raise a particular request to biometrics team for patient profiling.
2. The biometrics programming team accepted the request, then started preparing datasets and codes.
3. Biometrics team finished datasets and codes preparation, delivers figures to Physicians.

However, this traditional approach has several pain points. The first thing is it's adhoc to write codes per on a request basis, and it heavily depends on programmers' experience to guarantee quality in a timely manner. Adhoc tasks are less than ideal yet inevitable to support such exploratory activities in the traditional way. The datasets in use are SDTM datasets, and sometimes even raw datasets depending on visualization requirements and data sources for studies. This further pulls away programming resources to deliver ADaM datasets and further TLF(Table, Listing, and Figure)-s for submission purposes.

The second is that requests can be iterative from users, casting heavy communication overhead on the programming team. As a result, more time would be spent on fulfilling the requests from physicians and important subtleties and patterns may remain concealed.

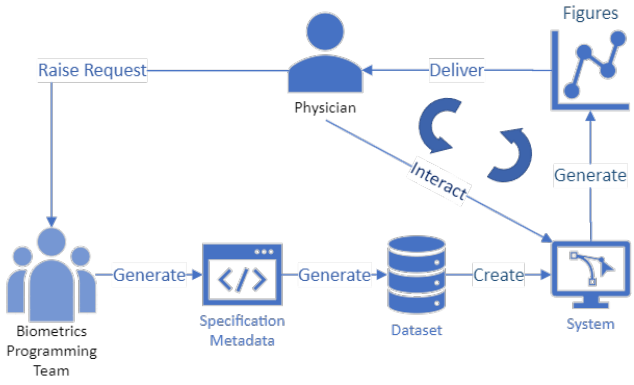


Figure 2. Our Proposed Workflow

OUR WORK: A NOVEL VISUALIZATION WORKFLOW

To overcome these obstacles, we introduce a concept of workflow for Physicians to interact and obtain the results they want directly without further communication with programming team. Our Proposed workflow (See Figure 2) would be following,

1. The biometrics programming team develops SDTM and ADaM Specification via Domain Specific Language (Zhao 2023).
2. With the help of Domain Specific Language, we can generate the ADaM Dataset once the ADaM Specification is ready.
3. We fully leverage the derivation rule information from ADaM Specification to build an "recommendation system" for Physicians to extract their potential interest points.
4. The system automatically generates the figures after Physicians picked up their interest points.

The novelty is in connecting ADaM specification and dataset with data visualization demands at the upstream. This can be achieved by a full automated data flow and accessible knowledges accumulated through all previous studies.

In the whole picture we proposed, this article will focus on how we could solve the **third element** and make the path for future research and try out.

METHODS

The methodology employed in our study is critical to understanding the enhancements proposed for patient data visualization. To further introduce the process, we need to bring up two assumptions for ADaM Specification.

1. The conditions in ADaM Specification include all the interest points that physicians may need.
2. The ADaM dataset includes all information we need for generating figures for Patient Profiling.

Based on these two assumptions, we started to explore the potential methodologies.

Figure 3. Simplified Metadata Flow

A simplified dataflow (See Figure 3) demonstrates how we could discover interest points from ADaM Specification. Now we will further illustrate the details.

EXTRACT FROM ADAM SPECIFICATION

Thanks to Domain Specific Language (DSL) (Zhao 2023), we can now easily extract conditions from the derivation rules in the ADaM Specification. Additionally, the DSL facilitates the computation necessary to filter the desired subject list once the conditions have been selected.

CONSTRUCT THE CLUSTERING MODELS

For Clustering Models, we need to firstly let the machine understand the plain text. With the help of Attention mechanism (Vaswani et al. 2017), machines can understand the context of a plain text nowadays. Based on the model we would use; we propose the following clustering workflow (See Figure 4).

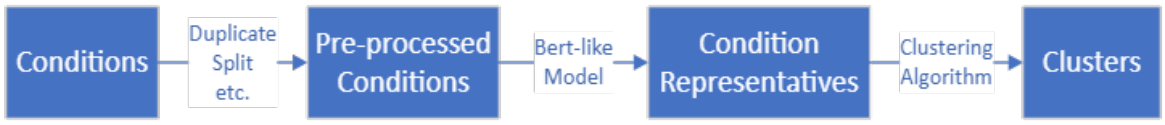


Figure 4. Clustering Model Workflow

The goal for our clustering models is to generate a cluster analysis that identifies potential interest point groups and let us compute the corresponding distance among those groups.

Preprocessing Conditions

Typically, a condition from the original derivation rule in the ADaM Specification may be either a simple, short sentence or a long and complex one with multiple instances of 'and,' as well as brackets and the 'or' conjunction. To address this complexity, we need to preprocess our conditions. We simplify the sentences by splitting them at conjunctions and brackets.

In addition, we substitute complex variable names (such as 'ADAE.AESTDTC') with their corresponding labels (such as 'Start Date/Time of Adverse Event') to enhance the machine's understanding of their meanings.

Leverage the Power of BERT-like Model

The BERT-like model serves as an effective tool for clustering analysis. Developed by Google, the BERT model (Devlin et al. 2018) is based on an encoder structure that leverages bidirectional context and self-attention mechanisms. These features provide a robust and nuanced representation of text, which is extremely beneficial for clustering analysis. This is because it captures a deep understanding of the structure and semantics of the text, enabling more precise and cohesive clustering of similar documents.

For this project, we use Sentence-Bert (Reimers and Gurevych 2019) to generate representative vectors for sentences. Typically, this involves using the first token, [CLS], from the final hidden layer as our feature representatives for the conditions.

Clustering Algorithms

When discussing clustering algorithms, there are numerous approaches to tackling the problem of clustering topics. For this project, we opted for HDBSCAN (Campello, Moulavi, and Sander 2013) , a hierarchical density-based algorithm derived from DBSCAN (Ester et al. 1996), as our preferred method.

To select the appropriate distance metrics combined with the dimension reduction method while being cautious of the curse of dimensionality, we selected UMAP (McInnes, Healy, and Melville 2018) dimension reduction method for our HDBSCAN clustering algorithm. Although cosine similarity is widely used sparse datasets like text data, in our experiments, Euclidean distance with UMAP dimension reduction demonstrated better performance in clustering when implemented through HDBSCAN. However, we did observe some anomalies in the clusters using cosine similarity. For instance, the conditions “Baseline Weight (Kg) <65” and “Baseline Weight (Kg) >90” were grouped into the same cluster, while the condition 'Baseline Weight (Kg) >= 65 and < 90' was not placed into the 'baseline weight' group. A possible explanation for this could be that our text analysis is magnitude-sensitive, meaning it captures the significance of certain terms, such as 'Baseline Weight', which could affect the clustering outcome.

Additionally, we employed an exhaustive cross-validation search to determine the optimal hyperparameters for HDBSCAN. We then selected the model that yields the highest Density-Based Clustering Validation (DBCV) score (Moulavi et al. 2014).

PICK UP THE INTEREST POINT

We identified two categories of interest points. The first is the Conditions of Interest (CoI), which are displayed in the frontend. The CoI enable us to generate the relevant variables through a variable dependency graph. This allows us to plot the desired graphs using those interest variables. When the user (Physician) interacts with the CoI, the backend reveals the second type of interest point, Patients of

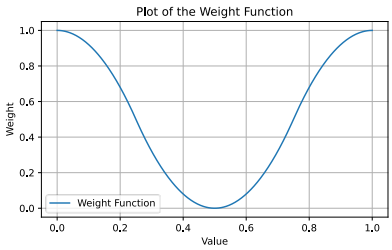
Interest (PoI). PoI acts as a subject filter, enabling us to refine the subject list using a pre-computed condition-subject 0-1 matrix, which has extracted condition items for the columns, and subjects/patients for the rows.

In our experiment, to automatically highlight points of interest, we must establish a starting point. Physicians will select an initial point that piques their interest from the frontend — this becomes the CoI. Based on the points they select, we then offer recommendations. In each iteration, we calculate a score for every potential data point, adhering to two core principles for score calculation:

1. Physicians are likely to seek out areas of knowledge distinct from their initial choices, such as points that are distantly related to their initial selections.
2. Physicians are inclined to focus on areas of knowledge adjacent to their initial choices, as these points might offer marginal enhancements to their initial selections.

With the above principles, we developed a weighted function to calculate every condition's score. Given a normalized distance x (ranging from 0 to 1), we can calculate the weighted distance between different data points. The weighting function is below.

$$W(d) = \begin{cases} -8d^2 + 1 & \text{If } d \in [0, 0.25) \\ 8(d - 0.5)^2 & \text{If } d \in (0.25, 0.75] \\ -8(d-1)^2 + 1 & \text{If } d \in [0.75, 1] \end{cases}$$



Then we assign a score α_i to each data point i,

$$\alpha_i = \sum_{s \in S, C_i \neq C_s} W(d_{is})$$

Where S stands for the set of selected condition, s is the point in set S, C_i is the cluster of point i, d_{is} is the normalized distance between point i and s.

Note that when two points fall within the same cluster, the calculated weight is set to 0. As the distance between points increases, the weight initially decreases and then subsequently increases.

RESULT

We conducted a series of trials and experiments. Utilizing the methodologies described above, we successfully performed cluster analysis using our ADaM Specification and obtained favorable results that illustrate the relationships between clusters. Below is a visualization with labels for the top 10 clusters, based on the safety section from the ADaM Specification.

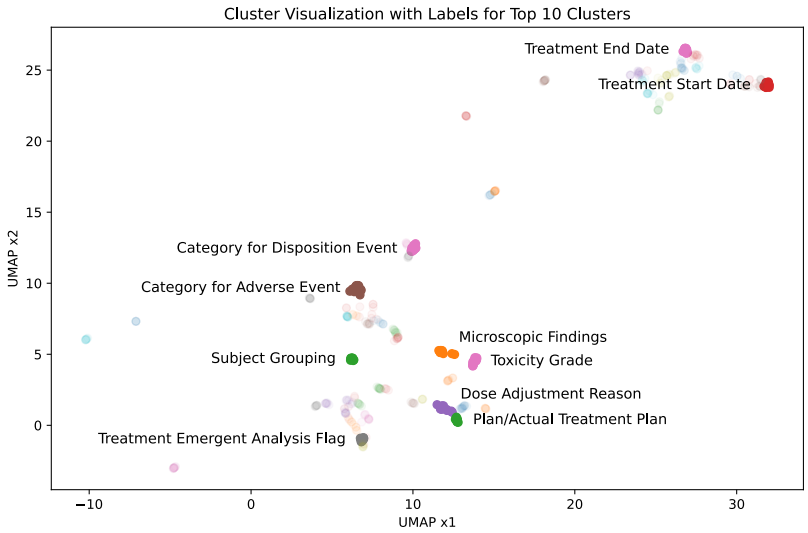


Figure 5. Cluster Visualization with Labels for Top 10 Clusters

Given the satisfactory clustering outcomes, we are now positioned to assign scores to each data point. We have developed a Proof of Concept (PoC) for Data Visualization using Streamlit, a Python package akin to RShiny.

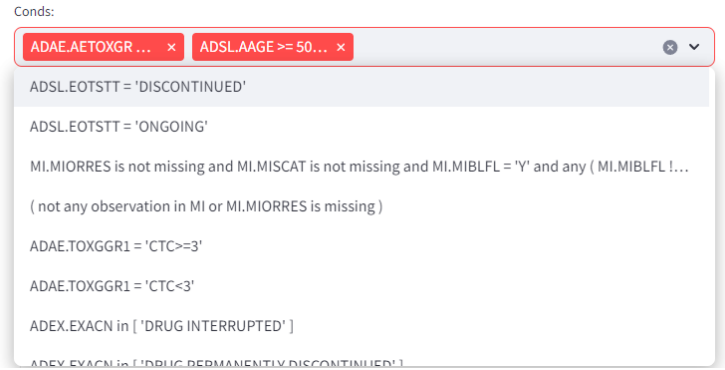


Figure 6. Condition Recommendation

As illustrated in Figure 6, the recommended condition will automatically be highlighted when a condition is selected. For instance, if we choose “ADAE.AETOXGR >= 3” and “ADSL.AAGE >= 50 and <65“, there is "ADSL.EOTSTT = 'DISCONTINUED'" on the top in the drop list, followed by other conditions of interested from previous clustering algorithm. Behind the scene, the system will automatically filter out subjects who meet the selected conditions.

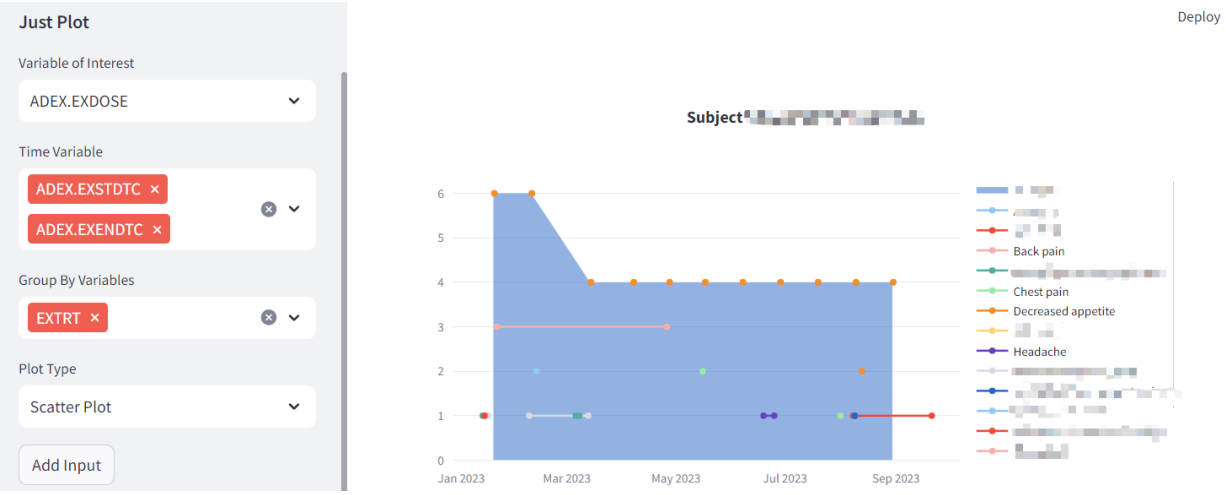


Figure 7. Generated Figure via our platform

Our subsequent step involves creating figures for our audience. In this scenario, we select three conditions: “ADAE.AETOXGR >= 3”, “ADSL.AAGE >= 50 and < 65”, and “EX.EXDOSE is greater than 0”. After selecting these conditions, we can reveal the relevant variables of interest through their downstream and upstream dependencies. This is displayed in Figure 7, where the blue area represents planned exposure with dosing date points. The remaining plots pertain to the adverse event toxicity grade change with line plots. This is a result from a few clicks on the UI.

FUTURE PROSPECT

For more intelligent visualization, we aim to fully utilize analysis-ready data with more physicians' feedback. Our goal is to make it useful for the physician group, and also to uncover potential relationships not only grounded in ADaM Specification but also influenced by the selections made by physicians.

Regarding potential risks, we must be cognizant of regulatory concerns and the possibility of unblinding. Given the sensitive nature of exposure data, its usage should be carefully monitored. In some cases, applying dummy data in blinded studies may be necessary to mitigate these risks.

CONCLUSION

The implementation of the clustering algorithm has provided us with valuable insights into physicians' needs. With a refined version of our tool, we will be able to precisely extract points of interest and offer immediate feedback with each click. This capability will significantly reduce the time spent redoing work. We also assisted in the establishment of standardized ADaM dataset derivation rules. Although the current version requires further enhancements, we have been able to effectively utilize the ADaM Specification and proof the concept. We believe this visualization represents an innovative and improved method for data presentation.

REFERENCES

Campello, Ricardo JGB, Davoud Moulavi, and Jörg Sander. 2013. "Density-based clustering based on hierarchical density estimates." In *Pacific-Asia conference on knowledge discovery and data mining*, 160-72. Springer.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. 'Bert: Pre-training of deep bidirectional transformers for language understanding', *arXiv preprint arXiv:1810.04805*.

Ester, Martin, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. "A density-based algorithm for discovering clusters in large

spatial databases with noise." In *kdd*, 226-31.
McInnes, Leland, John Healy, and James Melville. 2018. 'Umap: Uniform manifold approximation and projection for dimension reduction', *arXiv preprint arXiv:1802.03426*.
Moulavi, Davoud, Pablo Andretta Jaskowiak, Ricardo Campello, Arthur Zimek, and Joerg Sander. 2014. *Density-Based Clustering Validation*.
Reimers, Nils, and Iryna Gurevych. 2019. 'Sentence-bert: Sentence embeddings using siamese bert-networks', *arXiv preprint arXiv:1908.10084*.
Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. 'Attention is all you need', *Advances in neural information processing systems*, 30.
Zhao, Kang. 2023. "A Domain Specific Language for ADaM Dataset Generation in Regulatory Clinical Research." In *PharmaSUG China 2023*. Nanjing.

ACKNOWLEDGMENTS

Heartful thanks to China Biometrics Leader Team, Programming Team, and Data Transformation Team from AstraZeneca, Shanghai, China, for their supports on the project and this paper.

CONTACT INFORMATION

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

Zhengyuan (Roy) Wen
AstraZeneca
roy.wen@astrazeneca.com
https://www.linkedin.com/in/zhengyuan-wen-792b2b223
https://www.wennroy.com

Kang Zhao
AstraZeneca
kang.zhao@astrazeneca.com
https://www.linkedin.com/in/kangzhao-bbgw