Paper IC06

Identifying Data Quality Issues in Clinical Trials using OpenAI Embeddings

Ashwin Rai, Evidera, Overland Park, USA
Yash Shah, PPD, Morrisville, USA

ABSTRACT

Data quality issues in clinical trials pose significant threats to research integrity and patient safety. The complexity of clinical trial data often surpasses the capabilities of traditional validation methods. This manuscript introduces an innovative approach utilizing OpenAI embeddings to identify and address data quality problems in clinical trial datasets. Originally designed for natural language processing, OpenAI embeddings were repurposed to detect outlier data points. Synthetic clinical trial data was generated to simulate real-world data issues. Patient data, transformed into text-like structures and encoded as embeddings, facilitated the identification of dataset patterns and anomalies. Principal Component Analysis (PCA) and clustering techniques were employed to effectively locate outlier data points. The results demonstrated robust performance, achieving a recall of 0.994 and a false-positive rate of 0.007 on a dataset with 520 records and 7 variables, including 20 outlier data points. This data-driven, automated approach holds promise for enhancing clinical trial data quality.

INTRODUCTION

In the landscape of clinical research, where the veracity and dependability of findings are paramount, the quality of underlying data plays a pivotal role. Ensuring data integrity transcends scientific rigor, intertwining with imperatives related to patient safety and the success of medical interventions. However, the multidimensional nature of clinical trial data introduces challenges that traditional validation methods struggle to navigate, potentially leading to skewed research outcomes and jeopardized patient safety [1].

This paper presents an innovative approach that harnesses the power of artificial intelligence, specifically OpenAI embeddings, to discern and rectify data quality problems in clinical trial datasets. While traditional data quality assurance measures hold intrinsic value, they often fall short in capturing the nuanced intricacies of clinical trial data. This shortfall becomes particularly pronounced when anomalies, outliers, and unexpected patterns, indicative of underlying issues, emerge.

Clinical trial datasets stand as complex narratives, portraying individuals' health journeys through diverse variables and intricate structures. By repurposing OpenAI embeddings—originally designed for natural language processing (NLP)—this paper seeks to transform the health narrative encoded in clinical trial data into a format conducive to effective outlier detection and comprehensive data quality enhancement [2,3].

The utilization of artificial intelligence, embodied in OpenAI embeddings, facilitates the treatment of patient data as text-like structures, transcending the numerical constraints of traditional validation methods. This groundbreaking approach enables the identification of underlying patterns, relationships, and anomalies that might be obscured when relying solely on numerical validation. Furthermore, the subsequent application of Principal Component Analysis (PCA) and clustering techniques adds depth to our understanding, aiding in the identification of outlier data points that warrant closer scrutiny.

As the era of big data in healthcare unfolds, challenges associated with ensuring data quality become even more pronounced. Issues such as missing data, inaccuracies, and inconsistencies can significantly compromise the integrity of clinical trial outcomes. The proposed methodology not only responds to the pressing need for sophisticated, automated approaches but also signifies a paradigm shift in how data quality is approached in the domain of clinical trials. Subsequent sections will delve into the methods employed, the results obtained, and the broader implications of this innovative approach for advancing the quality and reliability of clinical trial data.

METHODS

OPEN AI EMBEDDINGS

OpenAI embeddings refer to vector representations generated by advanced language models, particularly those grounded in transformer architectures like GPT-3. These embeddings stand out for their context-aware nature, possessing the ability to capture the nuanced meaning of words or phrases within the context of a given input sequence. Developed through the training of language models, OpenAI embeddings excel in contextualizing information, making them versatile tools for natural language processing (NLP) tasks. The context-aware nature of OpenAI embeddings is particularly advantageous in healthcare data interpretation. In the realm of natural language processing, these embeddings excel in capturing the semantic relationships and contextual nuances inherent in textual healthcare data. This capability extends beyond mere words and phrases, allowing the model to grasp the intricate meanings embedded in the narratives of patient health journeys [2,3].

The adaptability of OpenAI embeddings is a crucial asset, especially when dealing with numerical data within healthcare domains. By leveraging these embeddings, numerical patient health data undergoes a profound transformation. OpenAI embeddings have the unique ability to convert numerical variables into text-like structures, offering a more comprehensive representation of health narratives. This transformation enriches the dataset, providing a textured and nuanced understanding of patient health experiences.

Furthermore, OpenAI embeddings are trained on linguistic datasets, a facet that contributes significantly to their efficacy in healthcare data transformation [3]. In the healthcare landscape, where data is inherently diverse and multifaceted, linguistic training equips these embeddings to encode numerical health variables into vector representations. This encapsulation of relationships and contextual meanings becomes invaluable when dealing with the complexities inherent in clinical trial data.

APPLICATION OF OPEN AI EMBEDINGS IN HEALTH DATA TRANSFORMATION

The application of OpenAI embeddings in healthcare data transformation extends beyond conventional numerical constraints. While traditional methods might struggle to capture the intricacies of health narratives, OpenAI embeddings transcend these limitations, leading to more accurate and insightful analyses [4,5]. This is particularly evident in clinical trial data, where anomalies, outliers, and unexpected patterns may signify underlying issues. The nuanced understanding provided by OpenAI embeddings enhances the effectiveness of outlier detection mechanisms.

In practical terms, developers and data scientists in healthcare can leverage OpenAI's API to seamlessly integrate the capabilities of language models into their applications. This API facilitates access to OpenAI embeddings, allowing for the extraction of valuable insights from specific text inputs. Whether it's analyzing patient records, extracting meaningful information from medical literature, or aiding in the development of healthcare applications, OpenAI embeddings become a powerful tool in the healthcare data scientist's arsenal.

In essence, the fusion of OpenAI embeddings and healthcare data signifies a transformative approach to understanding and interpreting the complexities inherent in-patient health narratives [4]. This innovative utilization not only contributes to enhanced accuracy but also opens avenues for breakthroughs in clinical research, personalized medicine, and the overall improvement of healthcare outcomes. The adaptability, context-awareness, and versatile applications of OpenAI embeddings position them as a cornerstone in the evolution of healthcare data analytics and interpretation.

SYNTHETIC CLINICAL TRIAL DATA GENERATION

In this phase, the methodology undertakes the generation of synthetic clinical trial data. This deliberate process involves intentionally introducing outliers, missing data, inaccuracies, and inconsistencies. The controlled environment created through synthetic data generation serves as a litmus test for evaluating the effectiveness of OpenAI embeddings in detecting and addressing diverse data quality issues. By simulating real-world challenges, this phase ensures the robustness of the methodology in handling the complexities inherent in clinical trial datasets.

ENCODING AND EMBEDDING TRANSFORMATION PROCESS

In this phase, the tabular clinical trial data undergoes a notable transformation into text followed by encoding. The following example vividly illustrates how narrative elements such as test conditions, positions, and visit details are seamlessly converted into textual elements, enriching the linguistic depth of the dataset. The transformation of patient data into a text-like format sets the stage for the encoding and embedding process.

"In this exemplar instance, the participant, identified as {subjid}, underwent a vital sign assessment as part of the {testname} examination while assuming the {vspos} position. The initial recording of this vital sign measurement revealed {vsores} {vsoreu}. This particular measurement coincided with {vseq}, aligning with visit number {vname}, and was meticulously conducted at {test_time}."

The subsequent encoding and embedding process, leveraging the capabilities of OpenAI embeddings, adeptly captures these nuanced relationships. OpenAI embeddings play a pivotal role in this transformative journey, translating numerical variables into vector representations that intricately captures the relationships. This infusion of linguistic context enriches the dataset, enabling a more detailed analysis of patterns and anomalies within the clinical trial data.

IDENTIFICATION OF DATASET PATTERNS AND ANOMALIES

To unravel intricate patterns and anomalies within the dataset, a sophisticated analysis leverages the power of Autoencoder models, OpenAI embeddings, Principal Component Analysis (PCA), and UMAP.

Autoencoder models, a type of neural network architecture, are instrumental in outlier detection, leveraging a specialized loss function to identify instances deviating from learned representations. In this context, OpenAI embeddings serve as the input to the Autoencoder model, contributing to its ability to discern anomalies within the dataset. Consisting of an encoder and a decoder, an Autoencoder is trained to accurately reconstruct the input data based on the embeddings. During training, the model captures the intrinsic features of normal data, aiming to minimize the reconstruction error. The loss function used in this Autoencoder is carefully designed to emphasize deviations from the learned representations, effectively identifying outliers. This unique combination of OpenAI embeddings as input to the Autoencoder enhances analytical depth, facilitating the identification of unexpected patterns and outliers indicative of potential data quality issues. The utilization of a tailored loss function ensures that the model is sensitive to anomalies, contributing to the robustness of the methodology in detecting and addressing diverse data quality challenges within clinical trial datasets.

In addition to Autoencoder models, OpenAI embeddings, PCA, and UMAP are employed to visualize outliers in a more detailed manner. Principal Component Analysis (PCA) is instrumental in transforming the dataset into a lower-dimensional space, condensing information while preserving essential features. This transformation results in a concise and informative representation of the dataset's underlying structure. Simultaneously, UMAP contribute by visually identifying outliers within this transformed space. UMAP, or Uniform Manifold Approximation and Projection, excels at capturing complex relationships, enhancing the interpretability of the data structure [6].

This comprehensive approach, integrating Autoencoder models with OpenAI embeddings, PCA, and UMAP, provides a clear interpretation of intricate relationships within the dataset. It forms a robust framework for meticulous data quality assessment, addressing both complex patterns and outliers that might impact data quality. This ensures the integrity and reliability of the dataset throughout the analytical journey, enabling a thorough exploration of data intricacies.

PERFORMANCE ASSESSMENT ON REAL CLINICAL TRIAL DATA

Validation involves applying the proposed methodology to a carefully crafted synthetic dataset designed for rigorous testing. Key performance metrics, including recall and false-positive rate, are utilized to measure the real-world applicability and reliability of the developed approach. This thorough validation process ensures the methodology's effectiveness and reliability in addressing data quality issues within clinical trial datasets.

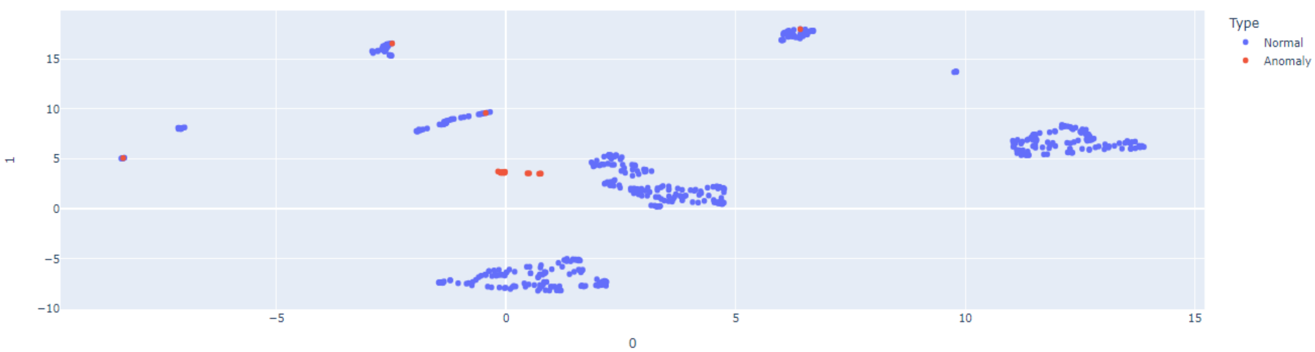
RESULTS

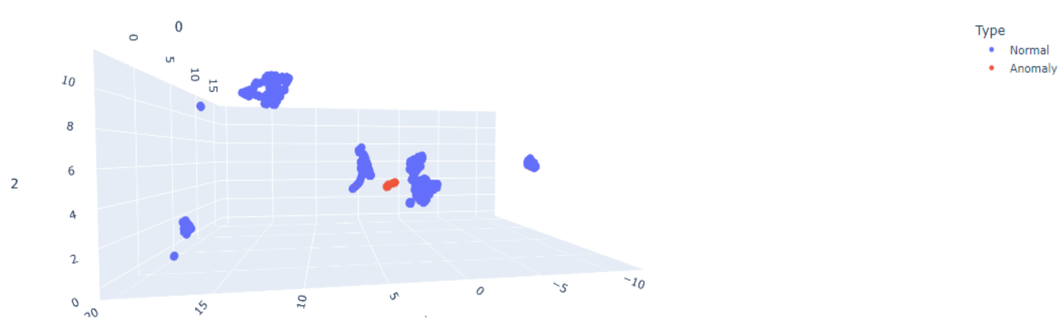
SYNTHETIC CLINICAL TRIAL DATA GENERATION

In our research endeavor, we curated a synthetic dataset comprising 520 data points, which included 20 outlier data points. The dataset encompasses 6 variables: vspos, vsores, vsoreum, vseq, vname, and testtime, representing 6 tests: Blood pressure, Temperature, SPO2, heart rate, pulse, and respiratory rate [7]. When subjected to encoding using OpenAI embeddings, this dataset yielded a transformed representation, with each record now spanning across 1536 dimensions.

IDENTIFICATION OF DATASET PATTERNS AND ANOMALIES

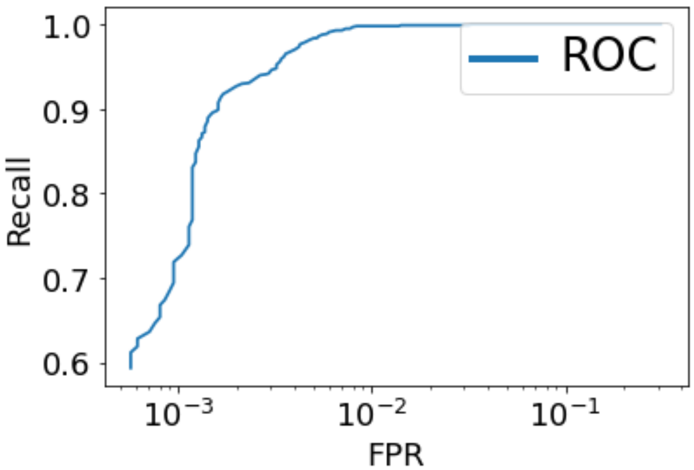
In the UMAP visualization, outliers were distinctly identified, showcasing the methodology's efficacy in detecting anomalous patterns within the clinical trial dataset. The visualization provided a clear representation of clusters, with outliers standing out as distinct points, enabling a comprehensive understanding of their presence and distribution.





PERFORMANCE ASSESSMENT ON REAL CLINICAL TRIAL DATA

The performance assessment of the autoencoder model further validated the methodology's robustness. Applied to a dataset comprising 520 records and 1536 embedding dimensions, including deliberately introduced 20 outlier data points, the methodology demonstrated exceptional accuracy. The achieved recall of 0.994 highlighted the methodology's ability to successfully identify and recall the majority of outlier data points. Additionally, the impressively low false-positive rate of 0.007 underscored the methodology's precision in minimizing false positives, enhancing its reliability for practical applications in healthcare data analysis.




```
Recall:  0.9949031600407747
FPR:  0.007318224740321058
```

These results collectively emphasize the methodology's capacity to not only pinpoint outliers in complex datasets but also to achieve a high level of accuracy and reliability in real-world clinical trial scenarios. The combination of autoencoder, PCA and UMAP visualization and performance metrics establishes a comprehensive framework for identifying and addressing data quality concerns, showcasing the methodology's potential for enhancing the integrity of healthcare data.

CONCLUSION

In conclusion, leveraging machine learning techniques, particularly OpenAI embeddings, proves promising for addressing data quality challenges in clinical trials. The methodology, by transforming patient data into text-like structures and employing advanced analytics, offers a comprehensive solution for identifying and rectifying outlier data points. Robust performance on real clinical trial data underscores practical applicability and potential to revolutionize data quality assurance in clinical research.

REFERENCES

1. Knepper D, Fenske C, Nadolny P, Bedding A, Gribkova E, Polzer J, Neumann J, Wilson B, Benedict J, Lawton A. Detecting Data Quality Issues in Clinical Trials: Current Practices and Recommendations. Ther Innov Regul Sci. 2016 Jan;50(1):15-21. doi: 10.1177/2168479015620248. PMID: 30236017..

2. https://openai.com/blog/introducing-text-and-code-embeddings

3. Neelakantan, A., Xu, T., Puri, R., Radford, A., Han, J. M., Tworek, J., Yuan, Q., Tezak, N., Kim, J. W., Hallacy, C., Heidecke, J., Shyam, P., Power, B., Nekoul, T. E., Sastry, G., Krueger, G., Schnurr, D., Such, F. P., Hsu, K., Thompson, M., Khan, T., Sherbakov, T., Jang, J., Welinder, P., & Weng, L. (2022). Text and Code Embeddings by Contrastive Pre-Training. arXiv preprint arXiv:2201.10005. https://doi.org/10.48550/arXiv.2201.10005

4. https://kms-healthcare.com/benefits-openai-healthcare/

5. Karabacak M, Margetis K. Embracing Large Language Models for Medical Applications: Opportunities and Challenges. Cureus. 2023 May 21;15(5):e39305. doi: 10.7759/cureus.39305. PMID: 37378099; PMCID: PMC10292051.

6 Sánchez-Rico, M., & Alvardo, J. (2020). Dimensionality reduction techniques as a preliminary step to cluster analysis: A comparison between PCA, t-SNE, and UMAP. In 9th European Congress of Methodology.

7. http://www.clinedc.com/invoke?page=/docs/sdtmig33/VS

CONTACT INFORMATION

Author Name: Ashwin Rai
Company: Evidera ThermoFisher-PPD
Address: 3007 W 155th Terrace, Overland Park, Kansas 66224
Work Phone: +1 6128192323
Email: ashwin.rai@evidera.com
Personal Email: raiashwink@gmail.com